SIAM J. NUMER. ANAL.
Vol. 38, No. 1, pp. 202-226

SECOND-ORDER RUNGE–KUTTA APPROXIMATIONS IN CONTROL CONSTRAINED OPTIMAL CONTROL*

A. L. DONTCHEV†, WILLIAM W. HAGER‡, AND VLADIMIR M. VELIOV§

Abstract. In this paper, we analyze second-order Runge–Kutta approximations to a nonlinear optimal control problem with control constraints. If the optimal control has a derivative of bounded variation and a coercivity condition holds, we show that for a special class of Runge–Kutta schemes, the error in the discrete approximating control is $O(h^2)$ where h is the mesh spacing.

Key words. optimal control, numerical solution, discretization, Runge–Kutta scheme, rate of convergence

AMS subject classifications. 49M25, 65L06

PII. S0036142999351765

1. Introduction.

Conditions are developed under which a Runge–Kutta discretization of an optimal control problem with control constraints yields a second-order approximation to the continuous control. When control constraints are active in an optimal control problem, the optimal solution is typically Lipschitz continuous at best, and at each point where a constraint changes between active and inactive, the derivative of the control is discontinuous. On the surface, one may think that Runge–Kutta approximations of second order are not possible. For example, when a function that is smooth except for a point of discontinuity in the derivative is approximated by a piecewise polynomial, the best possible approximation is of order $O(h^{3/2})$ in L^2, where h is the mesh spacing (without special choice of the mesh points). On the other hand, the schemes that we exhibit yield $O(h^2)$ approximations in a discrete L^∞ norm, regardless of how the mesh points fall relative to the point of discontinuity in the derivative. More precisely, we show that if the functions defining the control problem are smooth enough and a coercivity condition holds, then for Runge–Kutta schemes satisfying certain conditions, the error in the discrete approximation is $O(h)$ if the optimal control is Lipschitz continuous, $o(h)$ if the derivative of the optimal control is Riemann integrable, and $O(h^2)$ if the derivative of the optimal control has bounded variation.

This second-order convergence result exploits the fact that there are often a finite number of points where the control constraints change between active and inactive in an optimal control problem, and although the optimal control is only Lipschitz continuous, its derivative has bounded variation. For example, from a result of Brunovský [5], it follows that for a linear system with a strictly convex quadratic cost functional with analytic coefficient matrices and for a convex polyhedral constraint set, there are finitely many instants of time where the control constraint switches between active

*Received by the editors February 1, 1999; accepted for publication (in revised form) November 15, 1999; published electronically June 20, 2000. This work was supported by the National Science Foundation.

http://www.siam.org/journals/sinum/38-1/35176.html

†Mathematical Reviews, Ann Arbor, MI 48107-8604 (ald@ams.org).

‡Department of Mathematics, University of Florida, Gainesville, FL 32611 (hager@math.ufl.edu, http://www.math.ufl.edu/∼hager).

§Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria and Vienna University of Technology, Wiedner Hauptstr. 8–10/115, A-1040 Vienna, Austria (veliov@uranus.tuwien.ac.at).

and inactive. Moreover, the first derivative of the optimal control is piecewise analytic and has finitely many points of discontinuity. For a more general result on bounds for the number of switchings in solutions to piecewise analytic vector fields, see [50]. For regularity results concerning problems whose cost function satisfies a coercivity condition, see [32], [21], [24], and [17].

To illustrate the subtleties that arise in discrete approximations to control problems, let us consider the following example from [30, (P1)]:

$$\begin{aligned} &\text{minimize} \ \ \frac{1}{2}\int_0^1 u(t)^2 + 2x(t)^2 \, dt \\ &\text{subject to} \quad \dot{x}(t) = .5x(t) + u(t), \quad x(0) = 1, \end{aligned}$$

with the optimal solution

$$x^*(t) = \frac{2e^{3t} + e^3}{e^{3t/2}(2 + e^3)}, \quad u^*(t) = \frac{2(e^{3t} - e^3)}{e^{3t/2}(2 + e^3)}. \tag{1}$$

A very plausible two-stage Runge–Kutta discretization of this problem is the following:

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

Here $h = 1/N$ is the mesh size and x_k and u_k represent approximations to $x(kh)$ and $u(kh)$, respectively. 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 zero and a value around $-2/h$; hence the solution to the discrete problem diverges from the solution (1) to the continuous problem as h tends to zero.

Now let us replace the control variable u_k in the first stage by $u_{k+1/2}$ to obtain the following discretization:

$$\begin{aligned} &\text{minimize} \ \ \frac{h}{2}\sum_{k=0}^{N-1} u_{k+1/2}^2 + 2x_{k+1/2}^2 \\ &\text{subject to} \quad x_{k+1/2} = x_k + \tfrac{h}{2}(.5x_k + u_{k+1/2}), \\ &\qquad\qquad\quad\ x_{k+1} = x_k + h(.5x_{k+1/2} + u_{k+1/2}), \quad x_0 = 1. \end{aligned} \tag{3}$$

According to the theory developed in this paper, the solution to the discrete problem (3) not only converges to the solution u^* of the continuous problem, but the error is $O(h^2)$. Notice that in this convergent discretization, the dimension of the discrete control space has been reduced by identifying the control value u_k in the first stage of the Runge–Kutta scheme with the control value $u_{k+1/2}$ at the midpoint.

Convergence results for Runge–Kutta discretizations of optimal control problems are surprisingly scarce, although these methods are often used (for example, see [44],

[45], [48], [49]). To briefly summarize prior work on discrete approximations in optimal control, some of the initial efforts dealt with the convergence of the cost or controls for the discrete problem to the cost or controls for the continuous problem. For example, see [6], [8], [9], [10], [11], [12], [13], [14], [41], and the surveys in [42], [43], [18]. More recently Schwartz and Polak [46] consider a nonlinear optimal control problem with control and endpoint constraints and they analyze the consistency of explicit Runge–Kutta approximations. Convergence is proved for the global solution of the discrete problem to the global solution of the continuous problem. In [46] consistency and convergence are analyzed for schemes whose coefficients in the final stage of the Runge–Kutta scheme are all positive. In this paper, we analyze convergence rate and we show that coefficients in the final stage of the scheme can vanish if the dimension of the discrete control space is suitably reduced.

The early work dealing with convergence rates for discrete approximations to control problems includes [3], [4], [15], [29], [30], [31], and [35]. In the first paper [30] to consider the usual Runge–Kutta and multistep integration schemes, Hager studied an unconstrained optimal control problem and determined the relationship between the continuous dual variables and the Kuhn–Tucker multipliers associated with the discrete problem. It was observed that an order k integration scheme for differential equations did not always lead to an order k discrete approximation in optimal control; for related work following these results see [28]. In [15] (see also [16, Chap. 4]) Dontchev analyzed Euler's approximation to a constrained convex control problem obtaining an $O(h)$ error estimate in the L^2 norm. In [19] an $O(h)$ estimate in L^∞ is obtained for the error in the Euler discretization of a nonlinear optimal control problem with control constraints. More recently, in [20] an $O(h)$ estimate for the error in the Euler approximation to a general state constrained control problem is obtained. Results are obtained in [40] for the Euler discretization of a nonlinear problem with mixed control and state constraints. The underlying assumptions, however, exclude purely state constrained problems. In [51] an $O(h^2)$ approximation of the optimal cost is established for control constrained problems with linear dynamics, without assuming the regularity of the optimal control. In [52] this result is extended to systems that are nonlinear with respect to the state variable. In [39], $O(h^{1/2})$ and $O(h)$ error estimates are obtained for the optimal cost in Runge–Kutta discretizations of control systems with discontinuous right-hand side.

We also point out a companion paper [34] in which conditions are derived for the coefficients of a Runge–Kutta integration scheme that ensure a given order of accuracy in optimal control for orders up to four. The paper [34] focuses on Runge–Kutta schemes whose coefficients in the last stage are all positive, while here this positivity condition is removed by working in reduced dimension control spaces. In fact, we show that any second-order Runge–Kutta scheme for differential equations yields a second-order approximation in optimal control through an appropriate interpretation of the discrete controls.

The paper is organized in the following way: section 2 presents the Runge–Kutta discretization and the main theorem. Section 3 gives the abstract result [22, Thm. 3.1] on which the convergence theorem is based. In sections 4–8 we verify each of the hypotheses of the abstract theorem. Section 9 gives numerical illustrations, while section 10 shows how the optimal discrete control can be extended to a function in continuous time whose corresponding state trajectory has the same error at the grid points as that of the discrete state trajectory.

2. The problem and its discretization. We consider the following optimal control problem:

$$\begin{aligned} &\text{minimize} \;\; C(x(1)) \\ &\text{subject to} \quad \dot{x}(t) = f(x(t), u(t)), \quad u(t) \in U, \quad \text{almost everywhere (a.e.) } t \in [0,1], \\ &\qquad\qquad\quad x(0) = a, \quad x \in W^{1,\infty}, \quad u \in L^\infty, \end{aligned} \tag{4}$$

where the state $x(t) \in \mathbf{R}^n$, $\dot{x}$ stands for $\frac{d}{dt}x$, the control $u(t) \in \mathbf{R}^m$, $f : \mathbf{R}^n \times \mathbf{R}^m \mapsto \mathbf{R}^n$, $C : \mathbf{R}^n \mapsto \mathbf{R}$, and $U \subset \mathbf{R}^m$ is closed and convex. Note that an integral term in the cost function can be accommodated by adding another component to the state variable and putting the value of this new state variable component at $t = 1$ in place of the integral term.

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

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

where $|\cdot|$ is the Euclidean norm for vectors and the Frobenius norm for matrices. 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 $x : [0,1] \mapsto \mathbf{R}^n$ whose jth derivative lies in L^p for all $0 \le j \le m$ with the norm

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

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 A is A^{T}, and $B_a(x)$ is the closed ball centered at x with radius a.

We now present the assumptions that are employed in our analysis of Runge–Kutta discretizations of (4). The first assumption is related to the regularity of the solution and the problem functions.

Smoothness. The problem (4) has a local solution (x^*, u^*) which lies in $W^{2,\infty} \times W^{1,\infty}$. There exists an open set $\Omega \subset \mathbf{R}^n \times \mathbf{R}^m$ and $\rho > 0$ such that $B_\rho(x^*(t), u^*(t)) \subset \Omega$ for every $t \in [0,1]$, the first two derivatives of f are Lipschitz continuous in Ω, and the first two derivatives of C are Lipschitz continuous in $B_\rho(x^*(1))$.

Under this assumption, there exists an associated Lagrange multiplier $\psi^* \in W^{2,\infty}$ for which the following form of the first-order optimality conditions (minimum principle) is satisfied at (x^*, ψ^*, u^*):

$$\dot{x}(t) = f(x(t), u(t)) \quad \text{for all } t \in [0,1], \quad x(0) = a, \tag{5}$$

$$\dot{\psi}(t) = -\nabla_x H(x(t), \psi(t), u(t)) \quad \text{for all } t \in [0,1], \quad \psi(1) = \nabla C(x(1)), \tag{6}$$

$$u(t) \in U, \quad -\nabla_u H(x(t), \psi(t), u(t)) \in N_U(u(t)) \quad \text{for all } t \in [0,1]. \tag{7}$$

Here H is the Hamiltonian defined by

$$H(x(t), \psi(t), u(t)) = \psi(t) f(x(t), u(t)), \tag{8}$$

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

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

Let us define the following matrices:

$$A(t) = \nabla_x f(x^*(t), u^*(t)), \quad B(t) = \nabla_u f(x^*(t), u^*(t)), \quad V = \nabla C(x^*(1)), \tag{9}$$

$$Q(t) = \nabla_{xx} H(w^*(t)), \quad R(t) = \nabla_{uu} H(w^*(t)), \quad S(t) = \nabla_{xu} H(w^*(t)), \tag{10}$$

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

$$\mathcal{B}(x, u) = \frac{1}{2}\left(x(1)^\mathsf{T} V x(1) + \langle x, Qx \rangle + \langle u, Ru \rangle + 2\langle x, Su \rangle\right),$$

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

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

$$\mathcal{B}(x, u) \;\geq\; \alpha \|u\|_{L^2}^2 \quad \text{for all } \; (x, u) \in \mathcal{M},$$

where

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

Here the algebraic difference $U - U$ is defined by

$$U - U = \{r - s : r \in U \text{ and } s \in U\}.$$

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 (and in certain cases is equivalent to) Lipschitzian dependence of the solution and the multipliers with respect to parameters (see [20] and [25]). For recent work on second-order sufficient conditions, see [26] and [53].

If $U = \mathbf{R}^m$ the variational inequality (7) becomes an algebraic equation and the variational system (5)–(7) is a differential-algebraic equation. In this particular case the coercivity condition reduces to an index 1 condition for the differential-algebraic equation (for example, see [38, sect. 6.5]) and implies local solvability of the algebraic equation with respect to u. After expressing u in terms of x and ψ using (7), the variational system is converted to a boundary-value problem which is analyzed in [34]. On the other hand, the main focus of the present paper is on problems with nontrivial control constraints so that the mapping from (x, ψ) to a control u satisfying (7) is nonsmooth, leading to complications in the analysis.

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, the mesh is uniform of width $h = 1/N$, where N is a natural number.

(If the mesh is not uniform, then the parameter h in the error estimates should be replaced by the length of the largest mesh interval.) If $\mathbf{x}_k$ denotes the approximation to $\mathbf{x}(t_k)$ where $t_k = kh$, then an s-stage Runge–Kutta scheme [7] with coefficients a_{ij} and b_i, $1 \le i, j \le s$, is given by

$$\mathbf{x}_k' = \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \tag{11}$$

where

$$\mathbf{y}_i = \mathbf{x}_k + h \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \quad i = 1, \dots, s, \tag{12}$$

and the prime denotes the forward divided difference:

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

Throughout, we use bold letters for the discrete variables while the corresponding continuous variables are italic. Also, f and $\mathbf{f}$ are the same although we often use $\mathbf{f}$ in an equation involving discrete variables for consistency.

In (11) and (12), $\mathbf{y}_j$ and $\mathbf{u}_{kj}$ are the intermediate state and control variables on the interval $[t_k, t_{k+1}]$. Although there are different intermediate state variables for different intervals, this dependence on k is not explicit in our notation. The discrete variables $\mathbf{y}_i$ and $\mathbf{u}_{ki}$ can be regarded as approximations to the state and control at instants of time on the interval $[t_k, t_{k+1}]$. In particular, we view the value $\mathbf{u}_{ki}$ of the discrete control as an approximation to the value $u(t_k + \sigma_i h)$ of the continuous control at the point $t_k + \sigma_i h$. If $\sigma_i = \sigma_j$ for some $i \ne j$, then the discrete controls $\mathbf{u}_{ki}$ and $\mathbf{u}_{kj}$ are identical. We reduce the dimension of the discrete control space by requiring that intermediate controls be identical if the associated components of the vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_s)$ are equal. More precisely, let $\mathcal{N}_i$ be the indices for which the associated components of $\boldsymbol{\sigma}$ are equal to σ_i:

$$\mathcal{N}_i = \{ j \in [1, s] : \sigma_j = \sigma_i \}. \tag{13}$$

For any time interval, the set $\mathbf{U}$ of feasible discrete controls is the following:

$$\mathbf{U} = \{ (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_s) \in \mathbf{R}^{ms} : \mathbf{u}_i \in U \text{ for each } i \text{ and } \mathbf{u}_i = \mathbf{u}_j \text{ for every } j \in \mathcal{N}_i \}.$$

Throughout the paper, $\mathbf{u}_i$ and $\mathbf{u}_j \in \mathbf{R}^m$ denote components of the vector $\mathbf{u} \in \mathbf{R}^{ms}$ while $\mathbf{u}_k \in \mathbf{R}^{ms}$ is the entire vector at time level k:

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

Hence, the index k will always refer to the time level of the discrete problem. With this notation, the discrete control problem is the following:

$$\begin{aligned} &(14) \quad \text{minimize } C(\mathbf{x}_N) \\ &\text{subject to} \quad \mathbf{x}_k' = \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \quad \mathbf{x}_0 = \mathbf{a}, \quad \mathbf{u}_k \in \mathbf{U}, \\ &\qquad \mathbf{y}_i = \mathbf{x}_k + h \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \quad 1 \le i \le s, \quad 0 \le k \le N-1. \end{aligned}$$

Note that when the cost function in the continuous control problem contains an integral that is treated using an augmented state variable, we essentially employ the same discretization for the integral as that used for the differential equation.

For $\mathbf{x}_k$ near $x^*(t_k)$ and $\mathbf{u}_{kj}$, $1 \leq j \leq s$, near $u^*(t_k)$, it follows from the smoothness condition and the implicit function theorem that when h is small enough, the intermediate variables $\mathbf{y}_i$ in (12) are uniquely determined, smooth functions of $\mathbf{x}_k$ and $\mathbf{u}_k$. More precisely, the following holds (for example, see [7, Thm. 303A] and [1, Thm. 7.6]).

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

$$\mathbf{y}_i = \mathbf{x} + h \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}_j, \mathbf{u}_j), \quad 1 \leq i \leq s, \tag{15}$$

has a unique solution $\mathbf{y}_i \in B_\rho(x^*(t), u^*(t))$, $1 \leq i \leq s$. If $\mathbf{y}_i(\mathbf{x}, \mathbf{u})$, $1 \leq i \leq s$, denotes the solution of (15) associated with $(\mathbf{x}, \mathbf{u})$, then $\mathbf{y}_i(\mathbf{x}, \mathbf{u})$ is twice 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,

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

where $\mathbf{y}$ is the solution of (15) given by the uniqueness property. 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 (14):

$$\mathbf{x}'_k = \mathbf{f}^h(\mathbf{x}_k, \mathbf{u}_k), \quad \mathbf{x}_0 = \mathbf{a}, \tag{16}$$

$$\boldsymbol{\psi}'_k = -\nabla_x H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k), \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N), \tag{17}$$

$$\mathbf{u}_k \in \mathbf{U}, \quad -\sum_{j \in \mathcal{N}_i} \nabla_{u_j} H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k) \in N_U(\mathbf{u}_{ki}), \tag{18}$$

$1 \leq i \leq s$, $0 \leq k \leq N-1$, where $\boldsymbol{\psi}_k \in \mathbf{R}^n$. The sum over $\mathcal{N}_i$ in (18) arises since we are differentiating a function of s variables for which those variables associated with $j \in \mathcal{N}_i$ are identical. Hence, when we differentiate with respect to $\mathbf{u}_{ki}$, we obtain the sum of the partial derivatives with respect to all the variables associated with indices in $\mathcal{N}_i$.

We focus on second-order Runge–Kutta schemes in which cases the coefficients satisfy the following conditions:

(19)

$$\text{(a)} \ \sum_{i=1}^{s} b_i = 1, \ \ \text{(b)} \ \sum_{i=1}^{s} b_i c_i = \frac{1}{2}, \ \ c_i = \sum_{j=1}^{s} a_{ij}, \ \ \text{(c)} \ \sum_{i=1}^{s} b_i \sigma_i = \frac{1}{2}, \ \ 0 \leq \sigma_i \leq 1.$$

Conditions (a) and (b) are the standard conditions found in [7, p. 170] for a second-order Runge–Kutta scheme, while condition (c) ensures that if the discrete controls $\mathbf{u}_{ki}$ are replaced by the continuous control values $u(t_k + h\sigma_i)$, then the resulting Runge–Kutta scheme is second order. For the optimal control problem, additional conditions must be imposed on the coefficients. In particular, we assume that the following conditions hold for each integer $l \in [1, s]$:

(20)

$$\text{(a)} \sum_{i \in \mathcal{N}_l} b_i c_i = \sum_{i \in \mathcal{N}_l} b_i \sigma_i, \quad \text{(b)} \sum_{i=1}^{s} \sum_{j \in \mathcal{N}_l} b_i a_{ij} = \sum_{i \in \mathcal{N}_l} b_i (1 - \sigma_i), \quad \text{(c)} \sum_{i \in \mathcal{N}_l} b_i > 0.$$

These conditions are needed in our analysis of the residual obtained by substituting the continuous optimal solution into the discrete minimum principle (18). They imply that this residual is $O(h^2)$ under appropriate smoothness assumptions for the optimal control. Condition (20), part (b), is somewhat similar to the so-called simplified assumption $D(1)$ for Runge–Kutta schemes (see [37, p. 208]), but with the difference that c_i is replaced by σ_i. A trivial choice for σ_i that satisfies (19), part (c), and (20) is $\sigma_i = 1/2$ for each i, in which case

$$\mathcal{N}_l = \{1, 2, \ldots, s\}$$

for each l. For this choice, all the discrete controls associated with a given time level are equal.

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

$$\omega(v, J; t, h) = \sup\{|v(s_1) - v(s_2)| : s_1, s_2 \in [t - h/2, t + h/2] \cap J\}. \tag{21}$$

The averaged modulus of smoothness τ of v over $[0, 1]$ is the integral of the modulus of continuity:

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

Theorem 2.1. *If the coefficients of the Runge–Kutta integration scheme satisfy the conditions* (19) *and* (20) *and if the smoothness and coercivity conditions hold, then for all sufficiently small h, there exists a strict local minimizer* $(\mathbf{x}^h, \mathbf{u}^h)$ *of the discrete optimal control problem* (14) *and an associated adjoint variable* $\boldsymbol{\psi}^h$ *satisfying* (17) *and* (18) *such that*

$$\max_{\substack{0 \le k \le N \\ 1 \le i \le s}} |\mathbf{x}_k^h - x^*(t_k)| + |\boldsymbol{\psi}_k^h - \boldsymbol{\psi}^*(t_k)| + |\mathbf{u}_{ki}^h - u^*(t_k + \sigma_i h)| \le ch(h + \tau(\dot{u}^*; h)). \tag{22}$$

Since $\dot{u}^* \in L^\infty$, it follows from the properties [47, sect. 1.3] of the averaged modulus of smoothness that the error term in (22) is $O(h)$. Moreover, if $\dot{u}^*$ is Riemann integrable, then the error is $o(h)$, and if $\dot{u}^*$ has bounded variation, then the error is $O(h^2)$.

Remark 2.2. Let $\mathbf{u} \in \mathbf{R}^{smN}$ denote the vector of discrete 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. Any mathematical programming algorithm can be used to minimize $C(\mathbf{u})$ subject to the control constraint $\mathbf{u}_k \in \mathbf{U}$. Often these algorithms are much easier to implement when a formula is available for the cost gradient with respect to the control. If $b_i > 0$ for each i, then this gradient can be computed efficiently using the transformed adjoint equation as explained in [34]. When b_i vanishes for some i, the transformation in [34] cannot be applied. We now explain how the gradient computation is modified when one of the coefficients of b vanishes. As in [34], let us introduce a multiplier $\boldsymbol{\lambda}_i$ for the ith intermediate equation (12) in addition to the multiplier $\boldsymbol{\psi}_{k+1}$ for (11). Taking into account these additional multipliers, the first-order necessary conditions are the following:

$$\boldsymbol{\psi}_k - \boldsymbol{\psi}_{k+1} = \sum_{i=1}^{s} \boldsymbol{\lambda}_i, \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N), \tag{23}$$

$$h\left(b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij} \boldsymbol{\lambda}_i\right) \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) = \boldsymbol{\lambda}_j, \quad 1 \le j \le s, \tag{24}$$

$$\mathbf{u}_{kj} \in U, \quad -\left(b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij} \boldsymbol{\lambda}_i\right) \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}) \in N_U(\mathbf{u}_{kj}), \tag{25}$$

$0 \le k \le N-1$. Once again, the dual multipliers here are all treated as row vectors. Based on the analysis in [36], the gradient of the discrete cost is given by

$$\nabla_{u_{kj}} C(\mathbf{u}) = h\left(b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij} \boldsymbol{\lambda}_i\right) \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}),$$

where the intermediate values for the discrete state variables are obtained by solving the discrete equations (11) and (12), and where the multipliers are chosen to satisfy (23) and (24). For h sufficiently small, (24) is an invertible linear system for the $\boldsymbol{\lambda}_i$, $1 \le i \le s$, in terms of $\boldsymbol{\psi}_{k+1}$, while (23) yields $\boldsymbol{\psi}_k$ in terms of $\boldsymbol{\psi}_{k+1}$ and the $\boldsymbol{\lambda}_i$.

3. Abstract setting.

Our proof of Theorem 2.1 is based on the following abstract result, which is a corollary of [22, Thm. 3.1].

PROPOSITION 3.1. *Let $\mathcal{X}$ be a Banach space and let $\mathcal{Y}$ be a linear normed space with the norms in both spaces denoted $\|\cdot\|$. Let $\mathcal{F} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$, 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 ϵ, λ, 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 λ.*

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

$$\|w - w^*\| \le \frac{\lambda}{1 - \lambda\epsilon} \|\delta\|. \tag{26}$$

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

In applying Proposition 3.1, we utilize discrete analogues of various continuous spaces and norms. In particular, for a sequence $\mathbf{z}_0$, $\mathbf{z}_1$, ..., $\mathbf{z}_N$ whose ith element is a vector $\mathbf{z}_i \in \mathbf{R}^n$, the discrete analogues of the L^p and L^∞ 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 $\mathcal{X}$ is the discrete L^∞ space consisting of 3-tuples $w = (\mathbf{x}, \boldsymbol{\psi}, \mathbf{u})$, where

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

The mappings $\mathcal{T}$ and $\mathcal{F}$ of Proposition 3.1 are selected in the following way:

$$\mathcal{T}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = \begin{pmatrix} \mathbf{x}_k' - \mathbf{f}^h(\mathbf{x}_k, \mathbf{u}_k), \quad 0 \le k \le N-1, \\ \boldsymbol{\psi}_k' + \nabla_x H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k), \quad 0 \le k \le N-1, \\ -\displaystyle\sum_{j\in\mathcal{N}_i} \nabla_{u_j} H^h(\mathbf{x}_k, \boldsymbol{\psi}_{k+1}, \mathbf{u}_k), \quad 1 \le i \le s,\ 0 \le k \le N-1, \\ \boldsymbol{\psi}_N - \nabla C(\mathbf{x}_N) \end{pmatrix} \tag{27}$$

and

$$\mathcal{F}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = \begin{pmatrix} 0 \\ 0 \\ N_U(\mathbf{u}_{k1}) \times N_U(\mathbf{u}_{k2}) \times \cdots \times N_U(\mathbf{u}_{ks}), \quad 0 \le k \le N-1, \\ 0 \end{pmatrix}. \tag{28}$$

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 $\mathbf{x}_k^* = x^*(t_k)$, $\boldsymbol{\psi}_k^* = \psi^*(t_k)$, and $\mathbf{u}_{ki}^* = u^*(t_k + \sigma_i h)$ (obviously, for $k = N$ the $\mathbf{u}_k$ component of w_k should be removed). The operator $\mathcal{L}$ is obtained 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 other words, we choose

(29)

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

In the following sections, we verify the hypotheses of Proposition 3.1.

4. Analysis of the residual. In order to apply Proposition 3.1, we need an estimate for the distance from $\mathcal{T}(w^*)$ to $\mathcal{F}(w^*)$ for the specific $\mathcal{T}$ and $\mathcal{F}$ in (27) and (28), respectively. This distance emerges in several parts of the proposition. First, in (P1) the parameter δ is the perturbation of $\mathcal{T}(w^*)$ needed to reach the set $\mathcal{F}(w^*)$ and in (26), the distance from the solution w of the inclusion $\mathcal{T}(w) \in \mathcal{F}(w)$ to w^* is bounded in terms of the norm of δ. Also, δ needs to satisfy the additional conditions $\|\delta\| \le \sigma$ and $\|\delta\| \le (1-\lambda\epsilon)r/\lambda$. It is trivial to estimate the distance between the last components of $\mathcal{T}(w^*)$ and $\mathcal{F}(w^*)$ since $\boldsymbol{\psi}_N^* = \nabla C(\mathbf{x}_N^*)$ and the distance is simply zero. In this section, we focus on the analysis of the first three components, which we refer to as the state residual, the costate residual, and the control residual, respectively. The following result, proved in [47, Thm. 3.4], is used repeatedly in the analysis.

PROPOSITION 4.1. *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 ω *is the modulus of continuity defined in* (21). *Here* c *depends on the choice of* $\mathbf{b}$ *and* $\boldsymbol{\sigma}$, *but not on* ϕ *or* h.

Now let us proceed to analyze each of the first three components of $\mathcal{T}(w^*)$.

State residual. Suppose that $h \le \gamma$ and that h is small enough that $(\mathbf{x}_k^*, \mathbf{u}_{ki}^*) \in B_\beta(x^*(t_k), u^*(t_k))$ for each i and k. Let $\mathbf{y}_i^*$ denote $\mathbf{y}_i(\mathbf{x}_k^*, \mathbf{u}_k^*)$. Expanding $\mathbf{f}$ in (15) in a Taylor series around $(x^*(t_k), u^*(t_k))$, we have

$$\mathbf{y}_i^* = \mathbf{x}_k^* + h\sum_{j=1}^s a_{ij}\mathbf{f}(\mathbf{y}_j^*, \mathbf{u}_{kj}^*) = \mathbf{x}_k^* + h\sum_{j=1}^s a_{ij}\mathbf{f}_k + O(h^2) = \mathbf{x}_k^* + hc_i\mathbf{f}_k + O(h^2),$$

where $\mathbf{f}_k = f(x^*(t_k), u^*(t_k))$. A Taylor expansion of $x^*(t)$ around $t = t_k$ gives

$$x^*(t_k + \sigma_i h) = \mathbf{x}_k^* + h\sigma_i\mathbf{f}_k + O(h^2). \tag{30}$$

Combining these two expansions and utilizing (19) yields

$$\sum_{i=1}^s b_i\mathbf{y}_i^* = \sum_{i=1}^s b_i x^*(t_k + \sigma_i h) + O(h^2). \tag{31}$$

Let $\mathbf{x}_{ki}^*$ stand for $x^*(t_k + \sigma_i h)$. Since $\nabla_x\mathbf{f}$ is continuous, we have

$$\sum_{i=1}^s b_i(\mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) - \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*)) = \sum_{i=1}^s b_i\mathbf{F}_{ki}(\mathbf{y}_i^* - \mathbf{x}_{ki}^*), \tag{32}$$

where $\mathbf{F}_{ki}$ is the average of $\nabla_x\mathbf{f}(\cdot, \mathbf{u}_{ki}^*)$ along the line segment connecting $\mathbf{y}_i^*$ and $\mathbf{x}_{ki}^*$. Since $\nabla_x\mathbf{f}$ is Lipschitz continuous, it follows that

$$|\mathbf{F}_{ki} - \nabla_x\mathbf{f}_k| \le c(|\mathbf{y}_i^* - \mathbf{x}_k^*| + |\mathbf{x}_{ki}^* - \mathbf{x}_k^*| + |\mathbf{u}_{ki}^* - u^*(t_k)|). \tag{33}$$

By (15) and (30), $\mathbf{y}_i^* = \mathbf{x}_k^* + O(h)$ and $\mathbf{x}_{ki}^* = \mathbf{x}_k^* + O(h)$. And by the Lipschitz continuity of u^*, we have $\mathbf{u}_{ki}^* = u^*(t_k) + O(h)$. Hence, combining (32) and (33) gives

$$\begin{aligned}\sum_{i=1}^s b_i\mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) &= \sum_{i=1}^s b_i(\mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + \mathbf{F}_{ki}(\mathbf{y}_i^* - \mathbf{x}_{ki}^*)) \\ &= \sum_{i=1}^s b_i(\mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + \nabla_x\mathbf{f}_k(\mathbf{y}_i^* - \mathbf{x}_{ki}^*) + (\mathbf{F}_{ki} - \nabla_x\mathbf{f}_k)(\mathbf{y}_i^* - \mathbf{x}_{ki}^*)) \\ &= \sum_{i=1}^s b_i(\mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + \nabla_x\mathbf{f}_k(\mathbf{y}_i^* - \mathbf{x}_{ki}^*)) + O(h^2).\end{aligned}$$

And by (31) we have

$$\sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) = \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + O(h^2). \tag{34}$$

Finally, this relation along with Proposition 4.1 yields

$$\begin{aligned}(\mathbf{x}_k^*)' - \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) &= \frac{1}{h}\int_{t_k}^{t_{k+1}} \mathbf{f}(x^*(t), u^*(t))\, dt - \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + O(h^2) \\ &\le c\tau_k(\dot{f}(x^*, u^*); h) + O(h^2) \le c\tau_k(\dot{u}; h) + O(h^2),\end{aligned}$$

where

$$\tau_k(\phi; h) = \int_{t_k}^{t_{k+1}} \omega(\phi, [t_k, t_{k+1}]; t, h)\, dt.$$

Hence, the L^1 norm of the first component of $\mathcal{T}(w^*) - \mathcal{F}(w^*)$ satisfies the following bound:

$$\sum_{i=0}^{N-1} h \left| (\mathbf{x}_k^*)' - \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) \right| \le ch(h + \tau(\dot{u}^*; h)).$$

Costate residual. Letting $\boldsymbol{\psi}_{ki}^*$ denote $\boldsymbol{\psi}^*(t_k + \sigma_i h)$, a Taylor expansion yields

$$\boldsymbol{\psi}_{ki}^* = \boldsymbol{\psi}_k^* - h\sigma_i \nabla_x H_k + O(h^2) \quad \text{and} \quad \boldsymbol{\psi}_{k+1}^* = \boldsymbol{\psi}_k^* - h\nabla_x H_k + O(h^2),$$

where $\nabla_x H_k$ is the gradient of H evaluated at $(x^*(t_k), \psi^*(t_k), u^*(t_k))$. Utilizing (19), part (c), we have

$$\sum_{i=1}^{s} b_i (\boldsymbol{\psi}_{ki}^* - \boldsymbol{\psi}_{k+1}^*) = \frac{h}{2}\nabla_x H_k + O(h^2). \tag{35}$$

By exactly the same chain of equalities used to obtain (34), we deduce that $\nabla_x \mathbf{f}$ satisfies the same identity:

$$\sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) = \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + O(h^2). \tag{36}$$

By the definition of $\mathbf{y}_i(\mathbf{x}, \mathbf{u})$ in (15), it follows immediately that $\nabla_x \mathbf{y}_i(\mathbf{x}_k^*, \mathbf{u}_k^*) = \mathbf{I} + O(h)$. Furthermore, after differentiating the right side of (15), we see that

$$\begin{aligned}\nabla_x \mathbf{y}_i(\mathbf{x}_k^*, \mathbf{u}_k^*) &= \mathbf{I} + h\sum_{j=1}^{s} a_{ij} \nabla_x \mathbf{f}(\mathbf{y}_j^*, \mathbf{u}_{kj}^*) + O(h^2) \\ &= \mathbf{I} + h\sum_{j=1}^{s} a_{ij} \nabla_x \mathbf{f}_k + O(h^2). \qquad (37)\end{aligned}$$

Combining (36) and (37) and utilizing (19) yields

$$\begin{aligned}
&\sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) \nabla_x \mathbf{y}_i(\mathbf{x}_k^*, \mathbf{u}_k^*) \\
&\quad = \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) \left(\mathbf{I} + h \sum_{j=1}^{s} a_{ij} \nabla_x \mathbf{f}_k \right) + O(h^2) \\
&\quad = \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + h \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) \sum_{j=1}^{s} a_{ij} \nabla_x \mathbf{f}_k + O(h^2) \\
&\quad = \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + h \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}_k \sum_{j=1}^{s} a_{ij} \nabla_x \mathbf{f}_k + O(h^2) \\
&\quad = \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + \frac{h}{2} \nabla_x \mathbf{f}_k \nabla_x \mathbf{f}_k + O(h^2).
\end{aligned}$$

Multiplying this series of equalities on the left by $\boldsymbol{\psi}_{k+1}^*$ and referring to (35) and the definition of H^h, we have

$$\begin{aligned}
&\nabla_x H^h(\mathbf{x}_k^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_k^*) \\
&\quad = \boldsymbol{\psi}_{k+1}^* \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) \nabla_x \mathbf{y}_i(\mathbf{x}_k^*, \mathbf{u}_k^*) \\
&\quad = \sum_{i=1}^{s} b_i \nabla_x H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{ki}^*) + \frac{h}{2} \boldsymbol{\psi}_{k+1}^* \nabla_x \mathbf{f}_k \nabla_x \mathbf{f}_k + O(h^2) \\
&\quad = \sum_{i=1}^{s} b_i \nabla_x H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{ki}^*) + \frac{h}{2} \nabla_x H_k \nabla_x \mathbf{f}_k + O(h^2) \\
&\quad = \sum_{i=1}^{s} b_i \left(\nabla_x H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{ki}^*) + (\boldsymbol{\psi}_{ki}^* - \boldsymbol{\psi}_{k+1}^*) \nabla_x \mathbf{f}_k \right) + O(h^2) \\
&\quad = \sum_{i=1}^{s} b_i \left(\nabla_x H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{ki}^*) + (\boldsymbol{\psi}_{ki}^* - \boldsymbol{\psi}_{k+1}^*) \nabla_x \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) \right) + O(h^2) \\
&\quad = \sum_{i=1}^{s} b_i \nabla_x H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{ki}^*, \mathbf{u}_{ki}^*) + O(h^2).
\end{aligned}$$

When this relation is applied to the second component of $\mathcal{T}(w^*)$, we obtain, with the aid of Proposition 4.1,

$$\begin{aligned}
&|\boldsymbol{\psi}_k^{*\prime} + \nabla_x H^h(\mathbf{x}_k^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_k^*)| \\
&\quad = \left| \int_{t_k}^{t_{k+1}} \nabla_x H(x^*(t), \psi^*(t), u^*(t))\, dt - \sum_{i=1}^{s} b_i \nabla_x H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{ki}^*, \mathbf{u}_{ki}^*) \right| + O(h^2) \\
&\quad \le ch(h + \tau(\dot{u}^*; h)).
\end{aligned}$$

Hence, the L^1 norm of the second component of $\mathcal{T}(w^*) - \mathcal{F}(w^*)$ satisfies the following bound:

$$\sum_{i=0}^{N-1} |\boldsymbol{\psi}_k^{*\prime} + \nabla_x H^h(\mathbf{x}_k^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_k^*)| \le ch(h + \tau(\dot{u}^*; h)).$$

Control residual. Given any integer $l \in [1, s]$ and restricting the sum in (31) to $i \in \mathcal{N}_l$, it follows from (20), part (a), that

$$\sum_{i \in \mathcal{N}_l} b_i \mathbf{y}_i^* = \sum_{i \in \mathcal{N}_l} b_i x^*(t_k + \sigma_i h) + O(h^2). \tag{38}$$

Similarly, restricting the sum in (35) to $i \in \mathcal{N}_l$, we obtain

$$\sum_{i \in \mathcal{N}_l} b_i (\boldsymbol{\psi}_{ki}^* - \boldsymbol{\psi}_{k+1}^*) = h \sum_{i \in \mathcal{N}_l} b_i (1 - \sigma_i) \nabla_x H_k + O(h^2). \tag{39}$$

Restricting the sum in (34) to $i \in \mathcal{N}_l$ and utilizing (38) in place of (31), we obtain in the same fashion

$$\sum_{i \in \mathcal{N}_l} b_i \nabla_u \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) = \sum_{i \in \mathcal{N}_l} b_i \nabla_u \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) + O(h^2). \tag{40}$$

Using the implicit function theorem to evaluate $\nabla_{u_j} \mathbf{y}_i$ in (15), we have

$$\nabla_{u_j} \mathbf{y}_i(\mathbf{x}_k^*, \mathbf{u}_k^*) = h a_{ij} \nabla_u \mathbf{f}_k + O(h^2). \tag{41}$$

Combining (39)–(41) and utilizing (20), part (b), yields

$$\begin{aligned}
&\sum_{j \in \mathcal{N}_l} \nabla_{u_j} H^h(\mathbf{x}_k^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_k^*) \\
&\quad = \boldsymbol{\psi}_{k+1}^* \sum_{j \in \mathcal{N}_l} b_j \nabla_u \mathbf{f}(\mathbf{y}_j^*, \mathbf{u}_{kj}^*) + \boldsymbol{\psi}_{k+1}^* \sum_{j \in \mathcal{N}_l} \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) \nabla_{u_j} \mathbf{y}_i(\mathbf{x}_k^*, \mathbf{u}_k^*) \\
&\quad = \boldsymbol{\psi}_{k+1}^* \sum_{j \in \mathcal{N}_l} b_j \nabla_u \mathbf{f}(\mathbf{x}_j^*, \mathbf{u}_{kj}^*) + h \boldsymbol{\psi}_{k+1}^* \sum_{j \in \mathcal{N}_l} \sum_{i=1}^{s} b_i a_{ij} \nabla_x \mathbf{f}(\mathbf{y}_i^*, \mathbf{u}_{ki}^*) \nabla_u \mathbf{f}_k + O(h^2) \\
&\quad = \sum_{j \in \mathcal{N}_l} b_j \nabla_u H(\mathbf{x}_{kj}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{kj}^*) + h \sum_{j \in \mathcal{N}_l} \sum_{i=1}^{s} b_i a_{ij} \nabla_x H_k \nabla_u \mathbf{f}_k + O(h^2) \\
&\quad = \sum_{i \in \mathcal{N}_l} b_i \nabla_u H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{ki}^*) + h \sum_{i \in \mathcal{N}_l} b_i (1 - \sigma_i) \nabla_x H_k \nabla_u \mathbf{f}_k + O(h^2) \\
&\quad = \sum_{i \in \mathcal{N}_l} b_i \left(\nabla_u H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{ki}^*) + (\boldsymbol{\psi}_{ki}^* - \boldsymbol{\psi}_{k+1}^*) \nabla_u \mathbf{f}_k \right) + O(h^2) \\
&\quad = \sum_{i \in \mathcal{N}_l} b_i \left(\nabla_u H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_{ki}^*) + (\boldsymbol{\psi}_{ki}^* - \boldsymbol{\psi}_{k+1}^*) \nabla_u \mathbf{f}(\mathbf{x}_{ki}^*, \mathbf{u}_{ki}^*) \right) + O(h^2) \\
&\quad = \sum_{i \in \mathcal{N}_l} b_i \nabla_u H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{ki}^*, \mathbf{u}_{ki}^*) + O(h^2).
\end{aligned}$$

Finally, by (13), $\mathbf{x}_{ki}^* = \mathbf{x}_{kj}^*$, $\mathbf{u}_{ki}^* = \mathbf{u}_{kj}^*$, and $\boldsymbol{\psi}_{ki}^* = \boldsymbol{\psi}_{kj}^*$ for all i, $j \in \mathcal{N}_l$. Since

$$-\nabla_u H(\mathbf{x}_{ki}^*, \boldsymbol{\psi}_{ki}^*, \mathbf{u}_{ki}^*) \in N_U(\mathbf{u}_{ki}^*) \quad \text{for each } i \text{ and} \quad \sum_{i \in \mathcal{N}_l} b_i > 0,$$

we obtain the following estimate for the distance to $N_U(\mathbf{u}_{kl}^*)$:

$$\min \left\{ \left| \mathbf{y} + \sum_{j \in \mathcal{N}_l} \nabla_{u_j} H^h(\mathbf{x}_k^*, \boldsymbol{\psi}_{k+1}^*, \mathbf{u}_k^*) \right| : \mathbf{y} \in N_U(\mathbf{u}_{kl}^*) \right\} = O(h^2).$$

This analysis of the residual in the control problem is now pulled together.

LEMMA 4.2. *If the smoothness condition holds, the coefficients of the Runge–Kutta integration scheme satisfy conditions* (19) *and* (20), *and h is small enough that* $(\mathbf{x}_k^*, \mathbf{u}_{ki}^*) \in B_\beta(\mathbf{x}^*(t_k), \mathbf{u}^*(t_k))$ *for each k and i, where β appears in the uniqueness property, then for the $\mathcal{T}$ and $\mathcal{F}$ specified in section* 3 *and for*

$$w_k^* = (\mathbf{x}_k^*, \boldsymbol{\psi}_k^*, \mathbf{u}_k^*) \quad \text{where} \quad \mathbf{x}_k^* = x^*(t_k), \;\; \boldsymbol{\psi}_k^* = \psi^*(t_k), \;\; \text{and } \mathbf{u}_{ki}^* = u^*(t_k + \sigma_i h),$$

the distance from $\mathcal{T}(w^*)$ to $\mathcal{F}(w^*)$ is bounded by $ch(h+\tau(\dot{u}^*; h))$ in $L^1 \times L^1 \times L^\infty \times \mathbf{R}^n$.

5. Approximate stationarity. In this section we examine condition (P2) of Proposition 3.1. One can view this condition as an approximate stationarity condition in the sense that the derivative of $\mathcal{T} - \mathcal{L}$ almost vanishes at w^*.

LEMMA 5.1. *If the smoothness condition and* (19), *part* (a), *hold, then for the $\mathcal{T}$ and $\mathcal{L}$ specified in section* 3, *we have*

$$\|\nabla \mathcal{T}(w) - \mathcal{L}\| \le \|\nabla \mathcal{T}(w) - \mathcal{L}\|_{L^\infty} \le c(\|w - w^*\| + h) \tag{42}$$

for every $w \in B_\beta(w^*)$, where β appears in the uniqueness property.

Proof. For the last component of $\nabla \mathcal{T}(w) - \mathcal{L}$, the analysis is again trivial,

$$|\nabla^2 C(\mathbf{x}_N) - \nabla^2 C(\mathbf{x}_N^*)| \le c|\mathbf{x}_N - \mathbf{x}_N^*| \le c\|w - w^*\|,$$

when $w_k = (\mathbf{x}_k, \mathbf{u}_k, \boldsymbol{\psi}_k) \in B_\beta(w_k^*)$ for each k. For the first component, we need an estimate for the L^∞ norm of the vector sequence whose kth entry is

$$\begin{pmatrix} \sum_{i=1}^s b_i \nabla_x \mathbf{f}(\mathbf{y}_i(\mathbf{x},\mathbf{u}), \mathbf{u}_i) - \mathbf{A}_k \\ \sum_{i=1}^s b_i \nabla_u \mathbf{f}(\mathbf{y}_i(\mathbf{x},\mathbf{u}), \mathbf{u}_i) - \mathbf{B}_k \end{pmatrix} = \sum_{i=1}^s b_i \begin{pmatrix} \nabla_x \mathbf{f}(\mathbf{y}_i(\mathbf{x},\mathbf{u}), \mathbf{u}_i) - \mathbf{A}_k \\ \nabla_u \mathbf{f}(\mathbf{y}_i(\mathbf{x},\mathbf{u}), \mathbf{u}_i) - \mathbf{B}_k \end{pmatrix}, \tag{43}$$

where $\mathbf{u} \in \mathbf{R}^{sm}$ and $(\mathbf{x}, \mathbf{u}_i) \in B_\beta(\mathbf{x}_k^*, \mathbf{u}_k^*)$ for each i. By the chain rule,

$$\begin{aligned} \nabla_x \mathbf{f}(\mathbf{y}_i(\mathbf{x},\mathbf{u}), \mathbf{u}_i) &= \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_i)|_{\mathbf{y}_i = \mathbf{y}_i(\mathbf{x},\mathbf{u})} \nabla_x \mathbf{y}_i(\mathbf{x},\mathbf{u}) \\ &= \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_i)|_{\mathbf{y}_i = \mathbf{y}_i(\mathbf{x},\mathbf{u})} \left(\mathbf{I} + h \sum_{i=1}^s a_{ij} \nabla_x \mathbf{f}(\mathbf{y}_j(\mathbf{x},\mathbf{u}), \mathbf{u}_j)\right) \\ &= \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_i)|_{\mathbf{y}_i = \mathbf{y}_i(\mathbf{x},\mathbf{u})} + O(h). \end{aligned}$$

Subtracting $\mathbf{A}_k$ from each side of this equality gives

$$\begin{aligned} |\nabla_x \mathbf{f}(\mathbf{y}_i(\mathbf{x},\mathbf{u}), \mathbf{u}_i) - \mathbf{A}_k| &= |\nabla_x \mathbf{f}(\mathbf{y}_i(\mathbf{x},\mathbf{u}), \mathbf{u}_i) - \nabla_x \mathbf{f}(\mathbf{x}^*(t_k), \mathbf{u}^*(t_k))| \\ &\le c(|\mathbf{y}_i(\mathbf{x},\mathbf{u}) - \mathbf{x}^*(t_k)| + |\mathbf{u}_i - \mathbf{u}^*(t_k)| + h) \\ &\le c(|\mathbf{x} - \mathbf{x}^*(t_k)| + |\mathbf{u}_i - \mathbf{u}^*(t_k)| + h) \\ &\le c(\|w - w^*\| + h). \end{aligned}$$

The ∇_u component of (43) as well as the other components of $\mathcal{T}$ can be analyzed in exactly the same way to complete the proof. ∎

6. Lipschitz continuity. Focusing on condition (P3) of Proposition 3.1, we need to establish the Lipschitz continuity of the map $(\mathcal{F} - \mathcal{L})^{-1}$ in a ball around the point $\pi = (\mathcal{T} - \mathcal{L})(w^*)$ in $\mathcal{Y} = L^1 \times L^1 \times L^\infty \times \mathbf{R}^n$ where $\mathcal{F}$ and $\mathcal{L}$ are given in (28) and (29), respectively. In fact, we establish Lipschitz continuity over the entire space

$\mathcal{Y}$. That is, given a parameter $\pi = (\mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{s}) \in \mathcal{Y}$, we show that there exists a unique $w \in \mathcal{X}$ such that

$$\mathcal{L}(w) + \pi \in \mathcal{F}(w), \tag{44}$$

and this solution depends Lipschitz continuously on $\pi \in \mathcal{Y}$.

Our approach is the same one used in our earlier work (see [33], [19], [23], [20]). Namely, we write down an associated quadratic programming problem that has a unique solution, identical to that of the inclusion (44), depending Lipschitz continuously on the parameter. For the $\mathcal{L}$ appearing in section 3, the associated quadratic programming problem is the following:

$$\text{minimize } \mathcal{B}^h(\mathbf{x}, \mathbf{u}) + \mathbf{s}^{\mathsf{T}}\mathbf{x}_N + h \sum_{k=0}^{N-1} \left(\mathbf{q}_k^{\mathsf{T}}\mathbf{x}_k + \sum_{i=1}^{s} \mathbf{r}_{ki}^{\mathsf{T}}\mathbf{u}_{ki} \right) \tag{45}$$

$$\text{subject to} \quad \mathbf{x}_k' = \mathbf{A}_k\mathbf{x}_k + \mathbf{B}_k\mathbf{u}_k\mathbf{b} - \mathbf{p}_k, \quad \mathbf{x}_0 = \mathbf{a}, \quad \mathbf{u}_k \in \mathbf{U},$$

where

$$\mathcal{B}^h(\mathbf{x}, \mathbf{u}) = \frac{1}{2} \left(\mathbf{x}_N^{\mathsf{T}}\mathbf{V}\mathbf{x}_N + h \sum_{k=0}^{N-1} \left(\mathbf{x}_k^{\mathsf{T}}\mathbf{Q}_k\mathbf{x}_k + 2\mathbf{x}_k^{\mathsf{T}}\mathbf{S}_k\mathbf{u}_k\mathbf{b} + \sum_{i=1}^{s} b_i\mathbf{u}_{ki}^{\mathsf{T}}\mathbf{R}_k\mathbf{u}_{ki} \right) \right). \tag{46}$$

It can be verified that the first-order optimality condition for this problem is precisely the inclusion (44). According to the theory in [19], if the quadratic form $\mathcal{B}^h$ satisfies a discrete coercivity condition of the form

$$\mathcal{B}^h(\mathbf{x}, \mathbf{u}) \ \geq \ \bar{\alpha} \|\mathbf{u}\|_{L^2}^2 \quad \text{for all } (\mathbf{x}, \mathbf{u}) \in \mathcal{M}^h, \tag{47}$$

where $\bar{\alpha} > 0$ is independent of h and

$$\mathcal{M}^h = \{(\mathbf{x}, \mathbf{u}) : \mathbf{x}_k' = \mathbf{A}_k\mathbf{x}_k + \mathbf{B}_k\mathbf{u}_k\mathbf{b},\ \mathbf{x}_0 = \mathbf{0}, \quad \mathbf{u}_k \in \mathbf{U} - \mathbf{U}\}; \tag{48}$$

then the quadratic program (45) and the inclusion (44) have identical unique solutions, and these solutions depend Lipschitz continuously on the parameter π.

LEMMA 6.1. *If the smoothness and coercivity conditions,* (19)*, part* (a)*, and* (20)*, part* (c)*, all hold, then for* $\bar{h}$ *sufficiently small, there exists a constant* $\bar{\alpha} > 0$ *satisfying* (47) *for all* $h \leq \bar{h}$*. Moreover, the map* $(\mathcal{F} - \mathcal{L})^{-1}$ *with* $\mathcal{F}$ *and* $\mathcal{L}$ *defined in* (28) *and* (29)*, respectively, is Lipschitz continuous with a Lipschitz constant* λ *independent of* h *for* $h \leq \bar{h}$*.*

Proof. As explained above, the lemma follows immediately once we establish the existence of $\bar{\alpha} > 0$ satisfying (47). In [19, Lem. 11] we show that if the smoothness and coercivity conditions hold, then for h sufficiently small,

$$\bar{\mathcal{B}}^h(\mathbf{x}, \mathbf{v}) \ \geq \ \alpha/2 \sum_{k=0}^{N-1} h|\mathbf{v}_k|^2 \quad \text{for all } (\mathbf{x}, \mathbf{v}) \in \bar{\mathcal{M}}^h,$$

where

$$\bar{\mathcal{B}}^h(\mathbf{x}, \mathbf{v}) = \frac{1}{2} \left(\mathbf{x}_N^{\mathsf{T}}\mathbf{V}\mathbf{x}_N + h \sum_{k=0}^{N-1} \left(\mathbf{x}_k^{\mathsf{T}}\mathbf{Q}_k\mathbf{x}_k + 2\mathbf{x}_k^{\mathsf{T}}\mathbf{S}_k\mathbf{v}_k + \mathbf{v}_k^{\mathsf{T}}\mathbf{R}_k\mathbf{v}_k \right) \right) \tag{49}$$

and

$$\bar{\mathcal{M}}^h = \{(\mathbf{x}, \mathbf{v}) : \mathbf{x}'_k = \mathbf{A}_k \mathbf{x}_k + \mathbf{B}_k \mathbf{v}_k,\ \mathbf{x}_0 = \mathbf{0},\ \mathbf{v}_k \in U - U\}.$$

If $\mathbf{u}_k \in \mathbf{U} - \mathbf{U}$, then $\mathbf{v}_k = \mathbf{u}_k \mathbf{b} \in U - U$ since the b_i sum to one and the sum over $i \in \mathcal{N}_l$ is nonnegative for each l. In other words, $\mathbf{v}_k$ is a convex combination of points in $U - U$. Applying (49) with the specific choice $\mathbf{v}_k = \mathbf{u}_k \mathbf{b}$, it follows that

$$\begin{aligned} \mathcal{B}^h(\mathbf{x}, \mathbf{u}) &= \bar{\mathcal{B}}^h(\mathbf{x}, \mathbf{v}) + h \sum_{k=0}^{N-1} \left(\sum_{i=1}^{s} b_i \mathbf{u}_{ki}^{\mathsf{T}} \mathbf{R}_k \mathbf{u}_{ki} - (\mathbf{u}_k \mathbf{b})^{\mathsf{T}} \mathbf{R}_k (\mathbf{u}_k \mathbf{b}) \right) \\ &\geq h \sum_{k=0}^{N-1} \left(\frac{\alpha}{2} |\mathbf{u}_k \mathbf{b}|^2 + \sum_{i=1}^{s} b_i \mathbf{u}_{ki}^{\mathsf{T}} \mathbf{R}_k \mathbf{u}_{ki} - (\mathbf{u}_k \mathbf{b})^{\mathsf{T}} \mathbf{R}_k (\mathbf{u}_k \mathbf{b}) \right). \end{aligned} \tag{50}$$

As noted in [27] or [23, Lem. 2], for any $t \in [0, 1]$,

$$\mathbf{v}^{\mathsf{T}} R(t) \mathbf{v} \geq \alpha |\mathbf{v}|^2 \quad \text{for all } \ \mathbf{v} \in U - U. \tag{51}$$

(This is shown by choosing the control $\mathbf{u}(s)$ in the coercivity condition to be equal to $\mathbf{v}$ for s near t and to vanish elsewhere, and then letting the support of $\mathbf{u}$ tend to zero.) Hence, the functional $F(\mathbf{v}) = \mathbf{v}^{\mathsf{T}} R(t) \mathbf{v}$ is convex when restricted to $U - U$, which implies that

$$F(\mathbf{ub}) \leq \sum_{i=1}^{s} b_i F(\mathbf{u}_i)$$

for each $\mathbf{u} \in \mathbf{U}$. Utilizing this inequality, it follows that for each $\mathbf{u} \in \mathbf{U}$,

$$\begin{aligned} &\frac{\alpha}{2} |\mathbf{ub}|^2 + \sum_{i=1}^{s} b_i \mathbf{u}_i^{\mathsf{T}} R(t) \mathbf{u}_i - (\mathbf{ub})^{\mathsf{T}} R(t) (\mathbf{ub}) \\ &\quad = \frac{\alpha}{2} |\mathbf{ub}|^2 + \sum_{i=1}^{s} b_i F(\mathbf{u}_i) - F(\mathbf{ub}) \geq \frac{\alpha}{2} |\mathbf{ub}|^2 \geq 0, \end{aligned} \tag{52}$$

with equality achieved only when $\mathbf{ub} = \mathbf{0}$.

Since $\mathbf{0}$ lies in the relative interior of $\mathbf{U} - \mathbf{U}$, there exists a sphere $\mathcal{S}$ in the relative interior with center $\mathbf{0}$ and radius $\tau > 0$:

$$\mathcal{S} = \{\mathbf{u} \in \mathbf{R}^{ms} : |\mathbf{u}| = \tau, \quad \mathbf{u} \in \mathbf{U} - \mathbf{U}\}.$$

Since $\mathcal{S}$ is compact, the minimum of the expression (52) over $\mathbf{u} \in \mathcal{S}$ exists. If the minimum value is zero, then as noted previously, $\mathbf{ub} = \mathbf{0}$. But in this case, (52) reduces to the single sum

$$\sum_{i=1}^{s} b_i \mathbf{u}_i^{\mathsf{T}} R(t) \mathbf{u}_i.$$

Since the b_i sum to 1, $|\mathbf{u}| = \tau$, and (51) holds, this sum is positive. This contradicts our assumption that the minimum value in (52) is zero. Hence, the minimum of (52) over $\mathbf{s} \in \mathcal{S}$ is a positive number η:

$$\frac{\alpha}{2} |\mathbf{sb}|^2 + \sum_{i=1}^{s} b_i \mathbf{s}_i^{\mathsf{T}} R(t) \mathbf{s}_i - (\mathbf{sb})^{\mathsf{T}} R(t) (\mathbf{sb}) \geq \eta > 0 \quad \text{for all } \mathbf{s} \in \mathcal{S}. \tag{53}$$

Since $R(t)$ is a continuous function of t, it follows that η can be chosen so that (53) holds for all $t \in [0,1]$. Given $\mathbf{u}_k \in \mathbf{U} - \mathbf{U}$, we insert $\mathbf{s} = \tau \mathbf{u}_k / |\mathbf{u}_k|$ in (53) to obtain

$$\frac{\alpha}{2}|\mathbf{u}_k \mathbf{b}|^2 + \sum_{i=1}^{s} b_i \mathbf{u}_{ki}^{\mathsf{T}} \mathbf{R}_k \mathbf{u}_{ki} - (\mathbf{u}_k \mathbf{b})^{\mathsf{T}} \mathbf{R}_k (\mathbf{u}_k \mathbf{b}) \geq \frac{\eta}{\tau^2}|\mathbf{u}_k|^2 \quad \text{for all } \mathbf{u}_k \in \mathbf{U} - \mathbf{U}.$$

This lower bound for the terms in the sum (50) completes the proof. ∎

7. Local optimality. Given a solution w^h of the inclusion $\mathcal{T}(w^h) \in \mathcal{F}(w^h)$ corresponding to the first-order optimality system for the discrete control problem, we show in this section that w^h yields a local minimizer in (14) if $\|w^h - w^*\|$ is sufficiently small. Let $\mathcal{P}$ be the matrix sequence defined by

$$\mathcal{P} = (\mathbf{V}, \mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{S}, \mathbf{R}),$$

and let $\mathcal{B}^h(\mathcal{P}; \mathbf{x}, \mathbf{u})$ and $\mathcal{M}^h(\mathcal{P})$ be the quadratic form and set, defined in (46) and (48), respectively. Let $\mathcal{P}^\rho$ be any other matrix sequence with the property that

$$\|\mathcal{P} - \mathcal{P}^\rho\|_{L^\infty} \leq \rho.$$

LEMMA 7.1. *If* (47) *holds for some* $\bar{\alpha} > 0$, *then there exist positive constants* $\bar{\rho}$ *and* c, *independent of* h *and* $\bar{\alpha}$ *and depending only on* $\|\mathcal{P}\|_{L^\infty}$, *such that*

$$\mathcal{B}^h(\mathcal{P}^\rho; \mathbf{x}, \mathbf{u}) \geq (\bar{\alpha} - c\rho)\|\mathbf{u}\|_{L^2}^2 \quad \textit{for all } (\mathbf{x}, \mathbf{u}) \in \mathcal{M}^h(\mathcal{P}^\rho) \quad \textit{and} \quad 0 \leq \rho \leq \bar{\rho}. \tag{54}$$

Proof. Given any $(\mathbf{x}, \mathbf{u}) \in \mathcal{M}^h(\mathcal{P}^\rho)$, we have

$$\mathbf{x}_k' = \mathbf{A}_k^\rho \mathbf{x}_k + \mathbf{B}_k^\rho \mathbf{u}_k \mathbf{b}, \quad \mathbf{x}_0 = \mathbf{0}, \quad \mathbf{u}_k \in \mathbf{U} - \mathbf{U}. \tag{55}$$

Let $\mathbf{y}_k$ denote the solution to

$$\mathbf{y}_k' = \mathbf{A}_k \mathbf{y}_k + \mathbf{B}_k \mathbf{u}_k \mathbf{b}, \quad \mathbf{y}_0 = \mathbf{0}.$$

Hence, $(\mathbf{y}, \mathbf{u}) \in \mathcal{M}^h(\mathcal{P})$. Given any fixed $\bar{\rho} > 0$ and $\rho \leq \bar{\rho}$, we have

$$\|\mathbf{x}\|_{L^\infty} \leq c\|\mathbf{u}\|_{L^2}, \quad \|\mathbf{y}\|_{L^\infty} \leq c\|\mathbf{u}\|_{L^2}, \quad \text{and} \quad \|\mathbf{x} - \mathbf{y}\|_{L^\infty} \leq c\rho\|\mathbf{u}\|_{L^2}. \tag{56}$$

Since the proofs of these inequalities are similar, we focus on the first one. Taking the norm in (55) gives

$$\|\mathbf{x}_{k+1}\| \leq \|\mathbf{x}_k\| + ch(\|\mathbf{x}_k\| + \|\mathbf{u}_k\|).$$

Since $\mathbf{x}_0 = 0$, it follows that

$$\|\mathbf{x}_k\| \leq c \sum_{j=0}^{k-1} h\|\mathbf{u}_j\|.$$

Thinking of the last sum as a dot product between a vector whose components are all $\sqrt{h}$ and a vector whose jth component is $\sqrt{h}\|\mathbf{u}_j\|$, the Schwarz inequality implies that

$$\|\mathbf{x}_k\| \leq c \left(\sum_{j=0}^{k-1} h\|\mathbf{u}_j\|^2 \right)^2,$$

which yields the first inequality in (56).

Expanding $\mathcal{B}(\mathcal{P}^\rho)(\mathbf{x}, \mathbf{u})$ in a Taylor series around $\mathbf{x} = \mathbf{y}$ and utilizing (56), we have

$$\begin{aligned}
&\mathcal{B}^h(\mathcal{P}^\rho; \mathbf{x}, \mathbf{u}) \\
&\quad = \mathcal{B}^h(\mathcal{P}^\rho; \mathbf{y}, \mathbf{u}) + \nabla_x \mathcal{B}^h(\mathcal{P}^\rho; \mathbf{y}, \mathbf{u})(\mathbf{x} - \mathbf{y}) + \mathcal{B}^h(\mathcal{P}^\rho; \mathbf{x} - \mathbf{y}, \mathbf{u}) \\
&\quad = \mathcal{B}(\mathcal{P}; \mathbf{y}, \mathbf{u}) + \mathcal{B}(\mathcal{P}^\rho - \mathcal{P}; \mathbf{y}, \mathbf{u}) + \nabla_x \mathcal{B}(\mathcal{P}^\rho; \mathbf{y}, \mathbf{u})(\mathbf{x} - \mathbf{y}) + \mathcal{B}(\mathcal{P}^\rho; \mathbf{x} - \mathbf{y}, \mathbf{u}) \\
&\quad \geq \bar{\alpha} \|\mathbf{u}\|_{L^2}^2 - c\rho(\|\mathbf{y}\|_{L^\infty}^2 + \|\mathbf{u}\|_{L^2}^2) \\
&\quad \geq (\bar{\alpha} - c\rho)\|\mathbf{u}\|_{L^2}^2.
\end{aligned}$$

This completes the proof. ◻

LEMMA 7.2. *If the smoothness and coercivity conditions,* (19)*, part* (a)*, and* (20)*, part* (c)*, all hold, then there exist* $\bar{h}$ *and* $r > 0$ *with the property that any* w^h *satisfying* $\mathcal{T}(w^h) \in \mathcal{F}(w^h)$*, with* $\mathcal{T}$ *and* $\mathcal{F}$ *defined in* (27) *and* (28)*, is a strict local minimizer in* (14) *when* $\|w^h - w^*\|_{L^\infty} \leq r$ *and* $h \leq \bar{h}$.

Proof. Choose $\bar{h}$ according to Lemma 6.1 so that (47) holds for $\bar{\alpha} > 0$ and $h \leq \bar{h}$. Given w^h such that $\mathcal{T}(w^h) \in \mathcal{F}(w^h)$, the condition (47) almost implies that the second-order sufficient optimality condition (see [19, Cor. 6]) holds at w^h; the only discrepancy is that in the second-order sufficient optimality condition, the matrix sequence $\mathcal{P}$ associated with (47) is replaced by a nearby sequence $\mathcal{P}^h$ obtained by replacing $\mathbf{w}^*(t)$ in (9)–(10) with the components of w^h. Choose ρ small enough that $c\rho < \bar{\alpha}$ in (54) and choose $r \leq \beta$ and $\bar{h}$ smaller if necessary so that $\|\mathcal{P} - \mathcal{P}^h\|_{L^\infty} \leq \rho$ whenever $\|w^h - w^*\| \leq r$ and $h \leq \bar{h}$, in accordance with Lemma 5.1. This completes the proof. ◻

8. Proof of Theorem 2.1. We now collect results to prove Theorem 2.1 using Proposition 3.1 and the correspondence with the control problem described in section 3. Referring to Lemma 6.1, choose $\bar{h}$ small enough that $(\mathcal{F} - \mathcal{L})^{-1}$ is Lipschitz continuous with Lipschitz constant λ independent of $h \leq \bar{h}$. Choose ϵ small enough that $\epsilon\lambda < 1$. Choose r and $\bar{h}$ small enough that for the constant c in Lemma 5.1, we have $c(r + \bar{h}) \leq \epsilon$. Choose r and $\bar{h}$ smaller if necessary to satisfy the conditions of Lemma 7.2. Finally, choose $\bar{h}$ small enough that the distance estimated in Lemma 4.2 satisfies the condition

$$ch(h + \tau(\dot{u}^*; h)) \leq (1 - \lambda\epsilon)r/\lambda$$

whenever $h \leq \bar{h}$. All the conditions of Proposition 3.1 are satisfied and the estimate (26) is precisely the bound (22) of Theorem 2.1.

Remark 8.1. The proof techniques used in this paper are tailored to second-order convergence. In fact, in [34] where high-order convergence is established for unconstrained control problems, a slightly different approach is used involving a transformed adjoint system. Note though that for problems with control constraints, solutions often lose regularity at points where the constraints change from active to inactive, and the second-order convergence we obtain here is appropriate (and surprising as pointed out in the introduction) relative to the limited smoothness of the control.

9. Numerical examples. Some of the simplest Runge–Kutta schemes satisfying the conditions (19) and (20) are the implicit midpoint rule,

$$\mathbf{A} = [1/2], \quad \mathbf{b} = [1], \quad \boldsymbol{\sigma} = [1/2],$$

and the two-stage explicit midpoint rule,

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 1/2 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}. \tag{57}$$

Here $\mathbf{A}$ is the coefficient array for Runge–Kutta schemes, not the matrix $\mathbf{A}(t) = \nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t))$ in (9). The second scheme (57) is one member of the family of two-stage explicit schemes given by

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 1/(2\gamma) & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1-\gamma \\ \gamma \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}, \quad \gamma \in (0, 1].$$

In each of these schemes, we approximate one control value on each interval, the value at the midpoint of the interval. In the following two-stage explicit scheme, which satisfies (19) and (20), we obtain approximations to the values of the control at each grid point:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

An example of a very plausible two-stage scheme that is second-order accurate for ordinary differential equation, but which violates the condition (20), part (b), is the following explicit midpoint scheme:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 1/2 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}. \tag{58}$$

This scheme, like the previous example, tries to approximate the control at the grid points. As we saw in the introduction, this scheme (2) leads to discrete approximations that diverge from the solution to the continuous problem. It is interesting to note that for the scheme (58), one of the components of $\mathbf{b}$ vanishes. The transformation introduced in [30] and [34], to convert the discrete first-order optimality conditions (16)–(18) into a new system resembling a Runge–Kutta scheme applied to the continuous optimality conditions (5)–(7), also breaks down in exactly this same situation.

We also solve some test problems using the explicit midpoint scheme (57). The first test problem in [35] is

$$\begin{aligned} &\text{minimize} \quad \int_0^1 u(t)^2 + x(t)^2 \, dt \\ &\text{subject to} \quad \dot{x}(t) = u(t), \quad u(t) \le 1, \quad x(0) = \tfrac{1+3e}{2(1-e)}, \end{aligned} \tag{59}$$

with the optimal solution

$$u^*(t) = 1, \quad 0 \le t \le 1/2, \quad u^*(t) = \frac{e^t - e^{2-t}}{\sqrt{e}(1-e)}, \quad 1/2 \le t \le 1.$$

The L^∞ error for various choices of the mesh appears in Table 1. For the mesh on the left, the point of discontinuity lies at a grid point, while for the mesh at the right, the point of discontinuity is exactly between the grid points. Notice that the error decays to zero like h^2, according to Theorem 2.1, even though the optimal control lies in $W^{1,\infty}$, but not in $W^{2,p}$. More precisely, when we perform a least squares fit

TABLE 1
Results for test problem (59) *and the explicit midpoint rule* (57).

N	L^∞ Control error	N	L^∞ Control error
10	.001741757	15	.000268326
20	.000462070	25	.000103219
40	.000118823	45	.000033426
80	.000030130	85	.000009277
160	.000007000	165	.000002516
320	.000001717	325	.000000659

TABLE 2
Results for test problem (60) *and the explicit midpoint rule* (57).

N	L^∞ Control error $\kappa = 1$	L^∞ Control error $\kappa = 8$
10	.04897379	.01989437
20	.01448896	.01989437
40	.00389347	.01989437
80	.00100347	.01989437
160	.00025416	.01668540
320	.00006391	.00412669
640	.00001602	.00102489
1280	.00000401	.00025665

of the error to a function of the form ch^q, we obtain $q = 2.00$ for the left mesh and $q = 1.96$ for the right mesh. Normally, when we seek to approximate the solution to a problem with a discontinuous derivative, it is advantageous to place a grid point at the point of discontinuity. In this example, a smaller error is achieved when the point of discontinuity is between the grid points. Hence, the location of the grid points relative to the discontinuity in the optimal control is not very crucial.

The second test problem that we consider involves an integer parameter κ:

$$(60) \qquad \text{minimize } \; x(1) + \frac{1}{2}\int_0^1 u(t)^2 \, dt$$

$$\text{subject to} \quad \dot{x}_1(t) = x_2(t), \quad \dot{x}_2 = -(2\pi\kappa)^2 x_1(t) + u(t),$$
$$|u(t)| \le \tfrac{1}{4\pi\kappa}, \quad 0 \le t \le 1, \quad x_1(0) = x_2(0) = 0,$$

with the optimal solution

$$u^*(t) = \begin{cases} \frac{1}{2\pi\kappa}(\sin 2\pi\kappa t) & \text{if } |\sin(2\pi\kappa t)| \; \le \; \frac{1}{2}, \\ \frac{-1}{4\pi\kappa} & \text{if } \sin(2\pi\kappa t) \; < -\frac{1}{2}, \\ \frac{1}{4\pi\kappa} & \text{if } \sin(2\pi\kappa t) \; > \; \frac{1}{2}. \end{cases}$$

As κ increases, the number of oscillations and the total variation in the optimal solution increase. Moreover, the linearized operator $\mathcal{L}$ depends on the parameter κ, and the Lipschitz constant λ of $(\mathcal{F} - \mathcal{L})^{-1}$ is proportional to κ. Since the constant c of (22) is proportional to λ, due to (26), and since $\tau(\dot{u}^*; h)) \approx 4\kappa h$, the control error is proportional to κ^2 for small h. Hence, for large N, the error in Table 2 is about 64 times bigger for $\kappa = 8$, compared to the error for $\kappa = 1$.

10. Continuous extensions. The Runge–Kutta discretization (14) leads to an approximation to the continuous optimal control at a discrete set of points. We now show how to interpolate the discrete values in order to obtain an approximate control $u^I(t)$, $0 \le t \le 1$, for which the associated state variable $x^I(t)$ approximates the optimal state variable $x^*(t)$ at the grid points with an error similar to that of the discrete control.

If the vector $\boldsymbol{\sigma}$ contains both 0 and 1 as components, then u^I is obtained by continuous piecewise linear interpolation on each grid interval $[t_k, t_{k+1}]$ with a possible discontinuity in u^* at each grid point. If either 0 or 1 is not a component of $\boldsymbol{\sigma}$, then u^I is simply the continuous piecewise linear interpolant of the discrete values for the control. Since the discrete controls are all contained in U, it follows that $u^I(t) \in U$ for all $t \in [0,1]$. If the Runge–Kutta integration scheme is applied to the ordinary differential equation $\dot{x} = f(x,u)$ with the intermediate control values chosen to be those of u^I, then the resulting discrete state is precisely $\mathbf{x}^h$ since u^I has the same values as the discrete control $\mathbf{u}^h$ at the intermediate points in the integration scheme.

Returning to the analysis of the state residual in section 4, let us replace each superscript * with an I to obtain

$$\left|(\mathbf{x}_k^I)' - \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i^I, \mathbf{u}_{ki}^I)\right| \le ch(h + \tau_k(\dot{u}^I; h)), \tag{61}$$

where $\mathbf{u}_k^I$ is the same as $\mathbf{u}_k^h$, $\mathbf{x}_k^I$ denotes $x^I(t_k)$, and $\mathbf{y}_i^I$ denotes $\mathbf{y}_i(\mathbf{x}_k^I, \mathbf{u}_k^I)$. The constant c in (61) depends on the Lipschitz constant of u^I. Note though that this Lipschitz constant is bounded, independent of h, since u^* is Lipschitz continuous and the error estimate (22) holds.

It is well known (for example, see [7, Thm. 364B]) that in Runge–Kutta integration, the maximum error at the grid points on $[0,1]$ is bounded by h times the sum of the local errors (61). In other words,

$$\max_{0 \le k \le N} |\mathbf{x}_k^h - x^I(t_k)| \le ch\left(h + \sum_{k=0}^{N-1} \tau_k(\dot{u}^I; h)\right).$$

If the estimate

$$\sum_{k=0}^{N-1} \tau_k(\dot{u}^I; h) \le c(h + \tau(\dot{u}^*; h)) \tag{62}$$

holds, then the error in the continuous trajectory x^I has exactly the same form as the estimate (22) for the error in the discrete control.

We prove (62) in the case that both 0 and 1 are components of $\boldsymbol{\sigma}$, while the other case, in which either 0 or 1 are not components, is a small modification of this argument. Let us assume that the intermediate variables in the Runge–Kutta scheme have been rearranged so that

$$0 = \sigma_1 \le \sigma_2 \le \cdots \le \sigma_s = 1.$$

Obviously, $s \ge 2$ in this case. Since $\dot{u}^I$ is a piecewise constant function on $[t_k, t_{k+1}]$, we conclude that for any $t \in [t_k, t_{k+1}]$, there exists $i < j$ such that

$$\omega(\dot{u}^I, [t_k, t_{k+1}]; t, h) = \left|\frac{\mathbf{u}_{ki}^h - \mathbf{u}_{ki+}^h}{h(\sigma_i - \sigma_{i+})} - \frac{\mathbf{u}_{kj}^h - \mathbf{u}_{kj+}^h}{h(\sigma_j - \sigma_{j+})}\right|, \tag{63}$$

where $i+$ denotes the first $l > i$ for which $\sigma_l > \sigma_i$. Utilizing the estimate (22), we have for any $t \in [t_k, t_{k+1}]$,

$$\omega(\dot{u}^I, [t_k, t_{k+1}]; t, h) \le \left| \frac{\mathbf{u}^*_{ki} - \mathbf{u}^*_{ki+}}{h(\sigma_i - \sigma_{i+})} - \frac{\mathbf{u}^*_{kj} - \mathbf{u}^*_{kj+}}{h(\sigma_j - \sigma_{j+})} \right| + c(h + \tau(\dot{u}^*; h)).$$

Let t be a point where u^* is differentiable, and let us make the substitution

$$\mathbf{u}^*_{ki} - \mathbf{u}^*_{ki+} = h \int_{\sigma_i}^{\sigma_{i+}} \dot{u}^*(t_k + sh)\, ds$$

to obtain

$$\begin{aligned} &\left| \frac{\mathbf{u}^*_{ki} - \mathbf{u}^*_{ki+}}{h(\sigma_i - \sigma_{i+})} - \frac{\mathbf{u}^*_{kj} - \mathbf{u}^*_{kj+}}{h(\sigma_j - \sigma_{j+})} \right| \\ &\quad = \left| \tfrac{1}{\sigma_i - \sigma_{i+}} \int_{\sigma_i}^{\sigma_{i+}} (\dot{u}^*(t_k + sh) - \dot{u}^*(t))\, ds - \tfrac{1}{\sigma_j - \sigma_{j+}} \int_{\sigma_j}^{\sigma_{j+}} (\dot{u}^*(t_k + sh) - \dot{u}^*(t))\, ds \right| \\ &\quad \le c\omega(\dot{u}^*, [t_k, t_{k+1}]; t, 2h). \end{aligned}$$

After integrating over $t \in [t_k, t_{k+1}]$ and summing over k, we obtain

$$\sum_{k=0}^{N-1} \tau_k(\dot{u}^I; h) \le c(\tau(\dot{u}^*; 2h) + \tau(\dot{u}^*; h) + h) \le c(3\tau(\dot{u}^*; h) + h) \le c(\tau(\dot{u}^*; h) + h),$$

which completes the proof of (62). If either 0 or 1 is not a component of $\boldsymbol{\sigma}$, then the k indices on the right side of (63) may need to be replaced by either $k-1$ or $k+1$ since u^I is obtained by linear interpolating across the grid points. To summarize, we have the following theorem.

THEOREM 10.1. *If the hypotheses of Theorem* 2.1 *hold, then for* h *sufficiently small, the differential equation* $\dot{x} = f(x, u)$, $x(0) = a$, *has a solution* x^I *corresponding to the piecewise linear interpolant* u^I *of the discrete control* $\mathbf{u}^h$ *such that*

$$\max_{0 \le k \le N} |x^I(t_k) - x^*(t_k)| \le ch(h + \tau(\dot{u}^*; h)).$$

REFERENCES

[1] T. M. APOSTOL, *Mathematical Analysis*, Addison-Wesley, Reading, MA, 1957.

[2] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

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

[4] W. E. BOSARGE, JR., O. G. JOHNSON, R. S. MCKNIGHT, AND W. P. TIMLAKE, *The Ritz–Galerkin procedure for nonlinear control problems*, SIAM J. Numer. Anal., 10 (1973), pp. 94–111.

[5] P. BRUNOVSKÝ, *Regular synthesis for the linear quadratic optimal control problem with linear control constraints*, J. Differential Equations, 38 (1980), pp. 344–360.

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

[7] J. C. BUTCHER, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, New York, 1987.

[8] J. CULLUM, *Discrete approximations to continuous optimal control problems*, SIAM J. Control, 7 (1969), pp. 32–49.

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

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

[11] J. W. Daniel, *On the approximate minimization of functionals*, Math. Comp., 23 (1969), pp. 573–581.

[12] J. W. Daniel, *On the convergence of a numerical method in optimal control*, J. Optim. Theory Appl., 4 (1969), pp. 330–342.

[13] J. W. Daniel, *The Ritz–Galerkin method for abstract optimal control problems*, SIAM J. Control, 11 (1973), pp. 53–63.

[14] J. W. Daniel, *The Approximate Minimization of Functionals*, Wiley-Interscience, New York, 1983.

[15] A. L. Dontchev, *Error estimates for a discrete approximation to constrained control problems*, SIAM J. Numer. Anal., 18 (1981), pp. 500–514.

[16] A. L. Dontchev, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci. 52, Springer, New York, 1983.

[17] A. L. Dontchev, *An a priori estimate for discrete approximations in nonlinear optimal control*, SIAM J. Control Optim., 34 (1996), pp. 1315–1328.

[18] A. L. Dontchev, *Discrete approximations in optimal control*, in Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control, IMA Vol. Math. Appl. 78, Springer, New York, 1996, pp. 59–81.

[19] A. L. Dontchev and W. W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.

[20] A. L. Dontchev and W. W. Hager, *Lipschitzian stability for state constrained nonlinear optimal control*, SIAM J. Control Optim., 36 (1998), pp. 696–718.

[21] A. L. Dontchev and W. W. Hager, *A new approach to Lipschitz continuity in state constrained optimal control*, Systems Control Lett., 35 (1998), pp. 137–143.

[22] A. L. Dontchev and 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, and B. Yang, *Optimality, stability and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[24] A. L. Dontchev and I. Kolmanovsky, *On regularity of optimal control*, in Recent Developments in Optimization, Lecture Notes in Econom. and Math. Systems 429, Springer, Berlin, 1995, pp. 125–135.

[25] A. L. Dontchev and K. Malanowski, *A characterization of Lipschitzian stability in optimal control*, in Calculus of Variations and Optimal Control, Res. Notes Math. 411, Chapman and Hall/CRC, Boca Raton, FL, 2000, pp. 62–76.

[26] J. C. Dunn, *On L^2 sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.

[27] J. C. Dunn and T. Tian, *Variants of the Kuhn–Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.

[28] E. Farhi, *Runge–Kutta schemes applied to linear-quadratic optimal control problems*, in Mathematics and Mathematical Education, Bulgar. Akad. Nauk, Sofia, 1984, pp. 464–472.

[29] W. W. Hager, *The Ritz–Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.

[30] W. W. Hager, *Rate of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–471.

[31] W. W. Hager, *Convex control and dual approximations*, Control Cybernet., 8 (1979), pp. 1–22, 73–86.

[32] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.

[33] W. W. Hager, *Multiplier methods for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061–1080.

[34] W. W. Hager, *Runge–Kutta methods in optimal control and the transformed adjoint system*, Numer. Math., to appear.

[35] W. W. Hager and G. D. Ianculescu, *Dual approximations in optimal control*, SIAM J. Control Optim., 22 (1984), pp. 423–465.

[36] W. W. Hager and R. Rostamian, *Optimal coatings, bang-bang controls, and gradient techniques*, Optimal Control Appl. Methods, 8 (1987), pp. 1–20.

[37] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations,* I. *Nonstiff problems*, 2nd ed., Springer Ser. Comput. Math. 8, Springer-Verlag, Berlin, 1993.

[38] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations,* II. *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer Ser. Comput. Math., 14, Springer-Verlag, Berlin, 1996.

[39] F. Lempio and V. M. Veliov, *Discrete approximations to differential inclusions*, GAMM Mitteilungen, 21 (1998), pp. 105–135.

[40] K. Malanowski, C. Büskens, and H. Maurer, *Convergence of approximations to nonlinear optimal control problems*, in Mathematical Programming with Data Perturbations, A. V. Fiacco, ed., Lecture Notes in Pure and Appl. Math 195, Marcel Dekker, New York, 1997, pp. 253–284.

[41] B. Mordukhovich, *On difference approximations of optimal control systems*, J. Appl. Math. Mech., 42 (1978), pp. 452–461.

[42] E. Polak, *A historical survey of computations methods in optimal control*, SIAM Rev., 15 (1973), pp. 553–548.

[43] E. Polak, *Optimization: Algorithms and Consistent Approximation*, Springer, New York, 1997.

[44] R. Pytlak, *Runge–Kutta based procedure for the optimal control of differential-algebraic equations*, J. Optim. Theory Appl., 97 (1998), pp. 675–705.

[45] V. H. Schulz, *A direct PRSQP method for path planning of satellite mounted robots*, Z. Angew. Math. Mech., 76 (1996), pp. 17–20.

[46] A. Schwartz and E. Polak, *Consistent approximations for optimal control problems based on Runge–Kutta integration*, SIAM J. Control Optim., 34 (1996), pp. 1235–1269.

[47] B. Sendov and V. A. Popov, *The Averaged Moduli of Smoothness*, John Wiley, New York, 1988.

[48] M. Steinbach, *Stuctured interior point SQP methods in optimal control*, Z. Angew. Math. Mech., 76 (1996), pp. 59–62.

[49] O. von Stryk, *Numerical solution of optimal control problems by direct collocation*, in Optimal Control, Calculus of Variations, Optimal Control Theory and Numerical Methods, Internat. Ser. Numer. Math. 111, Birkhäuser, Basel, 1993, pp. 129–143.

[50] H. Sussmann, *Bounds on the number of switchings for trajectories of piecewise analytic vector fields*, J. Differential Equations, 43 (1982), pp. 399–418.

[51] V. Veliov, *Second-order discrete approximation to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439–451.

[52] V. M. Veliov, *On the time-discretization of control systems*, SIAM J. Control Optim., 35 (1997), pp. 1470–1486.

[53] V. Zeidan, *Sufficient conditions for variational problems with variable endpoints: Coupled points*, Appl. Math. Optim., 27 (1993), pp. 191–209.