Numerical Analysis in Optimal Control

William W. Hager

Abstract. In this paper we explain and exemplify how one goes about analyzing the convergence of algorithms and discrete approximations in optimal control.

1. Introduction

The techniques used to analyze the convergence of penalty methods, multiplier methods, sequential quadratic programming methods, and discrete approximations in optimal control are closely related. In each case, we wish to approximate a local minimizer $\mathbf{w}^*$ of an optimization problem, where the approximation is the solution to a problem of the form:

$$\text{Find } \mathbf{w} \in \mathcal{X} \text{ such that } \mathcal{T}(\mathbf{w}) \in \mathcal{F}(\mathbf{w}). \tag{1}$$

Here $\mathcal{X}$ is a Banach space, $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$, $\mathcal{Y}$ is a linear normed space, and $\mathcal{F} : \mathcal{X} \to 2^{\mathcal{Y}}$. Think of (1) as the first-order optimality system associated with the numerical approximating problem.

Typically, the solution $\mathbf{w}^*$ of the original optimization problem is not a solution of (1), rather it is an approximate solution. Let $\boldsymbol{\delta}$ be chosen as small as possible so that

$$\mathcal{T}(\mathbf{w}^*) + \boldsymbol{\delta} \in \mathcal{F}(\mathbf{w}^*). \tag{2}$$

Given that $\mathbf{w}^*$ is almost a solution of (1), we try to show that (1) has a solution $\mathbf{w}$ close to $\mathbf{w}^*$ which satisfies an estimate of the form $\|\mathbf{w} - \mathbf{w}^*\| \le c\|\boldsymbol{\delta}\|$, where c is a constant independent of $\boldsymbol{\delta}$ for $\boldsymbol{\delta}$ sufficiently small, and $\|\cdot\|$ denotes the norm in the appropriate space.

An existence result for (1), together with an estimate for the distance to $\mathbf{w}^*$, is gotten from a generalization of the implicit function theorem. In this generalization, the usual surjectivity property for the derivative of $\mathcal{T}$ at $\mathbf{w}^*$ is replaced by a Lipschitz property for an associated linearized problem of the form:

$$\text{Find } \mathbf{w} \in \mathcal{X} \text{ such that } \mathcal{L}(\mathbf{w}) + \boldsymbol{\pi} \in \mathcal{F}(\mathbf{w}). \tag{3}$$

Here $\mathcal{L}$ is a linear operator and $\boldsymbol{\pi} \in \mathcal{Y}$ stands for a "parameter."

This work, supported by the National Science Foundation, was presented at the Conference on Optimal Control of Complex Structures, June 4–10, 2000, Oberwolfach, Germany, organized by K.-H. Hoffmann, I. Lasiecka, G. Leugering, J. Sprekels and F. Troeltzsch.

In the typical first-order Taylor expansion, we would approximate $\mathcal{T}$ by $\mathcal{T}(\mathbf{w}^*) + \mathcal{T}'(\mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)$, in which case $\mathcal{L}$ would be $\mathcal{T}'(\mathbf{w}^*)$. If (3) has a unique solution for $\boldsymbol{\pi}$ near $\boldsymbol{\pi}^* = \mathcal{T}(\mathbf{w}^*) - \mathcal{T}'(\mathbf{w}^*)(\mathbf{w}^*)$ satisfying

$$\|\mathbf{w}_1 - \mathbf{w}_2\| \leq \lambda \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\|,$$

where $\mathbf{w} = \mathbf{w}_i$ is the solution of (3) corresponding to $\boldsymbol{\pi} = \boldsymbol{\pi}_i$, then under suitable assumptions, (1) has a solution $\mathbf{w}$, and the distance from $\mathbf{w}$ to $\mathbf{w}^*$ is very nearly bounded by $\lambda\|\boldsymbol{\delta}\|$. The precise estimate, given shortly, involves an additional factor $1/(1 - \lambda\epsilon)$ where ϵ is typically small. The bound $\lambda\|\boldsymbol{\delta}\|$ for the distance from $\mathbf{w}$ to $\mathbf{w}^*$ yields an error estimate for the numerical algorithm that (1) represents.

2. Abstract Estimate

The following result, given in a slightly more general form in [2, Thm. 3.1], is a version of the implicit function theorem for inclusions alluded to in the previous section. See [3, 4, 5, 10] for other related results.

Theorem 2.1. *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 $B_r(\mathbf{w}^*)$ for some $\mathbf{w}^* \in \mathcal{X}$ and $r > 0$, where $B_r(\mathbf{w}^*)$ is the ball with center $\mathbf{w}^*$ and radius r. Suppose that the following conditions hold for some $\boldsymbol{\delta} \in \mathcal{Y}$ and scalars ϵ, λ, and $\sigma > 0$:*

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

If $\epsilon\lambda < 1$, $\epsilon r \leq \sigma$, $\|\boldsymbol{\delta}\| \leq \sigma$, and

$$\|\boldsymbol{\delta}\| \leq (1 - \lambda\epsilon)r/\lambda,$$

then there exists a unique $\mathbf{w} \in B_r(\mathbf{w}^*)$ such that $\mathcal{T}(\mathbf{w}) \in \mathcal{F}(\mathbf{w})$. Moreover, we have the estimate

$$\|\mathbf{w} - \mathbf{w}^*\| \leq \frac{\lambda}{1 - \lambda\epsilon}\|\boldsymbol{\delta}\|. \tag{4}$$

Proof. Let us define $\Phi(\mathbf{w}) = (\mathcal{F} - \mathcal{L})^{-1}(\mathcal{T}(\mathbf{w}) - \mathcal{L}(\mathbf{w}))$. By a Taylor expansion around $\mathbf{w}^*$, with integral remainder term, we have

$$\mathcal{T}(\mathbf{w}) - \mathcal{L}(\mathbf{w}) = \mathcal{T}(\mathbf{w}^*) - \mathcal{L}(\mathbf{w}^*) + \int_0^1 (\nabla\mathcal{T}(s\mathbf{w} + (1-s)\mathbf{w}^*) - \mathcal{L})\, ds\, (\mathbf{w} - \mathbf{w}^*).$$

Hence, (P2) implies that $\|\mathcal{T}(\mathbf{w}) - \mathcal{L}(\mathbf{w}) - \boldsymbol{\pi}^*\| \le \epsilon r$ for all $\mathbf{w} \in B_r(\mathbf{w}^*)$. By (P3), it follows that for all $\mathbf{w}_1, \mathbf{w}_2 \in B_r(\mathbf{w}^*)$,

$$\begin{aligned} \|\Phi(\mathbf{w}_1) - \Phi(\mathbf{w}_2)\| &= \|(\mathcal{F} - \mathcal{L})^{-1}(\mathcal{T} - \mathcal{L})(\mathbf{w}_1) - (\mathcal{F} - \mathcal{L})^{-1}(\mathcal{T} - \mathcal{L})(\mathbf{w}_2)\| \\ &\le \lambda\|(\mathcal{T} - \mathcal{L})(\mathbf{w}_1) - (\mathcal{T} - \mathcal{L})(\mathbf{w}_2)\| \\ &\le \lambda\epsilon\|\mathbf{w}_1 - \mathbf{w}_2\|. \end{aligned}$$

Since $\lambda\epsilon < 1$, Φ is a contraction on $B_r(\mathbf{w}^*)$. Since $\|\boldsymbol{\delta}\| \le \sigma$, we conclude that $(\mathcal{T} - \mathcal{L})(\mathbf{w}^*) + \boldsymbol{\delta} \in B_\sigma(\boldsymbol{\pi}^*)$. By (P3), $(\mathcal{F} - \mathcal{L})^{-1}$ is single-valued on $B_\sigma(\boldsymbol{\pi}^*)$, and by (P1) we have

$$\mathbf{w}^* = (\mathcal{F} - \mathcal{L})^{-1}[(\mathcal{T} - \mathcal{L})(\mathbf{w}^*) + \boldsymbol{\delta}].$$

It follows from (P2) and (P3) that

$$\begin{aligned} \|\Phi(\mathbf{w}) - \mathbf{w}^*\| &= \|(\mathcal{F} - \mathcal{L})^{-1}[(\mathcal{T} - \mathcal{L})(\mathbf{w})] - (\mathcal{F} - \mathcal{L})^{-1}[(\mathcal{T} - \mathcal{L})(\mathbf{w}^*) + \boldsymbol{\delta}]\| \\ &\le \lambda\|(\mathcal{T} - \mathcal{L})(\mathbf{w}) - (\mathcal{T} - \mathcal{L})(\mathbf{w}^*) - \boldsymbol{\delta}\| \\ &\le \lambda(\epsilon\|\mathbf{w} - \mathbf{w}^*\| + \|\boldsymbol{\delta}\|) \qquad (5) \\ &\le \lambda(\epsilon r + \|\boldsymbol{\delta}\|) \end{aligned}$$

for all $\mathbf{w} \in B_r(\mathbf{w}^*)$. The condition $\lambda\|\boldsymbol{\delta}\|/(1-\epsilon\lambda) \le r$ implies that $\lambda(\epsilon r + \|\boldsymbol{\delta}\|) \le r$, and hence, $\|\Phi(\mathbf{w}) - \mathbf{w}^*\| \le r$. Since Φ maps $B_r(\mathbf{w}^*)$ into itself and Φ is a contraction on $B_r(\mathbf{w}^*)$, the contraction mapping principle yields the existence of a unique fixed point $\mathbf{w} \in B_r(\mathbf{w}^*)$. Since $\|\Phi(\mathbf{w}) - \mathbf{w}^*\| = \|\mathbf{w} - \mathbf{w}^*\|$ for this fixed point, (5) gives (4). □

Theorem 2.1 says roughly

$$\textbf{Consistency} + \textbf{Stability} \Rightarrow \textbf{Convergence},$$

where consistency is assumption (P1) and the bounds on the norm of $\boldsymbol{\delta}$, stability is assumption (P3) and the bound on the Lipschitz constant λ for the linearization, and convergence is (4).

3. Penalty Methods

We first illustrate the analysis using the penalty approximation to the following control problem:

$$\begin{aligned} &\text{minimize } C(\mathbf{x}, \mathbf{u}) = \int_0^1 \varphi(\mathbf{x}(t), \mathbf{u}(t))dt \qquad (6) \\ &\text{subject to } \quad \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{u}(t) \in U \quad \text{a. e. } t \in [0, 1], \\ &\qquad\qquad \mathbf{x}(0) = \mathbf{a}, \quad \mathbf{x} \in W^{1,\infty}, \quad \mathbf{u} \in L^\infty, \end{aligned}$$

where the state $\mathbf{x}(t) \in \mathbf{R}^n$, $\dot{\mathbf{x}}$ stands for $\frac{d}{dt}\mathbf{x}$, the control $\mathbf{u}(t) \in \mathbf{R}^m$, $\mathbf{f} : \mathbf{R}^n \times \mathbf{R}^m \mapsto \mathbf{R}^n$, $\varphi : \mathbf{R}^n \times \mathbf{R}^m \mapsto \mathbf{R}$, and $U \subset \mathbf{R}^m$ is closed and convex. Of course, L^p denotes the usual Lebesgue space of measurable functions with p-th power integrable, and $W^{m,p}$ is the Sobolev space consisting of vector-valued functions whose j-th

derivative lies in L^p for all $0 \le j \le m$. Assume that (6) has a local minimizer $(\mathbf{x}^*, \mathbf{u}^*)$ and that φ and $\mathbf{f}$ are twice continuously differentiable.

Enforcing the differential equation constraint with a quadratic penalty term involving a "large" penalty parameter τ, we are led to the following approximating problem:

$$\begin{aligned} &\text{minimize } C(\mathbf{x},\mathbf{u}) + \frac{\tau}{2}\langle \mathbf{f}(\mathbf{x},\mathbf{u}) - \dot{\mathbf{x}}, \mathbf{f}(\mathbf{x},\mathbf{u}) - \dot{\mathbf{x}}\rangle \qquad (7)\\ &\text{subject to} \quad \mathbf{u}(t) \in U \quad \text{a .e. } t \in [0,1],\\ &\qquad\qquad \mathbf{x}(0) = \mathbf{a}, \quad \mathbf{x} \in W^{1,\infty}, \quad \mathbf{u} \in L^\infty. \end{aligned}$$

Instead of studying (7) directly, we examine the first-order optimality system associated with (7):

$$\dot{\boldsymbol{\psi}} + \nabla_x H(\mathbf{x},\mathbf{u},\boldsymbol{\psi}) = \mathbf{0}, \quad \boldsymbol{\psi}(1) = \mathbf{0}, \qquad (8)$$

$$\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x},\mathbf{u}) + \boldsymbol{\psi}/\tau = \mathbf{0}, \quad \mathbf{x}(0) = \mathbf{a}, \qquad (9)$$

$$\nabla_u (H(\mathbf{x}(t),\mathbf{u}(t),\boldsymbol{\psi}(t))(\mathbf{v} - \mathbf{u}(t)) \ge 0 \quad \text{for all } \mathbf{v} \in U. \qquad (10)$$

Here H is the Hamiltonian defined by $H(\mathbf{x},\mathbf{u},\boldsymbol{\psi}) = \varphi(\mathbf{x},\mathbf{u}) + \boldsymbol{\psi}^{\mathsf{T}}\mathbf{f}(\mathbf{x},\mathbf{u})$. The first two equations combine to give the usual Euler equation describing a minimizer of (7) over $\mathbf{x}$, assuming $\mathbf{u}$ is fixed. The last inequality describes a minimizer over $\mathbf{u}$, assuming $\mathbf{x}$ is fixed. Letting $\mathbf{w}$ denote the triple $(\mathbf{x},\mathbf{u},\boldsymbol{\psi})$, the system (8)–(10) of equalities and inequalities corresponds to the abstract inclusion (1). Note that the inequality (10) ie equivalent to the inclusion $\nabla_u H(\mathbf{x},\mathbf{u},\boldsymbol{\psi}) \in \mathcal{N}(\mathbf{u})$ where

$$\mathcal{N}(\mathbf{u}) = \{\boldsymbol{\chi} \in L^\infty : \langle \boldsymbol{\chi}, \mathbf{v} - \mathbf{u}\rangle \ge 0 \text{ for all } \mathbf{v} \in L^\infty, \mathbf{v}(t) \in U \text{ a. e. } t \in [0,1]\}.$$

If $\boldsymbol{\psi}^*$ is the costate variable given by the Pontryagin minimum principle, then $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\psi}^*)$ does not satisfy (9) due to the $\boldsymbol{\psi}/\tau$ term. That is, $\mathbf{x}^*$ and $\mathbf{u}^*$ satisfy the state equation $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},\mathbf{u})$, not (9). Hence, the term $\boldsymbol{\psi}^*/\tau$ would be put in the $\boldsymbol{\delta}$ of (2): $\boldsymbol{\delta} = (\mathbf{0}, -\boldsymbol{\psi}^*, \mathbf{0})/\tau$. With this choice for $\boldsymbol{\delta}$, we have $\mathcal{T}(\mathbf{w}^*) + \boldsymbol{\delta} \in \mathcal{F}(\mathbf{w}^*)$.

To apply Theorem 2.1, we need to analyze a linearization of (8)–(10). The linearization is gotten by neglecting the penalty term (this term is small when the penalty is large), and differentiating the other terms. More precisely, the linearized problem is the following:

$$\begin{aligned} \dot{\boldsymbol{\psi}} + \mathbf{A}^{\mathsf{T}}\boldsymbol{\psi} + \mathbf{Q}\mathbf{x} + \mathbf{S}\mathbf{u} + \boldsymbol{\alpha} &= \mathbf{0}, \quad \boldsymbol{\psi}(1) = \mathbf{0},\\ L(\mathbf{x},\mathbf{u}) + \boldsymbol{\beta} &= \mathbf{0}, \quad \mathbf{x}(0) = \mathbf{a},\\ \mathbf{B}^{\mathsf{T}}\boldsymbol{\psi} + \mathbf{S}^{\mathsf{T}}\mathbf{x} + \mathbf{R}\mathbf{u} + \boldsymbol{\gamma} &\in \mathcal{N}(\mathbf{u}), \end{aligned}$$

where $\boldsymbol{\pi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the parameter, $L(\mathbf{x},\mathbf{u}) = \dot{\mathbf{x}} - \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{u}$ is the linearized system dynamics, and

$$\mathbf{A}(t) = \nabla_x \mathbf{f}(\mathbf{x}^*(t),\mathbf{u}^*(t)), \quad \mathbf{B}(t) = \nabla_u \mathbf{f}(\mathbf{x}^*(t),\mathbf{u}^*(t)),$$

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

To apply Theorem 2.1, we need to verify (P3), which amounts to proving that the linearized problem has a solution depending Lipschitz continuously on the parameter. A natural space for $\mathbf{w} = (\mathbf{x}, \mathbf{u}, \boldsymbol{\psi})$ is $\mathcal{X} = W_0^{1,\infty} \times L^\infty \times W_1^{1,\infty}$, where

$$W_0^{1,\infty} = \{\mathbf{x} \in W^{1,\infty} : \mathbf{x}(0) = \mathbf{a}\} \quad \text{and} \quad W_1^{1,\infty} = \{\boldsymbol{\psi} \in W^{1,\infty} : \boldsymbol{\psi}(1) = \mathbf{0}\}.$$

Hence, a natural space for the image of $\mathcal{T}$ is $\mathcal{Y} = L^\infty$, and (P3) amounts to a regularity property for the linearized problem: For each $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2 \in L^\infty$, there exist associated solutions, $(\mathbf{x}_1, \mathbf{u}_1, \boldsymbol{\psi}_1)$ and $(\mathbf{x}_2, \mathbf{u}_2, \boldsymbol{\psi}_2)$ respectively, of the linearized problem such that

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_{W^{1,\infty}} + \|\mathbf{u}_1 - \mathbf{u}_2\|_{L^\infty} + \|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2\|_{W^{1,\infty}} \leq \lambda \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\|_{L^\infty}.$$

It turns out that this Lipschitz property holds when the matrices in the linearized problem possess a coercivity property (that also arises in second-order sufficient optimality conditions): There exists a constant $\alpha > 0$ such that

$$\mathcal{B}(\mathbf{x}, \mathbf{u}) = \langle \mathbf{Q}\mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{S}\mathbf{u}, \mathbf{x} \rangle + \langle \mathbf{R}\mathbf{u}, \mathbf{u} \rangle \;\geq\; \alpha \|\mathbf{u}\|_{L^2}^2 \quad \text{for all } (\mathbf{x}, \mathbf{u}) \in \mathcal{M},$$

where

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

Notice that the coercivity condition is formulated in L^2 spaces while the original control problem is formulated in L^∞ spaces. In the literature, this difference in spaces is called the 2-norm discrepancy. We need to formulate the original problem in L^∞, to ensure continuity of the functions defining the problem, but the coercivity condition should be formulated in L^2, ensuring Lipschitz stability for the linearized problem. For a proof of Lipschitz stability for the linearized control problem, see [10]

Next, we verify the conditions of Theorem 2.1. First, choose ϵ small enough that $\epsilon\lambda < 1$; then choose r small enough and τ large enough that (P2) holds; finally, choose τ large enough that

$$\|\boldsymbol{\delta}\| = \|\boldsymbol{\psi}^*\|_{L^\infty}/\tau \leq (1 - \epsilon\lambda) r/\lambda.$$

Since σ is $+\infty$, all the assumption of Theorem 2.1 hold. Hence, (8)–(10) has a solution $(\mathbf{x}_\tau, \mathbf{u}_\tau, \boldsymbol{\psi}_\tau)$ and

$$\|\mathbf{x}_\tau - \mathbf{x}^*\|_{W^{1,\infty}} + \|\mathbf{u}_\tau - \mathbf{u}^*\|_{L^\infty} + \|\boldsymbol{\psi}_\tau - \boldsymbol{\psi}^*\|_{W^{1,\infty}} \leq \frac{\lambda}{\tau(1 - \epsilon\lambda)} \|\boldsymbol{\psi}^*\|_{L^\infty}.$$

As τ tends to infinity, the solution of the penalized problem approaches the original local minimizer.

In the final phase of the analysis, it should be shown that this solution of (8)–(10) is a local minimizer of (7). This is done by expanding the cost function in a Taylor series. The first-order terms either vanish by (8) or are nonnegative by (10), and the second-order term is positive when the coercivity condition holds

(see [10, Thm. 3] for the details). Penalty methods applied to terminal constraints are studied in [11].

4. SQP Methods

If $(\mathbf{x}_k, \mathbf{u}_k, \boldsymbol{\psi}_k)$ is an approximation to a solution of the control problem (6), then the next SQP iterate $(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}, \boldsymbol{\psi}_{k+1})$ is a solution, and the associated costate variable, for the linear-quadratic problem

$$\text{minimize } \langle \nabla_x \varphi_k, \mathbf{x} - \mathbf{x}_k \rangle + \langle \nabla_u \varphi_k, \mathbf{u} - \mathbf{u}_k \rangle + \frac{1}{2} \mathcal{B}_k(\mathbf{x} - \mathbf{x}_k, \mathbf{u} - \mathbf{u}_k) \qquad (11)$$

$$\text{subject to} \quad L_k(\mathbf{x} - \mathbf{x}_k, \mathbf{u} - \mathbf{u}_k) = \mathbf{f}_k - \dot{\mathbf{x}}_k, \quad \mathbf{u}(t) \in U \quad \text{a. e. } t \in [0,1],$$
$$\mathbf{x}(0) = \mathbf{a}, \quad \mathbf{x} \in W^{1,\infty}, \quad \mathbf{u} \in L^\infty,$$

where the k subscript means that the associated expression is evaluated at $\mathbf{x}_k$, $\mathbf{u}_k$, and $\boldsymbol{\psi}_k$. A bit more smoothness is needed in this section; for example, φ and $\mathbf{f} \in C^3$. As with the penalty method, we apply Theorem 2.1 to the first-order optimality conditions for (11). These conditions are the following:

$$\begin{aligned} \dot{\boldsymbol{\psi}} + \mathbf{A}_k^{\mathsf{T}} \boldsymbol{\psi} + \nabla_x \varphi_k + \mathbf{Q}_k(\mathbf{x} - \mathbf{x}_k) + \mathbf{S}_k(\mathbf{u} - \mathbf{u}_k) &= \mathbf{0}, & \boldsymbol{\psi}(1) &= \mathbf{0}, \\ L_k(\mathbf{x} - \mathbf{x}_k, \mathbf{u} - \mathbf{u}_k) &= \mathbf{f}_k - \dot{\mathbf{x}}_k, & \mathbf{x}(0) &= \mathbf{a}, \\ \mathbf{B}_k^{\mathsf{T}} \boldsymbol{\psi} + \nabla_u \varphi_k + \mathbf{S}_k^{\mathsf{T}}(\mathbf{x} - \mathbf{x}_k) + \mathbf{R}_k(\mathbf{u} - \mathbf{u}_k) &\in \mathcal{N}(\mathbf{u}). \end{aligned}$$

The matrices here are the same as those of Section 3 except that they are evaluated at $\mathbf{w}_k = (\mathbf{x}_k, \mathbf{u}_k, \boldsymbol{\psi}_k)$ instead of at $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\psi}^*)$.

The linearized problem is exactly the same as that used in the previous section; hence, when the coercivity property is satisfied, (P3) of Theorem 2.1 holds with $\sigma = +\infty$. By taking $\mathbf{w}_k$ close to $\mathbf{w}^*$, we can make $\|\nabla \mathcal{T}(\mathbf{w}) - \mathcal{L}\|$ in (P2) as small as we like. To evaluate the $\boldsymbol{\delta}$ of (2), we insert $\mathbf{w} = \mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\psi}^*)$ in the left side of the first-order optimality system; by inspection, we see that (P1) holds for the following choice:

$$\boldsymbol{\delta} = -\begin{pmatrix} \dot{\boldsymbol{\psi}}^* + \nabla_x H(\mathbf{x}_k, \mathbf{u}_k, \boldsymbol{\psi}^*) + \mathbf{Q}_k(\mathbf{x}^* - \mathbf{x}_k) + \mathbf{S}_k(\mathbf{u}^* - \mathbf{u}_k) \\ \dot{\mathbf{x}}^* - \mathbf{f}_k - \mathbf{A}_k(\mathbf{x}^* - \mathbf{x}_k) - \mathbf{B}_k(\mathbf{u}^* - \mathbf{u}_k) \\ \nabla_u H(\mathbf{x}_k, \mathbf{u}_k, \boldsymbol{\psi}^*) + \mathbf{S}_k^{\mathsf{T}}(\mathbf{x}^* - \mathbf{x}_k) + \mathbf{R}_k(\mathbf{u}^* - \mathbf{u}_k) - \nabla_u H(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\psi}^*) \end{pmatrix}$$

By Theorem 2.1, the first-order optimality system has a solution

$$\mathbf{w}_{k+1} = (\mathbf{x}_{k+1}, \mathbf{u}_{k+1}, \boldsymbol{\psi}_{k+1}),$$

and the distance from $\mathbf{w}_{k+1}$ to $\mathbf{w}^*$ is bounded by a constant times $\|\boldsymbol{\delta}\|$. Expanding the terms of $\boldsymbol{\delta}$ in a Taylor series around $\mathbf{x}^*$, $\mathbf{u}^*$, and $\boldsymbol{\psi}^*$, everything cancels but the quadratic terms to give us the following estimate:

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_{W^{1,\infty}} + \|\mathbf{u}_{k+1} - \mathbf{u}^*\|_{L^\infty} + \|\boldsymbol{\psi}_{k+1} - \boldsymbol{\psi}^*\|_{W^{1,\infty}} \le c \|\mathbf{w}_k - \mathbf{w}^*\|_{L^\infty}^2,$$

where c is independent of $\mathbf{w}_k = (\mathbf{x}_k, \mathbf{u}_k, \boldsymbol{\psi}_k)$ in a neighborhood of $\mathbf{w}^*$. In [7] we analyze problems that also include inequality control constraints and endpoint constraints on the state.

5. Discrete Approximations

For simplicity, we consider the discretization of the following unconstrained control problem:

$$\begin{aligned} &\text{minimize } C(\mathbf{x}(1)) \qquad (12)\\ &\text{subject to } \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \quad \text{a. e. } t \in [0,1],\\ &\qquad\qquad \mathbf{x}(0) = \mathbf{a}, \quad \mathbf{x} \in W^{1,\infty}, \quad \mathbf{u} \in L^\infty, \end{aligned}$$

where $C : \mathbf{R}^n \mapsto \mathbf{R}$. We study control constrained problems in [5, 8], state constrained problems in [2], and mixed control/state constraints in [6]. Suppose the differential equation in (12) is solved 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 [1] with coefficients a_{ij} and b_i, $1 \le i, j \le s$, is given by

$$\mathbf{x}'_k = \sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \qquad (13)$$

where

$$\mathbf{y}_i = \mathbf{x}_k + h \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \quad 1 \le i \le s, \qquad (14)$$

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

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

In (13) and (14), $\mathbf{y}_j$ 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:

$$\begin{aligned} &\text{minimize } C(\mathbf{x}_N) \qquad (15)\\ &\text{subject to } \mathbf{x}'_k = \textstyle\sum_{i=1}^{s} b_i \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki}), \quad \mathbf{x}_0 = \mathbf{a},\\ &\qquad\qquad \mathbf{y}_i = \mathbf{x}_k + h \textstyle\sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj}), \quad 1 \le i \le s, \quad 0 \le k \le N-1. \end{aligned}$$

We apply Theorem 2.1 to the first-order optimality system (Kuhn-Tucker conditions) associated with (15). Suppose that a multiplier $\boldsymbol{\lambda}_i$ is introduced for the i-th intermediate equation (14) in addition to the multiplier $\boldsymbol{\psi}_{k+1}$ for the

equation (13). Taking into account these additional multipliers, the Kuhn-Tucker conditions are the following:

$$\boldsymbol{\psi}_k - \boldsymbol{\psi}_{k+1} = \sum_{i=1}^{s} \boldsymbol{\lambda}_i, \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N), \tag{16}$$

$$h \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})^{\mathsf{T}} (b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij} \boldsymbol{\lambda}_i) = \boldsymbol{\lambda}_j, \tag{17}$$

$$\nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})^{\mathsf{T}} (b_j \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} a_{ij} \boldsymbol{\lambda}_i) = \mathbf{0}, \tag{18}$$

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

To apply Theorem 2.1, we should insert the continuous solution in these discrete first-order conditions and estimate a residual. Note though that the discrete first-order conditions seem to have no connection to the continuous first-order condition, the Pontryagin minimum principle. However, we first showed in [12] and more recently in [9], that when $b_j \neq 0$ for each j, the first-order conditions make more sense (and are more useful) when reformulated in terms of the variables $\boldsymbol{\chi}_j$ defined by

$$\boldsymbol{\chi}_j = \boldsymbol{\psi}_{k+1} + \sum_{i=1}^{s} \frac{a_{ij}}{b_j} \boldsymbol{\lambda}_i, \quad 1 \le j \le s. \tag{19}$$

With this definition, (16) and (17) are equivalent to the following scheme:

$$\boldsymbol{\psi}_{k+1} = \boldsymbol{\psi}_k - h \sum_{i=1}^{s} b_i \nabla_x \mathbf{f}(\mathbf{y}_i, \mathbf{u}_{ki})^{\mathsf{T}} \boldsymbol{\chi}_i, \quad \boldsymbol{\psi}_N = \nabla C(\mathbf{x}_N), \tag{20}$$

$$\boldsymbol{\chi}_i = \boldsymbol{\psi}_k - h \sum_{j=1}^{s} \bar{a}_{ij} \nabla_x \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})^{\mathsf{T}} \boldsymbol{\chi}_j, \quad \bar{a}_{ij} = \frac{b_i b_j - b_j a_{ji}}{b_i}. \tag{21}$$

This is a Runge-Kutta scheme applied to the adjoint equation, but the coefficients of this scheme are typically different from those of the original scheme.

This reformulation of the first-order optimality system is important not only for the analysis of the discretization, but also for numerical computations since it provides an efficient way to compute the gradient of the discrete cost function with respect to the control. 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 $C(\mathbf{x}_N)$ of the discrete cost function associated with these controls. From the results of [13], we have

$$\nabla_{u_{kj}} C(\mathbf{u}) = h b_j \nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})^{\mathsf{T}} \boldsymbol{\chi}_j, \tag{22}$$

where the intermediate values for the discrete state and costate variables are gotten by first solving the discrete state equations (13) and (14), for $k = 0, 1, \ldots, N-1$, using the given values for the controls, and then using these computed values for both the state and intermediate variables in (20) and (21) when computing the

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	$\sum b_i c_i^3 = \frac{1}{4}, \quad \sum d_i^3/b_i^2 = \frac{1}{4}, \quad \sum b_i c_i a_{ij} d_j/b_j = \frac{5}{24}, \quad \sum c_i d_i^2/b_i = \frac{1}{12},$
	$\sum d_i a_{ij} c_j = \frac{1}{24}, \quad \sum b_i c_i a_{ij} c_j = \frac{1}{8}, \quad \sum d_i c_i^2 = \frac{1}{12}, \quad \sum d_i a_{ij} d_j/b_j = \frac{1}{8}$

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

values of the discrete costate for $k = N-1, N-2, \ldots, 0$. Thus the discrete state equation is solved by marching forward from $k = 0$, while the discrete costate equation is solved by marching backward from $k = N-1$.

In applying Theorem 2.1 to the first-order order conditions, we need to estimate the residual $\boldsymbol{\delta}$, and we need to analyze a linearized problem. Our linearization corresponds to the choice

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

Here $\mathbf{V} = \nabla^2 C(\mathbf{x}^*(1))$, and the various matrices are the same as those introduced in Section 3 except that they are evaluated at $\mathbf{x}^*(t_k)$, $\mathbf{u}^*(t_k)$, and $\boldsymbol{\psi}^*(t_k)$. In [8, Lem. 6.1], we show that when the coercivity assumption holds, $b_j > 0$ for each j, and $\sum_{j=1}^s b_j = 1$, then the linearized problem is invertible, with norm of the inverse bounded by a constant independent of h for h sufficiently small.

To analyze the residual $\boldsymbol{\delta}$, we need to determine the order of the Runge-Kutta schemes (13), (14), (20), and (21), where $\mathbf{u}$ is chosen so that $\nabla_u \mathbf{f}(\mathbf{y}_j, \mathbf{u}_{kj})^{\mathsf{T}}\boldsymbol{\chi}_j = \mathbf{0}$. In Table 1 we give the order of these schemes. 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. *If an index range does not appear on a summation sign, then the summation is over each index, taking values from* 1 *to* s.

Notice that the order conditions of Table 1 are not the usual order conditions [1, p. 170] associated with a Runge-Kutta discretization of a differential equation. The conditions of Table 1 were gotten in [9] by checking the tree-based order conditions in [1]. However, it was pointed out by Peter Rentrop at the June 4–10, 2000, conference in Oberwolfach, Germany, that these conditions should also follow from the general theory developed for partitioned Runge-Kutta methods (see [14, II.15], [15]). In [14, Thm. 15.9] it is shown that a partitioned Runge-Kutta method is of order p if and only if certain equations hold for all P-trees of order up to p.

These order conditions for partitioned Runge-Kutta schemes when applied to (13), (14), (20), and (21), should also lead to the conditions of Table 1.

Finally, applying Theorem 2.1 as in [9], it follows that if $(\mathbf{x}^*, \mathbf{u}^*)$ is a local minimizer for (12), the Runge-Kutta scheme is order order p (see Table 1), $b_j > 0$ for each j, and the coercivity condition holds, then when $\mathbf{f}$ is sufficiently smooth, the discrete problem (15) has a local minimizer $(\mathbf{x}^h, \mathbf{u}^\mathbf{h})$, and we have

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

where $\boldsymbol{\psi}^h$ is the solution of the discrete costate equations (20)–(21) and $\mathbf{u}(\mathbf{x}, \boldsymbol{\psi})$ denotes a minimizer of the Hamiltonian $H(\mathbf{x}, \mathbf{u}, \boldsymbol{\psi})$ over $\mathbf{u}$ (not one of the discrete controls). The order of approximation of the discrete controls in (15) is typically less than p. To obtain an approximation to an optimal control with the same order as that of the Runge-Kutta scheme, the Hamiltonian should be minimized over the control, using the computed discrete state and costate at each time level.

In [9] we show that the following scheme is 3-rd order accurate for differential equations, but only second order accurate for optimal control:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{3}{4} & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \frac{2}{9} \\ \frac{1}{3} \\ \frac{4}{9} \end{bmatrix}.$$

The following scheme, with $b_1 = 0$, is 2-nd order accurate for differential equations, but divergent for optimal control:

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

Although it appears difficult to construct a 4-th order Runge-Kutta scheme (13 conditions in Table 1 must be satisfied), it is shown in [9, Prop. 6.1] that every 4-stage explicit 4-th order Runge-Kutta scheme for differential equations, with $b_j > 0$ for each j, is 4-th order accurate for optimal control. This surprising result is due to the following identity, established by Butcher [1, p. 178], for 4-stage explicit 4-th order Runge-Kutta schemes:

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

$j = 1, 2, 3, 4$.

References

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

[2] A. L. Dontchev and W. W. Hager, *The Euler approximation in state constrained optimal control*, Mathematics of Computation, 2000, 31 pages.

[3] A. L. Dontchev and W. W. Hager, *Lipschitzian stability for state constrained nonlinear optimal control*, SIAM Journal on Control and Optimization, **36** (1998), 696–718.

[4] A. Dontchev and W. W. Hager , *An inverse mapping theorem for set-valued maps*, Proceedings of the American Mathematical Society, **121** (1994), 481–489.

[5] A. Dontchev and W. W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM Journal on Control and Optimization, **31** (1993), 569–603.

[6] A. L. Dontchev, W. W. Hager, and K Malanowski, *Error bounds for Euler approximation of a state and control constrained optimal control problem*, Numerical Functional Analysis and Optimization, **21** (2000), 653–682.

[7] A. Dontchev, W. W. Hager, A. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Applied Mathematics and Optimization, **31** (1995), 297–326.

[8] A. L. Dontchev, W. W. Hager, and V. M. Veliov, *Second-order Runge-Kutta approximations in constrained optimal control*, SIAM Journal on Numerical Analysis, **38** (2000), 202–226.

[9] W. W. Hager, *Runge-Kutta methods in optimal control and the transformed adjoint system*, Numerische Mathematik, **87** (2000), pp. 247–282.

[10] W. W. Hager, *Multiplier methods for nonlinear optimal control*, SIAM Journal on Numerical Analysis, **27** (1990), 1061–1080.

[11] W. W. Hager, *Approximations to the multiplier method*, SIAM Journal on Numerical Analysis, **22** (1985), 16–46.

[12] W. W. Hager, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM Journal on Numerical Analysis, **13** (1976), 449–472.

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

[14] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I*, second revised edition, Springer-Verlag, Berlin, 1993.

[15] P. Rentrop, *Partitioned Runge-Kutta methods with stiffness detection and stepsize control*, Numerische Mathematik, **47** (1985), 545–564.

Department of Mathematics,
University of Florida,
358 Little Hall
Gainesville, FL 32611 USA
Web: `http://www.math.ufl.edu/~hager`
E-mail address: `hager@math.ufl.edu`