Numer. Math. (2000) 87: 247–282
Digital Object Identifier (DOI) 10.1007/s002110000178

Numerische Mathematik

# Runge-Kutta methods in optimal control and the transformed adjoint system*

**William W. Hager**

Department of Mathematics, University of Florida, Gainesville, FL 32611, USA;
e-mail: hager@math.ufl.edu, http://www.math.ufl.edu/∼hager

Received January 11, 1999 / Revised version received October 11, 1999 /
Published online July 12, 2000 – 

**Summary.** The convergence rate is determined for Runge-Kutta discretizations of nonlinear control problems. The analysis utilizes a connection between the Kuhn-Tucker multipliers for the discrete problem and the adjoint variables associated with the continuous minimum principle. This connection can also be exploited in numerical solution techniques that require the gradient of the discrete cost function.

*Mathematics Subject Classification (1991):* 49M25, 65L06

## 1. Introduction

We analyze the convergence rate of Runge-Kutta discretizations of control problems. Unless the coefficients in the final stage of the Runge-Kutta scheme are all positive, the solution to the discrete problem can diverge from the solution to the continuous problem. In the case that these final coefficients are all positive, Runge-Kutta schemes of orders 1 or 2 yield discretizations of optimal control problems of orders 1 or 2 respectively. A third-order Runge-Kutta scheme for differential equations must satisfy an additional condition to achieve third-order accuracy for optimal control problems, while a fourth-order Runge-Kutta scheme for differential equations must satisfy another four conditions to achieve fourth-order accuracy in optimal control. One particular family of integration schemes for differential equations, the 4-stage explicit fourth-order Runge-Kutta schemes, satisfy all the conditions needed for fourth-order accuracy in optimal control. For third and fourth-order Runge-Kutta schemes, the discrete controls

* This work was supported by the National Science Foundation.

often converge to the continuous solution more slowly than the discrete state and adjoint variables at the grid points. As a result, a better approximation to the continuous optimal control is obtained from an *a posteriori* computation involving the computed discrete state and adjoint variables.

The analysis exploits the tree-based expansions and order conditions developed by Butcher [8] for ordinary differential equations, and a transformation of the first-order necessary conditions for the discrete control problem presented by the author in [31]. This transformation leads to a Runge-Kutta scheme for the adjoint (costate) equation in the optimal control problem which is often different from the original Runge-Kutta discretization of the state equation. This discrepancy between the state and costate discretizations leads to additional conditions that the coefficients of the Runge-Kutta scheme must satisfy to achieve third or fourth-order accuracy in the control context. This local order-of-accuracy analysis can be extended to an $L^\infty$ error bound using previously developed theory (see [35], [20], and [22]). In a companion paper [24] it is shown for second-order Runge-Kutta schemes, the positivity restriction for the coefficients in the final stage can be removed through a reduction in the dimension of the discrete control space. Some of the earlier work on discrete approximations to problems in optimal control includes the following papers and books: [4]–[7], [9]–[22], [28], [30]–[32], [35]–[37], [41]–[47], [49], and [50].

The paper is organized in the following way: Section 2 presents the Runge-Kutta discretization and the main theorem for unconstrained control problems. Section 3 derives the transformed adjoint system, and relates it both to the continuous adjoint equation and to the original discretization. Section 4 analyzes the order of approximation of Runge-Kutta discretizations of optimal control problems. This analysis is local in nature and involves conditions that the Runge-Kutta coefficients must satisfy so that the Taylor expansion of the discrete and the continuous problem match to a given order. Section 5 uses the abstract framework in [22, Thm. 3.1] to convert the local analysis into an $L^\infty$ error estimate for the solution to the discrete problem. Section 6 gives specific illustrations of the theory, and proves that a 4-stage explicit fourth-order Runge-Kutta scheme for differential equations yields a fourth-order discretization in optimal control. Finally, Sect. 7 analyzes the effect of control constraints.

## 2. The problem and its discretization

We consider the following optimal control problem:

(1) minimize $C(\mathbf{x}(1))$

$$\text{subject to } \mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{u}(t) \in U, \quad \text{a. e. } t \in [0,1],$$

$$\mathbf{x}(0) = \mathbf{a}, \quad \mathbf{x} \in W^{1,\infty}, \quad \mathbf{u} \in L^\infty,$$

where the state $\mathbf{x}(t) \in \mathbf{R}^n$, $\mathbf{x}'$ stands for $\frac{d}{dt}\mathbf{x}$, the control $\mathbf{u}(t) \in \mathbf{R}^m$, $\mathbf{f} : \mathbf{R}^n \times \mathbf{R}^m \mapsto \mathbf{R}^n$, $C : \mathbf{R}^n \mapsto \mathbf{R}$, and $U \subset \mathbf{R}^m$ is closed and convex.

Throughout the paper, $L^p(\mathbf{R}^n)$ denotes the usual Lebesgue space of measurable functions $\mathbf{x} : [0,1] \mapsto \mathbf{R}^n$ with $|\mathbf{x}(\cdot)|^p$ integrable, equipped with its standard norm

$$\|\mathbf{x}\|_{L^p} = \left( \int_0^1 |\mathbf{x}(t)|^p dt \right)^{1/p},$$

where $|\cdot|$ is the Euclidean norm. Of course, $p = \infty$ corresponds to the space of essentially bounded, measurable functions equipped with the essential supremum norm. Further, $W^{m,p}(\mathbf{R}^n)$ is the Sobolev space consisting of vector-valued measurable functions $\mathbf{x} : [0,1] \mapsto \mathbf{R}^n$ whose $j$-th derivative lies in $L^p$ for all $0 \le j \le m$ with the norm

$$\|\mathbf{x}\|_{W^{m,p}} = \sum_{j=0}^{m} \|\mathbf{x}^{(j)}\|_{L^p}.$$

When the range $\mathbf{R}^n$ is clear from context, it is omitted. Throughout, $c$ is a generic constant, that has different values in different relations, and which is independent of time and the mesh spacing in the approximating problem. The transpose of a matrix $\mathbf{A}$ is $\mathbf{A}^\mathsf{T}$, and $B_a(\mathbf{x})$ is the closed ball centered at $\mathbf{x}$ with radius $a$.

We now present the assumptions that are employed in our analysis of Runge-Kutta discretizations of (1). The first assumption is related to the regularity of the solution and the problem functions.

**Smoothness.** *For some integer $\kappa \ge 2$, the problem (1) has a local solution $(\mathbf{x}^*, \mathbf{u}^*)$ which lies in $W^{\kappa,\infty} \times W^{\kappa-1,\infty}$. There exists an open set $\boldsymbol{\Omega} \subset \mathbf{R}^n \times \mathbf{R}^m$ and $\rho > 0$ such that $B_\rho(\mathbf{x}^*(t), \mathbf{u}^*(t)) \subset \boldsymbol{\Omega}$ for every $t \in [0,1]$, the first $\kappa$ derivatives of $\mathbf{f}$ are Lipschitz continuous in $\boldsymbol{\Omega}$, and the first $\kappa$ derivative of $C$ are Lipschitz continuous in $B_\rho(\mathbf{x}^*(1))$.*

Under this assumption, there exists an associated Lagrange multiplier $\boldsymbol{\psi}^* \in W^{\kappa,\infty}$ for which the following form of the first-order optimality conditions (minimum principle) is satisfied at $(\mathbf{x}^*, \boldsymbol{\psi}^*, \mathbf{u}^*)$:

(2) $\mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t))$ for all $t \in [0,1]$, $\mathbf{x}(0) = \mathbf{a}$,

(3) $\boldsymbol{\psi}'(t) = -\nabla_x H(\mathbf{x}(t), \boldsymbol{\psi}(t), \mathbf{u}(t))$ for all $t \in [0,1]$,
$\boldsymbol{\psi}(1) = \nabla C(\mathbf{x}(1))$,

(4) $\mathbf{u}(t) \in U$, $-\nabla_u H(\mathbf{x}(t), \boldsymbol{\psi}(t), \mathbf{u}(t)) \in N_U(\mathbf{u}(t))$ for all $t \in [0,1]$.

Here $H$ is the Hamiltonian defined by

$$H(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = \boldsymbol{\psi}\mathbf{f}(\mathbf{x}, \mathbf{u}), \tag{5}$$

where $\boldsymbol{\psi}$ is a row vector in $\mathbf{R}^n$. The normal cone mapping $N_U$ is the following: For any $\mathbf{u} \in U$,

$$N_U(\mathbf{u}) = \{\mathbf{w} \in \mathbf{R}^m : \mathbf{w}^{\mathsf{T}}(\mathbf{v} - \mathbf{u}) \le 0 \text{ for all } \mathbf{v} \in U\}.$$

Let us define the following matrices:

$$\mathbf{A}(t) = \nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t)),\ \mathbf{B}(t) = \nabla_u \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t)),\ \mathbf{V} = \nabla C(\mathbf{x}^*(1)),$$

$$\mathbf{Q}(t) = \nabla_{xx} H(\mathbf{w}^*(t)),\ \mathbf{R}(t) = \nabla_{uu} H(\mathbf{w}^*(t)),\ \mathbf{S}(t) = \nabla_{xu} H(\mathbf{w}^*(t)),$$

where $\mathbf{w}^* = (\mathbf{x}^*, \boldsymbol{\psi}^*, \mathbf{u}^*)$. Let $\mathcal{B}$ be the quadratic form defined by

$$\mathcal{B}(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\left(\mathbf{x}(1)^{\mathsf{T}}\mathbf{V}\mathbf{x}(1) + \langle \mathbf{x}, \mathbf{Q}\mathbf{x}\rangle + \langle \mathbf{u}, \mathbf{R}\mathbf{u}\rangle + 2\langle \mathbf{x}, \mathbf{S}\mathbf{u}\rangle\right),$$

where $\langle \cdot, \cdot \rangle$ denotes the usual $L^2$ inner product. Our second assumption is a growth condition:

**Coercivity.** *There exists a constant $\alpha > 0$ such that*

$$\mathcal{B}(\mathbf{x}, \mathbf{u}) \ \ge\ \alpha \|\mathbf{u}\|_{L^2}^2 \quad \textit{for all} \ \ (\mathbf{x}, \mathbf{u}) \in \mathcal{M},$$

*where*

$$\mathcal{M} = \{(\mathbf{x}, \mathbf{u}) : \mathbf{x} \in W^{1,2}, \mathbf{u} \in L^2,\ \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u},$$
$$\mathbf{x}(0) = \mathbf{0},\ \mathbf{u}(t) \in U - U \textit{ a. e. } t \in [0, 1]\}.$$

Coercivity is a strong form of a second-order sufficient optimality condition in the sense that it implies not only strict local optimality, but also Lipschitzian dependence of the solution and multipliers with respect to parameters (see [20], [23], [21]). For recent work on second-order sufficient conditions, see [26] and [51].

We consider the discrete approximation to this continuous problem that is obtained by solving the differential equation using a Runge-Kutta integration scheme. For convenience, we consider a uniform mesh of width $h = 1/N$ where $N$ is a natural number, and we let $\mathbf{x}_k$ denote the approximation to $\mathbf{x}(t_k)$ where $t_k = kh$. An $s$-stage Runge-Kutta scheme [8] with coefficients $a_{ij}$ and $b_i$, $1 \le i, j \le s$, is given by

$$\mathbf{x}_k' = \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \tag{6}$$

where

$$\mathbf{y}_i = \mathbf{x}_k + h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \quad 1 \le i \le s, \tag{7}$$

and prime denotes, in this discrete context, the forward divided difference:

$$\mathbf{x}'_k = \frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{h}.$$

In (6) and (7), $\mathbf{y}_j$ and $\mathbf{u}_{kj}$ are the intermediate state and control variables on the interval $[t_k, t_{k+1}]$. The dependence of the intermediate state variables on $k$ is not explicit in our notation even though these variables have different values on different intervals.

With this notation, the discrete control problem is the following:

$$\begin{aligned} &\text{minimize} \quad C(\mathbf{x}_N) \\ &\text{subject to} \quad \mathbf{x}'_k = \textstyle\sum_{i=1}^{s} b_i\mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \quad \mathbf{x}_0 = \mathbf{a}, \quad \mathbf{u}_{ki} \in U, \\ &\qquad\qquad \mathbf{y}_i = \mathbf{x}_k + h\textstyle\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \\ &\qquad\qquad 1 \le i \le s, \quad 0 \le k \le N-1. \end{aligned} \tag{8}$$

For $\mathbf{x}_k$ near $\mathbf{x}^*(t_k)$ and $\mathbf{u}_{kj}$, $1 \le j \le s$, near $\mathbf{u}^*(t_k)$, it follows from Smoothness and the implicit function theorem that when $h$ is small enough, the intermediate variables $\mathbf{y}_i$ in (7) are uniquely determined. More precisely, the following holds (for example, see [8, Thm. 303A] and [1, Thm. 13.7] or [29, Thm. 10.8]):

**State Uniqueness Property.** *There exist positive constants $\gamma$ and $\beta \le \rho$ such that whenever $h \le \gamma$ and $(\mathbf{x}, \mathbf{u}_j) \in B_\beta(\mathbf{x}^*(t), \mathbf{u}^*(t))$ for some $t \in [0,1]$, $j = 1, \cdots, s$, the system of equations*

$$\mathbf{y}_i = \mathbf{x} + h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{y}_j, \mathbf{u}_j), \quad 1 \le i \le s, \tag{9}$$

*has a unique solution $\mathbf{y}_i \in B_\rho(\mathbf{x}^*(t), \mathbf{u}^*(t))$, $1 \le i \le s$. If $\mathbf{y}(\mathbf{x}, \mathbf{u})$ denotes the solution of (9) associated with given $(\mathbf{x}, \mathbf{u}) \in \mathbf{R}^n \times \mathbf{R}^{sm}$, then $\mathbf{y}(\mathbf{x}, \mathbf{u})$ is $\kappa$ times continuously differentiable in $\mathbf{x}$ and $\mathbf{u}$.*

Let $\mathbf{f}^h : \mathbf{R}^n \times \mathbf{R}^{sm} \mapsto \mathbf{R}^n$ be defined by

$$\mathbf{f}^h(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{s} b_i\mathbf{f}(\mathbf{y}_i(\mathbf{x}, \mathbf{u}), \mathbf{u}_i).$$

In other words,

$$\mathbf{f}^h(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{s} b_i\mathbf{f}(\mathbf{y}_i, \mathbf{u}_i),$$

**Table 1.** Order of a Runge-Kutta discretization for optimal control

| Order | Conditions ($c_i = \sum_{j=1}^s a_{ij}, \quad d_j = \sum_{i=1}^s b_i a_{ij}$) |
|---|---|
| 1 | $\sum b_i = 1$ |
| 2 | $\sum d_i = \frac{1}{2}$ |
| 3 | $\sum c_i d_i = \frac{1}{6}, \quad \sum b_i c_i^2 = \frac{1}{3}, \quad \sum d_i^2/b_i = \frac{1}{3}$ |
| 4 | $\sum b_i c_i^3 = \frac{1}{4}, \quad \sum b_i c_i a_{ij} c_j = \frac{1}{8}, \sum d_i c_i^2 = \frac{1}{12}, \quad \sum d_i a_{ij} c_j = \frac{1}{24},$ |
| | $\sum c_i d_i^2/b_i = \frac{1}{12}, \quad \sum d_i^3/b_i^2 = \frac{1}{4}, \quad \sum b_i c_i a_{ij} d_j/b_j = \frac{5}{24}, \quad \sum d_i a_{ij} d_j/b_j = \frac{1}{8}$ |

where $\mathbf{y}$ is the solution of (9) given by the state uniqueness property and $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_s) \in \mathbf{R}^{sm}$. The corresponding discrete Hamiltonian $H^h : \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^{sm} \mapsto \mathbf{R}$ is defined by

$$H^h(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = \boldsymbol{\psi} \mathbf{f}^h(\mathbf{x}, \mathbf{u}).$$

We consider the following version of the first-order necessary optimality conditions associated with (8) (see [3]):

$$\text{(10)} \quad \mathbf{x}_k' = \mathbf{f}^h(\mathbf{x}_k, \mathbf{u}_k), \quad \mathbf{x}_0 = \mathbf{a},$$

$$\text{(11)} \quad \boldsymbol{\psi}_k' = -\nabla_x H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k), \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N),$$

$$\text{(12)} \quad \mathbf{u}_{ki} \in U, \quad -\nabla_{u_i} H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k) \in N_U(\mathbf{u}_{ki}), \quad 1 \le i \le s,$$

where $\boldsymbol{\psi}_k \in \mathbf{R}^n$, $0 \le k \le N-1$. Here $\mathbf{u}_k \in \mathbf{R}^{ms}$ is the entire discrete control vector at time level $k$:

$$\mathbf{u}_k = (\mathbf{u}_{k1}, \mathbf{u}_{k2}, \cdots, \mathbf{u}_{ks}) \in \mathbf{R}^{ms}.$$

Throughout the paper, the index $k$ refers to the time level in the discrete problem, while $\mathbf{u}_i$ and $\mathbf{u}_j \in \mathbf{R}^m$ denote components of the vector $\mathbf{u} \in \mathbf{R}^{sm}$

Our estimate for the error in the discrete approximation to the control problem depends both on the smoothness of the solution to the continuous problem and on the order-of-accuracy of the Runge-Kutta scheme used for the discretization. In Table 1 we give the order conditions for Runge-Kutta discretizations of control problems. The conditions for any given order are those listed in Table 1 for that specific order along with those for all lower orders. We employ the following summation convention:

**Summation Convention.** *If an index range does not appear on a summation sign, then the summation is over each index, taking values from* 1 *to* $s$.

This deviates slightly from the usual Einstein summation notation in which only repeated indices are summed over.

Notice that the order conditions of Table 1 are not the usual order conditions [8, p. 170] associated with a Runge-Kutta discretization of a differential

**Table 2.** Order of a Runge-Kutta discretization for differential equations

| Order | Conditions ($c_i = \sum_{j=1}^{s} a_{ij}, \quad d_j = \sum_{i=1}^{s} b_i a_{ij}$) |
|---|---|
| 1 | $\sum b_i = 1$ |
| 2 | $\sum d_i = \frac{1}{2}$ |
| 3 | $\sum c_i d_i = \frac{1}{6}, \quad \sum b_i c_i^2 = \frac{1}{3}$ |
| 4 | $\sum b_i c_i^3 = \frac{1}{4}, \quad \sum b_i c_i a_{ij} c_j = \frac{1}{8}, \quad \sum d_i c_i^2 = \frac{1}{12}, \quad \sum d_i a_{ij} c_j = \frac{1}{24}$ |

equation. For orders up to 4, these conditions appear in Table 2. Through order 2, the conditions in Tables 1 and 2 are identical. At order 3, one new condition emerges in the control context, and at order 4, four new conditions emerge.

Our main result is formulated in terms of the averaged modulus of smoothness of the optimal control. If $J$ is an interval and $\mathbf{v} : J \mapsto \mathbf{R}^n$, let $\omega(\mathbf{v}, J; t, h)$ denote the the modulus of continuity:

$$\omega(\mathbf{v}, J; t, h) = \sup\{|\mathbf{v}(s_1) - \mathbf{v}(s_2)| : s_1, s_2 \in [t - h/2, t + h/2] \cap J\}. \tag{13}$$

The averaged modulus of smoothness $\tau$ of $\mathbf{v}$ over $[0, 1]$ is the integral of the modulus of continuity:

$$\tau(\mathbf{v}; h) = \int_0^1 \omega(\mathbf{v}, [0, 1]; t, h)\, dt.$$

It is shown in [48, Sect. 1.3] that $\lim_{h\to 0} \tau(\mathbf{v}; h) = 0$ if and only if $\mathbf{v}$ is Riemann integrable, and $\tau(\mathbf{v}; h) \le ch$ if $\mathbf{v}$ has bounded variation. Our main result is stated below in the context of unconstrained control problems, while the generalization to constrained problems is given in Sect. 7.

**Theorem 2.1.** *If Smoothness and Coercivity hold, $b_i > 0$ for each $i$, the Runge-Kutta scheme is of order $\kappa$ for optimal control, and $U = \mathbf{R}^m$, then for all sufficiently small $h$, there exists a strict local minimizer $(\mathbf{x}^h, \mathbf{u}^h)$ of the discrete optimal control problem $(8)$ and an associated adjoint variable $\boldsymbol{\psi}^h$ satisfying $(11)$ and $(12)$ such that*

$$\begin{aligned} \max_{0\le k\le N} |\mathbf{x}_k^h - \mathbf{x}^*(t_k)| + |\boldsymbol{\psi}_k^h - \boldsymbol{\psi}^*(t_k)| + |\mathbf{u}(\mathbf{x}_k^h, \boldsymbol{\psi}_k^h) - \mathbf{u}^*(t_k)| \\ \le \quad ch^{\kappa-1}\left(h + \tau(\frac{d^{\kappa-1}}{dt^{\kappa-1}}\mathbf{u}^*; h)\right), \end{aligned} \tag{14}$$

*where $\mathbf{u}(\mathbf{x}_k^h, \boldsymbol{\psi}_k^h)$ is a local minimizer of the Hamiltonian $(5)$ corresponding to $\mathbf{x} = \mathbf{x}_k$ and $\boldsymbol{\psi} = \boldsymbol{\psi}_k$.*

*Remark 2.2.* Note that the estimate for the error in the discrete control in (14) is expressed in terms of $\mathbf{u}(\mathbf{x}_k^h, \boldsymbol{\psi}_k^h)$ not $\mathbf{u}_k^h$. For Runge-Kutta schemes of third or fourth order, the error in the discrete controls $\mathbf{u}_{kj}^h$ may be one or more orders larger than the error in the control approximation $\mathbf{u}(\mathbf{x}_k^h, \boldsymbol{\psi}_k^h)$ obtained by minimization of the Hamiltonian using the discrete state/costate pair. On the other hand, the control approximation obtained by minimization of the Hamiltonian has the same order of accuracy as that of the discrete state and costate.

## 3. The transformed adjoint system

We now rewrite the first-order conditions (10)–(12) in a way that is better suited for analysis and computation. Suppose that a multiplier $\boldsymbol{\lambda}_i$ is introduced for the $i$-th intermediate equation (7) in addition to the multiplier $\boldsymbol{\psi}_{k+1}$ for the equation (6). Taking into account these additional multipliers, the first-order necessary conditions are the following:

$$\boldsymbol{\psi}_k - \boldsymbol{\psi}_{k+1} = \sum_{i=1}^{s} \boldsymbol{\lambda}_i, \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N), \tag{15}$$

$$h(b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij} \boldsymbol{\lambda}_i) \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) = \boldsymbol{\lambda}_j, \tag{16}$$

$$\mathbf{u}_{kj} \in U, \quad -(b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij} \boldsymbol{\lambda}_i) \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \in N_U(\mathbf{u}_{kj}), \tag{17}$$

$1 \le j \le s$ and $0 \le k \le N-1$. Here and elsewhere the dual multipliers are treated as row vectors.

In the case that $b_j > 0$ for each $j$, we now reformulate the first-order conditions in terms of the variables $\boldsymbol{\chi}_j$ defined by

$$\boldsymbol{\chi}_j = \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} \frac{a_{ij}}{b_j} \boldsymbol{\lambda}_i, \quad 1 \le j \le s. \tag{18}$$

With this definition, (16) reduces to

$$h b_j \boldsymbol{\chi}_j \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) = \boldsymbol{\lambda}_j. \tag{19}$$

Multiplying (19) by $a_{ji}/b_i$, summing over $j$, and substituting from (18), we have

$$h \sum_{j=1}^{s} \frac{b_j a_{ji}}{b_i} \boldsymbol{\chi}_j \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) = \sum_{j=1}^{s} \frac{a_{ji}}{b_i} \boldsymbol{\lambda}_j = \boldsymbol{\chi}_i - \boldsymbol{\psi}_{k+1}. \tag{20}$$

Summing (19) over $j$ and utilizing (15) gives

$$h\sum_{j=1}^{s} b_j \boldsymbol{\chi}_j \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) = \sum_{j=1}^{s} \boldsymbol{\lambda}_j = \boldsymbol{\psi}_k - \boldsymbol{\psi}_{k+1}. \tag{21}$$

Finally, substituting (18) in (17) yields

$$\mathbf{u}_{kj} \in U, \quad -b_j \boldsymbol{\chi}_j \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \in N_U(\mathbf{u}_{kj}), \quad 1 \le j \le s. \tag{22}$$

Since $N_U$ is a cone, the positive factor $b_j$ in (22) can be removed and equations (20)–(22) yield the transformed first-order system:

$$\boldsymbol{\psi}_k = \boldsymbol{\psi}_{k+1} + h\sum_{i=1}^{s} b_i \boldsymbol{\chi}_i \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N), \tag{23}$$

$$\boldsymbol{\chi}_i = \boldsymbol{\psi}_{k+1} + h\sum_{j=1}^{s} \frac{b_j a_{ji}}{b_i} \boldsymbol{\chi}_j \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \tag{24}$$

$$\mathbf{u}_{ki} \in U, \quad -\boldsymbol{\chi}_i \nabla_u \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}) \in N_U(\mathbf{u}_{ki}), \tag{25}$$

$1 \le i \le s$ and $0 \le k \le N-1$.

Observe that conditions (23) and (24) are in essence a Runge-Kutta scheme applied to the continuous adjoint equation (4). Although the adjoint Runge Kutta scheme is generally not the same as the scheme (6) and (7) for the state equation, it is observed in [31] that some common Runge-Kutta schemes are symmetric in the sense that the state and adjoint schemes are the same.

**Proposition 3.1.** *If $b_j > 0$ for each $j$, then the first-order system* (15)–(17) *and the transformed first-order system* (23)–(25) *are equivalent. That is, if $\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_s$ satisfy* (15)–(17)*, then* (23)–(25) *hold for $\boldsymbol{\chi}_j$ defined in* (18)*. Conversely, if $\boldsymbol{\chi}_1, \cdots, \boldsymbol{\chi}_s$ satisfy* (23)–(25)*, then* (15)–(17) *hold for $\boldsymbol{\lambda}_j$ defined in* (19)*.*

*Proof.* We already derived the transformed first-order conditions starting from the original first-order conditions. Now suppose that $\boldsymbol{\chi}_1, \cdots, \boldsymbol{\chi}_s$ satisfy the transformed conditions (23)–(25). Summing over $j$ in (19), and utilizing (23) yields (15). To verify (16) and (17), we substitute for $\boldsymbol{\lambda}_i$ using (19) to obtain

$$\begin{aligned} b_j\boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij}\boldsymbol{\lambda}_i &= b_j\boldsymbol{\psi}_{k+1} + h\sum_{i=1}^{s} b_i a_{ij} \boldsymbol{\chi}_i \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}) \\ &= b_j\boldsymbol{\psi}_{k+1} + hb_j \sum_{i=1}^{s} \frac{a_{ij} b_i}{b_j} \boldsymbol{\chi}_i \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \\ &= b_j \boldsymbol{\chi}_j, \end{aligned} \tag{26}$$

where the last line comes from (24). Multiplying (26) on the right by $\nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})$ and substituting from (19) gives (16). Multiplying (26) on the right by $\nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})$ and utilizing (25) yields (17). □

*Remark 3.2.* Let $\mathbf{u} \in \mathbf{R}^{smN}$ denote the vector of intermediate control values for the entire interval $[0, 1]$, and let $C(\mathbf{u})$ denote the value $C(\mathbf{x}_N)$ for the discrete cost function associated with these controls. The transformed first-order system (23)–(25) provides a convenient way to compute the gradient of the discrete cost function (8). In particular, as seen in [37],

$$\nabla_{u_{kj}} C(\mathbf{u}) = h b_j \boldsymbol{\chi}_j \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \tag{27}$$

where the intermediate values for the discrete state and costate variables are gotten by first solving the discrete state equations (6) and (7), for $k = 0, 1, \cdots, N-1$, using the given values for the controls, and then using these computed values for both the state and intermediate variables in (23) and (24) when computing the values of the discrete costate for $k = N-1, N-2, \cdots, 0$. Thus the discrete state equation is solved by marching forward from $k = 0$ while the discrete costate equation is solved by marching backward from $k = N-1$.

We now observe that the multiplier $\boldsymbol{\psi}_k$ gotten by solving (24) for $\boldsymbol{\chi}$ and substituting into (23) is identical to the multiplier gotten from (11). Moreover, the condition (25) involving $\boldsymbol{\chi}_j$ satisfying (24) is equivalent to the condition (12).

**Proposition 3.3.** *Suppose that*

$$(\mathbf{x}_k, \mathbf{u}_{kj}) \in B_\beta(\mathbf{x}^*(t_k), \mathbf{u}^*(t_k)), \quad 1 \le j \le s,$$

*and for* $\mathbf{y} = \mathbf{y}(\mathbf{x}_k, \mathbf{u}_k)$, *let* $\mathbf{M}$ *be the* $s \times s$ *block matrix whose* $(i, j)$ *block is the* $n \times n$ *matrix* $a_{ij} \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})$. *If* $h$ *is small enough that* $\mathbf{I} - h\mathbf{M}$ *is invertible, then there exists a solution* $\boldsymbol{\chi}_1, \cdots, \boldsymbol{\chi}_s$ *to* (24), *and we have*

$$\nabla_x H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k) = \sum_{i=1}^{s} b_i \boldsymbol{\chi}_i \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}) = \sum_{i=1}^{s} b_i \nabla_x H(\mathbf{y}_i, \boldsymbol{\chi}_i, \mathbf{u}_{ki}) \tag{28}$$

*and*

$$\begin{aligned} \nabla_{u_j} H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k) &= b_j \boldsymbol{\chi}_j \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \\ &= b_j \nabla_u H(\mathbf{y}_j, \boldsymbol{\chi}_j, \mathbf{u}_{kj}). \end{aligned} \tag{29}$$

*Proof.* Our approach is to obtain identities in the vector $\boldsymbol{\lambda}$ which are then converted to identities in $\boldsymbol{\chi}$. Equation (16) has the form

$$\boldsymbol{\lambda}\mathbf{M} = h\boldsymbol{\psi}_{k+1}\mathbf{C},$$

where $\mathbf{C}$ is the $1 \times s$ block matrix whose $j$-th element is $b_j \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})$. Using the implicit function theorem and differentiating the solution $\mathbf{y}$ of (9) with respect to $\mathbf{x}$ and evaluating at $(\mathbf{x}_k, \mathbf{u}_k)$, we obtain a relation of the form

$$\mathbf{M}\nabla_x \mathbf{y} = \mathbf{D} \quad \text{or} \quad \nabla_x \mathbf{y} = \mathbf{M}^{-1}\mathbf{D}, \tag{30}$$

where $\mathbf{D}$ is the $s \times 1$ block matrix with each element an $n \times n$ identity matrix $\mathbf{I}$. Utilizing (11) and (30), we have

$$\begin{aligned} \nabla_x H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k) &= \boldsymbol{\psi}_{k+1} \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}) \nabla_x \mathbf{y}_i \\ &= \boldsymbol{\psi}_{k+1} \mathbf{C} \nabla_x \mathbf{y} = \boldsymbol{\psi}_{k+1} \mathbf{C}\mathbf{M}^{-1}\mathbf{D} \\ &= \frac{1}{h}\boldsymbol{\lambda}\mathbf{D} = \frac{1}{h}\sum \boldsymbol{\lambda}_i. \end{aligned} \tag{31}$$

Since (16) has a (unique) solution, it follows from Proposition 3.1 that (24) has a solution, and if $\boldsymbol{\chi}$ is a solution, then the unique solution to (16) is given by $\boldsymbol{\lambda}_j = h b_j \boldsymbol{\chi}_j \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})$. With this substitution in (31), we obtain (28).

Now consider the second relation (29). Differentiating (9) with respect to $\mathbf{u}_j$ and evaluating at $(\mathbf{x}_k, \mathbf{u}_k)$, we obtain the relation $\mathbf{M}\nabla_{u_j}\mathbf{y} = h\mathbf{D}_j \nabla_u \mathbf{f}$ $(\mathbf{y}_j, \mathbf{u}_{kj})$ where $\mathbf{D}_j$ is the $s \times 1$ block matrix whose $i$-th element is $a_{ij}\mathbf{I}$. In terms of the matrix $\mathbf{C}$ introduced above, we have

$$\begin{aligned} \nabla_{u_j} H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k) &= \boldsymbol{\psi}_{k+1}\Big( b_j \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \\ &\qquad + \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}) \nabla_{u_j} \mathbf{y}_i \Big) \\ &= \boldsymbol{\psi}_{k+1}\Big( b_j \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) + \mathbf{C}\nabla_{u_j}\mathbf{y} \Big) \\ &= \boldsymbol{\psi}_{k+1}\Big( b_j \mathbf{I} + h\mathbf{C}\mathbf{M}^{-1}\mathbf{D}_j \Big) \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \\ &= \Big( b_j \boldsymbol{\psi}_{k+1} + \boldsymbol{\lambda}\mathbf{D}_j \Big) \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \\ &= \Big( b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij}\boldsymbol{\lambda}_i \Big) \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \\ &= b_j \boldsymbol{\chi}_j \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}). \end{aligned}$$

This completes the proof. □

Since the boundary conditions for $\boldsymbol{\psi}_N$ are the same in both (11) and (15), it follows from (28) that when $h$ is sufficiently small and $(\mathbf{x}_k, \mathbf{u}_{kj}) \in B_\beta(\mathbf{x}^*(t_k), \mathbf{u}^*(t_k))$ for each $j$ and $k$, then the $\boldsymbol{\psi}_k$ given by (11) and by

(23)–(24) are the same. Moreover, the control gradient satisfies (12) if and only if the following relation holds in the transformed variables:

$$\mathbf{u}_{ki} \in U, \quad -\nabla_u H(\mathbf{y}_i, \boldsymbol{\chi}_i, \mathbf{u}_{ki}) \in N_U(\mathbf{u}_{ki}), \quad 1 \le i \le s.$$

The transformed discrete costate equations (23)–(24) march backwards in time while the discrete state equations (6)–(7) march forwards in time. To facilitate the error analysis, we now reverse the order of time in the costate equation. That is, we solve for $\boldsymbol{\psi}_{k+1}$ in (23) and substitute in (24) to obtain the following forward marching scheme:

$$\boldsymbol{\psi}_{k+1} = \boldsymbol{\psi}_k - h \sum_{i=1}^{s} b_i \boldsymbol{\chi}_i \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \tag{32}$$

$$\boldsymbol{\chi}_i = \boldsymbol{\psi}_k - h \sum_{j=1}^{s} \bar{a}_{ij} \boldsymbol{\chi}_j \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \quad \bar{a}_{ij} = \frac{b_i b_j - b_j a_{ji}}{b_i}. \tag{33}$$

We will now remove the control from the state equation and the transformed adjoint equation by use of the minimum principle. As noted in [27] or [23, Lem. 2], Coercivity implies that

$$\mathbf{v}^{\mathsf{T}} R(t) \mathbf{v} \ge \alpha |\mathbf{v}|^2 \quad \text{for all} \quad \mathbf{v} \in U - U \quad \text{and} \quad t \in [0, 1]. \tag{34}$$

It follows by Smoothness and [33, Thm. 4.1] that the Hamiltonian has a locally unique minimizer in the control and the following property holds:

**Control Uniqueness Property.** *There exist positive constants $\beta$ and $\sigma$, both smaller than $\rho$, such that whenever $(\mathbf{x}, \boldsymbol{\psi}) \in B_\beta(\mathbf{x}^*(t), \boldsymbol{\psi}^*(t))$ for some $t \in [0, 1]$, the problem*

$$\min_{\mathbf{u} \in B_\sigma(\mathbf{u}^*(t))} H(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) \tag{35}$$

*has a unique solution denoted $\mathbf{u}(\mathbf{x}, \boldsymbol{\psi})$ depending Lipschitz continuously on $\mathbf{x}$ and $\boldsymbol{\psi}$. Moreover, if $U = \mathbf{R}^m$, then (by the implicit function theorem) $\mathbf{u}(\mathbf{x}, \boldsymbol{\psi})$ is $\kappa - 1$ times Lipschitz continuously differentiable in $\mathbf{x}$ and $\boldsymbol{\psi}$.*

By the control uniqueness property, if $(\mathbf{x}, \boldsymbol{\psi})$ is sufficiently close to $(\mathbf{x}^*(t_k), \boldsymbol{\psi}^*(t_k))$, there exists a locally unique minimizer $\mathbf{u} = \mathbf{u}(\mathbf{x}, \boldsymbol{\psi})$ of the Hamiltonian in (35). Focusing on the situation where the control is uniquely determined by $(\mathbf{x}, \boldsymbol{\psi})$ through minimization of the Hamiltonian, let $\boldsymbol{\phi}$ denote the function defined by

$$\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\psi}) = -\nabla_x H(\mathbf{x}, \mathbf{u}, \boldsymbol{\psi})|_{\mathbf{u} = \mathbf{u}(\mathbf{x}, \boldsymbol{\psi})}.$$

And with some abuse of notation, let $\mathbf{f}(\mathbf{x}, \boldsymbol{\psi})$ denote the function $\mathbf{f}(\mathbf{x}, \mathbf{u}(\mathbf{x}, \boldsymbol{\psi}))$. In the case where the control has the special form $\mathbf{u}_{kj} = \mathbf{u}(\mathbf{y}_j,$

$\boldsymbol{\chi}_j)$, the Runge-Kutta discretization (6)–(7), coupled with the transformed, time reversed costate equations (32)–(33), can be expressed:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + h\sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i, \boldsymbol{\chi}_i), \quad \mathbf{x}_0 = \mathbf{a}, \tag{36}$$

$$\boldsymbol{\psi}_{k+1} = \boldsymbol{\psi}_k + h\sum_{i=1}^{s} b_i \boldsymbol{\phi}(\mathbf{y}_i, \boldsymbol{\chi}_i), \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N), \tag{37}$$

$$\mathbf{y}_i = \mathbf{x}_k + h\sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}_j, \boldsymbol{\chi}_j), \tag{38}$$

$$\boldsymbol{\chi}_i = \boldsymbol{\psi}_k + h\sum_{j=1}^{s} \bar{a}_{ij} \boldsymbol{\phi}(\mathbf{y}_j, \boldsymbol{\chi}_j), \tag{39}$$

where $\bar{a}_{ij}$ is defined in (33).

Since $\mathbf{u}(\mathbf{x}, \boldsymbol{\psi})$ depends Lipschitz continuously on $\mathbf{x}$ near $\mathbf{x}^*(t)$ and $\boldsymbol{\psi}$ near $\boldsymbol{\psi}^*(t)$, for any $t \in [0, 1]$, we have the following uniqueness property, analogous to the state uniqueness property:

**Costate Uniqueness Property.** *There exist positive constants $\gamma$ and $\beta \leq \rho$ such that whenever $h \leq \gamma$ and $(\mathbf{x}, \boldsymbol{\psi}) \in B_\beta(\mathbf{x}^*(t), \boldsymbol{\psi}^*(t))$ for some $t \in [0, 1]$, the system of equations*

$$\mathbf{y}_i = \mathbf{x} + h\sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}_j, \boldsymbol{\chi}_j), \tag{40}$$

$$\boldsymbol{\chi}_i = \boldsymbol{\psi} + h\sum_{j=1}^{s} \bar{a}_{ij} \boldsymbol{\phi}(\mathbf{y}_j, \boldsymbol{\chi}_j), \tag{41}$$

*has a unique solution $(\mathbf{y}_i, \boldsymbol{\chi}_i) \in B_\rho(\mathbf{x}^*(t), \boldsymbol{\psi}^*(t)), 1 \leq i \leq s$. The functions $\mathbf{f}$ and $\boldsymbol{\phi}$ are Lipschitz continuous in $B_\beta(\mathbf{x}^*(t), \boldsymbol{\psi}^*(t))$ for each $t \in [0, 1]$. Moreover, if $U = \mathbf{R}^m$, then $\mathbf{f}$ and $\boldsymbol{\phi}$ are $\kappa - 1$ times Lipschitz continuously differentiable in $B_\beta(\mathbf{x}^*(t), \boldsymbol{\psi}^*(t))$.*

The scheme (36)–(39) can be viewed as a discretization of the following two-point boundary-value problem:

$$\mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\psi}(t)), \quad \mathbf{x}(0) = \mathbf{a}, \tag{42}$$

$$\boldsymbol{\psi}'(t) = \boldsymbol{\phi}(\mathbf{x}(t), \boldsymbol{\psi}(t)), \quad \boldsymbol{\psi}(1) = \nabla C(\mathbf{x}(1)). \tag{43}$$

This two-point boundary-value problem is gotten by substituting in (2)–(4) the control obtained by solving (4) for $\mathbf{u}(t)$ in terms of $(\mathbf{x}(t), \boldsymbol{\psi}(t))$.

## 4. Order of approximation

Butcher [8] has devised an elegant theory for determining the order of accuracy of a Runge-Kutta integration scheme for a differential equation. If the continuous solution to the differential equation is substituted into the discrete equations, the residual is $O(h^k)$ where $k$ can be determined by checking the order conditions in Table 2. The theory developed by Butcher does not apply to the discretization (36)–(39) since the coefficient $\bar{a}_{ij}$ for the costate equation, given in (33), may not match the coefficient $a_{ij}$ of the state equation. In this section, we carry out an order analysis analogous to that of Butcher, but in the context of the special discretization (36)–(39) connected with optimal control. Conceptually, our approach applies to schemes of any order, however, the particular results that we present are for schemes of order less than 5.

Let $\mathbf{z}$ and $\mathbf{g}$ denote the following pairs:

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\psi} \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} \mathbf{f} \\ \boldsymbol{\phi} \end{pmatrix}.$$

With this notation, the differential equation (42)–(43) has the form $\mathbf{z}' = \mathbf{g}(\mathbf{z})$. There are two facets to Butcher's analysis. First, there is a tree-based formulation for the Taylor expansion of $\mathbf{z}$. Restricting the expansion to terms up to fourth order, [8, Thm. 302D] yields:

$$\begin{aligned} \mathbf{z}(t_{k+1}) &= \mathbf{z}(t_k) + \mathbf{g}h + \frac{1}{2}\mathbf{g}'\mathbf{g}h^2 + \frac{1}{6}\Big(\mathbf{g}''\mathbf{g}^2 + \mathbf{g}'\mathbf{g}'\mathbf{g}\Big)h^3 \\ &+\frac{1}{24}\Big(\mathbf{g}'''\mathbf{g}^3 + 3\mathbf{g}''\mathbf{g}\mathbf{g}'\mathbf{g} + \mathbf{g}'\mathbf{g}''\mathbf{g}^2 + (\mathbf{g}')^3\mathbf{g}\Big)h^4\Big|_{\mathbf{z}(t_k)} + O(h^5). \end{aligned} \tag{44}$$

In this expansion, $\mathbf{g}$ and its derivatives are all evaluated at $\mathbf{z}(t_k)$, and the various derivatives should be viewed in an operator context. That is, the first derivative $\mathbf{g}'$ of $\mathbf{g}$ operates on a vector to give a vector. Of course, the first derivative of a vector-valued function corresponds to the Jacobian matrix and the operation $\mathbf{g}'\mathbf{g}$ corresponds to multiplying the Jacobian matrix by the vector $\mathbf{g}$. The second derivative $\mathbf{g}''$ operates on a pair of vector to yield a vector; hence, the 4-th order term $\mathbf{g}'\mathbf{g}''\mathbf{g}^2$ means that the second derivative operates on the pair of vectors $(\mathbf{g}, \mathbf{g})$ to give a vector which is acted on by $\mathbf{g}'$.

The expansion (44) is the standard Taylor expansion for $\mathbf{z}(t)$ around $t = t_k$:

$$\begin{aligned} \mathbf{z}(t_{k+1}) &= \mathbf{z}(t_k) + \mathbf{z}'(t_k)h + \frac{1}{2}\mathbf{z}''(t_k)h^2 \\ &+ \cdots + \frac{1}{j!}\mathbf{z}^{(j)}(t_k)h^j + \frac{1}{j!}\int_{t_k}^{t_{k+1}} (t - t_k)^j \mathbf{z}^{(j+1)}(t)\, dt. \end{aligned} \tag{45}$$

In (44) the various derivatives of $\mathbf{z}$ are replaced by their equivalent representation in terms of $\mathbf{g}$ and its derivatives. In Theorem 302D, Butcher uses the integral form for the remainder term shown above. However, in optimal control, where the solutions may have limited smoothness, it is better to modify the expansion in the following way:

$$\begin{aligned}\mathbf{z}(t_{k+1}) &= \mathbf{z}(t_k) + \mathbf{z}'(t_k)h + \frac{1}{2}\mathbf{z}''(t_k)h^2 \\ &+ \cdots + \frac{1}{j!}\mathbf{z}^{(j)}(t_k)h^j + \frac{1}{(j-1)!}\int_{t_k}^{t_{k+1}} \\ &\times (t-t_k)^{j-1}(\mathbf{z}^{(j)}(t) - \mathbf{z}^{(j)}(t_k))\, dt. \end{aligned} \tag{46}$$

This form is gotten by stopping one term earlier in the Taylor series, and then adding and subtracting the $\mathbf{z}^{(j)}(t_k)$ term under the integral sign. The polynomials in $h$ appearing in (45) and (46) are identical, the expansions only differ in the form of the remainder term. The remainder term in (46) involves one less derivative of $\mathbf{z}$ than that in (45).

The second facet of Butcher's analysis is an analogous expansion for the next Runge-Kutta iterate $\mathbf{z}_{k+1}$ in terms of the current iterate $\mathbf{z}_k$. For given values of $\mathbf{x}_k$ and $\boldsymbol{\psi}_k$, the solution $\mathbf{y}_i$ and $\boldsymbol{\chi}_i$ to (38) and (39) are functions of $h$ that we denote $\mathbf{y}_i(h)$ and $\boldsymbol{\chi}_i(h)$. Let $\mathbf{x}_{k+1}(h)$ and $\boldsymbol{\psi}_{k+1}(h)$ denote the values $\mathbf{x}_{k+1}$ and $\boldsymbol{\psi}_{k+1}$ obtained by substituting $\mathbf{y}_i = \mathbf{y}_i(h)$ and $\boldsymbol{\chi}_i = \boldsymbol{\chi}_i(h)$ in (36) and (37), and let $\boldsymbol{\zeta}(h)$ be the vector of length $2n(s+1)$ given by

$$\boldsymbol{\zeta}_i(h) = \begin{pmatrix} \mathbf{y}_i(h) \\ \boldsymbol{\chi}_i(h) \end{pmatrix}, \quad 1 \le i \le s, \quad \boldsymbol{\zeta}_{s+1}(h) = \begin{pmatrix} \mathbf{x}_{k+1}(h) \\ \boldsymbol{\psi}_{k+1}(h) \end{pmatrix}.$$

With this notation, the system (36)–(39) can be expressed

$$\boldsymbol{\zeta}(h) = \boldsymbol{\zeta}(0) + h\mathbf{G}(\boldsymbol{\zeta}(h)),$$

where

$$\mathbf{G}_i(\boldsymbol{\zeta}) = \begin{pmatrix} \sum_{j=1}^{s} a_{ij}\mathbf{f}(\boldsymbol{\zeta}_j) \\ \sum_{j=1}^{s} \bar{a}_{ij}\boldsymbol{\phi}(\boldsymbol{\zeta}_j) \end{pmatrix}, \quad 1 \le i \le s+1,$$

with the convention that

$$a_{s+1,j} = \bar{a}_{s+1,j} = b_j, \quad 1 \le j \le s. \tag{47}$$

Expanding $\boldsymbol{\zeta}(h)$ in a Taylor series around $h = 0$, Butcher's result [8, Thm 303C] yields

$$\begin{aligned}\boldsymbol{\zeta}(h) &= \boldsymbol{\zeta}(0) + \mathbf{G}h + \frac{1}{2}(2\mathbf{G}'\mathbf{G})h^2 \\ &\frac{1}{6}\Big(3\mathbf{G}''\mathbf{G}^2 + 6\mathbf{G}'\mathbf{G}'\mathbf{G}\Big)h^3\end{aligned}$$

$$+\frac{1}{24}\Big(4\mathbf{G}'''\mathbf{G}^3+24\mathbf{G}''\mathbf{G}\mathbf{G}'\mathbf{G}+12\mathbf{G}'\mathbf{G}''\mathbf{G}^2$$

$$+24(\mathbf{G}')^3\mathbf{G}\Big)h^4\Big|_{\boldsymbol{\zeta}(0)}+O(h^5). \tag{48}$$

Here $\mathbf{G}$ and its derivatives are all evaluated at $\boldsymbol{\zeta}(0)$ where

$$\boldsymbol{\zeta}_i(0)=\mathbf{z}_k=\begin{pmatrix}\mathbf{x}_k\\ \boldsymbol{\psi}_k\end{pmatrix},\quad 1\le i\le s+1.$$

The error that results from stopping the Taylor expansion (48) at any term is again given by an integral as in either (45) or (46), but with $\mathbf{z}$ replaced by $\boldsymbol{\zeta}$ and with the integration performed over $h$ instead of $t$. (Note that in [8], Butcher utilizes a different representation for the error in the Taylor series for $\boldsymbol{\zeta}$, while here we utilize the integral representation appearing in (46)).

We say that the Runge-Kutta scheme (36)–(39) for the system (42)–(43) is of order $\nu$ if the expansion (44) and the $(s+1)$-st component $\boldsymbol{\zeta}_{s+1}(h)$ in the expansion (48) agree through terms of order $h^\nu$ when $\mathbf{f}$ and $\boldsymbol{\phi}$ have the necessary derivatives and $\mathbf{z}_k=\mathbf{z}(t_k)$.

**Theorem 4.1.** *For $\nu=1$, 2, 3, or 4, the Runge-Kutta scheme (36)–(39) is of order $\nu$ if the conditions of Table 1 are satisfied.*

*Proof.* Throughout the proof, no arguments are given for functions, which are all evaluated at $\mathbf{z}_k$. Also, we define

$$\bar{c}_i=\sum_{j=1}^{s}\bar{a}_{ij},\quad 1\le i\le s+1.$$

Due to the convention (47), $c_{s+1}=\bar{c}_{s+1}=\sum b_i$. Since $\mathbf{G}_{s+1}=\sum b_i\mathbf{g}$, we see immediately that if the order 1 condition of Table 1 holds, then the expansion (44) matches the corresponding term in (48) to first order.

By the definition of $\mathbf{c}$ in Table 1, we have

$$\mathbf{G}_i=\begin{pmatrix}c_i\mathbf{f}\\ \bar{c}_i\boldsymbol{\phi}\end{pmatrix}c_i\mathbf{f}_0+\bar{c}_i\boldsymbol{\phi}_0,\quad\text{where}\quad \mathbf{f}_0=\begin{pmatrix}\mathbf{f}\\ \mathbf{0}\end{pmatrix},$$

$$\boldsymbol{\phi}_0=\begin{pmatrix}\mathbf{0}\\ \boldsymbol{\phi}\end{pmatrix}. \tag{49}$$

The derivative of $\mathbf{G}$ can be viewed as a block matrix with the following elements:

$$(\mathbf{G}_i')_j=a_{ij}\mathbf{f}_0'+\bar{a}_{ij}\boldsymbol{\phi}_0',\quad (\mathbf{G}_{s+1}')_j=b_j\mathbf{g}'.$$

Hence, we have

$$\begin{aligned}\mathbf{G}'_{s+1}\mathbf{G} &= \sum_{i=1}^{s}(\mathbf{G}'_{s+1})_i\mathbf{G}_i \\ &= \sum b_i\mathbf{g}'(c_i\mathbf{f}_0 + \bar{c}_i\boldsymbol{\phi}_0). \qquad (50)\end{aligned}$$

On the other hand, the corresponding term in the expansion (45) is

$$\frac{1}{2}\mathbf{g}'\mathbf{g} = \frac{1}{2}\mathbf{g}'(\mathbf{f}_0 + \boldsymbol{\phi}_0).$$

This term is identical to (50) if

$$\sum b_i c_i = \frac{1}{2} = \sum b_i\bar{c}_i. \qquad (51)$$

The first equality is already contained in Table 1. For the second equality, we use the definition of $\bar{a}_{ij}$ to obtain the identity

$$\bar{c}_i = \sum_{j=1}^{s}\frac{b_ib_j - b_ja_{ji}}{b_i} = 1 - \sum_{j=1}^{s}\frac{b_ja_{ji}}{b_i} = 1 - d_i/b_i. \qquad (52)$$

Hence, by the conditions in Table 1 for orders 1 and 2, we have

$$\sum b_i\bar{c}_i = \sum b_i - d_i = \frac{1}{2},$$

which establishes the second equality in (51). Consequently, the second-order conditions of Table 1 imply that the Runge-Kutta scheme is second-order accurate for optimal control.

Now consider the third-order conditions. Due to the structure of $\mathbf{G}_i$, any mixed derivative vanishes:

$$\frac{\partial\mathbf{G}_i}{\partial\zeta_j\partial\zeta_k} = 0 \quad \text{for all } j \neq k.$$

As a result, the derivatives have a very special structure. In particular, for the second derivative, we have

$$\mathbf{G}''_i(\mathbf{v},\mathbf{w}) = \sum_{j=1}^{s}a_{ij}\mathbf{f}''_0(\mathbf{v}_j,\mathbf{w}_j) + \sum_{j=1}^{s}\bar{a}_{ij}\boldsymbol{\phi}''_0(\mathbf{v}_j,\mathbf{w}_j), \quad 1 \le i \le s+1.$$

In the special case $i = s+1$, this reduces to

$$\mathbf{G}''_{s+1}(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{s}b_i\mathbf{g}''(\mathbf{v}_i,\mathbf{w}_i).$$

Utilizing (49), we have

$$\mathbf{G}''_{s+1}(\mathbf{G},\mathbf{G}) = \sum b_i \mathbf{g}''(c_i \mathbf{f}_0 + \bar{c}_i \boldsymbol{\phi}_0)^2.$$

This is equal to the corresponding term in (44) if the following conditions hold:

$$\frac{1}{3} = \sum b_i c_i^2 = \sum b_i c_i \bar{c}_i = \sum b_i \bar{c}_i^2. \tag{53}$$

The first equality is contained in Table 1. For the second equality, we utilize the relation (52) and Table 1 to obtain

$$\sum b_i c_i \bar{c}_i = \sum c_i(b_i - d_i) = \sum d_i - c_i d_i = \frac{1}{3}.$$

For the third equality in (53), observe that

$$\sum b_i \bar{c}_i^2 = \sum (b_i - d_i)^2 / b_i = \sum b_i - 2d_i + d_i^2/b_i = \frac{1}{3},$$

which completes the proof of (53).

The final third-order term coming from (48) is

$$\mathbf{G}'_{s+1}\mathbf{G}'\mathbf{G} = \sum b_i \mathbf{g}'(a_{ij}\mathbf{f}'_0 + \bar{a}_{ij}\boldsymbol{\phi}'_0)(c_j \mathbf{f}_0 + \bar{c}_j \boldsymbol{\phi}_0).$$

This term is equal to the corresponding term in (44) if the following conditions hold:

$$\frac{1}{6} = \sum b_i a_{ij} c_j = \sum b_i a_{ij} \bar{c}_j = \sum b_i \bar{a}_{ij} c_j = \sum b_i \bar{a}_{ij} \bar{c}_j.$$

The first equality is contained in Table 1. For the second equality, Table 1 and (52) yield

$$\sum b_i a_{ij} \bar{c}_j = \sum d_j(1 - d_j/b_j) = \frac{1}{6}.$$

For the third equality, we have

$$\sum b_i \bar{a}_{ij} c_j = \sum (b_i b_j - b_j a_{ji}) c_j = \sum b_j c_j - b_j c_j^2 = \frac{1}{6}.$$

And for the fourth equality,

$$\sum b_i \bar{a}_{ij} \bar{c}_j = \sum (b_i b_j - b_j a_{ji})(1 - d_j/b_j) = \sum (1 - c_j)(b_j - d_j) = \frac{1}{6}.$$

At this point, we have checked the conditions of Table 1 up to third order. Observe that the conditions that need to be checked at each order correspond to all the ways of distributing bars on the various factors that appear in the

order conditions of Table 2. For example, the condition $\sum b_i c_i^3 = \frac{1}{4}$ in Table 2 will expand into the following set of relations:

$$\frac{1}{4} = \sum b_i c_i^3 = \sum b_i c_i^2 \bar{c}_i = \sum b_i c_i \bar{c}_i^2 = \sum b_i \bar{c}_i^3.$$

Altogether there are 26 separate conditions that must be satisfied to achieve fourth order accuracy, and besides the original 4 conditions in Table 2, 4 new conditions emerge in Table 1. We now check each of the 26 conditions associated with a fourth order scheme. There are 4 different terms that need to be checked, denoted 1, 2, 3, 4 below. The various ways of arranging the bars are denoted 1a, 1b, $\cdots$, 4g.

**1.** $\sum b_i c_i^3 = \frac{1}{4}$ (in Table 1)

**1a.** $\sum b_i c_i^2 \bar{c}_i = \sum c_i^2 (b_i - d_i) = \frac{1}{4}$

**1b.** $\sum b_i c_i \bar{c}_i^2 = \sum c_i (b_i - d_i)^2 / b_i = \sum c_i b_i - 2c_i d_i + c_i d_i^2 / b_i = \frac{1}{4}$

**1c.** $\sum b_i \bar{c}_i^3 = \sum (b_i - d_i)^3 / b_i^2 = \sum b_i - 3d_i + 3d_i^2 / b_i - d_i^3 / b_i^2 = \frac{1}{4}$

**2.** $\sum b_i c_i a_{ij} c_j = \frac{1}{8}$ (in Table 1)

**2a.** $\sum b_i \bar{c}_i a_{ij} c_j = \sum (b_i - d_i) a_{ij} c_j$
$= \sum d_j c_j - d_i a_{ij} c_j = \frac{1}{8}$

**2b.** $\sum b_i c_i \bar{a}_{ij} c_j = \sum c_i (b_i b_j - b_j a_{ji}) c_j$
$= (\sum c_i b_i)^2 - \sum b_i c_i a_{ij} c_j = \frac{1}{8}$

**2c.** $\sum b_i c_i a_{ij} \bar{c}_j = \sum b_i c_i a_{ij} (1 - d_j / b_j)$
$= \sum b_i c_i^2 - b_i c_i a_{ij} d_j / b_j = \frac{1}{8}$

**2d.** $\sum b_i \bar{c}_i \bar{a}_{ij} c_j = \sum (b_i - d_i) b_j (b_i - a_{ji}) c_j / b_i$
$= \sum (b_i - d_i) b_j c_j - b_j a_{ji} c_j + b_j d_i a_{ji} c_j / b_i$
$= \frac{1}{4} + \sum b_i c_i a_{ij} d_j / b_j - b_j c_j^2 = \frac{1}{8}$

**2e.** $\sum b_i \bar{c}_i a_{ij} \bar{c}_j = \sum (b_i - d_i) a_{ij} (b_j - d_j) / b_j$
$= \sum d_i - c_i d_i - d_i^2 / b_i + d_i a_{ij} d_j / b_j = \frac{1}{8}$

**2f.** $\sum b_i c_i \bar{a}_{ij} \bar{c}_j = \sum c_i (b_i - a_{ji}) (b_j - d_j)$
$= \sum b_i c_i - c_i d_i - b_i c_i d_j + d_i a_{ij} c_j = \frac{1}{8}$

**2g.** $\sum b_i \bar{c}_i \bar{a}_{ij} \bar{c}_j = (b_i - d_i)(b_i - a_{ji})(b_j - d_j) / b_i$
$= \sum (b_i - d_i)(b_j - d_j)$
$-(b_i - d_i) a_{ji} (b_j - d_j) / b_i$

$$= \sum b_i b_j - 2d_i + d_i d_j$$
$$\quad -a_{ji}(b_j - d_j) + d_i a_{ji}(b_j - d_j)/b_i$$
$$= \sum b_i b_j - 3d_i + d_i d_j$$
$$\quad +c_i d_i + d_i^2/b_i - d_i a_{ij} d_j/b_j = \tfrac{1}{8}$$

**3.** $\sum b_i a_{ij} c_j^2 = \frac{1}{12}$ (in Table 1)

**3a.** $\sum b_i \bar{a}_{ij} c_j^2 = \sum b_j (b_i - a_{ji}) c_j^2 = \sum b_j (c_j^2 - c_j^3) = \frac{1}{12}$

**3b.** $\sum b_i a_{ij} c_j \bar{c}_j = \sum b_i a_{ij} c_j (b_j - d_j)/b_j$
$$= \sum c_j d_j - c_j d_j^2/b_j = \tfrac{1}{12}$$

**3c.** $\sum b_i \bar{a}_{ij} c_j \bar{c}_j = \sum b_j \bar{a}_{ji} c_i \bar{c}_i = \sum (b_j - a_{ij}) c_i (b_i - d_i)$
$$= \sum b_i c_i - c_i d_i - b_i c_i^2 + c_i^2 d_i = \tfrac{1}{12}$$

**3d.** $\sum b_i a_{ij} \bar{c}_j^2 = \sum d_j (b_j - d_j)^2/b_j^2$
$$= \sum d_j - 2d_j^2/b_j + d_j^3/b_j^2 = \tfrac{1}{12}$$

**3e.** $\sum b_i \bar{a}_{ij} \bar{c}_j^2 = \sum b_j \bar{a}_{ji} \bar{c}_i^2 = \sum (b_j - a_{ij})(b_i - d_i)^2/b_i$
$$= \sum (1 - c_i)(b_i - d_i)^2/b_i$$
$$= \sum (1 - c_i)(b_i - 2d_i + d_i^2/b_i) = \tfrac{1}{12}$$

**4.** $\sum b_i a_{ij} a_{jk} c_k = \frac{1}{24}$ (in Table 1)

**4a.** $\sum b_i \bar{a}_{ij} a_{jk} c_k = \sum b_j (b_i - a_{ji}) a_{jk} c_k = \sum b_j (1 - c_j) a_{jk} c_k$
$$= \sum c_k d_k - b_j c_j a_{jk} c_k = \tfrac{1}{24}$$

**4b.** $\sum b_i a_{ij} \bar{a}_{jk} c_k = \sum d_j b_k (b_j - a_{kj}) c_k/b_j$
$$= d_j b_k c_k - d_j b_k a_{kj} c_k/b_j = \tfrac{1}{24}$$

**4c.** $\sum b_i a_{ij} a_{jk} \bar{c}_k = \sum d_j a_{jk} (b_k - d_k)/b_k$
$$= \sum c_j d_j - d_j a_{jk} d_k/b_k = \tfrac{1}{24}$$

**4d.** $\sum b_i \bar{a}_{ij} \bar{a}_{jk} c_k = \sum b_k (b_i - a_{ji})(b_j - a_{kj}) c_k$
$$= \sum b_k (1 - c_j)(b_j - a_{kj}) c_k$$
$$= \sum b_k c_k (b_j - a_{kj} - b_j c_j + c_j a_{kj})$$
$$= \sum \tfrac{1}{2} b_k c_k - b_k c_k^2 + b_k c_k a_{kj} c_j = \tfrac{1}{24}$$

**4e.** $\sum b_i \bar{a}_{ij} a_{jk} \bar{c}_k = \sum b_j (b_i - a_{ji}) a_{jk} (b_k - d_k)/b_k$

$$= \sum b_j a_{jk}(1-c_j)(b_k-d_k)/b_k$$
$$= \sum b_j a_{jk}(b_k-d_k-b_k c_j+c_j d_k)/b_k$$
$$= \sum d_k - d_k^2/b_k - b_j c_j^2 + b_j c_j a_{jk} d_k/b_k = \tfrac{1}{24}$$

**4f.** $\sum b_i a_{ij}\bar{a}_{jk}\bar{c}_k = \sum d_j(b_j-a_{kj})(b_k-d_k)/b_j$
$$= \sum d_j b_k - d_j d_k - d_j^2/b_j + d_k a_{kj} d_j/b_j = \tfrac{1}{24}$$

**4g.** $\sum b_i \bar{a}_{ij}\bar{a}_{jk}\bar{c}_k = (b_i - a_{ji})(b_j-a_{kj})(b_k-d_k)$
$$= (1-c_j)(b_j-a_{kj})(b_k-d_k)$$
$$= (b_j-a_{kj})(b_k-d_k) - c_j(b_j-a_{kj})(b_k-d_k)$$
$$= (\tfrac{1}{2}-c_k)(b_k-d_k) + a_{kj}c_j(b_k-d_k) = \tfrac{1}{24}$$

This completes the proof. □

## 5. Error estimate

Our proof of Theorem 2.1, as well as that of the constrained version in Sect. 7, are based on the following abstract result.

**Proposition 5.1.** *Let $\mathcal{X}$ be a Banach space and let $\mathcal{Y}$ be a linear normed space with the norm in both spaces denoted $\|\cdot\|$. Let $\mathcal{F} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ be a set-valued map, let $\mathcal{L} : \mathcal{X} \mapsto \mathcal{Y}$ be a bounded, linear operator, and let $\mathcal{T} : \mathcal{X} \mapsto \mathcal{Y}$ with $\mathcal{T}$ continuously Frechét differentiable in $B_r(w^*)$ for some $w^* \in \mathcal{X}$ and $r > 0$. Suppose that the following conditions hold for some $\delta \in \mathcal{Y}$ and scalars $\epsilon$, $\lambda$, and $\sigma > 0$:*

(P1) $\mathcal{T}(w^*) + \delta \in \mathcal{F}(w^*)$.
(P2) $\|\nabla\mathcal{T}(w) - \mathcal{L}\| \le \epsilon$ *for all* $w \in B_r(w^*)$.
(P3) *The map $(\mathcal{F}-\mathcal{L})^{-1}$ is single-valued and Lipschitz continuous in $B_\sigma(\pi)$, $\pi = (\mathcal{T}-\mathcal{L})(w^*)$, with Lipschitz constant $\lambda$.*

*If $\epsilon\lambda < 1$, $\epsilon r \le \sigma$, $\|\delta\| \le \sigma$, and $\|\delta\| \le (1-\lambda\epsilon)r/\lambda$, then there exists a unique $w \in B_r(w^*)$ such that $\mathcal{T}(w) \in \mathcal{F}(w)$. Moreover, we have the estimate*

$$\|w - w^*\| \le \frac{\lambda}{1-\lambda\epsilon}\|\delta\|. \tag{54}$$

*Proof.* This result is obtained from [22, Thm. 3.1] by identifying the set $\Pi$ of that theorem with the ball $B_\sigma(\pi)$. □

In applying Proposition 5.1, we utilize discrete analogues of various continuous spaces and norms. In particular, for a sequence $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_N$

whose $i$-th element is a vector $\mathbf{z}_i \in \mathbf{R}^n$, the discrete analogues of the $L^p$ and $L^\infty$ norms are the following:

$$\|\mathbf{z}\|_{L^p} = \left( \sum_{i=0}^{N} h|\mathbf{z}_i|^p \right)^p \quad \text{and} \quad \|\mathbf{z}\|_{L^\infty} = \sup_{0 \le i \le N} |\mathbf{z}_i|.$$

With this notation, the space $\mathcal{X}$ in the discrete control problem is the discrete $L^\infty$ space consisting of 3-tuples $w = (\mathbf{x}, \boldsymbol{\psi}, \mathbf{u})$ where

$$\begin{aligned} \mathbf{x} &= (\mathbf{a}, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N), \quad \mathbf{x}_k \in \mathbf{R}^n, \\ \boldsymbol{\psi} &= (\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_N), \quad \boldsymbol{\psi}_k \in \mathbf{R}^n, \\ \mathbf{u} &= (\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{N-1}), \quad \mathbf{u}_k \in \mathbf{R}^{sm}. \end{aligned}$$

The mappings $\mathcal{T}$ and $\mathcal{F}$ of proposition 5.1 are selected in the following way:

$$\mathcal{T}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = \begin{pmatrix} \mathbf{x}_k' - \mathbf{f}^h(\mathbf{x}_k, \mathbf{u}_k), \quad 0 \le k \le N-1 \\ \boldsymbol{\psi}_k' + \nabla_x H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k), \quad 0 \le k \le N-1 \\ \nabla_{u_j} H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k), \quad 1 \le j \le s, \ 0 \le k \le N-1 \\ \boldsymbol{\psi}_N - \nabla C(\mathbf{x}_N) \end{pmatrix} \tag{55}$$

and

$$\mathcal{F}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \Pi_{i=1}^s N_U(\mathbf{u}_{ki}), \quad 0 \le k \le N-1 \\ \mathbf{0} \end{pmatrix}.$$

The space $\mathcal{Y}$, associated with the four components of $\mathcal{T}$, is a space of 4-tuples of finite sequences in $L^1 \times L^1 \times L^\infty \times \mathbf{R}^n$. The reference point $w^*$ is the sequence with elements

$$w_k^* = (\mathbf{x}_k^*, \boldsymbol{\psi}_k^*, \mathbf{u}_k^*),$$

where $\mathbf{x}_k^* = \mathbf{x}^*(t_k)$, $\boldsymbol{\psi}_k^* = \boldsymbol{\psi}^*(t_k)$, and $\mathbf{u}_{ki}^* = \mathbf{u}(\mathbf{y}_i^*, \boldsymbol{\chi}_i^*)$. Here $\mathbf{y}_i^*$ and $\boldsymbol{\chi}_i^*$ are the solution to (40)–(41) corresponding to $\mathbf{x} = \mathbf{x}^*(t_k)$ and $\boldsymbol{\psi} = \boldsymbol{\psi}^*(t_k)$. Since $\mathbf{u}_{ki}^*$ is a solution to (35) associated with $\mathbf{x} = \mathbf{y}_i^*$ and $\boldsymbol{\psi} = \boldsymbol{\chi}_i^*$, we have

$$\mathbf{u}_{ki} \in U, \quad -\nabla_u H(\mathbf{y}_i, \boldsymbol{\chi}_i, \mathbf{u}_{ki}) \in N_U(\mathbf{u}_{ki}). \tag{56}$$

The operator $\mathcal{L}$ is gotten by linearizing around $w^*$, evaluating all variables on each interval at the grid point to the left, and dropping terms that vanish

at $h = 0$. In particular, we choose $\mathcal{L}(w) =$

$$= \begin{pmatrix} \mathbf{x}_k' - \mathbf{A}_k\mathbf{x}_k - \mathbf{B}_k\mathbf{u}_k\mathbf{b}, \quad 0 \le k \le N-1 \\ \boldsymbol{\psi}_k' + \boldsymbol{\psi}_{k+1}\mathbf{A}_k + (\mathbf{Q}_k\mathbf{x}_k + \mathbf{S}_k\mathbf{u}_k\mathbf{b})^{\mathsf{T}}, \quad 0 \le k \le N-1 \\ b_j(\mathbf{u}_{kj}^{\mathsf{T}}\mathbf{R}_k + \mathbf{x}_k^{\mathsf{T}}\mathbf{S}_k + \boldsymbol{\psi}_{k+1}\mathbf{B}_k), \quad 1 \le j \le s,\ 0 \le k \le N-1 \\ \boldsymbol{\psi}_N + \mathbf{V}\mathbf{x}_N \end{pmatrix}.$$

(57)

For these choices of the spaces and the functions, we now examine each of the hypotheses of Proposition 5.1. First, in [24, Lem. 5.1] we show that by Smoothness,

$$\|\nabla\mathcal{T}(w) - \mathcal{L}\| \le \|\nabla\mathcal{T}(w) - \mathcal{L}\|_{L^\infty} \le c(\|w - w^*\| + h) \tag{58}$$

for every $w \in B_\beta(w^*)$, where $\beta$ appears in the state uniqueness property. Moreover, by Smoothness, Coercivity, and [24, Lem. 6.1], the map $(\mathcal{F} - \mathcal{L})^{-1}$ is Lipschitz continuous with a Lipschitz constant $\lambda$ independent of $h$ for $h$ sufficiently small. Thus we can take $\sigma = \infty$ in Proposition 5.1.

To finish the analysis, we need an estimate for the distance from $\mathcal{T}(w^*)$ to $\mathcal{F}(w^*)$. In this section, we focus on the case where $U = \mathbf{R}^m$ and $\mathcal{F} = \mathbf{0}$, while Sect. 7 shows how the analysis must be changed to handle control constraints. When $\mathcal{F} = \mathbf{0}$, estimating the distance to $\mathcal{F}$ is equivalent to obtaining an estimate for $\|\mathcal{T}(w^*)\|$. By (4) we have $\boldsymbol{\psi}_N^* = \nabla C(\mathbf{x}_N^*)$. Consequently, the last component of $\mathcal{T}(w^*)$ vanishes. By the identities (29) and (56), the next-to-last component of $\mathcal{T}(w^*)$ also vanishes. The first two components of $\mathcal{T}(w^*)$ are estimated using the Taylor expansions of Sects. 4. Consistent with the notation of Sect. 4, we define

$$\mathbf{z}^* = \begin{pmatrix} \mathbf{x}^* \\ \boldsymbol{\psi}^* \end{pmatrix} \quad \text{and} \quad \mathbf{z}_k^* = \begin{pmatrix} \mathbf{x}^*(t_k) \\ \boldsymbol{\psi}^*(t_k) \end{pmatrix}.$$

Similarly, $\boldsymbol{\zeta}^*(h)$ is the vector whose first $s$ components are pairs $(\mathbf{y}_i^*(h), \boldsymbol{\chi}_i^*(h))$ satisfying (38) and (39) with $\mathbf{x}_k = \mathbf{x}_k^*$ and $\boldsymbol{\psi}_k = \boldsymbol{\psi}_k^*$, and whose last component is

$$\boldsymbol{\zeta}_{s+1}^*(h) = \mathbf{z}_k^* + h\sum_{i=1}^{s} b_i\mathbf{g}(\mathbf{y}_i^*(h), \boldsymbol{\chi}_i^*(h)).$$

Using this notation and exploiting the identity (28) of Proposition 3.3, the first two components of $\mathcal{T}(w^*)$, evaluated at time level $k$, can be expressed:

$$\begin{pmatrix} \mathbf{x}_k^{*\prime} - \mathbf{f}^h(\mathbf{x}_k^*, \mathbf{u}_k^*) \\ \boldsymbol{\psi}_k^{*\prime} + \nabla_x H^h(\mathbf{x}_k^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_k^*) \end{pmatrix} = \frac{1}{h}\left(\mathbf{z}_{k+1}^* - \boldsymbol{\zeta}_{s+1}(h)\right).$$

The order conditions of Table 1 were devised so that the terms in the Taylor expansions (44) for $\mathbf{z}^*_{k+1}$ and (48) for $\boldsymbol{\zeta}^*_{s+1}(h)$ match through order $\kappa$, leaving us with integral remainder terms:

$$\left|\mathbf{z}^*_{k+1} - \boldsymbol{\zeta}_{s+1}(h)\right| \le ch^{\kappa-1}\left( \int_{t_k}^{t_{k+1}} |\mathbf{z}^{*(\kappa)}(t) - \mathbf{z}^{*(\kappa)}(t_k)|\, dt \right.$$

$$\left. + \int_0^h |\boldsymbol{\zeta}^*_{s+1}{}^{(\kappa)}(t) - \boldsymbol{\zeta}^*_{s+1}{}^{(\kappa)}(0)|\, dt \right). \tag{59}$$

(Note that although we only assumed $\mathbf{z}^{*(\kappa)}$ lies in $L^\infty$, it follows from the smoothness properties of $\mathbf{f}$ and Control Uniqueness, that $\mathbf{z}^{*(\kappa)}$ is continuous when $U = \mathbf{R}^m$). By the chain rule, Smoothness, and Costate Uniqueness, the $\kappa$-derivative of $\mathbf{z}^*$ can be written

$$\mathbf{z}^{*(\kappa)}(t) = \mathbf{F}(t)\mathbf{u}^{*(\kappa-1)}(t) + \mathbf{H}(t),$$

where both $\mathbf{F}$ and $\mathbf{H}$ are Lipschitz continuous. Hence, for each $t \in [t_k, t_{k+1}]$, we have

$$\begin{aligned} |\mathbf{z}^{*(\kappa)}(t) - \mathbf{z}^{*(\kappa)}(t_k)| &\le |(\mathbf{F}(t_k) - \mathbf{F}(t))\mathbf{u}^{*(\kappa-1)}(t)| \\ &\quad + |\mathbf{F}(t_k)(\mathbf{u}^{*(\kappa-1)}(t) - \mathbf{u}^{*(\kappa-1)}(t_k))| + |\mathbf{H}(t_k) - \mathbf{H}(t)| \\ &\le c(h + \omega(\mathbf{u}^{*(\kappa-1)}, [t_k, t_{k+1}]; t, 2h)). \end{aligned}$$

After summing over $k$, we have

$$\begin{aligned} &\sum h \int_{t_k}^{t_{k+1}} |\mathbf{z}^{*(\kappa)}(t) - \mathbf{z}^{*(\kappa)}(t_k)|\, dt \\ &\quad \le ch\left( h + \sum \int_{t_k}^{t_{k+1}} \omega(\mathbf{u}^{*(\kappa-1)}, [t_k, t_{k+1}]; t, 2h)]\, dt \right) \\ &\quad \le ch(h + \tau(\mathbf{u}^{*(\kappa-1)}; 2h)) \\ &\quad \le ch(h + 2\tau(\mathbf{u}^{*(\kappa-1)}; h)), \end{aligned}$$

where the last inequality is found, for example, in [48, p. 11].

Now consider the last term in (59). By Costate Uniqueness, we know that the equation

$$\boldsymbol{\zeta}(h) = \boldsymbol{\zeta}(0) + h\mathbf{G}(\boldsymbol{\zeta}(h)) \tag{60}$$

has a locally unique solution whenever $h \le \gamma$. Moreover, by the implicit function theorem, $\boldsymbol{\zeta}(h)$ is $\kappa - 1$ times Lipschitz continuously differentiable

in $h$. Hence, the $\kappa$-th derivative of $\boldsymbol{\zeta}(h)$ lies in $L^\infty$. To see more precisely the structure of the $\kappa$-th derivative, we differentiate the identity (60) to obtain

$$\begin{aligned}\boldsymbol{\zeta}^{(\kappa)}(h) &= (\boldsymbol{\zeta}(0) + h\mathbf{G}(\boldsymbol{\zeta}(h)))^{(\kappa)} \\ &= \mathbf{G}(\boldsymbol{\zeta}(h))^{(\kappa-1)} + h\mathbf{G}(\boldsymbol{\zeta}(h))^{(\kappa)} \\ &= \bar{\mathbf{G}}(\boldsymbol{\zeta}(h),\ \cdots,\boldsymbol{\zeta}^{(\kappa-1)}(h)) + h\mathbf{G}(\boldsymbol{\zeta}(h))^{(\kappa)}, \end{aligned} \tag{61}$$

where $\bar{\mathbf{G}}$ is a function that involves various products of derivatives of $\mathbf{G}$ to order $\kappa - 1$. The last term in (61) is $O(h)$ since the $\kappa$-th derivatives are all bounded. Since the derivative of $\mathbf{G}$ to order $\kappa - 1$ are all Lipschitz continuous, it follows that

$$\begin{aligned}&|\boldsymbol{\zeta}^{(\kappa)}(h) - \boldsymbol{\zeta}^{(\kappa)}(0)| \\ &= O(h) + |\bar{\mathbf{G}}((\boldsymbol{\zeta}(h),\ \cdots,\boldsymbol{\zeta}^{(\kappa-1)}(h)) - \bar{\mathbf{G}}(\boldsymbol{\zeta}(0),\ \cdots,\boldsymbol{\zeta}^{(\kappa-1)}(0))| \\ &\le O(h) + c\sum_{i=0}^{\kappa-1} |\boldsymbol{\zeta}^{(i)}(h) - \boldsymbol{\zeta}^{(i)}(0)| = O(h).\end{aligned}$$

Hence, the last term in (59) is $O(h)$. To summarize, in $L^1 \times L^1 \times L^\infty \times \mathbf{R}^n$, we have

$$\begin{aligned}&\|\mathcal{T}(\mathbf{w}^*)\| \\ &= h\sum \left(|\mathbf{x}_k^{*\prime} - \mathbf{f}^h(\mathbf{x}_k^*, \mathbf{u}_k^*)| + |\boldsymbol{\psi}_k^{*\prime} + \nabla_x H^h(\mathbf{x}_k^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_k^*)|\right) \\ &\le ch^{\kappa-1}\left(h + \tau(\mathbf{u}^{*(\kappa-1)}; h)\right). \end{aligned} \tag{62}$$

To complete the proof of Theorem 2.1, using Proposition 5.1, let $\lambda$ be chosen large enough and let $\bar{h}$ be chosen small enough that the Lipschitz constant of $\mathcal{L}^{-1}$ is less than $\lambda$ for all $h \le \bar{h}$. Choose $\epsilon$ small enough that $\epsilon\lambda < 1$. Choose a small $r$ and choose $\bar{h}$ smaller if necessary so that $c(r+\bar{h}) \le \epsilon$ where $c$ is the constant appearing in (58). Finally, choose $\bar{h}$ smaller if necessary so that for the residual bound in (62), we have

$$c\bar{h}^{\kappa-1}\left(\bar{h} + \tau(\mathbf{u}^{*(\kappa-1)}; \bar{h})\right) \le (1 - \lambda\epsilon)r/\lambda.$$

Since the hypotheses of Proposition 5.1 are satisfied, we conclude that for each $h \le \bar{h}$, there exists $w^h = (\mathbf{x}^h, \boldsymbol{\psi}^h, \mathbf{u}^h) \in B_r(w^*)$ such that $\mathcal{T}(w^h) = 0$ and the estimate (54) holds, which establishes the bounds for the state and costate variables in (14). The estimate in (14) for the error in the control follows from the control uniqueness property and the fact that $\nabla_u H(\mathbf{x}^*(t_k), \boldsymbol{\psi}^*(t_k), \mathbf{u}^*(t_k)) = \mathbf{0}$. Finally, by [24, Lem. 7.2] $(\mathbf{x}^h, \mathbf{u}^h)$ is a strict local minimizer in (8) for $h$ sufficiently small.

## 6. Numerical illustrations

Through second order, the conditions in Tables 1 and 2 are the same, and at order three, one new condition emerges for the control problem. In [8, p. 174] Butcher shows that the set of third-order explicit Runge-Kutta schemes includes the following family involving the two parameters $c_2 \neq \frac{2}{3}, 0$, and $c_3 \neq c_2, 0$:

$$(63)\ \mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ c_2 & 0 & 0 \\ \frac{c_3(3c_2-c_3-3c_2^2)}{c_2(2-3c_2)} & \frac{c_3(c_3-c_2)}{c_2(2-3c_2)} & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \frac{2c_2c_3-c_3-c_2+2/3}{2c_2c_3} \\ \frac{3c_3-2}{6c_2(c_3-c_2)} \\ \frac{2-3c_2}{6c_3(c_3-c_2)} \end{bmatrix}.$$

There are also two one-parameter families of schemes when $c_2 = \frac{2}{3}$, however, it can be shown that neither of these families satisfies the condition

$$\sum d_i^2/b_i = 1/3 \tag{64}$$

of Table 1 needed for third-order accuracy in optimal control. Moreover, for the two-parameter family (63), the condition (64) is satisfied if and only if $c_3 = 1$ (a symbolic manipulation package like Maple facilitates the derivation of this result).

The following specific third-order schemes have appeared in the literature (for example, see [38, p. 402] and [39, p. 506]):

$$\text{(a)}\ \mathbf{A} = \begin{bmatrix} 0\ 0\ 0 \\ \frac{1}{2}\ 0\ 0 \\ -1\ 2\ 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \frac{1}{6} \\ \frac{2}{3} \\ \frac{1}{6} \end{bmatrix}, \quad \text{(b)}\ \mathbf{A} = \begin{bmatrix} 0\ 0\ 0 \\ \frac{1}{2}\ 0\ 0 \\ 0\ \frac{3}{4}\ 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \frac{2}{9} \\ \frac{1}{3} \\ \frac{4}{9} \end{bmatrix}.$$

The scheme (a) corresponds to $c_2 = 1/2$ and $c_3 = 1$ in (63), while (b) corresponds to $c_2 = 1/2$ and $c_3 = 3/4$. The first scheme satisfies (64) since $c_3 = 1$ while the second scheme does not satisfy (64). Let us consider the following simple test problem [31, (P1)]:

$$\text{minimize}\ \ \frac{1}{2}\int_0^1 u(t)^2 + 2x(t)^2\ dt \tag{65}$$

$$\text{subject to}\ \ x'(t) = .5x(t) + u(t), \quad x(0) = 1,$$

with the optimal solution

$$x^*(t) = \frac{2e^{3t} + e^3}{e^{3t/2}(2 + e^3)}, \quad u^*(t) = \frac{2(e^{3t} - e^3)}{e^{3t/2}(2 + e^3)}. \tag{66}$$

**Table 3.** Discrete state error in $L^\infty$ for problem (65) and the schemes (a) and (b)

| $N$ | (a) | (b) |
|---|---|---|
| 10 | 8.820781e−05 | 7.236809e−04 |
| 20 | 9.716458e−06 | 1.732318e−04 |
| 40 | 1.110740e−06 | 4.231934e−05 |
| 80 | 1.317159e−07 | 1.045581e−05 |
| 160 | 1.600043e−08 | 2.598415e−06 |
| 320 | 1.970437e−09 | 6.476597e−07 |

In Table 3 we give the $L^\infty$ error for the discrete state at the grid points for the schemes (a) and (b) and various choices of the mesh. When we perform a least squares fit of the errors in Table 3 to a function of the form $ch^q$, we obtain $q \approx 3.09$ for (a) and $q \approx 2.02$ for (b). The errors observed in this example are typical for Runge-Kutta discretizations of the form (63). If $c_3 = 1$ and condition (64) holds, then the control discretization is third-order accurate, and if $c_3 \neq 1$ so that (64) is violated, then the control discretization is second-order accurate.

In Theorem 2.1, we require that $b_i > 0$ for each $i$. If $b_i$ vanishes, then the solution of the discrete problem may not converge to the solution of the continuous problem, as the following discretization of (65) illustrates:

$$
\begin{aligned}
(67) \qquad \text{minimize} \quad & \frac{h}{2}\sum_{k=0}^{N-1} u_{k+1/2}^2 + 2x_{k+1/2}^2 \\
\text{subject to} \quad & x_{k+1/2} = x_k + \tfrac{h}{2}(.5x_k + u_k), \\
& x_{k+1} \quad = x_k + h(.5x_{k+1/2} + u_{k+1/2}), \quad x_0 = 1.
\end{aligned}
$$

This scheme is second-order accurate for differential equations. The first stage of the Runge-Kutta scheme approximates $x$ at the midpoint of the interval $[kh, (k+1)h]$, and the second stage gives a second-order approximation to $x((k+1)h)$. Obviously, zero is a lower bound for the cost function. A discrete control that achieves this lower bound is $u_k = -\frac{4+h}{2h}x_k$ and $u_{k+1/2} = 0$ for each $k$, in which case $x_{k+1/2} = 0$ and $x_k = 1$ for each $k$. This optimal discrete control oscillates back and forth between 0 and a value around $-2/h$; hence the solution to the discrete problem diverges from the solution (66) to the continuous problem as $h$ tends to zero. In [24] we show that this divergent scheme can be fixed by replacing the control $u_k$ in the first stage by $u_{k+1/2}$.

Next, we illustrate the observation contained in Remark 2.2: For third or fourth-order Runge-Kutta schemes, the discrete controls $u_{kj}^h$ often converge to the continuous solution more slowly than $\mathbf{u}(\mathbf{x}_k^h, \boldsymbol{\psi}_k^h)$. To see this property,

**Table 4.** Discrete control errors in $L^\infty$ for test problem (68) and scheme (a)

| $N$ | $\mathbf{u}^h_{k1}$ | $\mathbf{u}^h_{k2}$ | $\mathbf{u}^h_{k3}$ | $\mathbf{u}(\mathbf{x}^h_k, \boldsymbol{\psi}^h_k)$ |
|---|---|---|---|---|
| 10 | 2.581245e−03 | 1.285116e−03 | 2.639595e−03 | 1.933271e−05 |
| 20 | 7.999243e−04 | 3.605417e−04 | 6.481638e−04 | 2.699320e−06 |
| 40 | 2.191605e−04 | 9.455063e−05 | 1.594966e−04 | 3.569218e−07 |
| 80 | 5.715833e−05 | 2.415251e−05 | 3.948989e−05 | 4.589058e−08 |
| 160 | 1.458317e−05 | 6.099989e−06 | 9.820382e−06 | 5.817758e−09 |
| 320 | 3.682317e−06 | 1.532569e−06 | 2.448334e−06 | 7.323739e−10 |

we need a slightly more complicated example than (65). We consider the following quadratic problem [31, (P2)] which includes an $xu$ term:

$$\text{(68)} \qquad \text{minimize } \tfrac{1}{2}\int_0^1 u(t)^2 + x(t)u(t) + \tfrac{5}{4}x(t)^2\, dt$$

$$\text{subject to } x'(t) = .5x(t) + u(t), \quad x(0) = 1,$$

with the optimal solution

$$x^*(t) = \frac{\cosh(1-t)}{\cosh(1)}, \quad u^*(t) = -\frac{(\tanh(1-t) + .5)\cosh(1-t)}{\cosh(1)}.$$

In Table 4 we give the $L^\infty$ error in the discrete controls $\mathbf{u}^h_{ki}$, $1 \le i \le 3$, for scheme (a) and problem (68), while the last column gives the error in $\mathbf{u}(\mathbf{x}^h_k, \boldsymbol{\psi}^h_k)$. Note that the errors in the last column of Table 4 are much smaller than the errors in the preceding columns. The error in each of the discrete controls is $O(h^2)$ while the error in the approximation $\mathbf{u}(\mathbf{x}^h_k, \boldsymbol{\psi}^h_k)$ generated by the discrete state and costate variables is $O(h^3)$, in accordance with Theorem 2.1. More precisely, if we perform a least squares fit of the errors in Table 4 to a function of the form $ch^q$, we obtain $q \approx 1.90$, 1.95, 2.01, and 2.94 for the respective columns of Table 4.

For 3-stage explicit third-order Runge-Kutta schemes, there are 6 nonzero coefficients to be specified: $a_{21}$, $a_{31}$, $a_{32}$, $b_1$, $b_2$, and $b_3$. In Table 1, there are 5 conditions to be satisfied in order to achieve third-order accuracy. Hence, we might anticipate a one-parameter family satisfying these 5 conditions. This family of solutions corresponds to (63) and $c_3 = 1$. Proceeding to 4-stage explicit fourth-order Runge-Kutta schemes, there are 10 nonzero coefficients to be specified and 13 conditions in Table 1 to be satisfied. Hence, by the same reasoning used for third-order schemes, one may think that a 4-stage explicit fourth-order method is impossible in optimal control. Quite to the contrary, we have

**Proposition 6.1.** *Every 4-stage explicit Runge-Kutta scheme with $b_i > 0$ for every $i$ that satisfies all the conditions of Table 2 also satisfies all the conditions of Table 1.*

In other words, any 4-stage explicit Runge-Kutta scheme with $b_i > 0$ for every $i$ that is fourth-order accurate for differential equations is also fourth-order accurate for optimal control.

*Proof.* In [8, p. 178] it is shown that in any 4-stage explicit fourth-order Runge-Kutta scheme, the following identity holds:

$$\sum_i b_i a_{ij} = b_j(1 - c_j),$$

$j = 1, 2, 3, 4$. Thus $d_j = b_j(1 - c_j)$ for each $j$. With this substitution, each of the 5 conditions in Table 1, not appearing in Table 2, can be deduced directly from the conditions in Table 2. □

For a practical illustration, we consider the orbit transfer problem presented in [6, pp. 66–68]. Given a constant-thrust rocket engine with thrust $T$ operating for a given length of time $t_f$, we wish to find the thrust-direction history $\phi(t)$ that transfers a spacecraft from a given initial circular orbit to the largest possible circular orbit. The notation is the following:

$r$ = radial distance of spacecraft from attracting center
$u$ = radial component of velocity
$v$ = tangential component of velocity
$m_0$ = initial mass of spacecraft
$\dot{m}$ = fuel consumption rate (assumed constant)
$\phi$ = thrust direction angle
$\mu$ = gravitational constant of attracting center

The problem of maximizing the radius of the final orbit can be expressed:

$$\begin{aligned}
&\text{maximize } r(t_f)\\
&\text{subject to } r' = u, && r(0) = r_0,\\
&\qquad u' = \frac{v^2}{r} - \frac{\mu}{r^2} + \frac{T\sin\phi}{m_0 - |\dot{m}|t}, && u(0) = 0, \quad u(t_f) = 0,\\
&\qquad v' = -\frac{uv}{r} + \frac{T\cos\phi}{m_0 - |\dot{m}|t}, && v(0) = \sqrt{\frac{\mu}{r_0}},\ v(t_f) = \sqrt{\frac{\mu}{r(t_f)}}.
\end{aligned}$$

We have solved the following instance of this problem stated in [6] (also see [40]): $m_0 = 10,000$ kg, $\dot{m} = 12.9$ kg/day, $r_0 = 149.6 \times 10^9$ m (distance from Sun to Earth), $T = 8.336$ N, $\mu = 1.3273310^{20}\text{m}^3/\text{s}^2$ (gravitational constant for the Sun), and $t_f = 193$ days. The trajectory, appearing in Fig. 1, takes the spacecraft from an Earth orbit around the Sun to a Mars orbit.

The terminal constraints on $u$ and $v$ at $t_f$ were treated using penalty/multiplier techniques (see [2] and [34]). We discretized the problem using the 3-stage methods (a) and (b). To estimate the errors associated with each

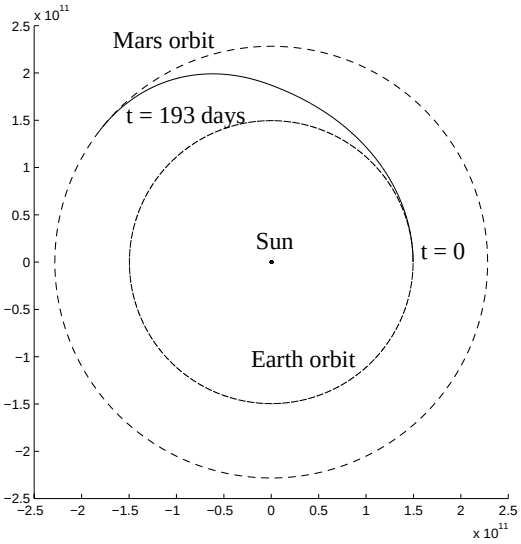


**Fig. 1.** Transfer spacecraft from Earth orbit to Mars orbit.

**Table 5.** Discrete state errors for orbit transfer problem and schemes (a) and (b)

| | (a) | (b) | (a) | (b) |
|---|---|---|---|---|
| $N$ | $L^\infty$ error | $L^\infty$ error | $L^2$ error | $L^2$ error |
| 500 ($r$) | 1.2e+03 | 2.3e+05 | 2.3e+06 | 6.4e+08 |
| 1000 ($r$) | 1.5e+02 | 5.8e+04 | 2.8e+05 | 1.6e+08 |
| 2000 ($r$) | 1.9e+01 | 1.4e+04 | 3.5e+04 | 4.0e+07 |
| | | | | |
| 500 ($u$) | 6.9e−04 | 2.2e−01 | 5.1e−01 | 1.3e+02 |
| 1000 ($u$) | 8.5e−05 | 5.6e−02 | 6.4e−02 | 3.2e+01 |
| 2000 ($u$) | 1.0e−05 | 1.4e−02 | 7.9e−03 | 7.9e+00 |
| | | | | |
| 500 ($v$) | 7.6e−04 | 1.3e−01 | 9.8e−01 | 1.0e+02 |
| 1000 ($v$) | 9.4e−05 | 3.2e−02 | 1.2e−01 | 2.6e+01 |
| 2000 ($v$) | 1.2e−05 | 8.1e−03 | 1.5e−02 | 6.4e+00 |

discretization, solutions were obtained for three meshes corresponding to $N = 500$, 1000, and 2000, and Aitken's extrapolation was used to estimate the exact solution at each grid point corresponding to the $N = 500$ mesh. The error in the discrete approximations in both the $L^2$ and $L^\infty$ norms appears in Table 5. Observe that the discrete errors are several orders of magnitude smaller for scheme (a) compared to scheme (b).

The orbit transfer problem was solved using a software package called optcon_xrk. To apply this collection of Fortran programs, the user provides

the coefficients of an explicit Runge-Kutta scheme along with subroutines to evaluate the right side of the differential equation, the terminal cost, and their gradients. The optimization is performed using steepest descent followed by the conjugate gradient method, where the transformed adjoint system (23)–(24) and the formula (27) are used to compute the gradient of the cost function. The software package optcon_xrk, along with a sample program based on the orbit transfer problem, is available at the following web site:

http://www.math.ufl.edu/~hager

## 7. Control constraints

When control constraint are present, adjustments are needed in the way we estimate the distance from $\mathcal{T}(w^*)$ to $\mathcal{F}(w^*)$ since $\mathbf{g}$ is often at best Lipschitz continuous when control constraints are present. All the other analysis in Sect. 5 remains unchanged in the control constrained case. Let us define the following quantities:

$$\mathbf{x}^*_{ki} = \mathbf{x}^*(t_k + c_i h), \quad \boldsymbol{\psi}^*_{ki} = \boldsymbol{\psi}^*(t_k + c_i h), \quad \bar{\boldsymbol{\psi}}^*_{ki} = \boldsymbol{\psi}^*(t_k + \bar{c}_i h).$$

Also, we set $\mathbf{z}^*_{ki} = (\mathbf{x}^*_{ki}, \boldsymbol{\psi}^*_{ki})$ and $\bar{\mathbf{z}}^*_{ki} = (\mathbf{x}^*_{ki}, \bar{\boldsymbol{\psi}}^*_{ki})$. By (60), we have $\boldsymbol{\zeta}_i(h) = \mathbf{z}^*_k + O(h)$. Consequently, the Lipschitz continuity of $\mathbf{g}$ yields

$$\begin{aligned}\boldsymbol{\zeta}_i(h) &= \mathbf{z}^*_k + h\mathbf{G}_i(\boldsymbol{\zeta}(h)) = \mathbf{z}^*_k + h\sum_{j=1}^{s} a_{ij}\mathbf{f}_0(\boldsymbol{\zeta}_j(h)) + h\sum_{j=1}^{s} \bar{a}_{ij}\boldsymbol{\phi}_0(\boldsymbol{\zeta}_j(h)) \\ &= \mathbf{z}^*_k + hc_i\mathbf{f}_0(\mathbf{z}^*_k) + h\bar{c}_i\boldsymbol{\phi}_0(\mathbf{z}_k) + O(h^2) = \bar{\mathbf{z}}^*_{ki} + O(h^2).\end{aligned}$$

With this substitution for $\boldsymbol{\zeta}_i(h)$, the Lipschitz continuity of $\mathbf{g}$ also yields:

$$\boldsymbol{\zeta}_{s+1}(h) = \mathbf{z}^*_k + h\sum b_i\mathbf{g}(\boldsymbol{\zeta}_i(h)) = \mathbf{z}^*_k + h\sum b_i\mathbf{g}(\bar{\mathbf{z}}^*_{ki}) + O(h^3).$$

On the the other hand, the fundamental theorem of calculus gives

$$\mathbf{z}^*_{k+1} = \mathbf{z}^*_k + \int_{t_k}^{t_{k+1}} \mathbf{g}(\mathbf{z}^*(t))\, dt.$$

Combining these last two identities, we obtain

$$\mathbf{z}^*_{k+1} - \boldsymbol{\zeta}_{s+1}(h) = \int_{t_k}^{t_{k+1}} \mathbf{g}(\mathbf{z}^*(t))\, dt - h\sum b_i\mathbf{g}(\bar{\mathbf{z}}^*_{ki}) + O(h^3). \tag{69}$$

In the case that $c_i = \bar{c}_i$ for each $i$, the difference (69) can be estimated by the following formula for the error in quadrature ([48, Thm. 3.4]):

**Proposition 7.1.** *For any* $\mathbf{b}$ *and* $\boldsymbol{\sigma} \in \mathbf{R}^s$ *such that*

$$\sum_{i=1}^{s} b_i = 1, \quad \sum_{i=1}^{s} b_i \sigma_i = \frac{1}{2}, \quad \textit{and} \quad 0 \le \sigma_i \le 1, \quad 1 \le i \le s,$$

*and for all* $\phi \in W^{1,\infty}$*, we have*

$$\left| \int_0^h \phi(s)\, ds - h \sum_{i=1}^{s} b_i \phi(\sigma_i h) \right| \le ch \int_0^h \omega(\dot{\phi}, [0,h]; s, h)\, ds,$$

*where* $\omega$ *is the modulus of continuity defined in* (13)*. Here c depends on the choice of* $\mathbf{b}$ *and* $\boldsymbol{\sigma}$*, but not on* $\phi$ *or* $h$*.*

Suppose that the Runge-Kutta scheme is at least second-order accurate, and that $0 \le c_i \le 1$. If $c_i = \bar{c}_i$ for each $i$, then $\mathbf{z}^*_{ki} = \bar{\mathbf{z}}^*_{ki}$ for each $i$, and applying Proposition 7.1 with $\sigma_i = c_i$, we have

$$\left|\mathbf{z}^*_{k+1} - \boldsymbol{\zeta}_{s+1}(h)\right| = \left| \int_{t_k}^{t_{k+1}} \mathbf{g}(\mathbf{z}^*(t))\, dt \right.$$

$$\left. - h \sum b_i \mathbf{g}(\mathbf{z}^*_{ki}) \right| + O(h^3) \tag{70}$$

$$\le ch \int_{t_k}^{t_{k+1}} \omega(\mathbf{g}(\mathbf{z}^*)', [t_k, t_{k+1}]; t, h)\, dt + O(h^3). \tag{71}$$

The function $\mathbf{g}$, evaluated at $\mathbf{z}^*$, has the following special form:

$$\mathbf{g}(\mathbf{z}^*) = \begin{pmatrix} \mathbf{f}(\mathbf{x}^*, \mathbf{u}(\mathbf{x}^*, \boldsymbol{\psi}^*)) \\ \nabla_x H(\mathbf{x}^*, \boldsymbol{\psi}^*, \mathbf{u}(\mathbf{x}^*, \boldsymbol{\psi}^*)) \end{pmatrix} = \begin{pmatrix} \mathbf{f}(\mathbf{x}^*, \mathbf{u}^*) \\ \nabla_x H(\mathbf{x}^*, \boldsymbol{\psi}^*, \mathbf{u}^*) \end{pmatrix}.$$

For any $t \in [t_k, t_{k+1}]$, Smoothness yields

$$\omega(\mathbf{g}(\mathbf{z}^*)', [t_k, t_{k+1}]; t, h) \le \omega(\mathbf{u}^{*\prime}, [t_k, t_{k+1}]; t, h)\, dt + O(h).$$

Hence, (71) yields

$$\left|\mathbf{z}^*_{k+1} - \boldsymbol{\zeta}_{s+1}(h)\right| \le ch \int_{t_k}^{t_{k+1}} \omega(\mathbf{u}^{*\prime}, [t_k, t_{k+1}]; t, h)\, dt + O(h^3). \tag{72}$$

After multiplying this inequality by $h$ and summing over $k$, we again obtain (62) in the case $\kappa = 2$. This shows that if $c_i = \bar{c}_i$ for each $i$, then Theorem 2.1 is valid in the control constrained case.

If $c_i \ne \bar{c}_i$ for some $i$, then $\bar{\mathbf{z}}^*_i$ may not equal $\mathbf{z}^*_i$. In this case, we write the difference (69) in the following way:

$$\mathbf{z}^*_{k+1} - \boldsymbol{\zeta}_{s+1}(h) = \left( \int_{t_k}^{t_{k+1}} \mathbf{g}(\mathbf{z}^*(t))\, dt - \sum b_i \mathbf{g}(\mathbf{z}^*_{ki}) \right)$$

$$+ \sum b_i (\mathbf{g}(\mathbf{z}^*_{ki}) - \mathbf{g}(\bar{\mathbf{z}}^*_{ki})) + O(h^3). \tag{73}$$

The term in (73) involving the integral is again bounded by the expression on the right side of (72). For the second term, we use fundamental theorem of calculus to estimate the difference $\mathbf{g}(\mathbf{z}^*_{ki}) - \mathbf{g}(\bar{\mathbf{z}}^*_{ki})$. Since $\mathbf{g}$ is the sum of two terms $\mathbf{f}_0$ and $\phi_0$, and both terms can be analyzed in similar ways, we focus on the $\mathbf{f}_0$ term:

$$\mathbf{f}(\mathbf{z}^*_{ki}) - \mathbf{f}(\bar{\mathbf{z}}^*_{ki}) = \mathbf{f}(\mathbf{x}^*_{ki}, \mathbf{u}(\mathbf{z}^*_{ki})) - \mathbf{f}(\mathbf{x}^*_{ki}, \mathbf{u}(\bar{\mathbf{z}}^*_{ki})) = \mathbf{F}^i_u(\mathbf{u}(\mathbf{z}^*_{ki}) - \mathbf{u}(\bar{\mathbf{z}}^*_{ki})),$$

where $\mathbf{F}^i_u$ is the average of the $u$-gradient of $\mathbf{f}$ evaluated along a line segment connecting $\mathbf{u}(\mathbf{z}^*_i)$ and $\mathbf{u}(\bar{\mathbf{z}}^*_i)$. Since the control function $\mathbf{u}(\mathbf{x}, \boldsymbol{\psi})$ is a Lipschitz continuous function of its arguments, we have

$$\mathbf{F}^i_u(\mathbf{u}(\mathbf{z}^*_{ki}) - \mathbf{u}(\bar{\mathbf{z}}^*_{ki})) = \mathbf{F}_{uk}(\mathbf{u}(\mathbf{z}^*_{ki}) - \mathbf{u}(\bar{\mathbf{z}}^*_{ki})) + O(h^2)$$

where $\mathbf{F}_{uk} = \nabla_u \mathbf{f}(\mathbf{z}^*_k)$. By the Lipschitz continuity of $\mathbf{u}$, we can write

$$\mathbf{u}(\mathbf{z}^*_{ki}) - \mathbf{u}(\bar{\mathbf{z}}^*_{ki}) = \int_{t_k + \bar{c}_i h}^{t_k + c_i h} \frac{d}{dt}\mathbf{u}(\mathbf{x}^*_{ki}, \boldsymbol{\psi}^*(t))\, dt. \tag{74}$$

To bound this term, we need to utilize a modulus of continuity $\hat{\omega}$ for a function of two variables $\mathbf{v}(s,t)$ defined in the following way:

$$\begin{aligned}\hat{\omega}(\mathbf{v}, J; t, h) = \sup\{&|\mathbf{v}(s_1, t_1) \\ &-\mathbf{v}(s_2, t_2)| : s_1, s_2, t_1, t_2 \in [t - h/2, t + h/2] \cap J\}.\end{aligned}$$

The identity $\sum b_i c_i = \sum b_i \bar{c}_i$ implies that

$$\sum b_i \int_{t_k + c_i h}^{t_k + \bar{c}_i h} 1\, dt = 0.$$

Utilizing this relation, we can subtract any fixed value for $\mathbf{v}$ under the integral in (74). With the choice

$$\mathbf{v}^*(s,t) = \frac{d}{dt}\mathbf{u}(\mathbf{x}^*(s), \boldsymbol{\psi}^*(t)), \tag{75}$$

we have

$$\left|\sum b_i\big(\mathbf{u}(\mathbf{z}^*_i) - \mathbf{u}(\bar{\mathbf{z}}^*_i)\big)\right| \le c \int_{t_k}^{t_{k+1}} \hat{\omega}(\mathbf{v}^*, [t_k, t_{k+1}]; t, h)\, dt.$$

This bound for the second term in (73) coupled with the bound (72) for the first term leads to the following analogue of Theorem 2.1 in the case of control constraints:

**Theorem 7.2.** *If Coercivity and Smoothness with $\kappa = 2$ hold, $b_i > 0$ and $0 \le c_i \le 1$ for each $i$, and the Runge-Kutta scheme is second-order accurate, then for all sufficiently small h, there exists a strict local minimizer*

*$(\mathbf{x}^h, \mathbf{u}^h)$ of the discrete optimal control problem $(8)$ and an associated adjoint variable $\boldsymbol{\psi}^h$ satisfying $(11)$ and $(12)$ such that*

$$\max_{0\le k\le N} |\mathbf{x}_k^h - \mathbf{x}^*(t_k)| + |\boldsymbol{\psi}_k^h - \boldsymbol{\psi}^*(t_k)| + |\mathbf{u}(\mathbf{x}_k^h, \boldsymbol{\psi}_k^h) - \mathbf{u}^*(t_k)|$$

$$\le \quad ch\left(h + \tau(\frac{d}{dt}\mathbf{u}^*; h) + \hat{\tau}(\mathbf{v}^*; h)\right), \tag{76}$$

*where $\mathbf{v}^*$ is defined in $(75)$, and $\hat{\tau}$ is given by*

$$\hat{\tau}(\mathbf{v}^*; h) = \int_0^1 \hat{\omega}(\mathbf{v}^*, [0,1]; t, h)\, dt.$$

*In the case that $c_i = \bar{c}_i$ for each $i$, or equivalently $d_i = b_i(1 - c_i)$, the $\hat{\tau}$ term in $(76)$ can be dropped.*

With regard to the condition $d_i = b_i(1-c_i)$ of Theorem 7.2, we noted in the proof of Proposition 6.1 that this is satisfied by every 4-stage explicit fourth-order Runge-Kutta scheme for differential equations. Also, it can be shown, using (63) with $c_3 = 1$, that this holds for any 3-stage explicit Runge-Kutta scheme *that is third-order accurate for optimal control*.

## References


1. T. M. Apostol, Mathematical Analysis, 2nd ed. Addison-Wesley, Reading, MA, 1974
2. D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods. Academic Press, New York, 1982
3. D. P. Bertsekas, Nonlinear Programming. Athena Scientific, Belmont, MA, 1995
4. W. E. Bosarge, Jr., O. G. Johnson, Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximations. SIAM J. Control **9**, 15–28 (1971)
5. W. E. Bosarge, Jr., O. G. Johnson, R. S. McKnight, W. P. Timlake, The Ritz-Galerkin procedure for nonlinear control problems. SIAM J. Numer. Anal. **10**, 94–111 (1973)
6. A. E. Bryson, Jr., Y.-C. Ho, Applied Optimal Control. Blaisdell, Waltham, MA, 1969
7. B. M. Budak, E. M. Berkovich, E. N. Solov'eva, Difference approximations in optimal control problems. SIAM J. Control **7**, 18–31 (1969)
8. J. C. Butcher, The Numerical Analysis of Ordinary Differential Equations. John Wiley, New York, 1987
9. J. Cullum, Discrete approximations to continuous optimal control problems. SIAM J. Control **7**, 32–49 (1969)
10. J. Cullum, An explicit procedure for discretizing continuous, optimal control problems. J. Optimization Theory Appl. **8**, 15–34 (1971)
11. J. Cullum, Finite-dimensional approximations of state-constrained continuous optimal control problems. SIAM J. Control **10**, 649–670 (1972)
12. J. W. Daniel, On the approximate minimization of functionals. Math. Comp. **23**, 573–581 (1969)
13. J. W. Daniel, On the convergence of a numerical method in optimal control. J. Optimization Theory Appl. **4**, 330–342 (1969)

14. J. W. Daniel, The Ritz-Galerkin method for abstract optimal control problems. SIAM J. Control **11**, 53–63 (1973)
15. J. W. Daniel, The Approximate Minimization of Functionals. John Wiley-Interscience. New York 1983
16. A. L. Dontchev, Error estimates for a discrete approximation to constrained control problems. SIAM J. Numer. Anal. **18**, 500–514 (1981)
17. A. L. Dontchev, Perturbations, approximations and sensitivity analysis of optimal control systems. Lecture Notes in Control and Information Sciences, **52**. Springer, New York 1983
18. A. L. Dontchev, An a priori estimate for discrete approximations in nonlinear optimal control. SIAM J. Control Optim. **34**, 1315–1328 (1996)
19. A. L. Dontchev, Discrete approximations in optimal control. in Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control (Minneapolis, MN, 1993), IMA Vol. Math. Appl. **78**, 59–81 (1996)
20. A. L. Dontchev, W. W. Hager, Lipschitzian stability in nonlinear control and optimization. SIAM J. Control Optim. **31**, 569-603 (1993)
21. A. L. Dontchev, W. W. Hager, Lipschitzian stability for state constrained nonlinear optimal control. SIAM J. Control Optim. **36**, 696–718 (1998)
22. A. L. Dontchev, W. W. Hager, The Euler approximation in state constrained optimal control. Math. Comp. (to appear)
23. A. L. Dontchev, W. W. Hager, A. B. Poore, B. Yang, Optimality, stability and convergence in nonlinear control. Appl. Math. Optim. **31**, 297–326 (1995)
24. A. L. Dontchev, W. W. Hager, V. M. Veliov, Second-order Runge-Kutta approximations in constrained optimal control. SIAM J. Numer. Anal. (to appear)
25. A. L. Dontchev, K. Malanowski, A characterization of Lipschitzian stability in optimal control. in Proceedings of the Conference on Calculus of Variations, Haifa, 1998, (to appear)
26. J. C. Dunn, On $L^2$ sufficient conditions and the gradient projection method for optimal control problems. SIAM J. Control Optim. **34**, 1270–1290 (1996)
27. J. C. Dunn, T. Tian, Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions. SIAM J. Control Optim. **30**, 1361–1384 (1992)
28. E. Farhi, Runge-Kutta schemes applied to linear-quadratic optimal control problems.in Mathematics and Mathematical Education (Sunny Beach, 1984), pp. 464–472, Bulgarian Akad. Nauk., Sofia, 1984
29. L. Flatto, Advanced Calculus. Waverly Press, Baltimore, MD, 1976
30. W. W. Hager, The Ritz-Trefftz method for state and control constrained optimal control problems. SIAM J. Numer. Anal. **12**, 854–867 (1975)
31. W. W. Hager, Rate of convergence for discrete approximations to unconstrained control problems. SIAM J. Numer. Anal. **13**, 449–471 (1976)
32. W. W. Hager, Convex control and dual approximations. Control and Cybernetics **8**, 1–22, 73–86 (1979)
33. W. W. Hager, Inequalities and approximation. in Constructive Approaches to Mathematical Models. Academic Press, New York, 1979, pp. 189–202
34. W. W. Hager, Approximations to the multiplier method. SIAM J. Numer. Anal. **22**, 16–46 (1985)
35. W. W. Hager, Multiplier methods for nonlinear optimal control. SIAM J. Numer. Anal. **27** , 1061–1080 (1990)
36. W. W. Hager, G. D. Ianculescu, Dual approximations in optimal control. SIAM J. Control Optim. **22**, 423–465 (1984)
37. W. W. Hager, R. Rostamian, Optimal coatings, bang-bang controls, and gradient techniques. Optimal Control Applications and Methods **8**, 1–20 (1987)

38. E. Isaacson, H. B. Keller, Analysis of Numerical Methods. John Wiley, New York 1966
39. D. Kincaid, W. Cheney, Numerical Analysis. Brooks/Cole, Pacific Grove, CA, 1991
40. R. E. Kopp, R. McGill, Several trajectory optimization techniques, in Computing Methods in Optimization Problems, A. V. Balakrishnan and L. W. Neustadt (eds.) Academic Press, New York, 1964, pp. 65–89
41. F. Lempio, V. M. Veliov, Discrete approximations to differential inclusions. GAMM Mitteilungen **21**, 105–135 (1998)
42. K. Malanowski, C. Büskens, H. Maurer, Convergence of approximations to nonlinear optimal control problems. in Mathematical Programming with Data Perturbations, Ed. A. V. Fiacco, Lecture Notes in Pure and Appl. Math, vol. 195, pp. 253–284. Marcel Dekker, New York 1997
43. B. Mordukhovich, On difference approximations of optimal control systems. J. Appl. Math. Mech. **42**, 452–461 (1978)
44. E. Polak, A historical survey of computations methods in optimal control. SIAM Review **15**, 553–548 (1973)
45. E. Polak, Optimization: Algorithms and Consistent Approximation. Springer, New York 1997
46. R. Pytlak, Runge-Kutta based procedure for the optimal control of differential-algebraic equations. J. Optim. Theory Appl. **97**, 675–705 (1998)
47. A. Schwartz, E. Polak, Consistent approximations for optimal control problems based on Runge-Kutta integration. SIAM J. Control Optim. **34**, 1235–1269 (1996)
48. B. Sendov, V. A. Popov, The averaged moduli of smoothness. John Wiley 1988
49. V. M. Veliov, Second-order discrete approximations to linear differential inclusions. SIAM J. Numer. Anal. **29**, 439–451 (1992)
50. V. M. Veliov, On the time-discretization of control systems. SIAM J. Control Optim. **35**, 1470–1486 (1997)
51. V. Zeidan, Sufficient conditions for variational problems with variable endpoints: coupled points. Appl. Math. Optim. **27**, 191–209 (1993)