

Convergence Rate for a Gauss Collocation Method Applied to Unconstrained Optimal Control

William W. Hager¹ · Hongyan Hou² · Anil V. Rao³

Received: 1 June 2015 / Accepted: 19 March 2016 / Published online: 30 March 2016
© Springer Science+Business Media New York 2016

Abstract A local convergence rate is established for an orthogonal collocation method based on Gauss quadrature applied to an unconstrained optimal control problem. If the continuous problem has a smooth solution and the Hamiltonian satisfies a strong convexity condition, then the discrete problem possesses a local minimizer in a neighborhood of the continuous solution, and as the number of collocation points increases, the discrete solution converges to the continuous solution at the collocation points, exponentially fast in the sup-norm. Numerical examples illustrating the convergence theory are provided.

Keywords Gauss collocation method · Convergence rate · Optimal control · Orthogonal collocation

Mathematics Subject Classification 49M25 · 49M37 · 65K05 · 90C30

✉ William W. Hager
hager@ufl.edu

Hongyan Hou
hongyan388@gmail.com

Anil V. Rao
anilvrao@ufl.edu

¹ Department of Mathematics, University of Florida, P.O. Box 118105, Gainesville, FL 32611-8105, USA

² Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

³ Department of Mechanical and Aerospace Engineering, University of Florida, P.O. Box 116250, Gainesville, FL 32611-6250, USA

1 Introduction

The convergence rate is determined for the orthogonal collocation method developed in [1, 2] where the collocation points are the Gauss quadrature abscissas, or equivalently, the roots of a Legendre polynomial. Other sets of collocation points that have been studied in the literature include the Lobatto quadrature points [3, 4], the Chebyshev quadrature points [5, 6], the Radau quadrature points [7–10], and extrema of Jacobi polynomials [11]. We show that if the continuous problem has a smooth solution and the Hamiltonian satisfies a strong convexity assumption, then the discrete problem has a local minimizer that converges exponentially fast at the collocation points to a solution of the continuous control problem. In other work [12, 13], Kang considers control systems in feedback linearizable normal form and shows that when the Lobatto discretized control problem is augmented with bounds on the states and control and on certain Legendre polynomial expansion coefficients, then the objectives in the discrete problem converge to the optimal objective of the continuous problem at an exponential rate. Kang's analysis does not involve coercivity assumptions, but instead employs bounds in the discrete problem. Also, in [14] a consistency result is established for a scheme based on Lobatto collocation.

These collocation schemes based on the roots of orthogonal polynomials or their derivatives all employ approximations to the state based on global polynomials. Previously, convergence rates have been obtained when the approximating space consists of piecewise polynomials as in [15–21]. In this earlier work, convergence is achieved by letting the mesh spacing tend to zero, while keeping fixed the degree of the approximating polynomials on each mesh interval. In contrast, for schemes based on global polynomials, convergence is achieved by letting the degree of the approximating polynomials and the number of collocation points tend to infinity. This leads to an exponential convergence rate for problems with a smooth solution.

2 Preliminaries

A convergence rate is established for an orthogonal collocation method applied to an unconstrained control problem of the form

$$\left. \begin{array}{l} \text{minimize } C(\mathbf{x}(1)) \\ \text{subject to } \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad t \in [-1, 1], \\ \mathbf{x}(-1) = \mathbf{x}_0, \quad (\mathbf{x}, \mathbf{u}) \in \mathcal{C}^1(\mathbb{R}^n) \times \mathcal{C}^0(\mathbb{R}^m), \end{array} \right\} \quad (1)$$

where the state $\mathbf{x}(t) \in \mathbb{R}^n$, $\dot{\mathbf{x}} := \frac{d}{dt}\mathbf{x}$, the control $\mathbf{u}(t) \in \mathbb{R}^m$, $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, $C : \mathbb{R}^n \rightarrow \mathbb{R}$, and \mathbf{x}_0 is the initial condition, which we assume is given; \mathcal{C}^k denotes the space of k times continuously differentiable functions. Although there is no integral (Lagrange) cost term appearing in (1), it could be included in the objective by augmenting the state with an $(n + 1)$ st component x_{n+1} with initial condition $x_{n+1}(-1) = 0$ and with dynamics equal to the integrand of the Lagrange term. It is assumed that \mathbf{f} and C are at least continuous. Let \mathcal{P}_N denote the space of polynomials

of degree at most N defined on the interval $[-1, +1]$, and let \mathcal{P}_N^n denote the n -fold Cartesian product $\mathcal{P}_N \times \cdots \times \mathcal{P}_N$. We analyze a discrete approximation to (1) of the form

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

The collocation points $\tau_i, 1 \leq i \leq N$, are where the equation should be satisfied, and \mathbf{u}_i is the control approximation at time τ_i . The dimension of \mathcal{P}_N is $N + 1$, while there are $N + 1$ equations in (2) corresponding to the collocated dynamics at N points and the initial condition. In this paper, we collocate at the Gauss quadrature points, which are symmetric about $t = 0$ and satisfy

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

In addition, the analysis utilizes two noncollocated points $\tau_0 = -1$ and $\tau_{N+1} = +1$.

To state our convergence results in a precise way, we need to introduce a function space setting. Let $\mathcal{C}^k(\mathbb{R}^n)$ denote the space of k times continuously differentiable functions $\mathbf{x} : [-1, +1] \rightarrow \mathbb{R}^n$. For the space of continuous functions, we use the sup-norm $\|\cdot\|_\infty$ given by

$$\|\mathbf{x}\|_\infty = \sup\{|\mathbf{x}(t)| : t \in [-1, +1]\}, \tag{3}$$

where $|\cdot|$ is the Euclidean norm. For $\mathbf{y} \in \mathbb{R}^n$, let $\mathcal{B}_\rho(\mathbf{y})$ denote the ball with center \mathbf{y} and radius ρ :

$$\mathcal{B}_\rho(\mathbf{y}) = \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{y}| \leq \rho\}.$$

The following regularity assumption is assumed to hold throughout the paper.

Smoothness. *The problem (1) has a local minimizer $(\mathbf{x}^*, \mathbf{u}^*)$ in $\mathcal{C}^1(\mathbb{R}^n) \times \mathcal{C}^0(\mathbb{R}^m)$, and the first two derivatives of f and C are continuous on the closure of an open set Ω and on $\mathcal{B}_\rho(\mathbf{x}^*(1))$ respectively, where $\Omega \subset \mathbb{R}^{m+n}$ has the property that for some $\rho > 0$,*

$$\mathcal{B}_\rho(\mathbf{x}^*(t), \mathbf{u}^*(t)) \subset \Omega \text{ for all } t \in [-1, +1].$$

Let $\boldsymbol{\lambda}^*$ denote the solution of the linear costate equation

$$\dot{\boldsymbol{\lambda}}^*(t) = -\nabla_x H(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \quad \boldsymbol{\lambda}^*(1) = \nabla C(\mathbf{x}^*(1)), \tag{4}$$

where H is the Hamiltonian defined by $H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^\top \mathbf{f}(\mathbf{x}, \mathbf{u})$. Here ∇C denotes the gradient of C . By the first-order optimality conditions (Pontryagin’s minimum principle), we have

$$\nabla_u H(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) = \mathbf{0} \tag{5}$$

for all $t \in [-1, +1]$.

Since the discrete collocation problem (2) is finite dimensional, the first-order optimality conditions (Karush–Kuhn–Tucker conditions) imply that when a constraint qualification holds [22], the gradient of the Lagrangian vanishes. By the analysis in [2], the gradient of the Lagrangian vanishes if and only if there exists $\lambda \in \mathcal{P}_N^n$ such that

$$\dot{\lambda}(\tau_i) = -\nabla_x H(\mathbf{x}(\tau_i), \mathbf{u}_i, \lambda(\tau_i)), \quad 1 \leq i \leq N, \tag{6}$$

$$\lambda(1) = \nabla C(\mathbf{x}(1)), \tag{7}$$

$$\mathbf{0} = \nabla_u H(\mathbf{x}(\tau_i), \mathbf{u}_i, \lambda(\tau_i)), \quad 1 \leq i \leq N. \tag{8}$$

The assumptions that play a key role in the convergence analysis are the following:

- (A1) The matrix $\nabla^2 C(\mathbf{x}^*(1))$ is positive semidefinite and for some $\alpha > 0$, the smallest eigenvalue of the Hessian matrix

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

is greater than α , uniformly for $t \in [-1, +1]$.

- (A2) The Jacobian of the dynamics satisfies

$$\|\nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t))\|_\infty \leq \beta < 1/2 \quad \text{and} \quad \|\nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t))^\top\|_\infty \leq \beta$$

for all $t \in [-1, +1]$ 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 $(\nabla_x f_i)^\top$.

The coercivity assumption (A1) ensures that the solution of the discrete problem is a stationary point. The condition (A2) arises in the theoretical convergence analysis since the solution of the problem (1) is approximated by polynomials defined on the entire time interval $[-1, +1]$. In contrast, when the solution is approximated by piecewise polynomials as in [15–21], this condition is not needed. To motivate why this condition arises, suppose that the system dynamics \mathbf{f} has the form $\mathbf{f}(\mathbf{x}, \mathbf{u}) = \mathbf{A}\mathbf{x} + \mathbf{g}(\mathbf{u})$ where \mathbf{A} is an n by n matrix. Since the dynamics are linear in the state, it follows that for any given \mathbf{u} with $\mathbf{g}(\mathbf{u})$ absolutely integrable, we can invert the continuous dynamics to obtain the state \mathbf{x} as a function of the control \mathbf{u} . The dynamics in the discrete approximation (2) are also a linear function of the discrete state evaluated at the collocation points; however, the invertibility of the matrix in the discrete dynamics is not obvious. If the system matrix satisfies $\|\mathbf{A}\|_\infty < 1/2$, then we will show that the matrix for the discrete dynamics is invertible with its norm uniformly bounded as a function of N , the degree of the polynomials. Consequently, it is possible to solve for the state as a function of the control in the discrete problem (2). When $\|\mathbf{A}\|_\infty > 1/2$, the matrix for the discrete dynamics could be singular for some choice of N . In general, (A2) could be replaced by any condition that ensures the solvability of the linearized dynamics for the state in terms of the control, and the stability of this solution under perturbations in the dynamics.

When the dynamics of the control problem are nonlinear, the convergence analysis leads us to study the linearized dynamics around $(\mathbf{x}^*, \mathbf{u}^*)$, and (A2) implies that the

linearized dynamics in the discrete problem (2) are invertible with respect to the state. In contrast, if the global polynomials are replaced by piecewise polynomials on a uniform mesh of width h , then (A2) is replaced by

$$\|\nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t))\|_\infty < 1/h \quad \text{and} \quad \|\nabla_x \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t))^\top\|_\infty < 1/h,$$

which is satisfied when h is sufficiently small. In other words, when h is sufficiently small, it is possible to solve for the discrete states in terms of the discrete controls, independent of the degree of the polynomials on each mesh interval. These schemes where both the mesh width h and the degree of the polynomials p on each mesh interval are free parameters have been referred to as hp -collocation methods [9, 10].

In addition to the two assumptions, the analysis employs two properties of the Gauss collocation scheme. Let ω_j , $1 \leq j \leq N$, denote the Gauss quadrature weights, and for $1 \leq i \leq N$ and $0 \leq j \leq N$, define

$$D_{ij} = \dot{L}_j(\tau_i), \quad \text{where } L_j(\tau) := \prod_{\substack{k=0 \\ k \neq j}}^N \frac{\tau - \tau_k}{\tau_j - \tau_k}. \tag{9}$$

\mathbf{D} is a differentiation matrix in the sense that $(\mathbf{D}\mathbf{p})_i = \dot{p}(\tau_i)$, $1 \leq i \leq N$, where $p \in \mathcal{P}_N$ is the polynomial that satisfies $p(\tau_j) = p_j$ for $0 \leq j \leq N$. The submatrix $\mathbf{D}_{1:N}$ consisting of the tailing N columns of \mathbf{D} has the following properties:

- (P1) $\mathbf{D}_{1:N}$ is invertible and $\|\mathbf{D}_{1:N}^{-1}\|_\infty \leq 2$.
- (P2) If \mathbf{W} is the diagonal matrix containing the Gauss quadrature weights ω on the diagonal, then the rows of the matrix $[\mathbf{W}^{1/2}\mathbf{D}_{1:N}]^{-1}$ have Euclidean norm bounded by $\sqrt{2}$.

The fact that $\mathbf{D}_{1:N}$ is invertible is established in [2, Prop.1]. The bounds on the norms in (P1) and (P2), however, are more subtle. We refer to (P1) and (P2) as properties rather than assumptions since the matrices are readily evaluated, and we can check numerically that (P1) and (P2) are always satisfied. In fact, numerically we find that $\|\mathbf{D}_{1:N}^{-1}\|_\infty = 1 + \tau_N$ where the last Gauss quadrature abscissa τ_N approaches $+1$ as N tends to ∞ . On the other hand, we do not yet have a general proof for the properties (P1) and (P2). A prize for obtaining a proof is explained on William Hager’s Web site (Google William Hager 10,000 yen). In contrast to these properties, conditions (A1) and (A2) are assumptions that are only satisfied by certain control problems.

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

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

The following convergence result relative to the vector ∞ -norm (largest absolute element) is established:

Theorem 2.1 *If $(\mathbf{x}^*, \mathbf{u}^*)$ is a local minimizer for the continuous problem (1) with $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{C}^\eta$ for some $\eta \geq 4$ and both (A1)–(A2) and (P1)–(P2) hold, then for N sufficiently large, the discrete problem (2) has a stationary point $(\mathbf{X}^N, \mathbf{U}^N)$ and an associated discrete costate $\boldsymbol{\Lambda}^N$, and there exists a constant c independent of N and η such that*

$$\begin{aligned} & \max \left\{ \|\mathbf{X}^N - \mathbf{X}^*\|_\infty, \|\mathbf{U}^N - \mathbf{U}^*\|_\infty, \|\boldsymbol{\Lambda}^N - \boldsymbol{\Lambda}^*\|_\infty \right\} \\ & \leq \left(\frac{c}{N} \right)^{p-3} \left(\|\mathbf{x}^{*(p)}\|_\infty + \|\boldsymbol{\lambda}^{*(p)}\|_\infty \right), \quad p = \min\{\eta, N + 1\}. \end{aligned} \quad (10)$$

Moreover, if $\nabla^2 C(\mathbf{x}^*(1))$ is positive definite, then $(\mathbf{X}^N, \mathbf{U}^N)$ is a local minimizer of (2). Here the superscript (p) denotes the (p) th derivative.

Although the discrete problem only possesses discrete controls at the collocation points $-1 < \tau_i < +1$, $1 \leq i \leq N$, an estimate for the discrete control at $t = -1$ and $t = +1$ could be obtained from the minimum principle (5) since we have estimates for the discrete state and costate at the end points. Alternatively, polynomial interpolation could be used to obtain estimates for the control at the end points of the interval.

The paper is organized as follows. In Sect. 3, the discrete optimization problem (2) is reformulated as a nonlinear system of equations obtained from the first-order optimality conditions, and a general approach to convergence analysis is presented. Section 4 obtains an estimate for how closely the solution to the continuous problem satisfies the first-order optimality conditions for the discrete problem. Section 5 proves that the linearization of the discrete control problem around a solution of the continuous problem is invertible. Section 6 establishes an L^2 stability property for the linearization, while Sect. 7 strengthens the norm to L^∞ . This stability property is the basis for the proof of Theorem 2.1. Numerical examples illustrating the exponential convergence result are given in Sect. 8.

Notation. The meaning of the norm $\|\cdot\|_\infty$ is based on context. If $\mathbf{x} \in \mathcal{C}^0(\mathbb{R}^n)$, then $\|\mathbf{x}\|_\infty$ denotes the maximum of $|\mathbf{x}(t)|$ over $t \in [-1, +1]$, where $|\cdot|$ is the Euclidean norm. If $\mathbf{A} \in \mathbb{R}^m \times n$, then $\|\mathbf{A}\|_\infty$ is the largest absolute row sum (the matrix norm induced by the vector sup-norm). For a vector $\mathbf{v} \in \mathbb{R}^m$, $\|\mathbf{v}\|_\infty$ is the maximum of $|v_i|$ over $1 \leq i \leq m$. We let $|\mathbf{A}|$ denote the matrix norm induced by the Euclidean vector norm. The dimension of the identity matrix \mathbf{I} is often clear from context; when necessary, the dimension of \mathbf{I} is specified by a subscript. For example, \mathbf{I}_n is the n by n identity matrix. ∇C denotes the gradient, a column vector, while $\nabla^2 C$ denotes the Hessian matrix. Throughout the paper, c denotes a generic constant which has different values in different equations. The value of this constant is always independent of N . $\mathbf{1}$ denotes a vector whose entries are all equal to one, while $\mathbf{0}$ is a vector whose entries are all equal to zero, and their dimension should be clear from context. Finally, $\bar{\mathbf{X}} \in \mathbb{R}^{nN}$ is the vector obtained by vertically stacking $\mathbf{X}_i \in \mathbb{R}^n$, $1 \leq i \leq N$.

3 Abstract Setting

As shown in [2], the discrete problem (2) can be reformulated as the nonlinear programming problem

$$\begin{aligned} &\text{minimize } C(\mathbf{X}_{N+1}) \\ &\text{subject to } \sum_{j=0}^N D_{ij} \mathbf{X}_j = \mathbf{f}(\mathbf{X}_i, \mathbf{U}_i), \quad 1 \leq i \leq N, \quad \mathbf{X}_0 = \mathbf{x}_0, \\ &\quad \mathbf{X}_{N+1} = \mathbf{X}_0 + \sum_{j=1}^N \omega_j \mathbf{f}(\mathbf{X}_j, \mathbf{U}_j). \end{aligned} \tag{11}$$

As indicated before Theorem 2.1, \mathbf{X}_i corresponds to $\mathbf{x}^N(\tau_i)$. Also, Garg et al. [2] show that the equations obtained by setting the gradient of the Lagrangian to zero are equivalent to the system of equations

$$\sum_{j=1}^{N+1} D_{ij}^\dagger \Lambda_j = -\nabla_x H(\mathbf{X}_i, \mathbf{U}_i, \Lambda_i), \quad 1 \leq i \leq N, \tag{12}$$

$$\Lambda_{N+1} = \nabla C(\mathbf{X}_{N+1}), \tag{13}$$

$$\mathbf{0} = \nabla_u H(\mathbf{X}_i, \mathbf{U}_i, \Lambda_i), \quad 1 \leq i \leq N, \tag{14}$$

where

$$D_{ij}^\dagger = -\left(\frac{\omega_j}{\omega_i}\right) D_{ji}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq N, \tag{15}$$

$$D_{i,N+1}^\dagger = -\sum_{j=1}^N D_{ij}^\dagger, \quad 1 \leq i \leq N. \tag{16}$$

Here Λ_i corresponds to $\lambda^N(\tau_i)$. The relationship between the discrete costate Λ_i , the KKT multipliers λ_i , $1 \leq i \leq N$, associated with the discrete dynamics, and the multiplier λ_{N+1} associated with the equation for \mathbf{X}_{N+1} is

$$\omega_i \Lambda_i = \lambda_i + \omega_i \lambda_{N+1} \quad \text{when } 1 \leq i \leq N, \quad \text{and } \Lambda_{N+1} = \lambda_{N+1}. \tag{17}$$

The first-order optimality conditions for the nonlinear program (11) consist of the Eqs. (12)–(14) and the constraints in (11). This system can be written as $\mathcal{T}(\mathbf{X}, \mathbf{U}, \Lambda) = \mathbf{0}$ where

$$(\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_5)(\mathbf{X}, \mathbf{U}, \Lambda) \in \mathbb{R}^{nN} \times \mathbb{R}^n \times \mathbb{R}^{nN} \times \mathbb{R}^n \times \mathbb{R}^{mN}.$$

The five components of \mathcal{T} are defined as follows:

$$\mathcal{T}_{1i}(\mathbf{X}, \mathbf{U}, \Lambda) = \left(\sum_{j=0}^N D_{ij} \mathbf{X}_j \right) - \mathbf{f}(\mathbf{X}_i, \mathbf{U}_i), \quad 1 \leq i \leq N,$$

$$\mathcal{T}_2(\mathbf{X}, \mathbf{U}, \Lambda) = \mathbf{X}_{N+1} - \mathbf{X}_0 - \sum_{j=1}^N \omega_j \mathbf{f}(\mathbf{X}_j, \mathbf{U}_j),$$

$$\begin{aligned} \mathcal{T}_{3i}(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) &= \left(\sum_{j=1}^{N+1} D_{ij}^\dagger \mathbf{\Lambda}_j \right) + \nabla_x H(\mathbf{X}_i, \mathbf{U}_i, \mathbf{\Lambda}_i), \quad 1 \leq i \leq N, \\ \mathcal{T}_4(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) &= \mathbf{\Lambda}_{N+1} - \nabla_x C(\mathbf{X}_{N+1}), \\ \mathcal{T}_{5i}(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) &= \nabla_u H(\mathbf{X}_i, \mathbf{U}_i, \mathbf{\Lambda}_i), \quad 1 \leq i \leq N. \end{aligned}$$

Note that in formulating \mathcal{T} , we treat \mathbf{X}_0 as a constant whose value is the given starting condition \mathbf{x}_0 . Alternatively, we could treat \mathbf{X}_0 as an unknown and then expand \mathcal{T} to have a sixth component $\mathbf{X}_0 - \mathbf{x}_0$. With this expansion of \mathcal{T} , we need to introduce an additional multiplier $\mathbf{\Lambda}_0$ for the constraint $\mathbf{X}_0 - \mathbf{x}_0$; it turns out that $\mathbf{\Lambda}_0$ approximates $\lambda^*(-1)$. To achieve a slight simplification in the analysis, we employ a five-component \mathcal{T} and treat \mathbf{X}_0 as a constant, not an unknown.

The proof of Theorem 2.1 reduces to a study of solutions to $\mathcal{T}(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) = \mathbf{0}$ in a neighborhood of $(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$. Our analysis is based on Dontchev et al. [17, Proposition 3.1], which we simplify below to take into account the structure of our \mathcal{T} . Other results like this are contained in Theorem 3.1 of [16], in Proposition 5.1 of Hager [19], and in Theorem 2.1 of Hager [23].

Proposition 3.1 *Let \mathcal{X} be a Banach space and \mathcal{Y} be a linear normed space with the norms in both spaces denoted $\|\cdot\|$. Let $\mathcal{T}: \mathcal{X} \mapsto \mathcal{Y}$ with \mathcal{T} continuously Fréchet differentiable in $\mathcal{B}_r(\theta^*)$ for some $\theta^* \in \mathcal{X}$ and $r > 0$. Suppose that*

$$\|\nabla \mathcal{T}(\theta) - \nabla \mathcal{T}(\theta^*)\| \leq \varepsilon \text{ for all } \theta \in \mathcal{B}_r(\theta^*)$$

where $\nabla \mathcal{T}(\theta^*)$ is invertible, and define $\mu := \|\nabla \mathcal{T}(\theta^*)^{-1}\|$. If $\varepsilon\mu < 1$ and $\|\mathcal{T}(\theta^*)\| \leq (1 - \mu\varepsilon)r/\mu$, then there exists a unique $\theta \in \mathcal{B}_r(\theta^*)$ such that $\mathcal{T}(\theta) = \mathbf{0}$. Moreover, we have the estimate

$$\|\theta - \theta^*\| \leq \frac{\mu}{1 - \mu\varepsilon} \|\mathcal{T}(\theta^*)\| \leq r. \tag{18}$$

We apply Proposition 3.1 with $\theta^* = (\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ and $\theta = (\mathbf{X}^N, \mathbf{U}^N, \mathbf{\Lambda}^N)$. The key steps in the analysis are the estimation of the residual $\|\mathcal{T}(\theta^*)\|$, the proof that $\nabla \mathcal{T}(\theta^*)$ is invertible, and the derivation of a bound for $\|\nabla \mathcal{T}(\theta^*)^{-1}\|$ that is independent of N . In our context, we use the ∞ -norm for both \mathcal{X} and \mathcal{Y} . In particular,

$$\|\theta\| = \|(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda})\|_\infty = \max\{\|\mathbf{X}\|_\infty, \|\mathbf{U}\|_\infty, \|\mathbf{\Lambda}\|_\infty\}. \tag{19}$$

For this norm, the left side of (10) and the left side of (18) are the same. The norm on \mathcal{Y} enters into the estimation of both the residual $\|\mathcal{T}(\theta^*)\|$ in (18) and the parameter $\mu := \|\nabla \mathcal{T}(\theta^*)^{-1}\|$. In our context, we think of an element of \mathcal{Y} as a vector with components \mathbf{y}_i , $1 \leq i \leq 3N + 2$, where $\mathbf{y}_i \in \mathbb{R}^n$ for $1 \leq i \leq 2N + 2$ and $\mathbf{y}_i \in \mathbb{R}^m$ for $i > 2N + 2$. For example, $\mathcal{T}_1(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) \in \mathbb{R}^{nN}$ corresponds to the components $\mathbf{y}_i \in \mathbb{R}^n$, $1 \leq i \leq N$.

4 Analysis of the Residual

We now establish a bound for the residual.

Lemma 4.1 *If $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{C}^\eta$ and $p = \min\{\eta, N + 1\} \geq 4$, then there exists a constant c independent of N and η such that*

$$\|\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*)\|_\infty \leq \left(\frac{c}{N}\right)^{p-3} \left(\|\mathbf{x}^{*(p)}\|_\infty + \|\boldsymbol{\lambda}^{*(p)}\|_\infty\right). \tag{20}$$

Proof By the definition of \mathcal{T} , $\mathcal{T}_4(\mathbf{X}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*) = \mathbf{0}$ since \mathbf{x}^* and $\boldsymbol{\lambda}^*$ satisfy the boundary condition in (4). Likewise, $\mathcal{T}_5(\mathbf{X}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*) = \mathbf{0}$ since (5) holds for all $t \in [-1, +1]$; in particular, (5) holds at the collocation points.

Now let us consider \mathcal{T}_1 . By Garg et al. [2, Eq. (7)],

$$\sum_{j=0}^N D_{ij} \mathbf{X}_j^* = \dot{\mathbf{x}}^I(\tau_i), \quad 1 \leq i \leq N,$$

where $\mathbf{x}^I \in \mathcal{P}_N^n$ is the (interpolating) polynomial that passes through $\mathbf{x}^*(\tau_i)$ for $0 \leq i \leq N$. Since \mathbf{x}^* satisfies the dynamics of (1), it follows that $\mathbf{f}(\mathbf{X}_i^*, \mathbf{U}_i^*) = \dot{\mathbf{x}}^*(\tau_i)$. Hence, we have

$$\mathcal{T}_1(\mathbf{X}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*) = \dot{\mathbf{x}}^I(\tau_i) - \dot{\mathbf{x}}^*(\tau_i). \tag{21}$$

By Hager et al. [24, Prop. 2.1], we have

$$\|\dot{\mathbf{x}}^I - \dot{\mathbf{x}}^*\|_\infty \leq \left(1 + 2N^2\right) \inf_{\mathbf{q} \in \mathcal{P}_N^n} \|\dot{\mathbf{x}} - \dot{\mathbf{q}}\|_\infty + N^2(1 + L_N) \inf_{\mathbf{q} \in \mathcal{P}_N^n} \|\mathbf{x} - \mathbf{q}\|_\infty, \tag{22}$$

where L_N is the Lebesgue constant of the point set τ_i , $0 \leq i \leq N$. In [24, Thm. 4.1], we show that $L_N = O(\sqrt{N})$. It follows from Elschner [25, Prop. 3.1] that for some constant c , independent of N , we have

$$\inf_{\mathbf{q} \in \mathcal{P}_N^n} \|\mathbf{x} - \mathbf{q}\|_\infty \leq \left(\frac{c}{N + 1}\right)^\eta \|\mathbf{x}^{*(\eta)}\|_\infty \tag{23}$$

and

$$\inf_{\mathbf{q} \in \mathcal{P}_N^n} \|\dot{\mathbf{x}} - \dot{\mathbf{q}}\|_\infty \leq \left(\frac{c}{N}\right)^{\eta-1} \|\mathbf{x}^{*(\eta)}\|_\infty \tag{24}$$

whenever $N + 1 \geq \eta$. In the case that $N + 1 < \eta$, the smoothness requirement η can be relaxed to $N + 1$ to obtain similar bounds. Since the first bound (23) is dominated by (24), it follows that the first term in (22) dominates the second term, and we have

$$\|\dot{\mathbf{x}}^I - \dot{\mathbf{x}}^*\|_\infty \leq \left(\frac{c}{N}\right)^{p-3} \|\mathbf{x}^{*(p)}\|_\infty \tag{25}$$

where $p = \min\{\eta, N + 1\}$. The bound (25) is valid in both cases $N + 1 \geq \eta$ and $N + 1 < \eta$. By (21) and (25), $\mathcal{T}_1(\mathbf{X}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*)$ complies with the bound (20).

Next, let us consider

$$\mathcal{T}_2(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*) = \mathbf{x}^*(1) - \mathbf{x}^*(-1) - \sum_{j=1}^N \omega_j \mathbf{f}(\mathbf{x}^*(\tau_j), \mathbf{u}^*(\tau_j)). \tag{26}$$

By the fundamental theorem of calculus and the fact that N -point Gauss quadrature is exact for polynomials of degree up to $2N - 1$, we have

$$\mathbf{0} = \mathbf{x}^I(1) - \mathbf{x}^I(-1) - \int_{-1}^1 \dot{\mathbf{x}}^I(t) dt = \mathbf{x}^I(1) - \mathbf{x}^*(-1) - \sum_{j=1}^N \omega_j \dot{\mathbf{x}}^I(\tau_j) \tag{27}$$

since $\mathbf{x}^I(-1) = \mathbf{x}^*(-1)$. Subtract (27) from (26) to obtain

$$\mathcal{T}_2(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*) = \mathbf{x}^*(1) - \mathbf{x}^I(1) + \sum_{j=1}^N \omega_j \left(\dot{\mathbf{x}}^I(\tau_j) - \dot{\mathbf{x}}^*(\tau_j) \right). \tag{28}$$

Since $\omega_j > 0$ and their sum is 2,

$$\left| \sum_{j=1}^I \omega_j \left(\dot{\mathbf{x}}^I(\tau_j) - \dot{\mathbf{x}}^*(\tau_j) \right) \right| \leq 2 \|\dot{\mathbf{x}}^I - \dot{\mathbf{x}}^*\|_\infty. \tag{29}$$

Likewise,

$$|\mathbf{x}^*(1) - \mathbf{x}^I(1)| = \left| \int_{-1}^{+1} \left(\dot{\mathbf{x}}^*(t) - \dot{\mathbf{x}}^I(t) \right) dt \right| \leq 2 \|\dot{\mathbf{x}}^* - \dot{\mathbf{x}}^I\|_\infty. \tag{30}$$

We combine (28)–(30) and (25) to see that $\mathcal{T}_2(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ complies with the bound (20).

Finally, let us consider \mathcal{T}_3 . By Garg et al. [2, Thm. 1],

$$\sum_{j=1}^{N+1} D_{ij}^\dagger \mathbf{\Lambda}_j^* = \dot{\boldsymbol{\lambda}}^I(\tau_i), \quad 1 \leq i \leq N,$$

where $\boldsymbol{\lambda}^I \in \mathcal{P}_N^n$ is the (interpolating) polynomial that passes through $\mathbf{\Lambda}_j^* = \boldsymbol{\lambda}(\tau_j)$ for $1 \leq j \leq N + 1$. Since $\boldsymbol{\lambda}^*$ satisfies (4), it follows that $\dot{\boldsymbol{\lambda}}^*(\tau_i) = -\nabla_x H(\mathbf{X}_i^*, \mathbf{U}_i^*, \mathbf{\Lambda}_i^*)$. Hence, we have

$$\mathcal{T}_{3i}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*) = \dot{\boldsymbol{\lambda}}^I(\tau_i) - \dot{\boldsymbol{\lambda}}^*(\tau_i).$$

In exactly the same way that \mathcal{T}_1 in (21) was handled, we conclude that $\mathcal{T}_3(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ complies with the bound (20). This completes the proof. \square

5 Invertibility

In this section, we show that the derivative $\nabla\mathcal{T}(\theta^*)$ is invertible. This is equivalent to showing that for each $\mathbf{y} \in \mathcal{Y}$, there is a unique $\theta \in \mathcal{X}$ such that $\nabla\mathcal{T}(\theta^*)[\theta] = \mathbf{y}$. In our application, $\theta^* = (\mathbf{X}^*, \mathbf{U}^*, \Lambda^*)$ and $\theta = (\mathbf{X}, \mathbf{U}, \Lambda)$. To simplify the notation, we let $\nabla\mathcal{T}^*[\mathbf{X}, \mathbf{U}, \Lambda]$ denote the derivative of \mathcal{T} evaluated at $(\mathbf{X}^*, \mathbf{U}^*, \Lambda^*)$ operating on $(\mathbf{X}, \mathbf{U}, \Lambda)$. This derivative involves the following six matrices:

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

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

$$\begin{aligned} \nabla\mathcal{T}_{1i}^*[\mathbf{X}, \mathbf{U}, \Lambda] &= \left(\sum_{j=1}^N D_{ij} \mathbf{X}_j \right) - \mathbf{A}_i \mathbf{X}_i - \mathbf{B}_i \mathbf{U}_i, \quad 1 \leq i \leq N, \\ \nabla\mathcal{T}_2^*[\mathbf{X}, \mathbf{U}, \Lambda] &= \mathbf{X}_{N+1} - \sum_{j=1}^N \omega_j (\mathbf{A}_j \mathbf{X}_j + \mathbf{B}_j \mathbf{U}_j), \\ \nabla\mathcal{T}_{3i}^*[\mathbf{X}, \mathbf{U}, \Lambda] &= \left(\sum_{j=1}^{N+1} D_{ij}^\dagger \Lambda_j \right) + \mathbf{A}_i^\top \Lambda_i + \mathbf{Q}_i \mathbf{X}_i + \mathbf{S}_i \mathbf{U}_i, \quad 1 \leq i \leq N, \\ \nabla\mathcal{T}_4^*[\mathbf{X}, \mathbf{U}, \Lambda] &= \Lambda_{N+1} - \mathbf{T} \mathbf{X}_{N+1}, \\ \nabla\mathcal{T}_{5i}^*[\mathbf{X}, \mathbf{U}, \Lambda] &= \mathbf{S}_i^\top \mathbf{X}_i + \mathbf{R}_i \mathbf{U}_i + \mathbf{B}_i^\top \Lambda_i, \quad 1 \leq i \leq N. \end{aligned}$$

Notice that \mathbf{X}_0 does not appear in $\nabla\mathcal{T}^*$ since \mathbf{X}_0 is treated as a constant whose gradient vanishes.

The analysis of invertibility starts with the first component of $\nabla\mathcal{T}^*$.

Lemma 5.1 *If (P1) and (A2) hold, then for each $\mathbf{q} \in \mathbb{R}^n$ and $\mathbf{p} \in \mathbb{R}^{nN}$ with $\mathbf{p}_i \in \mathbb{R}^n$, $1 \leq i \leq N$, the linear system*

$$\left(\sum_{j=1}^N D_{ij} \mathbf{X}_j \right) - \mathbf{A}_i \mathbf{X}_i = \mathbf{p}_i \quad 1 \leq i \leq N, \tag{31}$$

$$\mathbf{X}_{N+1} - \sum_{j=1}^N \omega_j \mathbf{A}_j \mathbf{X}_j = \mathbf{q}, \tag{32}$$

has a unique solution $\mathbf{X}_j \in \mathbb{R}^n$, $1 \leq j \leq N + 1$. This solution has the bound

$$\|\mathbf{X}_j\|_\infty \leq 2\|\mathbf{p}\|_\infty / (1 - 2\beta), \quad 1 \leq j \leq N, \tag{33}$$

$$\|\mathbf{X}_{N+1}\|_\infty \leq \|\mathbf{q}\|_\infty + 4\beta\|\mathbf{p}\|_\infty / (1 - 2\beta). \tag{34}$$

Proof Let $\bar{\mathbf{X}}$ be the vector obtained by vertically stacking \mathbf{X}_1 through \mathbf{X}_N , let \mathbf{A} be the block diagonal matrix with i -th diagonal block \mathbf{A}_i , $1 \leq i \leq N$, and define $\bar{\mathbf{D}} = \mathbf{D}_{1:N} \otimes \mathbf{I}_n$ where \otimes is the Kronecker product. With this notation, the linear system (31) can be expressed $(\bar{\mathbf{D}} - \mathbf{A})\bar{\mathbf{X}} = \mathbf{p}$. By (P1) $\mathbf{D}_{1:N}$ is invertible which implies that $\bar{\mathbf{D}}$ is invertible with $\bar{\mathbf{D}}^{-1} = \mathbf{D}_{1:N}^{-1} \otimes \mathbf{I}_n$. Moreover, $\|\bar{\mathbf{D}}^{-1}\|_\infty = \|\mathbf{D}_{1:N}^{-1}\|_\infty \leq 2$ by (P1). By (A2) $\|\mathbf{A}\|_\infty \leq \beta$ and $\|\bar{\mathbf{D}}^{-1}\mathbf{A}\|_\infty \leq \|\bar{\mathbf{D}}^{-1}\|_\infty\|\mathbf{A}\|_\infty \leq 2\beta$. By Horn and Johnson [26, p. 351], $\mathbf{I} - \bar{\mathbf{D}}^{-1}\mathbf{A}$ is invertible since $\beta < 1/2$ and $\|(\mathbf{I} - \bar{\mathbf{D}}^{-1}\mathbf{A})^{-1}\|_\infty \leq 1/(1 - 2\beta)$. Consequently, $\bar{\mathbf{D}} - \mathbf{A} = \bar{\mathbf{D}}(\mathbf{I} - \bar{\mathbf{D}}^{-1}\mathbf{A})$ is invertible, and

$$\|(\bar{\mathbf{D}} - \mathbf{A})^{-1}\|_\infty \leq \|(\mathbf{I} - \bar{\mathbf{D}}^{-1}\mathbf{A})^{-1}\|_\infty\|\bar{\mathbf{D}}^{-1}\|_\infty \leq 2/(1 - 2\beta).$$

Thus there exists a unique $\bar{\mathbf{X}}$ such that $(\bar{\mathbf{D}} - \mathbf{A})\bar{\mathbf{X}} = \mathbf{p}$, and (33) holds. By (32), we have

$$\|\mathbf{X}_{N+1}\|_\infty \leq \|\mathbf{q}\|_\infty + \sum_{j=1}^N \omega_j \|\mathbf{A}_j\|_\infty \|\mathbf{X}_j\|_\infty.$$

Hence, (34) follows from (33) and the fact that the ω_j are positive and sum to 2 and $\|\mathbf{A}_j\|_\infty \leq \beta$ by (A2). □

Next, we establish the invertibility of ∇T^* .

Proposition 5.1 *If (A1), (A2), and (P1) hold, then ∇T^* is invertible.*

Proof We formulate a strongly convex quadratic programming problem whose first-order optimality conditions reduce to $\nabla T^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y}$. Due to the strong convexity of the objective function, the quadratic programming has a solution and there exists $\mathbf{\Lambda}$ such that $\nabla T^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y}$. Since T^* is square and $\nabla T^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y}$ has a solution for each choice of \mathbf{y} , it follows that ∇T^* is invertible.

The quadratic program is

$$\left. \begin{aligned} &\text{minimize } \frac{1}{2}\mathcal{Q}(\mathbf{X}, \mathbf{U}) + \mathcal{L}(\mathbf{X}, \mathbf{U}) \\ &\text{subject to } \sum_{j=1}^N D_{ij}\mathbf{X}_j = \mathbf{A}_i\mathbf{X}_i + \mathbf{B}_i\mathbf{U}_i + \mathbf{y}_{1i}, \quad 1 \leq i \leq N, \\ &\quad \mathbf{X}_{N+1} = \mathbf{y}_2 + \sum_{j=1}^N \omega_j (\mathbf{A}_j\mathbf{X}_j + \mathbf{B}_j\mathbf{U}_j), \end{aligned} \right\} \quad (35)$$

where the quadratic and linear terms in the objective are

$$\mathcal{Q}(\mathbf{X}, \mathbf{U}) = \mathbf{X}_{N+1}^\top \mathbf{T} \mathbf{X}_{N+1} + \mathcal{Q}_0(\bar{\mathbf{X}}, \mathbf{U}) \tag{36}$$

$$\mathcal{Q}_0(\bar{\mathbf{X}}, \mathbf{U}) = \sum_{i=1}^N \omega_i \left(\mathbf{x}_i^\top \mathbf{Q}_i \mathbf{x}_i + 2\mathbf{x}_i^\top \mathbf{S}_i \mathbf{u}_i + \mathbf{u}_i^\top \mathbf{R}_i \mathbf{u}_i \right),$$

$$\mathcal{L}(\mathbf{X}, \mathbf{U}) = \mathbf{y}_4^\top \mathbf{X}_{N+1} + \mathcal{L}_0(\bar{\mathbf{X}}, \mathbf{U}),$$

$$\mathcal{L}_0(\bar{\mathbf{X}}, \mathbf{U}) = - \sum_{i=1}^N \omega_i \left(\mathbf{y}_{3i}^\top \mathbf{x}_i + \mathbf{y}_{5i}^\top \mathbf{u}_i \right). \tag{37}$$

The first-order optimality conditions for (35) reduce to $\nabla T^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y}$. See Garg et al. [2] for the manipulations needed to obtain the first-order optimality conditions in this form.

Notice that the state variable \mathbf{X}_{N+1} in the quadratic program (35) can be expressed as a function of $\bar{\mathbf{X}}, \mathbf{U}$, and the parameter \mathbf{y}_2 . After making this substitution, we obtain a new quadratic \bar{Q} and a new linear term \bar{L} in the objective:

$$\begin{aligned} \bar{Q}(\bar{\mathbf{X}}, \mathbf{U}) &= Q_0(\bar{\mathbf{X}}, \mathbf{U}) + Q_1(\bar{\mathbf{X}}, \mathbf{U}), \\ Q_1(\bar{\mathbf{X}}, \mathbf{U}) &= \left(\sum_{i=1}^N \omega_i (\mathbf{A}_i \mathbf{X}_i + \mathbf{B}_i \mathbf{U}_i) \right)^T \mathbf{T} \left(\sum_{i=1}^N \omega_i (\mathbf{A}_i \mathbf{X}_i + \mathbf{B}_i \mathbf{U}_i) \right), \\ \bar{L}(\bar{\mathbf{X}}, \mathbf{U}) &= L_0(\bar{\mathbf{X}}, \mathbf{U}) + \sum_{i=1}^N \omega_i (\mathbf{y}_4 + 2\mathbf{T}\mathbf{y}_2)^T (\mathbf{A}_i \mathbf{X}_i + \mathbf{B}_i \mathbf{U}_i). \end{aligned}$$

Hence, the elimination of \mathbf{X}_{N+1} from the quadratic program (35) leads to the reduced problem

$$\left. \begin{aligned} &\text{minimize } \frac{1}{2} \bar{Q}(\bar{\mathbf{X}}, \mathbf{U}) + \bar{L}(\bar{\mathbf{X}}, \mathbf{U}) \\ &\text{subject to } \sum_{j=1}^N D_{ij} \mathbf{X}_j = \mathbf{A}_i \mathbf{X}_i + \mathbf{B}_i \mathbf{U}_i + \mathbf{y}_{1i}, \quad 1 \leq i \leq N. \end{aligned} \right\} \tag{38}$$

By (A1), we have

$$\bar{Q}(\bar{\mathbf{X}}, \mathbf{U}) \geq \alpha \left(\sum_{i=1}^N \omega_i (|\mathbf{X}_i|^2 + |\mathbf{U}_i|^2) \right). \tag{39}$$

Since α and ω are strictly positive, the objective of (38) is strongly convex with respect to $\bar{\mathbf{X}}$ and \mathbf{U} , and by Lemma 5.1, the quadratic programming problem is feasible. Hence, there exists a unique solution to both (35) and (38) for any choice of \mathbf{y} , and since the constraints are linear, the first-order necessary optimality conditions hold. Consequently, $\nabla T^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y}$ has a solution for any choice of \mathbf{y} and the proof is complete. \square

6 ω -Norm Bounds

In this section, we obtain a bound for the (\mathbf{X}, \mathbf{U}) component of the solution to $\nabla T^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y}$ in terms of \mathbf{y} . We bound the Euclidean norm of \mathbf{X}_{N+1} , while the other components of the state and the control are bounded in the ω -norm defined by

$$\|\bar{\mathbf{X}}\|_\omega^2 = \sum_{i=1}^N \omega_i |\mathbf{X}_i|^2 \quad \text{and} \quad \|\mathbf{U}\|_\omega^2 = \sum_{i=1}^N \omega_i |\mathbf{U}_i|^2. \tag{40}$$

This defines a norm since the Gauss quadrature weight $\omega_i > 0$ for each i . Since the (\mathbf{X}, \mathbf{U}) component of the solution to $\nabla T^*[\mathbf{X}, \mathbf{U}, \boldsymbol{\Lambda}] = \mathbf{y}$ is a solution of the quadratic program (35), we will bound the solution to the quadratic program.

First, let us think more abstractly. Let π be a symmetric, continuous bilinear functional defined on a Hilbert space \mathcal{H} , let ℓ be a continuous linear functional, let $\phi \in \mathcal{H}$, and consider the quadratic program

$$\min \left\{ \frac{1}{2} \pi(v + \phi, v + \phi) + \ell(v + \phi) : v \in \mathcal{V} \right\},$$

where \mathcal{V} is a subspace of \mathcal{H} . If w is a minimizer, then by the first-order optimality conditions, we have

$$\pi(w, v) + \pi(\phi, v) + \ell(v) = 0 \quad \text{for all } v \in \mathcal{V}.$$

Inserting $v = w$ yields

$$\pi(w, w) = -(\pi(w, \phi) + \ell(w)). \tag{41}$$

We apply this observation to the quadratic program (38). We identify ℓ with the linear functional $\tilde{\mathcal{L}}$, and π with the bilinear form associated with the quadratic $\tilde{\mathcal{Q}}$. The subspace \mathcal{V} is the null space of the linear operator in (38), and ϕ is a particular solution of the linear system. For the particular solution of the linear system in (38), we take $\bar{\mathbf{X}} = \boldsymbol{\chi}$ and $\mathbf{U} = \mathbf{0}$, where $\boldsymbol{\chi}$ denotes the solution to (31) given by Lemma 5.1 for $\mathbf{p} = \mathbf{y}_1$. The complete solution of (38) is the particular solution $(\boldsymbol{\chi}, \mathbf{0})$ plus the minimizer over the null space, which we denote $(\mathbf{X}^{\mathcal{N}}, \mathbf{U}^{\mathcal{N}})$.

In the relation (41) describing the null space component $(\mathbf{X}^{\mathcal{N}}, \mathbf{U}^{\mathcal{N}})$ of the solution, $\pi(w, w)$ corresponds to $\tilde{\mathcal{Q}}(\bar{\mathbf{X}}^{\mathcal{N}}, \mathbf{U}^{\mathcal{N}})$. The terms on the right side of (41) correspond to

$$\begin{aligned} \pi(w, \phi) &= \sum_{i=1}^N \omega_i \left[(\boldsymbol{\chi}_i^{\top} \mathbf{Q}_i + \mathbf{z}^{\top} \mathbf{A}_i) \mathbf{X}_i^{\mathcal{N}} + (\boldsymbol{\chi}_i^{\top} \mathbf{S}_i + \mathbf{z}^{\top} \mathbf{B}_i) \mathbf{U}_i^{\mathcal{N}} \right], \quad \text{and} \\ \ell(w) &= \sum_{i=1}^N \omega_i \left[\left((\mathbf{y}_4 + 2\mathbf{T}\mathbf{y}_2)^{\top} \mathbf{A}_i - \mathbf{y}_{3i}^{\top} \right) \mathbf{X}_i^{\mathcal{N}} + \left((\mathbf{y}_4 + 2\mathbf{T}\mathbf{y}_2)^{\top} \mathbf{B}_i - \mathbf{y}_{5i}^{\top} \right) \mathbf{U}_i^{\mathcal{N}} \right], \end{aligned}$$

where

$$\mathbf{z} = \mathbf{T} \left(\sum_{i=1}^N \omega_i \mathbf{A}_i \boldsymbol{\chi}_i \right).$$

(A1) implies that the quadratic term has the lower bound

$$\pi(w, w) = \tilde{\mathcal{Q}}(\bar{\mathbf{X}}^{\mathcal{N}}, \mathbf{U}^{\mathcal{N}}) \geq \alpha (\|\bar{\mathbf{X}}^{\mathcal{N}}\|_{\omega}^2 + \|\mathbf{U}^{\mathcal{N}}\|_{\omega}^2). \tag{42}$$

The terms corresponding to $\pi(w, \phi)$ and $\ell(w)$ can all be bounded in terms of the ω -norms of $\bar{\mathbf{X}}$ and \mathbf{U} and $\|\mathbf{y}\|$. For example, by the Schwarz inequality, we have

$$\begin{aligned} \sum_{i=1}^N \omega_i \mathbf{y}_{3i}^T \mathbf{X}_i^{\mathcal{N}} &\leq \left(\sum_{i=1}^N \omega_i |\mathbf{y}_{3i}|^2 \right)^{1/2} \left(\sum_{i=1}^N \omega_i |\mathbf{X}_i^{\mathcal{N}}|^2 \right)^{1/2} \\ &\leq \sqrt{2} \|\mathbf{y}_3\|_{\infty} \left(\|\bar{\mathbf{X}}^{\mathcal{N}}\|_{\omega}^2 + \|\mathbf{U}^{\mathcal{N}}\|_{\omega}^2 \right)^{1/2}. \end{aligned} \tag{43}$$

The last inequality exploits the fact that the ω_i sum to 2 and $|\mathbf{y}_{3i}| \leq \|\mathbf{y}_3\|_{\infty}$. To handle the terms involving χ in $\pi(w, \phi)$ and $\ell(w)$, we utilize the upper bound $|\chi_j| \leq 2\|\mathbf{y}\|_{\infty}/(1 - 2\beta)$ given in (33). Hence, we have

$$\begin{aligned} |\mathbf{z}| &\leq |\mathbf{T}| \sum_{i=1}^N \omega_i |\mathbf{A}_i| [2\|\mathbf{y}\|_{\infty}/(1 - 2\beta)] \leq [2\beta/(1 - 2\beta)] |\mathbf{T}| \|\mathbf{y}\|_{\infty} \sum_{i=1}^N \omega_i \\ &= [4\beta/(1 - 2\beta)] |\mathbf{T}| \|\mathbf{y}\|_{\infty}, \end{aligned}$$

where we used (A2) to obtain the bound

$$|\mathbf{A}_i| = \sqrt{|\mathbf{A}_i^T \mathbf{A}_i|} \leq \sqrt{\|\mathbf{A}_i^T \mathbf{A}_i\|_{\infty}} \leq \sqrt{\|\mathbf{A}_i^T\|_{\infty} \|\mathbf{A}_i\|_{\infty}} \leq \beta. \tag{44}$$

Combine upper bounds of the form (43) with the lower bound like (42) to conclude that $\|\bar{\mathbf{X}}^{\mathcal{N}}\|_{\omega}$ and $\|\mathbf{U}^{\mathcal{N}}\|_{\omega}$ are bounded by a constant times $\|\mathbf{y}\|_{\infty}$. The complete solution of (38) is the null space component that we just bounded plus the particular solution $(\chi, \mathbf{0})$. Again, since $\|\chi_j\|_{\infty} \leq 2\|\mathbf{y}\|_{\infty}/(1 - 2\beta)$, we deduce that the complete solution $(\bar{\mathbf{X}}, \mathbf{U})$, null space component plus particular solution, has $\|\bar{\mathbf{X}}\|_{\omega}$ and $\|\mathbf{U}\|_{\omega}$ bounded by a constant times $\|\mathbf{y}\|_{\infty}$. Finally, the equation for \mathbf{X}_{N+1} in (35) yields

$$|\mathbf{X}_{N+1}| \leq |\mathbf{y}_2| + \sum_{i=1}^N \omega_i [|\mathbf{A}_i| |\mathbf{X}_i| + |\mathbf{B}_i| |\mathbf{U}_i|].$$

Again, the Schwarz inequality gives bounds such as

$$\sum_{i=1}^N \omega_i |\mathbf{A}_i| |\mathbf{X}_i| \leq \left(\sum_{i=1}^N \omega_i |\mathbf{A}_i|^2 \right)^{1/2} \|\bar{\mathbf{X}}\|_{\omega} \leq \sqrt{2}\beta \|\bar{\mathbf{X}}\|_{\omega},$$

where the last inequality is due to (44) and the fact that the ω_i sum to 2. Since $\|\bar{\mathbf{X}}\|_{\omega}$ and $\|\mathbf{U}\|_{\omega}$ are bounded by a constant times $\|\mathbf{y}\|_{\infty}$, so is $|\mathbf{X}_{N+1}|$. This yields the following:

Lemma 6.1 *If (A1), (A2), and (P1) hold, then there exists a constant c , independent of N , such that the solution (\mathbf{X}, \mathbf{U}) of (35) satisfies $|\mathbf{X}_{N+1}| \leq c\|\mathbf{y}\|_{\infty}$, $\|\bar{\mathbf{X}}\|_{\omega} \leq c\|\mathbf{y}\|_{\infty}$, and $\|\mathbf{U}\|_{\omega} \leq c\|\mathbf{y}\|_{\infty}$.*

7 Infinity Norm Bounds

We now need to convert these ω -norm bounds for \mathbf{X} and \mathbf{U} into ∞ -norm bounds and, at the same time, obtain an ∞ -norm estimate for $\mathbf{\Lambda}$. By Lemma 5.1, the solution to the dynamics in (35) can be expressed

$$\bar{\mathbf{X}} = (\mathbf{I} - \bar{\mathbf{D}}^{-1}\mathbf{A})^{-1}\bar{\mathbf{D}}^{-1}\mathbf{B}\mathbf{U} + (\bar{\mathbf{D}} - \mathbf{A})^{-1}\mathbf{y}_1, \tag{45}$$

where \mathbf{B} is the block diagonal matrix with i th diagonal block \mathbf{B}_i . Taking norms and utilizing the bounds $\|(\bar{\mathbf{D}} - \mathbf{A})^{-1}\mathbf{y}_1\|_\infty \leq 2\|\mathbf{y}_1\|_\infty/(1 - 2\beta)$ and $\|(\mathbf{I} - \bar{\mathbf{D}}^{-1}\mathbf{A})^{-1}\|_\infty \leq 1/(1 - 2\beta)$ from Lemma 5.1, we obtain

$$\|\bar{\mathbf{X}}\|_\infty \leq \left(\|\bar{\mathbf{D}}^{-1}\mathbf{B}\mathbf{U}\|_\infty + 2\|\mathbf{y}_1\|_\infty \right) / (1 - 2\beta). \tag{46}$$

We now write

$$\bar{\mathbf{D}}^{-1}\mathbf{B}\mathbf{U} = [\mathbf{D}_{1:N}^{-1} \otimes \mathbf{I}_n]\mathbf{B}\mathbf{U} = [(\mathbf{W}^{1/2}\mathbf{D}_{1:N})^{-1} \otimes \mathbf{I}_n]\mathbf{B}\mathbf{U}_\omega, \tag{47}$$

where \mathbf{W} is the diagonal matrix with the quadrature weights on the diagonal and \mathbf{U}_ω is the vector whose i th element is $\sqrt{\omega_i}\mathbf{U}_i$. Note that the $\sqrt{\omega_i}$ factors in (47) cancel each other. An element of the vector $\bar{\mathbf{D}}^{-1}\mathbf{B}\mathbf{U}$ is the dot product between a row of $(\mathbf{W}^{1/2}\mathbf{D}_{1:N})^{-1} \otimes \mathbf{I}_n$ and the column vector $\mathbf{B}\mathbf{U}_\omega$. By (P2) the rows of $(\mathbf{W}^{1/2}\mathbf{D}_{1:N})^{-1} \otimes \mathbf{I}_n$ have Euclidean length bounded by $\sqrt{2}$. By the properties of matrix norms induced by vector norms, we have

$$|\mathbf{B}\mathbf{U}_\omega| \leq |\mathbf{B}|\|\mathbf{U}_\omega\| = |\mathbf{B}|\|\mathbf{U}\|_\omega. \tag{48}$$

Thus thinking of $\bar{\mathbf{D}}^{-1}\mathbf{B}\mathbf{U}$ in (47) as being the dot product of $\mathbf{B}\mathbf{U}_\omega$ with a vector of length at most $\sqrt{2}$, where the Euclidean length of $\mathbf{B}\mathbf{U}_\omega$ is estimated in (48), we have

$$\|\bar{\mathbf{D}}^{-1}\mathbf{B}\mathbf{U}\|_\infty \leq \sqrt{2}|\mathbf{B}|\|\mathbf{U}\|_\omega. \tag{49}$$

Combine Lemma 6.1 with (46) and (49) to deduce that $\|\bar{\mathbf{X}}\|_\infty \leq c\|\mathbf{y}\|_\infty$, where c is independent of N . Since $|\mathbf{X}_{N+1}| \leq c\|\mathbf{y}\|_\infty$ by Lemma 6.1, it follows that

$$\|\mathbf{X}\|_\infty \leq c\|\mathbf{y}\|_\infty. \tag{50}$$

Next, we use the third and fourth components of the linear system

$$\nabla\mathcal{T}^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y} \tag{51}$$

to obtain bounds for $\mathbf{\Lambda}$. These equations can be written

$$\bar{\mathbf{D}}^\dagger \bar{\mathbf{\Lambda}} + \bar{\mathbf{D}}_{N+1}^\dagger \mathbf{\Lambda}_{N+1} + \mathbf{A}^\top \bar{\mathbf{\Lambda}} + \mathbf{Q}\bar{\mathbf{X}} + \mathbf{S}\mathbf{U} = \mathbf{y}_3 \tag{52}$$

and

$$\mathbf{\Lambda}_{N+1} - \mathbf{T}\mathbf{X}_{N+1} = \mathbf{y}_4, \tag{53}$$

where $\bar{\mathbf{D}}^\dagger = \mathbf{D}_{1:N}^\dagger \otimes \mathbf{I}_n$, $\bar{\mathbf{D}}_{N+1}^\dagger = \mathbf{D}_{N+1}^\dagger \otimes \mathbf{I}_n$ and \mathbf{D}_{N+1}^\dagger is the $(N + 1)$ th column of \mathbf{D}^\dagger , $\bar{\mathbf{\Lambda}}$ is obtained by vertically stacking $\mathbf{\Lambda}_1$ through $\mathbf{\Lambda}_N$, and \mathbf{Q} and \mathbf{S} are block diagonal matrices with i th diagonal blocks \mathbf{Q}_i and \mathbf{S}_i , respectively.

We show in Proposition 10.1 of the ‘‘Appendix’’ that $\mathbf{D}_{1:N} = -\mathbf{J}\mathbf{D}_{1:N}^\dagger\mathbf{J}$, where \mathbf{J} is the exchange matrix with ones on its counterdiagonal and zeros elsewhere. It follows that $\mathbf{D}_{1:N}^{-1} = -\mathbf{J}(\mathbf{D}_{1:N}^\dagger)^{-1}\mathbf{J}$. Consequently, the elements in $\mathbf{D}_{1:N}^{-1}$ are the negative of the elements in $(\mathbf{D}_{1:N}^\dagger)^{-1}$, but rearranged as in (60). As a result, $(\mathbf{D}_{1:N}^\dagger)^{-1}$ also possesses properties (P1) and (P2), and the analysis of the discrete costate closely parallels the analysis of the state. The main difference is that the costate equation contains the additional $\mathbf{\Lambda}_{N+1}$ term along with the additional equation (53). By (53) and the previously established bound $\|\mathbf{X}\|_\infty \leq c\|\mathbf{y}\|_\infty$, it follows that

$$\|\mathbf{\Lambda}_{N+1}\|_\infty \leq c\|\mathbf{y}\|_\infty, \tag{54}$$

where c is independent of N . Since $\mathbf{D}^\dagger\mathbf{1} = \mathbf{0}$, we deduce that $(\mathbf{D}_{1:N}^\dagger)^{-1}\bar{\mathbf{D}}_{N+1}^\dagger = -\mathbf{1}$. It follows that

$$(\bar{\mathbf{D}}^\dagger)^{-1}\bar{\mathbf{D}}_{N+1}^\dagger = [(\mathbf{D}_{1:N}^\dagger)^{-1} \otimes \mathbf{I}_n][\mathbf{D}_{N+1}^\dagger \otimes \mathbf{I}_n] = -\mathbf{1} \otimes \mathbf{I}_n.$$

Exploiting this identity, the analog of (45) is

$$\bar{\mathbf{\Lambda}} = (\mathbf{I} + (\bar{\mathbf{D}}^\dagger)^{-1}\mathbf{A}^\top)^{-1}[(\mathbf{1} \otimes \mathbf{I}_n)\mathbf{\Lambda}_{N+1} + (\bar{\mathbf{D}}^\dagger)^{-1}(\mathbf{y}_3 - \mathbf{Q}\bar{\mathbf{X}} - \mathbf{S}\mathbf{U})].$$

Hence, we have

$$\|\bar{\mathbf{\Lambda}}\|_\infty \leq \|(\mathbf{1} \otimes \mathbf{I}_n)\mathbf{\Lambda}_{N+1} + (\bar{\mathbf{D}}^\dagger)^{-1}(\mathbf{y}_3 - \mathbf{Q}\bar{\mathbf{X}} - \mathbf{S}\mathbf{U})\|_\infty / (1 - 2\beta). \tag{55}$$

Moreover, $\|(\mathbf{1} \otimes \mathbf{I}_n)\mathbf{\Lambda}_{N+1}\|_\infty \leq c\|\mathbf{y}\|_\infty$ by (54), $\|(\bar{\mathbf{D}}^\dagger)^{-1}\mathbf{y}_3\|_\infty \leq 2\|\mathbf{y}_3\|_\infty$, and $\|(\bar{\mathbf{D}}^\dagger)^{-1}\mathbf{Q}\bar{\mathbf{X}}\|_\infty \leq 2\|\mathbf{Q}\|_\infty\|\bar{\mathbf{X}}\|_\infty$ where $\|\bar{\mathbf{X}}\|_\infty$ is bounded by (50). The term $\|(\bar{\mathbf{D}}^\dagger)^{-1}\mathbf{S}\mathbf{U}\|_\infty$ is handled exactly as $\|\bar{\mathbf{D}}^{-1}\mathbf{B}\mathbf{U}\|_\infty$ was handled in the state equation (45). Combine (54) with (55) to conclude that $\|\bar{\mathbf{\Lambda}}\|_\infty \leq c\|\mathbf{y}\|_\infty$ where c is independent of N .

Finally, let us examine the fifth component of the linear system (51). These equations can be written

$$\mathbf{S}_i^\top \mathbf{X}_i + \mathbf{R}_i \mathbf{U}_i + \mathbf{B}_i^\top \mathbf{\Lambda}_i = \mathbf{y}_{5i}, \quad 1 \leq i \leq N.$$

By (A1) the smallest eigenvalue of \mathbf{R}_i is greater than $\alpha > 0$. Consequently, the bounds $\|\mathbf{X}\|_\infty \leq c\|\mathbf{y}\|_\infty$ and $\|\mathbf{\Lambda}\|_\infty \leq c\|\mathbf{y}\|_\infty$ imply the existence of a constant c , independent of N , such that $\|\mathbf{U}\|_\infty \leq c\|\mathbf{y}\|_\infty$. In summary, we have the following result:

Lemma 7.1 *If (A1)–(A2) and (P1)–(P2) hold, then there exists a constant c , independent of N , such that the solution of $\nabla T^*[\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}] = \mathbf{y}$ satisfies*

$$\max \{ \|\mathbf{X}\|_\infty, \|\mathbf{U}\|_\infty, \|\mathbf{\Lambda}\|_\infty \} \leq c \|\mathbf{y}\|_\infty.$$

Let us now prove Theorem 2.1 using Proposition 3.1. By Lemma 7.1, $\mu = \|\nabla T(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)^{-1}\|_\infty$ is bounded uniformly in N . Choose ε small enough that $\varepsilon\mu < 1$. When we compute the difference $\nabla T(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) - \nabla T(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ for $(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda})$ near $(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$ in the ∞ -norm, the \mathbf{D} and \mathbf{D}^\dagger constant terms cancel, and we are left with terms involving the difference of derivatives of \mathbf{f} or C up to second order at nearby points. By the smoothness assumption, these second derivatives are uniformly continuous on the closure of Ω and on a ball around $\mathbf{x}^*(1)$. Hence, for r sufficiently small, we have

$$\|\nabla T(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) - \nabla T(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)\|_\infty \leq \varepsilon$$

whenever

$$\max\{\|\mathbf{X} - \mathbf{X}^*\|_\infty, \|\mathbf{U} - \mathbf{U}^*\|_\infty, \|\mathbf{\Lambda} - \mathbf{\Lambda}^*\|_\infty\} \leq r. \tag{56}$$

Since the smoothness $\eta \geq 4$ in Theorem 2.1, let us choose $\eta = 4$ in Lemma 4.1 and then take \bar{N} large enough that $\|\mathcal{T}(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)\| \leq (1 - \mu\varepsilon)r/\mu$ for all $N \geq \bar{N}$. Hence, by Proposition 3.1, there exists a solution to $\mathcal{T}(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) = \mathbf{0}$ satisfying (56). Moreover, by (18) and (20), the estimate (10) holds.

Notice that the smoothness parameter η does not enter into the discretization. We chose $\eta = 4$ to establish the existence of a unique solution to $\mathcal{T}(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) = \mathbf{0}$ satisfying (56). Once we know that the solution exists, larger values for η can be inserted in the error bound (10) if the problem solution has more than 4 continuous derivatives. In particular, if the problem solution has an infinite number of continuous derivatives that are nicely bounded, then we might take $\eta = N + 1$ in (10).

The analysis shows the existence of a solution to $\mathcal{T}(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) = \mathbf{0}$, which corresponds to the first-order necessary optimality conditions for either (11) or (2). To complete the proof, we need to show that this stationary point is a local minimizer when the assumption that $\nabla^2 C(\mathbf{x}^*(1))$ is positive semidefinite is strengthened to positive definite. After replacing the KKT multipliers by the transformed quantities given by (17), the Hessian of the Lagrangian is a block diagonal matrix whose i th diagonal block, $1 \leq i \leq N$, is $\omega_i \nabla_{(x,u)}^2 H(\mathbf{X}_i, \mathbf{U}_i, \mathbf{\Lambda}_i)$, where H is the Hamiltonian, and whose $(N + 1)$ st diagonal block is $\nabla^2 C(\mathbf{X}_{N+1})$. In computing the Hessian, we assume that the \mathbf{X} and \mathbf{U} variables are arranged in the following order: $\mathbf{X}_1, \mathbf{U}_1, \mathbf{X}_2, \mathbf{U}_2, \dots, \mathbf{X}_N, \mathbf{U}_N, \mathbf{X}_{N+1}$. By the strengthened version of (A1), the Hessian is positive definite when evaluated at $(\mathbf{X}^*, \mathbf{U}^*, \mathbf{\Lambda}^*)$. By continuity of the second derivative of C and \mathbf{f} and by the convergence result (10), we conclude that the Hessian of the Lagrangian, evaluated at the solution of $\mathcal{T}(\mathbf{X}, \mathbf{U}, \mathbf{\Lambda}) = \mathbf{0}$ satisfying (56), is positive definite for N sufficiently large. Hence, by the second-order sufficient optimality condition [22, Thm.12.6], (\mathbf{X}, \mathbf{U}) is a strict local minimizer of (11). This completes the proof of Theorem 2.1.

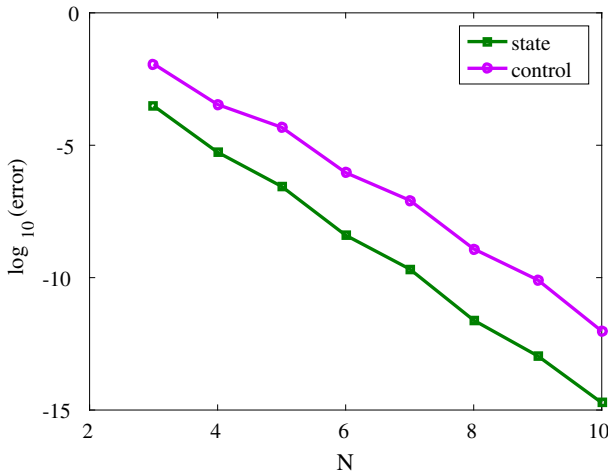


Fig. 1 The base 10 logarithm of the error in the sup-norm at the collocation points as a function of the polynomial degree for example (57)

8 Numerical Illustrations

The first example we study is

$$\begin{aligned}
 &\text{minimize } \frac{1}{2} \int_0^1 \left[u(t)^2 + x(t)u(t) + \frac{5}{4}x(t)^2 \right] dt && (57) \\
 &\text{subject to } x'(t) = .5x(t) + u(t), \quad x(0) = 1,
 \end{aligned}$$

with the optimal solution

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

To put this problem in the form of (1), we could introduce a new state variable y with dynamics

$$y'(t) = \frac{1}{2} \left(u(t)^2 + x(t)u(t) + 1.25x(t)^2 \right), \quad y(0) = 0,$$

in which case the objective is $y(1)$. Finally, we make the change of variable $t = (1 + \tau)/2$ to obtain the form (1). For this problem, (A1)–(A2) are satisfied so we expect the error to decay at least as fast as the bound given in Theorem 2.1. Since the derivatives of the hyperbolic functions are nicely bounded, exponential convergence is expected. Figure 1 plots the logarithm of the error in the sup-norm at the collocation points. Fitting the data of Fig. 1 by straight lines, we find that the error is $O(10^{-1.5N})$ roughly.

Although the assumptions (A1)–(A2) are sufficient for exponential convergence, the following example indicates that these assumptions are conservative. Let us consider the unconstrained control problem

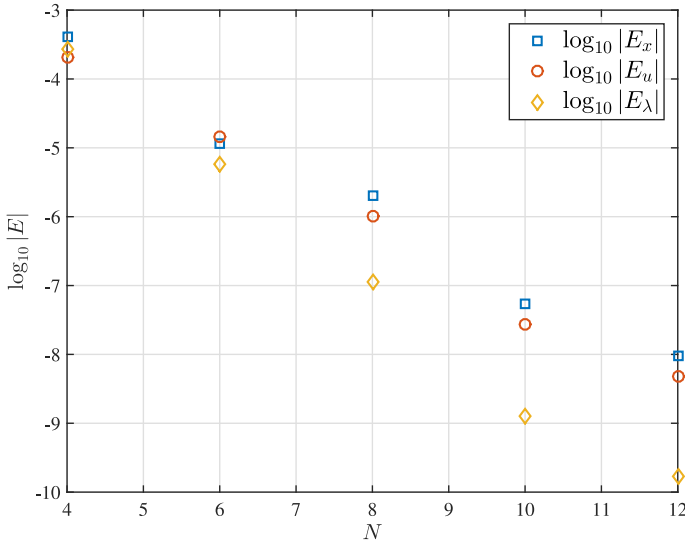


Fig. 2 The base 10 logarithm of the error in the sup-norm as a function of the number of collocation points for example (58)

$$\min \left\{ -x(2) : \dot{x}(t) = \frac{5}{2}(-x(t) + x(t)u(t) - u(t)^2), \quad x(0) = 1 \right\}. \quad (58)$$

The optimal solution and associated costate are

$$\begin{aligned} x^*(t) &= 4/a(t), \quad a(t) = 1 + 3 \exp(2.5t), \\ u^*(t) &= x^*(t)/2, \\ \lambda^*(t) &= -a^2(t) \exp(-2.5t) / [\exp(-5) + 9 \exp(5) + 6]. \end{aligned}$$

This violates (A2) since $\|\nabla_x f(\mathbf{x}^*(t), \mathbf{u}^*(t))\|_\infty$ is around 5/2 for t near 2. Nonetheless, as shown in Fig. 2, the logarithm of the error decays nearly linearly; the error behaves like $c10^{-0.6N}$ for either the state or the control and $c10^{-0.8N}$ for the costate.

A more complex problem with a known solution is

$$\min \int_0^1 \left[2x_1^2 x_2^2 + 1.25/x_2^2 + u_2/x_2 + u_1^2 + u_2^2 \right] dt, \quad (59)$$

subject to the dynamics

$$\begin{aligned} \dot{x}_1 &= x_1 + u_1/x_2 + u_2 x_1 x_2, & x_1(0) &= 1, \\ \dot{x}_2 &= -x_2(0.5 + u_2 x_2), & x_2(0) &= 1. \end{aligned}$$

The argument “(t)” on the states and controls was suppressed. The solution of the problem is

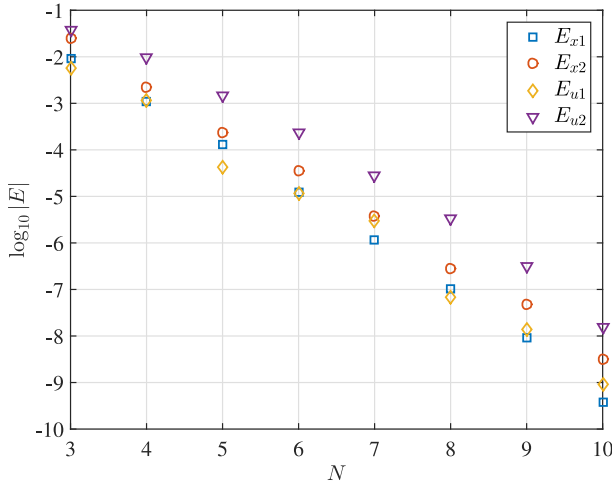


Fig. 3 The base 10 logarithm of the error in the sup-norm as a function of the number of collocation points for example (59)

$$\begin{aligned}
 x_1^*(t) &= \frac{\cosh(1-t)(2\exp(3t) + \exp(3))}{(2 + \exp(3))\exp(3t/2)\cosh(1)}, \\
 x_2^*(t) &= \frac{\cosh(1)}{\cosh(1-t)}, \\
 u_1^*(t) &= \frac{2(\exp(3t) - \exp(3))}{(2 + \exp(3))\exp(3t/2)}, \\
 u_2^*(t) &= \frac{-\cosh(1-t)(\tanh(1-t) + 0.5)}{\cosh(1)}.
 \end{aligned}$$

Figure 3 plots the logarithm of the sup-norm error in the state and control as a function of the number of collocation points. The convergence is again exponentially fast,

9 Conclusions

A Gauss collocation scheme is analyzed for an unconstrained control problem. When the problem has a smooth solution with η continuous derivatives and with a Hamiltonian that satisfies a strong convexity assumption, we show that the discrete problem has a local minimizer in a neighborhood of the continuous solution, and as the degree N of the approximating polynomials increases, the error in the sup-norm at the collocation point is $O(N^{3-p})$, where $p = \min\{\eta, N + 1\}$. Numerical examples confirm the exponential convergence.

Acknowledgments The authors gratefully acknowledge support by the Office of Naval Research under Grants N00014-11-1-0068 and N00014-15-1-2048 and by the National Science Foundation under Grants DMS-1522629 and CBET-1404767. Comments and suggestions from the reviewers are gratefully acknowledged. In particular, in the initial draft of this paper, it was assumed that $\nabla^2 C(\mathbf{x}^*(1))$ was positive definite,

while one of the reviewers correctly pointed out that this assumption could be relaxed to positive semidefinite without effecting the convergence results for a stationary point of the discrete problem.

10 Appendix

In (15), we define a new matrix \mathbf{D}^\dagger in terms of the differentiation matrix \mathbf{D} . The following proposition shows that the elements of \mathbf{D}^\dagger can be gotten by rearranging the elements of \mathbf{D} .

Proposition 10.1 *The entries of the matrices \mathbf{D} and \mathbf{D}^\dagger satisfy*

$$D_{ij} = -D_{N+1-i, N+1-j}^\dagger, \quad 1 \leq i \leq N, \quad 1 \leq j \leq N. \tag{60}$$

In other words, $\mathbf{D}_{1:N} = -\mathbf{J}\mathbf{D}_{1:N}^\dagger\mathbf{J}$ where \mathbf{J} is the exchange matrix with ones on its counterdiagonal and zeros elsewhere. Equivalently, $\mathbf{D}_{1:N}^\dagger = -\mathbf{J}\mathbf{D}_{1:N}\mathbf{J}$.

Proof By (9) the elements of \mathbf{D} can be expressed in terms of the derivatives of a set of Lagrange basis functions evaluated at the collocation points:

$$D_{ij} = \dot{L}_j(\tau_i) \quad \text{where } L_j \in \mathcal{P}_N, \quad L_j(\tau_k) = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } 0 \leq k \leq N, k \neq j. \end{cases}$$

In (9) we give an explicit formula for the Lagrange basis functions, while here we express the basis function in terms of polynomials L_j that equal one at τ_j and vanish at τ_k where $0 \leq k \leq N, k \neq j$. These $N + 1$ conditions uniquely define $L_j \in \mathcal{P}_N$. Similarly, by Garg et al. [2, Thm. 1], the entries of $\mathbf{D}_{1:N}^\dagger$ are given by

$$D_{ij}^\dagger = \dot{M}_j(\tau_i) \quad \text{where } M_j \in \mathcal{P}_N, \quad M_j(\tau_k) = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } 1 \leq k \leq N + 1, k \neq j. \end{cases}$$

Observe that $M_{N+1-j}(t) = L_j(-t)$ due to the symmetry of the quadrature points around $t = 0$:

- (a) Since $-\tau_{N+1-j} = \tau_j$, we have $L_j(-\tau_{N+1-j}) = L_j(\tau_j) = 1$.
- (b) Since $\tau_{N+1} = 1$ and $\tau_0 = -1$, we have $L_j(-\tau_{N+1}) = L_j(\tau_0) = 0$.
- (c) Since $-\tau_i = \tau_{N+1-i}$, we have $L_j(-\tau_i) = L_j(\tau_{N+1-i}) = 0$ if $i \neq N + 1 - j$.

Since $M_{N+1-j}(t)$ is equal to $L_j(-t)$ at $N + 1$ distinct points, the two polynomials are equal everywhere. Replacing $M_{N+1-j}(t)$ by $L_j(-t)$, we have

$$D_{N+1-i, N+1-j}^\dagger = -\dot{L}_j(-\tau_{N+1-i}) = -\dot{L}_j(\tau_i) = -D_{ij}. \tag{□}$$

Tables 1 and 2 illustrate properties (P1) and (P2) for the differentiation matrix \mathbf{D} . In Table 1, we observe that $\|\mathbf{D}_{1:N}^{-1}\|_\infty$ monotonically approaches the upper limit 2. More precisely, it is found that $\|\mathbf{D}_{1:N}^{-1}\|_\infty = 1 + \tau_N$, where the final collocation point τ_N approaches one as N tends to infinity. In Table 2, we give the maximum 2-norm of the rows of $[\mathbf{W}^{1/2}\mathbf{D}_{1:N}]^{-1}$. It is found that the maximum is attained by the last row of $[\mathbf{W}^{1/2}\mathbf{D}_{1:N}]^{-1}$, and the maximum monotonically approaches $\sqrt{2}$.

Table 1 $\|\mathbf{D}_{1:N}^{-1}\|_\infty$

N	25	50	75	100	125	150
Norm	1.995557	1.998866	1.999494	1.999714	1.999816	1.999872
N	175	200	225	250	275	300
Norm	1.999906	1.999928	1.999943	1.999954	1.999962	1.999968

Table 2 Maximum Euclidean norm for the rows of $[\mathbf{W}^{1/2}\mathbf{D}_{1:N}]^{-1}$

N	25	50	75	100	125	150
Norm	1.412201	1.413703	1.413985	1.414085	1.414131	1.414156
N	175	200	225	250	275	300
Norm	1.414171	1.414181	1.414188	1.414193	1.414196	1.414199

References

- Benson, D.A., Huntington, G.T., Thorvaldsen, T.P., Rao, A.V.: Direct trajectory optimization and costate estimation via an orthogonal collocation method. *J. Guid. Control Dyn.* **29**(6), 1435–1440 (2006)
- Garg, D., Patterson, M.A., Hager, W.W., Rao, A.V., Benson, D.A., Huntington, G.T.: A unified framework for the numerical solution of optimal control problems using pseudospectral methods. *Automatica* **46**, 1843–1851 (2010)
- Elnagar, G., Kazemi, M., Razzaghi, M.: The pseudospectral Legendre method for discretizing optimal control problems. *IEEE Trans. Automat. Control* **40**(10), 1793–1796 (1995)
- Fahroo, F., Ross, I.M.: Costate estimation by a Legendre pseudospectral method. *J. Guid. Control Dyn.* **24**(2), 270–277 (2001)
- Elnagar, G.N., Kazemi, M.A.: Pseudospectral Chebyshev optimal control of constrained nonlinear dynamical systems. *Comput. Optim. Appl.* **11**(2), 195–217 (1998)
- Fahroo, F., Ross, I.M.: Direct trajectory optimization by a Chebyshev pseudospectral method. *J. Guid. Control Dyn.* **25**, 160–166 (2002)
- Fahroo, F., Ross, I.M.: Pseudospectral methods for infinite-horizon nonlinear optimal control problems. *J. Guid. Control Dyn.* **31**(4), 927–936 (2008)
- Garg, D., Patterson, M.A., Darby, C.L., Françolin, C., Huntington, G.T., Hager, W.W., Rao, A.V.: Direct trajectory optimization and costate estimation of finite-horizon and infinite-horizon optimal control problems using a Radau pseudospectral method. *Comput. Optim. Appl.* **49**(2), 335–358 (2011)
- Liu, F., Hager, W.W., Rao, A.V.: Adaptive mesh refinement method for optimal control using non-smoothness detection and mesh size reduction. *J. Frankl. Inst.* **352**, 4081–4106 (2015)
- Patterson, M.A., Hager, W.W., Rao, A.V.: A *ph* mesh refinement method for optimal control. *Optim. Control Appl. Meth.* **36**, 398–421 (2015)
- Williams, P.: Jacobi pseudospectral method for solving optimal control problems. *J. Guid. Control Dyn.* **27**(2), 293–297 (2004)
- Kang, W.: The rate of convergence for a pseudospectral optimal control method. In: *Proceeding of the 47th IEEE Conference on Decision and Control*, IEEE, pp. 521–527 (2008)
- Kang, W.: Rate of convergence for the Legendre pseudospectral optimal control of feedback linearizable systems. *J. Control Theory Appl.* **8**, 391–405 (2010)
- Gong, Q., Ross, I.M., Kang, W., Fahroo, F.: Connections between the covector mapping theorem and convergence of pseudospectral methods for optimal control. *Comput. Optim. Appl.* **41**(3), 307–335 (2008)
- Dontchev, A.L., Hager, W.W.: Lipschitzian stability in nonlinear control and optimization. *SIAM J. Control Optim.* **31**, 569–603 (1993)

16. Dontchev, A.L., Hager, W.W.: The Euler approximation in state constrained optimal control. *Math. Comput.* **70**, 173–203 (2001)
17. Dontchev, A.L., Hager, W.W., Veliov, V.M.: Second-order Runge–Kutta approximations in constrained optimal control. *SIAM J. Numer. Anal.* **38**, 202–226 (2000)
18. Dontchev, A.L., Hager, W.W., Malanowski, K.: Error bounds for Euler approximation of a state and control constrained optimal control problem. *Numer. Funct. Anal. Optim.* **21**, 653–682 (2000)
19. Hager, W.W.: Runge-Kutta methods in optimal control and the transformed adjoint system. *Numer. Math.* **87**, 247–282 (2000)
20. Kameswaran, S., Biegler, L.T.: Convergence rates for direct transcription of optimal control problems using collocation at Radau points. *Comput. Optim. Appl.* **41**(1), 81–126 (2008)
21. Reddien, G.W.: Collocation at Gauss points as a discretization in optimal control. *SIAM J. Control Optim.* **17**(2), 298–306 (1979)
22. Nocedal, J., Wright, S.J.: *Numerical Optimization*, 2nd edn. Springer, New York (2006)
23. Hager, W.W.: Numerical analysis in optimal control. In: Hoffmann, K.H., Lasiecka, I., Leugering, G., Sprekels, J., Tröltzsch, F. (eds.) *International Series of Numerical Mathematics*, vol. 139, pp. 83–93. Birkhauser Verlag, Basel (2001)
24. Hager, W.W., Hou, H., Rao, A.V.: Lebesgue constants arising in a class of collocation methods (2015). [arXiv:1507.08316](https://arxiv.org/abs/1507.08316)
25. Elschner, J.: The h-p-version of spline approximation methods for Mellin convolution equations. *J. Integral Equ. Appl.* **5**(1), 47–73 (1993)
26. Horn, R.A., Johnson, C.R.: *Matrix Analysis*. Cambridge University Press, Cambridge (2013)