2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

A pseudospectral method for optimal control based on collocation at the Gauss points*

William W. Hager[1], Jun Liu[2], Subhashree Mohapatra,[3] Anil V. Rao,[4] and Xiang-Sheng Wang[5]

***Abstract*— A Gauss collocation method is developed for solving optimal control problems with convex control constraints. The method has a local exponential convergence rate when the solution of the continuous problem is smooth and the Hamiltonian possesses a convexity property.**

I. INTRODUCTION

This paper is an abbreviated version of results that just appeared in [1] along with some more recent extensions developed in [2]. The focus of the work is on optimal control problems of the form

$$\begin{array}{ll}\text{minimize} & C(\mathbf{x}(1))\\ \text{subject to} & \dot{\mathbf{x}}(t)=\mathbf{f}(\mathbf{x}(t),\mathbf{u}(t)), \quad \mathbf{u}(t)\in\mathcal{U},\\ & t\in\Omega, \quad \mathbf{x}(-1)=\mathbf{x}_0,\\ & (\mathbf{x},\mathbf{u})\in\mathcal{C}^1(\Omega;\ \mathbb{R}^n)\times\mathcal{C}^0(\Omega;\ \mathbb{R}^m),\end{array} \tag{1}$$

where $\Omega=[-1,1]$, the control constraint set $\mathcal{U}\subset\mathbb{R}^m$ is closed and convex with nonempty interior, the state $\mathbf{x}(t)\in\mathbb{R}^n$, $\dot{\mathbf{x}}$ denotes the derivative of $\mathbf{x}$ with respect to t, $\mathbf{x}_0$ is the initial condition which we assume is given, $\mathbf{f}:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^n$, $C:\mathbb{R}^n\to\mathbb{R}$, and $\mathcal{C}^l(\Omega;\ \mathbb{R}^n)$ is the space of l times continuously differentiable functions mapping Ω to $\mathbb{R}^n$. It is assumed that $\mathbf{f}$ and C are at least continuous. The problem domain is chosen to be $[-1,1]$ since the Gauss points lie on this interval, which facilitates the formulation of the collocation scheme. A general interval $[T_1,T_2]$ can be transformed to $[-1,1]$ by the change of variables $s=(2t-T_1-T_2)/(T_2-T_1)$.

We approximate each component of the state $\mathbf{x}$ by a polynomial in $\mathcal{P}_N$, the space of polynomials of degree at most N. Let $\mathcal{P}_N^n$ denote the n-fold Cartesian product $\mathcal{P}_N\times\ldots\times\mathcal{P}_N$. In a collocation scheme, the dynamics are enforced at a set of points τ_i, $1\le i\le N$. This leads to the following discretization of (1) (see [3], [4]):

$$\begin{array}{ll}\text{minimize} & C(\mathbf{x}(1))\\ \text{subject to} & \dot{\mathbf{x}}(\tau_i)=\mathbf{f}(\mathbf{x}(\tau_i),\mathbf{u}_i), \quad \mathbf{u}_i\in\mathcal{U},\\ & \quad 1\le i\le N,\\ & \mathbf{x}(-1)=\mathbf{x}_0, \quad \mathbf{x}\in\mathcal{P}_N^n.\end{array} \tag{2}$$

The parameter $\mathbf{u}_i$ represents an approximation to the control at time τ_i. The dimension of $\mathcal{P}_N$ is $N+1$, and there are $N+1$ equations in (2) corresponding to the collocated dynamics at N points and the initial condition. This paper considers collocation at the Gauss quadrature points, which are symmetric about $t=0$ and satisfy

$$-1<\tau_1<\tau_2<\ldots<\tau_N<+1.$$

It is also convenient to introduce two noncollocated points

$$\tau_0=-1 \quad \text{and} \quad \tau_{N+1}=+1.$$

An earlier paper [5] considers collocation at the Gauss points, but without control constraints. In this paper, we introduce a control constraint $\mathbf{u}(t)\in\mathcal{U}$, which leads to profound changes in the analysis. When control constraints are present, the solution of (1) is typically nonsmooth, and the assumption in [5] that the problem solution has at least four derivatives is too strong. The earlier analysis in [5] employed sup-norms and a bound for a Lebesgue constant established in [6]. The new analysis for the control constrained problem employs the 2-norm and Sobolev best approximation results such as those in [7], [8].

The convergence analysis for (2) entrails an analysis of the relationship between the Pontryagin minimum principle associated with a solution of the continuous problem (1) and the Karush-Kuhn-Tucker (KKT) conditions associated with a solution of the finite dimensional programming problem (2). Let $(\mathbf{x}^*,\mathbf{u}^*)$ denote a local minimizer for (1) and let $\boldsymbol{\lambda}^*$ denote the solution of the linear costate equation

$$\dot{\boldsymbol{\lambda}}^*(t) \;=\; -\nabla_x H(\mathbf{x}^*(t),\mathbf{u}^*(t),\boldsymbol{\lambda}^*(t)), \tag{3}$$

$$\boldsymbol{\lambda}^*(1) \;=\; \nabla C(\mathbf{x}^*(1)), \tag{4}$$

where H is the Hamiltonian defined by $H(\mathbf{x},\mathbf{u},\boldsymbol{\lambda})=\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{f}(\mathbf{x},\mathbf{u})$. Under suitable assumptions, the minimum principle asserts that

$$-\nabla_u H(\mathbf{x}^*(t),\mathbf{u}^*(t),\boldsymbol{\lambda}^*(t))\in N_{\mathcal{U}}(\mathbf{u}^*(t)) \tag{5}$$

for all $t\in\Omega$, where ∇ denotes gradient and $N_{\mathcal{U}}$ is the normal cone. For any $\mathbf{u}\in\mathcal{U}$,

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

*The authors gratefully acknowledge support by the Office of Naval Research under grants N00014-15-1-2048 and N00014-18-1-2100, by the National Science Foundation under grants DMS-1522629, CBET-1404767, and DMS-1819002, and by the U.S. Air Force Research Laboratory under contract FA8651-08-D-0108/0054.

[1] William W. Hager is with the Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 hager@ufl.edu

[2] Jun Liu is with the Department of Mathematics and Statistics, Southern Illinois University Edwardsville, Edwardsville, IL 62026 juliu@siue.edu

[3] Subhashree Mohapatra is with the Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 subhashree3mohapatra@gmail.com

[4] Anil V. Rao is with the Department of Mechanical and Aerospace Engineering, Gainesville, FL 32611-6250 anilvrao@ufl.edu

[5] Xiang-Sheng Wang is with the Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70503 xswang@louisiana.edu

while $N_{\mathcal{U}}(\mathbf{u}) = \emptyset$ if $\mathbf{u} \notin \mathcal{U}$.

It can be shown that the KKT conditions for (2) are equivalent to the existence of a polynomial $\boldsymbol{\lambda} \in \mathcal{P}_N^n$ such that

$$\begin{aligned} \dot{\boldsymbol{\lambda}}(\tau_i) &= -\nabla_x H\left(\mathbf{x}(\tau_i), \mathbf{u}_i, \boldsymbol{\lambda}(\tau_i)\right), 1 \le i \le N, && (6)\\ \boldsymbol{\lambda}(1) &= \nabla C(\mathbf{x}(1)), && (7)\\ N_{\mathcal{U}}(\mathbf{u}_i) &\ni -\nabla_u H\left(\mathbf{x}(\tau_i), \mathbf{u}_i, \boldsymbol{\lambda}(\tau_i)\right), 1 \le i \le N. && (8) \end{aligned}$$

Thus the KKT conditions are equivalent to the continuous minimum principle enforced at the collocation points.

The following assumptions are utilized in the convergence analysis.

A1. For some $\alpha > 0$, the smallest eigenvalue of the Hessian matrices $\nabla^2 C(\mathbf{x}^*(1))$ and

$$\nabla^2_{(x,u)} H(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t))$$

are greater than α, uniformly for $t \in [0, 1]$.

A2. For some $\beta < 1/2$, the Jacobian of the dynamics satisfies

$$\|\nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t))\|_\infty \le \beta$$

and

$$\|\nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t))^{\mathsf{T}}\|_\infty \le \beta$$

for all $t \in \Omega$ where $\|\cdot\|_\infty$ is the matrix sup-norm (largest absolute row sum), and the Jacobian $\nabla_x \mathbf{f}$ is an n by n matrix whose i-th row is $(\nabla_x f_i)$.

The condition (A2) seems to be a strong requirement for a solution of the continuous control problem. However, when this condition is violated, it may not be possible to solve for the discrete state in terms of the discrete control in (2). In fact, the discrete dynamics could be infeasible even when the continuous dynamics is feasible.

To motivate the need for this condition, suppose that $\mathbf{f}(\mathbf{x}, \mathbf{u}) = \mathbf{A}\mathbf{x} + \mathbf{g}(\mathbf{u})$ where $\mathbf{A}$ is an n by n matrix. Since the dynamics are linear in the state, it is possible to solve for the continuous state as a function of the control whenever $\mathbf{g}(\mathbf{u})$ is absolutely integrable, regardless of the choice for $\mathbf{A}$. Although the dynamics in the discrete approximation (2) are also linear in the state, it may not be possible to solve for the discrete state in terms of the control, as we will show. On the other hand, the condition $\|\mathbf{A}\|_\infty < 1/2$ ensures that the discrete state is uniquely determined from the control, independent of the degree of the polynomials used in the discrete problem.

In general, (A2) could be replaced by any condition that ensures the solvability in the discrete linearized dynamics for the state in terms of the control, and the stability of this solution under perturbations in the dynamics. In more recent work [2] on *hp*-collocation schemes, (A2) is essentially removed by partitioning the original domain Ω into subdomains with a different polynomial on each subdomain. More precisely, the condition $\beta < 1/2$ of (A2) is replaced by $h\beta < 1/2$ where h is the subdomain width over 2, and when $h\beta < 1/2$, convergence occurs when the degree of the polynomials increase on each subdomain.

To better understand the dynamics in the discrete problem (2), we need to introduce the differentiation matrix $\mathbf{D}$. This is the N by $N+1$ matrix defined by

$$D_{ij} = \dot{L}_j(\tau_i), \text{ where } L_j(\tau) := \prod_{\substack{l=0\\ l \ne j}}^{N} \frac{\tau - \tau_l}{\tau_j - \tau_l}. \qquad (9)$$

Given a vector $\mathbf{p} \in \mathbb{R}^{N+1}$, let $p \in \mathcal{P}_N$ be the polynomial that satisfies the $N+1$ conditions $p(\tau_i) = p_i$, $0 \le i \le N$. The differentiation matrix has the property that $(\mathbf{D}\mathbf{p})_i = \dot{p}(\tau_i)$, $1 \le i \le N$. Our differentiation matrix is nonsquare since the derivative at $\tau_0 = -1$ does not appear in the discrete problem (2). The collocation conditions are only required to hold at the N Gauss points.

Given $\mathbf{x} \in \mathcal{P}_N^n$, let us define $\mathbf{X}_j = \mathbf{x}(\tau_j)$. By the definition of the differentiation matrix,

$$\dot{\mathbf{x}}(\tau_j) = \sum_{j=0}^{N} D_{ij}\mathbf{X}_j.$$

Hence, the dynamics in the discrete problem (2) can be reformulated as

$$\sum_{j=0}^{N} D_{ij}\mathbf{X}_j = \mathbf{f}(\mathbf{X}_i, \mathbf{U}_i),$$

where for consistency we let $\mathbf{U}_i$ denote the discrete control approximation at τ_i.

Suppose that the continuous variables $\mathbf{x}$ and $\mathbf{u}$ are scalars and the continuous dynamics has the form $\mathbf{f}(\mathbf{x}, \mathbf{u}) = \lambda x + g(u)$. If $\mathbf{D}_i$ is the i-th column of $\mathbf{D}$, $\mathbf{D}_{1:N}$ is the submatrix of $\mathbf{D}$ corresponding to columns 1 through N, $\mathbf{X}_{1:N}$ is the vector with components X_i, $1 \le i \le N$, and $g(\mathbf{U})$ is the vector with components $g(U_i)$, $1 \le i \le N$, then the discrete dynamics in (2) has the form

$$\mathbf{D}_{1:N}\mathbf{X}_{1:N} = \lambda\mathbf{X}_{1:N} + g(\mathbf{U}) - \mathbf{D}_0 X_0.$$

If λ is an eigenvalue of $\mathbf{D}_{1:N}$, then the coefficient matrix $\mathbf{D}_{1:N} - \lambda\mathbf{I}$ of $\mathbf{X}_{1:N}$ is singular, and generally it is not possible to solve for $\mathbf{X}_{1:N}$ in terms of $g(\mathbf{U})$ and X_0.

Nonetheless, the differentiation matrix has two important properties which are established in Appendix 1 of [2]:

(P1) $\mathbf{D}_{1:N}$ is invertible and $\|\mathbf{D}_{1:N}^{-1}\|_\infty \le 2$.

(P2) If $\mathbf{W}$ is the diagonal matrix containing the Gauss quadrature weights ω_i, $1 \le i \le N$, on the diagonal, then the rows of the matrix $[\mathbf{W}^{1/2}\mathbf{D}_{1:N}]^{-1}$ have Euclidean length bounded by $\sqrt{2}$.

By (P1), we have

$$\mathbf{D}_{1:N} - \lambda\mathbf{I} = \mathbf{D}_{1:N}(\mathbf{I} - \lambda\mathbf{D}_{1:N}^{-1}),$$

which is invertible when $|\lambda| < 1/2$ since $|\lambda|\|\mathbf{D}_{1:N}^{-1}\|_\infty < 1$. For the dynamics $\mathbf{f}(\mathbf{x}, \mathbf{u}) = \lambda x + g(u)$, (A2) reduces to the condition $|\lambda| < 1/2$. In general, when (A2) holds, the dynamics is locally, uniquely solvable for $\mathbf{X}_{1:N}$ in terms of the discrete control $\mathbf{U}$ and the initial condition $\mathbf{X}_0$.

If $\mathbf{x}^N \in \mathcal{P}_N^n$ is a solution of (2) associated with the discrete controls $\mathbf{u}_i$, $1 \le i \le N$, and if $\boldsymbol{\lambda}^N \in \mathcal{P}_N^n$ satisfies (6)–(8), then we define

$$\begin{array}{lllllll}
\mathbf{X}^N & = [\mathbf{x}^N(-1), & \mathbf{x}^N(\tau_1), & \ldots, & \mathbf{x}^N(\tau_N), & \mathbf{x}^N(+1)], \\
\mathbf{X}^* & = [\mathbf{x}^*(-1), & \mathbf{x}^*(\tau_1), & \ldots, & \mathbf{x}^*(\tau_N), & \mathbf{x}^*(+1)], \\
\mathbf{U}^N & = [& \mathbf{u}_1, & \ldots, & \mathbf{u}_N &], \\
\mathbf{U}^* & = [& \mathbf{u}^*(\tau_1), & \ldots, & \mathbf{u}^*(\tau_N) &], \\
\boldsymbol{\Lambda}^N & = [\boldsymbol{\lambda}^N(-1), & \boldsymbol{\lambda}^N(\tau_1), & \ldots, & \boldsymbol{\lambda}^N(\tau_N), & \boldsymbol{\lambda}^N(+1)], \\
\boldsymbol{\Lambda}^* & = [\boldsymbol{\lambda}^*(-1), & \boldsymbol{\lambda}^*(\tau_1), & \ldots, & \boldsymbol{\lambda}^*(\tau_N), & \boldsymbol{\lambda}^*(+1)].
\end{array}$$

The following convergence result relative to the vector ∞-norm (largest absolute element) is essentially established in [1]. In the statement of the theorem, $|\cdot|_{\mathcal{H}^p(\Omega;\,\mathbb{R}^n)}$ denotes the seminorm defined by

$$|\mathbf{x}|_{\mathcal{H}^p(\Omega;\,\mathbb{R}^n)} = \left(\int_{-1}^{1} \left| \frac{d^p \mathbf{x}(t)}{dt^p} \right|^2 \, dt \right)^{1/2}.$$

Theorem 1. *Suppose* $(\mathbf{x}^*, \mathbf{u}^*)$ *is a local minimizer for the continuous problem* (1) *with* $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{H}^\eta(\Omega; \mathbb{R}^n)$ *for some* $\eta \ge 2$. *If* (A1)–(A2) *hold, then for* N *sufficiently large, the discrete problem* (2) *has a local minimizer* $\mathbf{x}^N \in \mathcal{P}_N^n$ *and* $\mathbf{u} \in \mathbb{R}^{mN}$*, and an associated multiplier* $\boldsymbol{\lambda}^N \in \mathcal{P}_N^n$ *satisfying* (6)–(8)*; moreover, there exists a constant* c *independent of* N *and* η *such that*

$$\begin{aligned}
\max \left\{ \|\mathbf{X}^N - \mathbf{X}^*\|_\infty, \|\mathbf{U}^N - \mathbf{U}^*\|_\infty, \|\boldsymbol{\Lambda}^N - \boldsymbol{\Lambda}^*\|_\infty \right\} \\
\le \left(\frac{c}{N} \right)^{p-3/2} \left(|\mathbf{x}^*|_{\mathcal{H}^p(\Omega;\,\mathbb{R}^n)} + |\boldsymbol{\lambda}^*|_{\mathcal{H}^p(\Omega;\,\mathbb{R}^n)} \right), \quad (10)
\end{aligned}$$

where $p := \min\{\eta, N+1\}$.

One difference between this statement of the convergence rate bound and the statement of the analogous result in [1] is that the seminorm here is replaced by the full norm $\|\cdot\|_{\mathcal{H}^p(\Omega;\,\mathbb{R}^n)}$ in [1]. The full norm involves all the derivatives up to order p, while the seminorm only involves the p-th order derivative. In order to analyze convergence in an hp-setting where the domain Ω is partitioned into subdomains as in [2], it is important to state the error bound in a more precise way using seminorms since the error depends on both the width of the subdomains and the degree of the polynomials, and the impact of the subdomain size on the error is connected with all the derivatives appearing in the error bound. All the lower order derivatives should be eliminated from the error bound in order to describe more precisely the dependence of the error on the subdomain size. This is accomplished using a result such as Proposition 3.1 of [8] where the error in best approximation is expressed using a seminorm.

Also, when [1] was published, (P2) had not yet been proved. Hence, (P1) and (P2) appear as assumptions in Theorem 1.1 of [1], while these assumptions are removed in Theorem 1 above since these properties have now been established in general.

II. ERROR ANALYSIS

The derivation of Theorem 1 proceeds as follows.

A. Reformulate the KKT conditions for the discrete problem (2) as a generalized equation of the form
$$\mathcal{T}(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) \in \mathcal{F}(\mathbf{U}), \qquad (11)$$
where $(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda})$ correspond to the discrete state, control, and costate evaluated at the collocation points or at the end points of the interval $[-1, 1]$.

B. Estimate the distance d^* from $\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*)$ to $\mathcal{F}(\mathbf{U}^*)$. The bound we obtain for d^* has the same form as the right side of (10).

C. Linearize (11) around $(\mathbf{X}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*)$ and use (A1) and (A2) to show that the linearization has a unique solution which depends Lipschitz continuously on perturbations.

D. Combine these results to obtain (10). Further analysis based on (A1) shows that the solution of the generalized equation is a local minimizer of (2).

A. The Generalized Equation

The generalized equation corresponds to (6)–(8) along with the constraints of (2). The $\mathcal{T}$ in (11) has 7 components and $\mathcal{F}$ is composed of 7 sets. The sets forming $\mathcal{F}$ are

$$\mathcal{F}_0 = \mathcal{F}_1 = \ldots = \mathcal{F}_5 = \{\mathbf{0}\},$$

and

$$\mathcal{F}_{6i}(\mathbf{U}) = N_{\mathcal{U}}(\mathbf{U}_i), \quad 1 \le i \le N.$$

The 7 components of $\mathcal{T}$ are as follows:

$$\begin{array}{rcl}
\mathcal{T}_0(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) & = & \mathbf{X}_0 - \mathbf{x}_0, \\
\mathcal{T}_{1i}(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) & = & \left(\displaystyle\sum_{j=0}^{N} D_{ij} \mathbf{X}_j \right) - \mathbf{f}(\mathbf{X}_i, \mathbf{U}_i), \\
\mathcal{T}_2(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) & = & \mathbf{X}_{N+1} - \mathbf{X}_0 - \displaystyle\sum_{j=1}^{N} \omega_j \mathbf{f}(\mathbf{X}_j, \mathbf{U}_j), \\
\mathcal{T}_3(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) & = & \boldsymbol{\Lambda}_{N+1} - \boldsymbol{\Lambda}_0 \\
 & & \quad + \displaystyle\sum_{i=1}^{N} \omega_i \nabla_x H(\mathbf{X}_i, \mathbf{U}_i, \boldsymbol{\Lambda}_i), \\
\mathcal{T}_{4i}(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) & = & \left(\displaystyle\sum_{j=1}^{N+1} D_{ij}^{\dagger} \boldsymbol{\Lambda}_j \right) + \nabla_x H(\mathbf{X}_i, \mathbf{U}_i, \boldsymbol{\Lambda}_i), \\
\mathcal{T}_5(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) & = & \boldsymbol{\Lambda}_{N+1} - \nabla C(\mathbf{X}_{N+1}), \\
\mathcal{T}_{6i}(\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}) & = & -\nabla_u H(\mathbf{X}_i, \mathbf{U}_i, \boldsymbol{\Lambda}_i),
\end{array}$$

where $1 \le i \le N$ and $\mathbf{D}^\dagger$ is defined by

$$D_{ij} = -\left(\frac{\omega_j}{\omega_i} \right) D_{ji}^{\dagger}, \quad D_{i,N+1}^{\dagger} = -\sum_{j=1}^{N} D_{ij}^{\dagger},$$

$1 \le i \le N$. $\mathcal{T}_0$ corresponds to the initial condition, $\mathcal{T}_1$ yields the discrete state dynamics, $\mathcal{T}_2$ is a quadrature formula giving the discrete state at $t = 1$, $\mathcal{T}_3$ is a quadrature formula giving the discrete costate at $t = -1$, $\mathcal{T}_4$ yields the discrete costate dynamics, $\mathcal{T}_5$ corresponds to the terminal condition for the costate, and $\mathcal{T}_6$ corresponds to the minimum principle.

B. Estimate for the Residual

Next, we start with a continuous solution $(\mathbf{x}^*, \mathbf{u}^*)$ of (1) and the associated costate $\boldsymbol{\lambda}^*$, and evaluate them at the collocation points to form the vectors $(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$. The goal is to obtain a bound for the distance d^* from $\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ to $\mathcal{F}(\mathbf{U}^*)$. By (5), it follows that

$$-\nabla_u H(\mathbf{X}_i^*, \mathbf{U}_i^*, \mathbf{\Lambda}_i^*) \in N_{\mathcal{U}}(\mathbf{U}_i^*),$$

which implies that the associated residual vanishes. The components of the generalized equation where the residual is nontrivial correspond to the differential equations. Let $\dot{\mathbf{X}}^*$ denote the vector whose i-th component is $\dot{\mathbf{x}}^*(\tau_i)$, $0 \le i \le N$. Since $\mathbf{x}^*$ satisfies the dynamics in (1), it follows that

$$\dot{\mathbf{X}}_i^* = f(\mathbf{X}_i^*, \mathbf{U}_i^*) \tag{12}$$

for $1 \le i \le N$.

Let $\mathbf{x}^I$ denote the polynomial in $\mathcal{P}_N^n$ that passes through $\mathbf{x}^*(\tau_j)$ for $0 \le j \le N$, let $\mathbf{X}^I$ denote the vector with components $\mathbf{X}_j^I = \mathbf{x}^*(\tau_j)$, $0 \le j \le N$, and let $\dot{\mathbf{X}}^I$ denote the vector with components $\dot{\mathbf{X}}_j^I = \dot{\mathbf{x}}^I(\tau_j)$, $1 \le j \le N$. With these definitions and due to the property of the differentiation matrix, we have

$$\left(\sum_{j=0}^{N} D_{ij}\mathbf{X}_j^*\right) = \left(\sum_{j=0}^{N} D_{ij}\mathbf{X}_j^I\right) = \dot{\mathbf{X}}_i^I, \tag{13}$$

$1 \le i \le N$. Combine (12) and (13) to obtain

$$\mathcal{T}_1(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*) = \dot{\mathbf{X}}^I - \dot{\mathbf{X}}^*.$$

Since the derivative of a function is typically not equal to the derivative of its interpolant, the residual $\mathcal{T}_1(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ is nonzero in general.

To measure the size of the residual associated with the dynamics, we use the ω-norm defined by

$$\|\mathbf{z}\|_\omega^2 = \left(\sum_{i=1}^{N} \omega_i |\mathbf{z}_i|^2\right)^{1/2}, \quad \mathbf{z} \in \mathbb{R}^{nN}, \tag{14}$$

where $|\cdot|$ is the Euclidean norm. Hence, we have

$$\|\mathcal{T}_1(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)\|_\omega = \|\dot{\mathbf{X}}^I - \dot{\mathbf{X}}^*\|_\omega.$$

Let $(\dot{\mathbf{x}}^*)^J \in \mathcal{P}_{N-1}^n$ denote the interpolant that passes through $\dot{\mathbf{x}}^*(\tau_i)$ for $1 \le i \le N$. Since both $\dot{\mathbf{x}}^I$ and $(\dot{\mathbf{x}}^*)^J$ are polynomials of degree $N-1$ and Gaussian quadrature is exact for polynomials of degree $2N-1$, it follows that

$$\begin{aligned} \|\dot{\mathbf{X}}^I - \dot{\mathbf{X}}^*\|_\omega &= \|\dot{\mathbf{x}}^I - (\dot{\mathbf{x}}^*)^J\|_{\mathcal{L}^2(\Omega)} \\ &\le \|\dot{\mathbf{x}}^I - \dot{\mathbf{x}}^*\|_{\mathcal{L}^2(\Omega)} + \|\dot{\mathbf{x}}^* - (\dot{\mathbf{x}}^*)^J\|_{\mathcal{L}^2(\Omega)}. \end{aligned}$$

The L^2 error in polynomial interpolation has classic bounds such as

$$\|\dot{\mathbf{x}}^* - (\dot{\mathbf{x}}^*)^J\|_{\mathcal{L}^2(\Omega)} \le (c/N)^{p-1} |\mathbf{x}^*|_{\mathcal{H}^p(\Omega)},$$

(see (5.4.33) in [9]). The estimation of the error in the derivative of the interpolant, however, is more subtle. In [1] we establish the following bound, where again the full norm is replaced by the seminorm:

Lemma 1. *If $u \in \mathcal{H}^\eta(\Omega)$ for some $\eta \ge 1$, then there exists a constant c, independent of N and η, such that*

$$|u - u^I|_{\mathcal{H}^1(\Omega)} \le (c/N)^{p-3/2} |u|_{\mathcal{H}^p(\Omega)}, \tag{15}$$

where $p = \min\{\eta, N+1\}$ and $u^I \in \mathcal{P}_N$ is the interpolant of u satisfying $u^I(\tau_i) = u(\tau_i)$, $0 \le i \le N$, and $N > 0$.

Lemma 1 relies both on results of [7] and an observation of Yvon Maday given in Appendix 2 of [1]. Applying Lemma 1 to the dynamics of both the state and the costate, we obtain the following estimate for the distance from $\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ to $\mathcal{F}(\mathbf{U}^*)$.

Corollary 1. *If $\mathbf{x}^*$ and $\boldsymbol{\lambda}^* \in \mathcal{H}^\eta(\Omega;\, \mathbb{R}^n)$ for some $\eta \ge 2$, then there exists a constant c, independent of N and η, such that*

$$\begin{aligned} &\text{dist}[\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*),\ \mathcal{F}(\mathbf{U}^*)] \\ &\le \left(\frac{c}{N}\right)^{p-3/2} \left(|\mathbf{x}^*|_{\mathcal{H}^p(\Omega;\,\mathbb{R}^n)} + |\boldsymbol{\lambda}^*|_{\mathcal{H}^p(\Omega;\,\mathbb{R}^n)}\right), \end{aligned}$$

where $p = \min\{\eta, N+1\}$.

C. Stability of the Generalized Equation

By Corollary 1, the distance d^* from $\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ to $\mathcal{F}(\mathbf{U}^*)$ tends to zero as N tends to infinity. To obtain Theorem 1, we need to analyze the stability of the generalized equation with respect to a small parameter. The analysis entails a study of a linearized version of the generalized equation (11): Given a parameter $\mathbf{Y}$, the linearized problem is to find $(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda})$ such that

$$\nabla\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] + \mathbf{Y} \in \mathcal{F}(\mathbf{U}). \tag{16}$$

Here $\nabla\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}]$ denotes the derivative of $\mathcal{T}$ evaluated at $(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ operating on $[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}]$. The term $\nabla\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ will be abbreviated $\nabla\mathcal{T}^*$.

Let us introduce the following matrices:

$$\begin{aligned} \mathbf{A}(t) &= \nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t)), \\ \mathbf{B}(t) &= \nabla_u \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t)), \\ \mathbf{Q}(t) &= \nabla_{xx} H\left(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)\right), \\ \mathbf{S}(t) &= \nabla_{ux} H\left(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)\right), \\ \mathbf{R}(t) &= \nabla_{uu} H\left(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)\right), \\ \mathbf{T} &= \nabla^2 C(\mathbf{x}^*(1)). \end{aligned}$$

Subscripts are used to denote the value of these matrices at the collocation points:

$$\begin{gathered} \mathbf{A}_i = \mathbf{A}(\tau_i), \quad \mathbf{B}_i = \mathbf{B}(\tau_i), \quad \mathbf{Q}_i = \mathbf{Q}(\tau_i) \\ \mathbf{S}_i = \mathbf{S}(\tau_i), \quad \mathbf{R}_i = \mathbf{R}(\tau_i). \end{gathered}$$

With this notation, the 7 components of $\nabla\mathcal{T}^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}]$ are as follows:

$$
\begin{aligned}
\nabla\mathcal{T}_0^*[\mathbf{X},\mathbf{U},\mathbf{\Lambda}] &= \mathbf{X}_0, \\
\nabla\mathcal{T}_{1i}^*[\mathbf{X},\mathbf{U},\mathbf{\Lambda}] &= \left(\sum_{j=1}^{N} D_{ij}\mathbf{X}_j\right) - \mathbf{A}_i\mathbf{X}_i - \mathbf{B}_i\mathbf{U}_i, \\
\nabla\mathcal{T}_2^*[\mathbf{X},\mathbf{U},\mathbf{\Lambda}] &= \mathbf{X}_{N+1} - \mathbf{X}_0 \\
&\quad - \sum_{j=1}^{N}\omega_j(\mathbf{A}_j\mathbf{X}_j + \mathbf{B}_j\mathbf{U}_j), \\
\nabla\mathcal{T}_3^*[\mathbf{X},\mathbf{U},\mathbf{\Lambda}] &= \mathbf{\Lambda}_{N+1} - \mathbf{\Lambda}_0 \\
&\quad + \sum_{j=1}^{N}\omega_j(\mathbf{A}_j^{\mathsf{T}}\mathbf{\Lambda}_j + \mathbf{Q}_j\mathbf{X}_j + \mathbf{S}_j\mathbf{U}_j), \\
\nabla\mathcal{T}_{4i}^*[\mathbf{X},\mathbf{U},\mathbf{\Lambda}] &= \left(\sum_{j=1}^{N+1} D_{ij}^{\dagger}\mathbf{\Lambda}_j\right) \\
&\quad + \mathbf{A}_i^{\mathsf{T}}\mathbf{\Lambda}_i + \mathbf{Q}_i\mathbf{X}_i + \mathbf{S}_i\mathbf{U}_i, \\
\nabla\mathcal{T}_5^*[\mathbf{X},\mathbf{U},\mathbf{\Lambda}] &= \mathbf{\Lambda}_{N+1} - \mathbf{T}\mathbf{X}_{N+1}, \\
\nabla\mathcal{T}_{6i}^*[\mathbf{X},\mathbf{U},\mathbf{\Lambda}] &= -(\mathbf{S}_i^{\mathsf{T}}\mathbf{X}_i + \mathbf{R}_i\mathbf{U}_i + \mathbf{B}_i^{\mathsf{T}}\mathbf{\Lambda}_i),
\end{aligned}
$$

where $1 \le i \le N$.

We now need to show that the linearized problem (16) has a solution, and then analyze its stability with respect to the parameter $\mathbf{Y}$. To carry out this analysis, a norm must be chosen for $\mathbf{Y}$. There is one component of $\mathbf{Y}$, denoted $\mathbf{y}_i$ where $0 \le i \le 6$, for each component of $\mathcal{T}$. The following norm is employed in the analysis:

$$\|\mathbf{y}\|_{\mathcal{Y}} = |\mathbf{y}_0| + |\mathbf{y}_2| + |\mathbf{y}_3| + |\mathbf{y}_5| + \|\mathbf{y}_6\|_\infty + \|\mathbf{y}_1\|_\omega + \|\mathbf{y}_4\|_\omega.$$

Again, $|\cdot|$ denotes the Euclidean norm, while the ω-norm used for $\mathbf{y}_1$ (state dynamics) and $\mathbf{y}_4$ (costate dynamics) is defined in (14).

Part of the linearized problem corresponds to the linearized dynamics for the state and the costate. These are analyzed using assumption (A2) and properties (P1) and (P2). In particular, for each $\mathbf{q}_0$ and $\mathbf{q}_1 \in \mathbb{R}^n$ and $\mathbf{p} \in \mathbb{R}^{nN}$ with $\mathbf{p}_i \in \mathbb{R}^n$, $1 \le i \le N$, the linear system

$$
\begin{aligned}
\left(\sum_{j=0}^{N} D_{ij}\mathbf{X}_j\right) - \mathbf{A}_i\mathbf{X}_i &= \mathbf{p}_i \quad 1 \le i \le N, \\
\mathbf{X}_{N+1} - \mathbf{X}_0 - \sum_{j=1}^{N}\omega_j\mathbf{A}_j\mathbf{X}_j &= \mathbf{q}_1, \quad \mathbf{X}_0 = \mathbf{q}_0,
\end{aligned}
$$

has a unique solution $\mathbf{X} \in \mathbb{R}^{n(N+2)}$. Moreover, there exists a constant c, independent of N, such that

$$\|\mathbf{X}\|_\infty \le c(|\mathbf{q}_0| + |\mathbf{q}_1| + \|\mathbf{p}\|_\omega).$$

The costate dynamics have an analogous bound; namely, for each $\mathbf{q}_0$ and $\mathbf{q}_1 \in \mathbb{R}^n$ and $\mathbf{p} \in \mathbb{R}^{nN}$ with $\mathbf{p}_i \in \mathbb{R}^n$, $1 \le i \le N$, the linear system

$$
\begin{aligned}
\left(\sum_{j=1}^{N+1} D_{ij}^{\dagger}\mathbf{\Lambda}_j\right) + \mathbf{A}_i^{\mathsf{T}}\mathbf{\Lambda}_i &= \mathbf{p}_i \quad 1 \le i \le N, \\
\mathbf{\Lambda}_{N+1} - \mathbf{\Lambda}_0 + \sum_{j=1}^{N}\omega_j\mathbf{A}_j^{\mathsf{T}}\mathbf{\Lambda}_j &= \mathbf{q}_0, \quad \mathbf{\Lambda}_{N+1} = \mathbf{q}_1,
\end{aligned}
$$

has a unique solution $\mathbf{\Lambda} \in \mathbb{R}^{n(N+2)}$. Moreover, there exists a constant c, independent of N, such that

$$\|\mathbf{\Lambda}\|_\infty \le c(|\mathbf{q}_0| + |\mathbf{q}_1| + \|\mathbf{p}\|_\omega).$$

These stability results for the linearized state dynamics and the linearized costate dynamics enter into the complete analysis of the linearized problem (16). The analysis of the linearized problem is accomplished through an analysis of the following related quadratic programming problem:

$$
\begin{aligned}
\text{minimize} \quad & \tfrac{1}{2}\mathcal{Q}(\mathbf{X},\mathbf{U}) + \mathcal{L}(\mathbf{X},\mathbf{U},\mathbf{Y}) \\
\text{subject to} \quad & \textstyle\sum_{j=1}^{N} D_{ij}\mathbf{X}_j = \mathbf{A}_i\mathbf{X}_i + \mathbf{B}_i\mathbf{U}_i - \mathbf{y}_{1i}, \\
& \mathbf{U}_i \in \mathcal{U}, \quad 1 \le i \le N, \\
& \mathbf{X}_0 = -\mathbf{y}_0, \\
& \mathbf{X}_{N+1} = \mathbf{X}_0 - \mathbf{y}_2 \\
& \quad + \textstyle\sum_{j=1}^{N}\omega_j\left(\mathbf{A}_j\mathbf{X}_j + \mathbf{B}_j\mathbf{U}_j\right).
\end{aligned}
$$

Here the quadratic and linear terms in the objective are

$$
\begin{aligned}
\mathcal{Q}(\mathbf{X},\mathbf{U}) &= \mathbf{X}_{N+1}^{\mathsf{T}}\mathbf{T}\mathbf{X}_{N+1} \\
&+ \textstyle\sum_{i=1}^{N}\omega_i\left(\mathbf{X}_i^{\mathsf{T}}\mathbf{Q}_i\mathbf{X}_i + 2\mathbf{X}_i^{\mathsf{T}}\mathbf{S}_i\mathbf{U}_i + \mathbf{U}_i^{\mathsf{T}}\mathbf{R}_i\mathbf{U}_i\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathcal{L}(\mathbf{X},\mathbf{U},\mathbf{Y}) &= \mathbf{X}_0^{\mathsf{T}}\left(\mathbf{y}_3 - \textstyle\sum_{i=1}^{N}\omega_i\mathbf{y}_{4i}\right) \\
&- \mathbf{y}_5^{\mathsf{T}}\mathbf{X}_{N+1} + \textstyle\sum_{i=1}^{N}\omega_i\left(\mathbf{y}_{4i}^{\mathsf{T}}\mathbf{X}_i - \mathbf{y}_{6i}^{\mathsf{T}}\mathbf{U}_i\right).
\end{aligned}
$$

It can be shown that the first-order optimality conditions for the quadratic program reduce to (16); and conversely, when the convexity condition (A1) holds, a solution of (16) is also the solution of the quadratic program. Using (A1), (A2), (P1), and (P2) we can then analyze the effect of $\mathbf{Y}$ on the solution of (16). More precisely, the change $(\Delta\mathbf{X}, \Delta\mathbf{U}, \Delta\mathbf{\Lambda})$ in the solution of (16) corresponding to a change $\Delta\mathbf{Y}$ in $\mathbf{Y}$ satisfies

$$\max\{\|\Delta\mathbf{X}\|_\infty, \|\Delta\mathbf{U}\|_\infty, \|\Delta\mathbf{\Lambda}\|_\infty\} \le c\|\Delta\mathbf{Y}\|_{\mathcal{Y}},$$

where c is independent of N.

D. Final results

The final step is to combine the bound for the residual with the stability of the linearized problem to obtain Theorem 1. Existing results in the literature that lead to Theorem 1 include [10, Proposition 3.1], [11, Thm. 3.1], [12, Thm. 1], [13, Prop. 5.1], and [14, Thm. 2.1]. The proof that the solution to the discrete problem (2) is a local minimizer is a small modification of the analysis in [5, Thm. 2.1].

III. Conclusions

If the derivatives of a solution $(\mathbf{x}^*, \mathbf{u}^*)$ of (1) are uniformly bounded, then we can take $\eta = \infty$ in Theorem 1 to see that the error in the solution to the discrete problem (2) decays exponentially fast like $1/N^{N-1/2}$. For control problems where the optimal control is Lipschitz continuous in time, we often have $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{H}^2(\Omega; \mathbb{R}^n)$. In this case, Theorem 1 shows that the error decays like $1/\sqrt{N}$. Hence, there is convergence in a case where our earlier theory in [5] did not yield convergence. Our most recent work [2] develops a convergence theory for collocation at the Radau quadrature points, and treats *hp*-methods where the problem domain is partitioned into subdomains and a different polynomial approximates the state variable on each subdomain.

References

[1] W. W. Hager, J. Liu, S. Mohapatra, A. V. Rao, and X.-S. Wang, "Convergence rate for a Gauss collocation method applied to constrained optimal control," *SIAM J. Control Optim.*, vol. 56, pp. 1386–1411, 2018.

[2] W. W. Hager, H. Hou, S. Mohapatra, A. V. Rao, and X.-S. Wang, "Convergence rate for an *hp*-collocation method applied to constrained optimal control," 2017, arXiv: 1605.02121.

[3] D. A. Benson, G. T. Huntington, T. P. Thorvaldsen, and A. V. Rao, "Direct trajectory optimization and costate estimation via an orthogonal collocation method," *J. Guid. Control Dyn.*, vol. 29, no. 6, pp. 1435–1440, November-December 2006.

[4] D. Garg, M. A. Patterson, W. W. Hager, A. V. Rao, D. A. Benson, and G. T. Huntington, "A unified framework for the numerical solution of optimal control problems using pseudospectral methods," *Automatica*, vol. 46, pp. 1843–1851, 2010.

[5] W. W. Hager, H. Hou, and A. V. Rao, "Convergence rate for a Gauss collocation method applied to unconstrained optimal control," *J. Optim. Theory Appl.*, vol. 169, pp. 801–824, 2016.

[6] ——, "Lebesgue constants arising in a class of collocation methods," *IMA Journal of Numerical Analysis*, vol. 37, pp. 1884–1901, 2017.

[7] C. Bernardi and Y. Maday, "Polynomial interpolation results in Sobolev spaces," *J. Comput. Appl. Math.*, vol. 43, pp. 53–82, 1992.

[8] J. Elschner, "The h-p-version of spline approximation methods for Melin convolution equations," *J. Integral Equations Appl.*, vol. 5, no. 1, pp. 47–73, 1993.

[9] C. Canuto, M. Hussaini, A. Quarteroni, and T. Zang, *Spectral methods, Fundamentals in single domains*. Springer, 2006.

[10] A. L. Dontchev, W. W. Hager, and V. M. Veliov, "Second-order Runge-Kutta approximations in constrained optimal control," *SIAM J. Numer. Anal.*, vol. 38, pp. 202–226, 2000.

[11] A. L. Dontchev and W. W. Hager, "The Euler approximation in state constrained optimal control," *Math. Comp.*, vol. 70, pp. 173–203, 2001.

[12] W. W. Hager, "Multiplier methods for nonlinear optimal control," *SIAM J. Numer. Anal.*, vol. 27, pp. 1061–1080, 1990.

[13] ——, "Runge-Kutta methods in optimal control and the transformed adjoint system," *Numer. Math.*, vol. 87, pp. 247–282, 2000.

[14] ——, "Numerical analysis in optimal control," in *International Series of Numerical Mathematics*, K.-H. Hoffmann, I. Lasiecka, G. Leugering, J. Sprekels, and F. Tröltzsch, Eds., vol. 139. Basel/Switzerland: Birkhauser Verlag, 2001, pp. 83–93.