Contractivity of the Method of Successive Approximations for Optimal Control

Kevin D. Smith and Francesco Bullo

Abstract—Strongly contracting dynamical systems have numerous properties (e.g., incremental ISS), find widespread applications (e.g., in controls and learning), and their study is receiving increasing attention. This letter starts with the simple observation that, given a strongly contracting system, its adjoint dynamical system is also strongly contracting, with the same rate, with respect to the dual norm, under time reversal. As main implication of this dual contractivity, we show that the classic Method of Successive Approximations (MSA), an indirect method in optimal control, is a contraction mapping for short optimization intervals or large contraction rates. Consequently, we establish new convergence conditions for the MSA algorithm, which further imply uniqueness of the optimal control and sufficiency of Pontryagin’s minimum principle under additional assumptions.

Index Terms—Optimal control, contraction theory, iterative methods.

I. INTRODUCTION

Optimal control is generally a difficult problem, and with the exception of some analytically tractable cases, it must be solved numerically. Numerical approaches broadly fall into two categories: direct and indirect methods. Direct methods, like direct collocation and direct shooting methods [3], [22], [23], discretize and approximate the state and/or control to encode the problem as a nonlinear program. Due to their relative simplicity, robustness, and the wide availability of software implementations, direct methods tend to be favored in modern times [3, Sec. 4.3], [8].

Indirect methods are an older class of methods based on Pontryagin’s minimum principle (PMP), which gives a necessary condition for optimality of a control signal. PMP states that the optimal trajectory must solve a two-point boundary problem, together with a costate, and that the optimal control minimizes a Hamiltonian function at each point in time. Indirect methods search for an input, state trajectory, and costate trajectory that satisfy PMP. Many direct methods, including shooting and collocation, can also be applied as indirect methods to the PMP boundary value problem [12]. Another approach is the Method of Successive Approximations (MSA) [7], also called the Forward-Backward-Sweep algorithm [17], which is the main topic of this letter.

MSA [1], [13], [16] and its variants [7], [21] are classic approaches that have received renewed attention in the machine learning community [4], [18], [19] as alternatives to gradient descent for training residual neural networks (ResNets). Indeed, a new thrust of machine learning research is to apply control-theoretic techniques to the training of ResNets by viewing these models as forward Euler discretizations of continuous-time control systems [10], [24], [25]. Within this framework, training the ResNet can be viewed as an optimal control problem. As argued in [18], [19], MSA (and its variants) allow for error and convergence analysis and can lead to better training dynamics than gradient descent.

Unfortunately, MSA does not always converge, a problem that is still the subject of ongoing research. In [20], the authors prove convergence criteria based on boundedness and Lipschitz assumptions. Similar bounds are established in [18], [19]. This letter provides a new set of convergence criteria when MSA is applied to strongly contracting dynamical systems.

The contributions of this letter are as follows. First, in Section III, we study the adjoints of nonlinear systems that arise in optimal control theory. We show that adjoints of contracting systems under time reversal are also contracting with the same rate, albeit with respect to the dual norm. This property allows us to prove Grönwall-like and ISS-like bounds on the adjoint dynamics. Section IV applies these bounds to analyze MSA. Assuming Lipschitz continuity of all relevant maps in the optimal control problem, we obtain a bound on the Lipschitz constant of each MSA iteration. This Lipschitz constant becomes arbitrarily small in the limits of short optimization intervals and large contraction rates, thereby establishing conditions for when the iteration is a contraction mapping. With an additional assumption of pointwise uniqueness of the minimizer of the Hamiltonian, we show that these conditions also lead to uniqueness of the optimal control and sufficiency of PMP. Finally, in an additional section available on arXiv, we provide an illustrative example.

II. PRELIMINARIES

A. Contracting Dynamics Over Normed Vector Spaces

Let $\|\cdot\|_x : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a norm. The dual norm $\|\cdot\|_* : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is the norm $\|x\|_* = \sup_{\|y\|\leq 1} y^\top x$. Given a matrix $A \in \mathbb{R}^{n \times n}$, the induced norm of $A$ is $\|A\| = \sup_{\|x\|\leq 1} \|Ax\|$. The dual of $\|\cdot\|_x$ is $\|\cdot\|_* = \sup_{\|y\|\leq 1} y^\top x$. Given a matrix $A \in \mathbb{R}^{n \times n}$, the induced norm of $A$ is $\|A\| = \sup_{\|x\|\leq 1} \|Ax\|$.
exists such that $d$ dynamical system if the map $d$ dynamical system $-$ not consider such systems in this letter.

B. Optimal Control

[6, Sec. 2.4].

\begin{align*}
\sup_{\|x\|=1} \|Ax\| \quad \text{and the induced logarithmic norm of } A \text{ is } \\
\mu(A) = \lim_{\alpha \to 0^+} \frac{\|I_n + \alpha A\| - 1}{\alpha}.
\end{align*}

Explicit formulas for the induced (logarithmic) norms are known for the standard $p \in \{1, 2, \infty\}$ norms on $\mathbb{R}^n$.

A map $T: X \to Y$ between normed spaces $(X, \| \cdot \|_X)$ and $(Y, \| \cdot \|_Y)$ is Lipschitz continuous if a constant $\ell \geq 0$ exists such that $\|T(x) - T(y)\|_Y \leq \ell \|x - y\|_X$ for all $x, y \in X$. The minimal Lipschitz constant $\text{Lip}(T)$ is the infimum over $\ell$ that satisfy this inequality. If $T$ is continuously differentiable, then $\text{Lip}(T) = \sup_{x \in X} \|D_T(x)\|$, where $D_T(x)$ denotes the Jacobian matrix of $T$. Furthermore, if $X = Y = \mathbb{R}^n$, then the one-sided Lipschitz constant of $T$ is $\text{osL}(T) = \sup_{x \in \mathbb{R}^n} \mu(D_T(x))$. A dynamical system $\dot{x} = f(t, x, \ldots)$ with a continuously differentiable vector field $f: \mathbb{R}^n \to \mathbb{R}^n$ is said to be strongly infinitesimally contracting with rate $c > 0$ if the map $x \mapsto f(t, x, \ldots)$ is uniformly one-sided Lipschitz with constant $-c$ for all $t$ and for all inputs.

Strongly contracting systems enjoy numerous properties. As a useful example, we state the following lemma without proof (as it slightly generalizes [6, Th. 3.15, Corollary 3.16]).

**Lemma 1 (Grönwall Comparison Lemma)**: Consider a dynamical system

$$\dot{x}(t) = f(t, x(t), u_1(t), \ldots, u_m(t)), \quad \forall t \geq 0, \quad (1)$$

with $x(t) \in \mathbb{R}^n$ and inputs $u_i \in U_i \subseteq \mathbb{R}^{k_i}$ for $i \in \{1, 2, \ldots, m\}$. Let $\| \cdot \|$ be a norm on $\mathbb{R}^n$, and let $\| \cdot \|_{U_i}$ be norms on $U_i$. Assume that

(i) the system (1) is strongly infinitesimally contracting with rate $c > 0$, and

(ii) for each $i \in \{1, 2, \ldots, m\}$, the maps $u_i \mapsto f(t, x, u_1, \ldots, u_i, \ldots, u_m)$ are uniformly Lipschitz continuous with constant $\ell_{f, U_i}$ for all $t \geq 0, x \in \mathbb{R}^n$, and $u_i \in U_i$ with $j \neq i$.

Let $(u_1, \ldots, u_m)$ and $(\bar{u}_1, \ldots, \bar{u}_m)$ be input signals, and let $x, \bar{x}$ be the corresponding trajectories of (1). For all $t \geq 0$,

$$\|x(t) - \bar{x}(t)\| \leq e^{-ct} \|x(0) - \bar{x}(0)\| + \sum_{i=1}^m \ell_{f, U_i} \int_0^t e^{-c(t-\tau)} \|u_i(\tau) - \bar{u}_i(\tau)\|_{U_i} \, d\tau. \quad (2)$$

Note that (2) still holds when $c \leq 0$, i.e., for expansive systems with a bounded rate of expansion; however, we do not consider such systems in this letter.

B. Optimal Control

We study the following optimal control problem:

**Problem 1 (Optimal Control Problem)**: Consider a dynamical system

$$\dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0 \in \mathbb{R}^n, \quad (3)$$

where $f$ is continuous in all arguments and continuously differentiable in the second and third arguments. Further consider a cost functional

$$J[u] = \int_0^T \phi(t, x(t), u(t)) \, dt + \psi(x(T)), \quad (4)$$

where $\phi: [0, T] \times \mathbb{R}^n \times U \to \mathbb{R}$ is a running cost that is differentiable in the second argument, and $\psi: \mathbb{R}^n \to \mathbb{R}$ is a differentiable terminal cost. Let $U \equiv \{u : [0, T] \to U \text{ s.t. } u \text{ measurable} \}$ be a space of permissible control signals, where $T > 0$ and $U \subseteq \mathbb{R}^k$ is a compact set containing $\Omega_u$. The optimal control problem is to find $u^* \in U$ that minimizes $J[u^*]$.

An elementary necessary condition for the optimality of a control is Pontryagin’s minimum principle (PMP) [2, Ths. 5.10 and 5.11], [5, Ths. 6.3.1 and 6.5.1]:

**Theorem 1 (Pontryagin’s Minimum Principle)**: Let $u^* \in U$ be an optimal control for Problem 1 (if one exists), and let $x: [0, T] \to \mathbb{R}^n$ be the corresponding trajectory of (3). For all $t \in [0, T]$,

$$u^*(t) \in \arg\min_{\bar{u} \in U} H(t, x(t), \lambda(t), \bar{u}), \quad (5)$$

where $H: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times U \to \mathbb{R}$ is the Hamiltonian

$$H(t, x, \lambda, u) = \lambda^T f(t, x, u) + \phi(t, x, u) \quad (6)$$

and $\lambda: [0, T] \to \mathbb{R}^n$ is the costate trajectory

$$\dot{\lambda}(t) = -\ell f(t, x(t), u(t))^T \lambda(t) - \phi_x(t, x(t), u(t)) \quad (7)$$

with the boundary condition $\lambda(T) = \psi_x(x(T))$

C. Method of Successive Approximations

The Method of Successive Approximations (MSA) [7], also called the Forward-Backward Sweep algorithm [17], is a basic approach to computing an input that satisfies PMP. The method iteratively solves the PMP two-point boundary value problem, then updates the control to minimize the new Hamiltonian at each time, as outlined in Algorithm 1.

The algorithm can run for a fixed number of iterations; alternatively, it may terminate when the difference between successive iterates $u^{(i-1)}$, $u^{(i)}$ is within a specified tolerance. Note that each iteration of the algorithm maps a control $u^{(i-1)}$ to a new control $u^{(i)}$, so that each iteration can be thought of as an operator $\text{MSA} : U \to U$.

**Definition 1 (MSA Operator)**: Given a control $u \in U$, let $x: [0, T] \to \mathbb{R}^n$ be the corresponding trajectory of (3), and let $\lambda: [0, T] \to \mathbb{R}^n$ be the trajectory of (7) from $\lambda(T) = \psi_x(x(T))$. Then $\text{MSA}(u)$ is the control that satisfies (5) with respect to $x(t)$ and $\lambda(t)$ for all $t \in [0, T]$, with ties broken in an arbitrary deterministic manner.

Definition 1 is well-posed if the signal of Hamiltonian-minimizing controls from (5) is measurable. When we analyze the MSA algorithm in Section IV, we will impose Lipschitz continuity assumptions that forbid any edge cases where $\text{MSA}(u)$ is not measurable.

D. Adjoint

Adjointss are familiar from linear systems theory. Given input and output Hilbert spaces $\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{out}}$ and a linear system
G : $S_{in} \rightarrow S_{out}$, the adjoint of $G$ is the unique linear system $G^\ast : S_{out} \rightarrow S_{in}$ such that $(Gu, y)_{S_{out}} = (u, G^\ast y)_{S_{in}}$ for all $u \in S_{in}$ and $y \in S_{out}$. For an LTV system with the usual $(A, B, C, D)$ representation, the adjoint dynamics are

$$\dot{\lambda}(t) = -A(t)^\top \lambda(t) - C(t)^\top v(t)$$

$$z(t) = B(t)^\top \lambda(t) + D(t)^\top v(t)$$

with $v \in S_{out}$ and $z \in S_{in}$ [11, 3.2.4]. The theory of adjoints leads to the duality of controllability and observability and of linear quadratic regulators and estimators [15].

### III. Contractivity of the Adjoint

This section examines adjoints of nonlinear systems. We first explain how the notion of “adjoint” frequently used in the optimal control literature relates to the adjoint from linear systems. We then prove a simple yet powerful result: that the adjoint of a strongly infinitesimally contracting system is itself strongly infinitesimally contracting, with respect to the dual norm, when integrated backwards in time. This dual contractivity property leads to useful bounds for the evolution of costates, to later be employed in Section IV.

#### A. Adjoint of Nonlinear Systems

Nonlinear systems do not properly have adjoints according to the definition in Section II-D. Instead, the adjoint of the system’s linearized variational dynamics is often referred to as its adjoint [9], [13]. Consider the nonlinear system (3) with output $y(t) = x(t)$. Let $u(t)$ be an input signal corresponding to a nominal trajectory $x(t)$, let $\tilde{x}(t)$ be the trajectory from $\tilde{u}(t)$. Linearizing the dynamics of $\delta x(t) = \tilde{x}(t) - x(t)$ from $\delta u(t) = \tilde{u}(t) - u(t)$,

$$(\delta x)(t) = D_x f(t, \lambda(t), u(t)) \delta x(t) + D_{u,f}(t, x(t), u(t)) \delta u(t)$$

$$(\delta y)(t) = \delta x(t)$$

so by (8a), the adjoint dynamics are

$$\dot{\lambda}(t) = -D_x f(t, \lambda(t), u(t))^\top \lambda(t) - v(t)$$

$$z(t) = D_{u,f}(t, x(t), u(t))^\top \lambda(t)$$

where $v(t) \in V \subseteq \mathbb{R}^n$. Not coincidentally, the costate dynamics (7) from PMP are of the form (9a), with a forcing term $v(t) = \phi_0(t, x(t), u(t))$ from the running cost. Indeed, PMP can be derived from the variational linearization described above; see [5, Ths. 2.3.1 and 6.1.1].

#### B. Contractivity of the Adjoint

We now examine the adjoints of strongly contracting systems. When the original system is contracting with respect to a norm $\|\cdot\|$, it is natural to study the adjoint system using the dual norm $\|\cdot\|^\ast$, as the following lemma suggests.

**Lemma 2 (Dual Lipschitz Constants):** Let $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ be a norm, and let $\|\cdot\|^\ast$ be its dual norm. Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a pair of continuously differentiable vector fields, such that $D_x f(x) = D_x g(x)^\top$ for all $x \in \mathbb{R}^n$. Then

(i) $\text{Lip}_f(\|\cdot\|) = \text{Lip}_{g}(\|\cdot\|^\ast)$, and

(ii) $\text{osLip}_f(\|\cdot\|) = \text{osLip}_g(\|\cdot\|^\ast)$.

The following is an immediate consequence of Lemma 2:

**Theorem 2 (Dual Contraction):** Consider the pair of dynamical systems (3) and (9a). Let $T > 0$ and $c > 0$, let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, and let $\lambda^\ast(t) = \lambda(T - t)$ be the time-reversed trajectory of (9a) (where we study the time-reversed dynamics due to the minus sign in the vector field). The following are equivalent:

(i) the $x(t)$ system is strongly infinitesimally contracting with respect to $\|\cdot\|$ with rate $c$, and

(ii) the $\lambda^\ast(t)$ system is strongly infinitesimally contracting with respect to $\|\cdot\|^\ast$, with rate $c$.

**Proof:** Let $g$ be the function

$$g(t, \tilde{x}, \tilde{u}) \triangleq \frac{dx^\ast(t)}{dt} = D_x f(t, \tilde{x}, \tilde{u})^\top \lambda + v(t)$$

For all fixed $t$, $\tilde{x}$, and $\tilde{u}$, we have $D_x g(t, \tilde{x}, \tilde{u}) = D_x f(t, \tilde{x}, \tilde{u})^\top$. Hence, applying Lemma 2 to the maps $\tilde{g}(\lambda) = g(t, \lambda, \tilde{u})$ and $\tilde{f}(\lambda) = f(t, \lambda, \tilde{u})$, we obtain $\text{osLip}_f(\|\cdot\|) \leq c$ and only if the maps $x \mapsto f(t, \lambda, \tilde{u})$ have the same property with respect to $\|\cdot\|^\ast$. $\blacksquare$

#### C. Bounds on Adjoint Dynamics

Theorem 2 establishes that $\lambda^\ast(t)$ is strongly contracting so long as the original system is strongly contracting, so we can exploit standard bounds on contracting systems to bound the evolution of $\lambda(t)$. Before stating these bounds, we impose the following two assumptions:

**Assumption 1 (Strong Contractivity):** The system (3) is strongly infinitesimally contracting with rate $c > 0$, i.e., $\text{osLip}_f(\|\cdot\|) \leq c$ for all $t \in [0, T]$ and $\tilde{u} \in U$. Furthermore, the trajectory of (3) on the interval $[0, T]$ with $u(t) = \emptyset_k$ is bounded.

**Assumption 2 (Lipschitz Continuity, Pt. I):** For all fixed $t \in [0, T]$ and $x \in \mathbb{R}^n$, the map $u \mapsto f(t, \lambda, u)$ from $(\mathbb{R}^n, \|\cdot\|)$ into $(\mathbb{R}^n, \|\cdot\|^\ast)$ is Lipschitz with constant $c_f$.

Strong contractivity is a fairly strong assumption. For example, if some $x^\ast \in \mathbb{R}^n$ is an equilibrium point of the unforced system for all $t$, strong contractivity implies that $x^\ast$ is globally exponentially stable (due to Lemma 1). Due to Theorem 2, the assumption also implies that the adjoint dynamics are also strongly contracting. Consequently, we can prove that all state and costate trajectories remain bounded.

**Lemma 3 (Boundedness of State and Costate):** Consider the system (3) and its adjoint (7). If the input spaces $U \subseteq (\mathbb{R}^n, \|\cdot\|)$ and $V \subseteq (\mathbb{R}^n, \|\cdot\|^\ast)$ are bounded, then under Assumptions 1 and 2, there exist bounded sets $X \subseteq (\mathbb{R}^n, \|\cdot\|)$ and $\Delta \subseteq (\mathbb{R}^n, \|\cdot\|^\ast)$ such that $x(t) \in X$ and $\lambda(t) \in \Delta$ for all $t \in [0, T]$ and measurable $u : [0, T] \rightarrow U$ and $v : [0, T] \rightarrow V$.

In the remainder of this letter, we will let $X, \Delta \subseteq \mathbb{R}^n$ be the bounded sets guaranteed by Lemma 3. In particular, the boundedness of $\lambda(t)$ allows us to impose additional Lipschitz continuity assumptions:

**Assumption 3 (Lipschitz Continuity, Pt. II):** For all fixed $t \in [0, T]$, $\tilde{x} \in \mathbb{R}^n$, $\tilde{u} \in U$, and $\tilde{\lambda} \in \Delta$,

(i) the map $x \mapsto D_x f(t, \tilde{x}, \tilde{u})^\top \tilde{\lambda}$ from $(\mathbb{R}^n, \|\cdot\|)$ into $(\mathbb{R}^n, \|\cdot\|^\ast)$ is Lipschitz with constant $c_{f,x}$, and

(ii) the map $u \mapsto D_{u,f}(t, \tilde{x}, \tilde{u})^\top \tilde{\lambda}$ from $(\mathbb{R}^n, \|\cdot\|)$ into $(\mathbb{R}^n, \|\cdot\|^\ast)$ is Lipschitz with constant $c_{f,u}$.

We are now ready to state the first bound on the evolution of the adjoint trajectories.

**Theorem 3 (Grönwall Comparison of Costates):** Consider the system (3) and its adjoint (9a) with Assumptions 1–3. Let $u, \tilde{u} : [0, T] \rightarrow U$ and $v, \tilde{v} : [0, T] \rightarrow V$ be two pairs of measurable input signals, and let $\lambda, \tilde{\lambda} : [0, T] \rightarrow \mathbb{R}^n$ be the...
corresponding adjoint trajectories. Then for all $t \geq 0$, \[
\|\lambda(t) - \tilde{\lambda}(t)\| \leq e^{-c(T-t)}\|\lambda(0) - \tilde{\lambda}(0)\| + \int_t^T e^{-c(T-s)}\|\psi(s) - \tilde{\psi}(s)\|\,ds + \int_t^T e^{-c(T-s)}\|u(s) - \tilde{u}(s)\|\,ds.
\]

Theorem 3 provides a somewhat unwieldy bound. We can sacrifice its sharpness to obtain a much simpler incremental ISS property.

Corollary 1 (Incremental ISS of Adjoint Systems): Under the same hypotheses as Theorem 3,
\[
\sup_{t \in [0,T]} \|\lambda(t) - \tilde{\lambda}(t)\| \leq \|\lambda(0) - \tilde{\lambda}(0)\| + \kappa \sup_{t \in [0,T]} \|\psi(t) - \tilde{\psi}(t)\|_c + \left(\ell_{f,\kappa} + \ell_{f,\kappa}\ell_{f,\kappa}^2\right) \sup_{t \in [0,T]} \|u(t) - \tilde{u}(t)\|_U,
\]
where \[
\kappa = c^{-1}(1 - e^{-cT}).
\]

IV. APPLICATIONS TO OPTIMAL CONTROL

Here we show how the contractivity of the adjoint system leads to the contractivity of the MSA iteration, under additional Lipschitz continuity assumptions.

Assumption 4 (Lipschitz Continuity of Cost Gradients): For all fixed $t \in [0,T]$, $\bar{x} \in X$, and $\bar{u} \in \mathcal{U}$,
(i) the map $x \mapsto \phi(t, x, \bar{u})$ from $(\mathbb{R}^n, \|\cdot\|)$ into $(\mathbb{R}^n, \|\cdot\|)$ is Lipschitz with constant $\ell_{\phi, x}$,
(ii) the map $u \mapsto \phi(t, \bar{x}, u)$ from $(\mathbb{R}^k, \|\cdot\|_U)$ into $(\mathbb{R}^n, \|\cdot\|)$ is Lipschitz with constant $\ell_{\phi, u}$, and
(iii) the map $x \mapsto \psi(x)$ from $(\mathbb{R}^n, \|\cdot\|)$ into $(\mathbb{R}^n, \|\cdot\|)$ is Lipschitz with constant $\ell_{\psi, x}$.

Assumption 5 (Lipschitz Continuity of the Optimum): There exists a continuous map $h : [0, T] \times X \times \Lambda \to \mathcal{U}$ such that
\[
h(t, x, \lambda) \in \arg\min_{u \in \mathcal{U}} H(t, x, \lambda, u)
\]
for all $t \in [0, T]$, $x \in X$, and $\lambda \in \Lambda$, with ties broken in an identical manner as the MSA operator, where for all fixed $t \in [0, T]$, $\bar{x} \in X$, and $\bar{\lambda} \in \Lambda$,
(i) the map $x \mapsto h(t, x, \bar{\lambda})$ from $(\mathbb{R}^n, \|\cdot\|)$ into $(\mathbb{R}^k, \|\cdot\|_U)$ is Lipschitz with constant $\ell_{h,x}$, and
(ii) the map $\lambda \mapsto h(t, \bar{x}, \lambda)$ from $(\mathbb{R}^n, \|\cdot\|)$ into $(\mathbb{R}^k, \|\cdot\|_U)$ is Lipschitz with constant $\ell_{h,\lambda}$.

Notice that Assumption 5 implies that MSA($u$) is measurable for any $u \in \mathcal{U}$. With these Lipschitz assumptions, we can finally bound the Lipschitz constant of the MSA operator.

Theorem 4 (Contractivity of MSA): Suppose that Problem 1 is nonsingular and satisfies Assumptions 1–5, and consider the norm $\|\cdot\|_U : \mathcal{U} \to \mathbb{R}_{\geq 0}$ given by
\[
\|u\|_U = \sup_{t \in [0,T]} \|u(t)\|_U.
\]
The following are true:
(i) The Lipschitz constant of an MSA iteration with respect to the $\|\cdot\|_U$ norm is bounded by
\[
\text{Lip}(\text{MSA}) \leq b_1 \kappa + b_2 \kappa^2
\]
where
\[
b_1 = \ell_{h,x}\ell_f,_{u} + \ell_{h,\lambda}(\ell_{\psi, x} \ell_{f, u} + \ell_{\phi, u} + \ell_{f, u}) \quad (16a)
b_2 = \ell_{h,\lambda}\ell_f,_{x}(\ell_{\psi, x} + \ell_{f, x}) \quad (16b)
\]
(ii) If $b_1 \kappa + b_2 \kappa^2 < 1$, then the MSA operator is a contraction; hence it has a unique fixed point $\hat{u} \in \mathcal{U}$, the MSA iterates $u^{(i)} = \text{MSA}(u^{(i)})$ converge to $\hat{u}$ from any initial guess $u^{(0)} \in \mathcal{U}$, and
\[
\|u^{(i)}(t) - \hat{u}(t)\|_U \leq \frac{(b_1 \kappa + b_2 \kappa^2)^i}{1 - b_1 \kappa - b_2 \kappa^2} \|u^{(0)}(t) - \hat{u}(t)\|_U
\]
for all $t \in [0, T]$.

Corollary 2 (Uniqueness and Sufficiency): Under the same hypotheses as Theorem 4, if additionally
(i) the Hamiltonian has a unique minimizer for all $t \in [0, T]$, $x \in X$, and $\lambda \in \Lambda$,
(ii) an optimal control $u^*$ exists, and
(iii) the time horizon $T$ is sufficiently small or the contraction rate $c$ is sufficiently large that Lip(MSA) < 1, then $u^*$ is the unique optimal control, and PMP is a sufficient condition for optimality.

V. CONCLUSION

In this letter, we have examined an indirect method for the optimal control of strongly contracting systems. We have observed that the time-reversed adjoints of such systems are also contracting with the same rate, with respect to the dual norm, leading to useful bounds on the costate trajectory from PMP. Based on this observation, we bounded the Lipschitz constant of each iteration of MSA, demonstrating that the iteration is actually a contraction mapping for sufficiently strongly contracting systems or for sufficiently short horizons. In these cases, MSA is guaranteed to converge to a unique control that satisfies PMP. With an additional assumption on pointwise uniqueness of the minimizer of the Hamiltonian, we showed that this control is indeed the unique optimal control.

The main approach of this letter, namely using ISS properties of the adjoint to bound the Lipschitz constant of a $\mathcal{U} \to \mathcal{U}$ operator, is quite general and could be applied to many other indirect methods in optimal control. Several variants of MSA, both older [7], [21] and newer [18], [19], could be studied with this type of analysis in future work, possibly with more general convergence criteria. Another practical future direction would be the study of discretized implementations of the forward and backward integration steps, as in [20]. Of course, one could also analyze indirect methods for extensions of the optimal control problem, such as constraints on the terminal state or in the infinite time horizon. An additional interesting direction would be the application of convergence guarantees to model predictive control of contractive nonlinear systems.
APPENDIX A
PROOFS

A. Proof of Lemma 2

For a general matrix \( A \in \mathbb{R}^{n \times n} \), from the definitions of induced norms and dual norms we have

\[
\|A\|_* = \sup_{\|z\|_2 = 1} \langle A^T z, z \rangle = \sup_{\|z\|_2 = 1} \|A^T z\|_2^2.
\]

Swapping the order of the suprema and applying the definition of the dual norm once again yields

\[
\|A\|_* = \sup_{\|z\|_2 = 1} \sup_{\|x\|_2 = 1} \langle x^T A^T z, z \rangle = \sup_{\|z\|_2 = 1} \|A^T z\|_*.
\]

But \( \mathbb{R}^n \) with any norm is a reflexive Banach space, so \( \|\cdot\|_* = \|\cdot\| \), and thus \( \|A\|_* = \|A^T\|_* \). We use this fact to prove both statements. Because \( g \) is continuously differentiable and \( D_x g(x) = D_f(x)^T \),

\[
\text{Lip}_{\|\cdot\|_*} (g) = \sup_{x \in X} \|D_x g(x)\|_* = \sup_{x \in X} \|D_f(x)\| = \text{Lip}_{\|\cdot\|} (f)
\]

and

\[
\text{oS} \|\cdot\|_* (g) = \sup_{x \in X \ a \to 0^+} \frac{\| h_a(x) + \alpha D_x g(x) \|_* - 1}{\alpha} = \sup_{x \in X \ a \to 0^+} \frac{\| h_a(x) + \alpha D_f(x) \| - 1}{\alpha} = \text{oS} \|\cdot\| (f)
\]

B. Proof of Lemma 3

Let \( \tilde{x}(t) \) be the trajectory of (3) corresponding to input \( \tilde{u}(t) = 0 \). Since (3) is strongly infinitesimally contracting, we can use Lemma 1 to compare a trajectory \( x(t) \) with \( \tilde{x}(t) \):

\[
\|x(t) - \tilde{x}(t)\| \leq \frac{\ell_{fu}}{c} (1 - e^{-cT}) \sup_{t \in [0,T]} \|u(\tau)\| U
\]

for all \( t \in [0, T] \). Since \( U \) and \( \tilde{x}(t) \) are bounded, \( x(t) \) is bounded as well. Similarly, the time-reversed costate dynamics (9) have an equilibrium point at the origin when \( v(t) = 0 \), regardless of \( x(t) \) and \( u(t) \), and (due to Theorem 2) they are strongly contracting with rate \( c > 0 \). Again, we can use Lemma 1 to compare \( \lambda^{-}(t) \) with the trajectory of:

\[
\|\lambda^{-}(t)\|_* \leq \|\lambda^{-}(0)\|_* + \frac{1}{c} (1 - e^{-cT}) \sup_{t \in [0,T]} \|v(\tau)\|_*
\]

for all \( t \in [0, T] \). Since \( V \) is bounded, \( \lambda^{-}(t) \) is confined to a ball \( \Lambda \) about the origin.

C. Proof of Theorem 3

As in Theorem 2, let \( \lambda^{-}(t) = \lambda(T - t) \), so that

\[
\frac{d\lambda^{-}(t)}{dt} = D_f(T - t, x(T - t), u(T - t)) \lambda^{-}(t) - v(t - T)
\]

At any fixed \( t \), the \( \lambda^{-} \) vector field has the Jacobian matrix \( D_f(T - t, x(T - t), u(T - t)) \), which is transpose the Jacobian matrix of \( f(T - t, \cdot) \). By Assumption 1, \( \text{oS}(f(T - t, \cdot), u(T - t)) \leq -c \), so Lemma 2 implies that the \( \lambda^{-}(t) \) vector field is also one-sided Lipschitz with constant \( c \), with respect to \( \|\cdot\| \). Then we apply Lemma 1 to bound \( \|\lambda^{-}(t) - \tilde{\lambda}(t)\|_* \)

with respect to the inputs \( u(t), x(t) \), and \( v(t) \), resulting in the following bound on \( \|\lambda(t) - \tilde{\lambda}(t)\|_* \):

\[
\|\lambda(t) - \tilde{\lambda}(t)\|_* \leq e^{-c(T-t)} \|\lambda(T) - \tilde{\lambda}(T)\|_*
\]

\[
+ \ell_{fu} \int_t^T e^{-c(t-\tau)} \|u(\tau) - \tilde{u}(\tau)\| U d\tau
\]

\[
+ \ell_{fx} \int_t^T e^{-c(T-\tau)} \|x(\tau) - \tilde{x}(\tau)\|_X d\tau
\]

\[
+ \int_t^T e^{-c(T-\tau)} \|v(\tau) - \tilde{v}(\tau)\|_* d\tau
\]

We apply Lemma 1 once more to remove explicit dependence on \( x \), via the bound

\[
\int_t^T e^{-c(T-\tau)} \|x(\tau) - \tilde{x}(\tau)\|_X d\tau
\]

\[
\leq \ell_{fu} \int_t^T \int_0^T e^{-c(t-\tau)} e^{-c(T-\tau)} \|u(\tau') - \tilde{u}(\tau')\| U d\tau d\tau'
\]

We then swap the order of integration:

\[
\int_t^T \int_0^T e^{-c(t-\tau)} e^{-c(T-\tau)} \|u(\tau') - \tilde{u}(\tau')\| U d\tau d\tau'
\]

\[
= \int_t^T e^{-c(T-t)} \int_0^T \frac{e^{-c(t-\tau)} \|u(\tau') - \tilde{u}(\tau')\| U d\tau}{c}
\]

\[
+ \frac{e^{-c(T-t)}}{c} \int_0^T \sinh(c(T-\tau)) \|u(\tau) - \tilde{u}(\tau)\| U d\tau
\]

D. Proof of Corollary 1

The first three terms are obvious upper bounds on the first three terms in (10), and

\[
\frac{\sinh(c(T-t))}{c} \int_0^T e^{-c(T-\tau)} d\tau + e^{-c(T-t)} \int_t^T \sinh(c(T-\tau)) d\tau
\]

\[
= \int_t^T \int_0^T e^{-c(t-\tau)} e^{-c(T-\tau)} d\tau d\tau' \leq \kappa \int_t^T e^{-c(T-\tau)} d\tau \leq \kappa^2
\]

E. Proof of Theorem 4

Let \( u, \tilde{u} \in U \), and let \( x, \tilde{x} \) and \( \lambda, \tilde{\lambda} \) be the corresponding state and costate trajectories. Then for all \( t \in [0, T] \),

\[
\|\text{MSA}(u)(t) - \text{MSA}(\tilde{u})(t)\|_U
\]

\[
= \|h(t, x(t), \lambda(t)) - h(t, \tilde{x}(t), \tilde{\lambda}(t))\|_U
\]

\[
\leq \ell_{hx} \|x(t) - \tilde{x}(t)\| + \ell_{h\lambda} \|\lambda(t) - \tilde{\lambda}(t)\|_*
\]

(17)

The costate dynamics are (9) with \( v(t) = -\phi_x(t, x(t), u(t)) \), which is bounded in \( \mathbb{R}^n \) by the boundedness of \( x(t) \) and \( u(t) \) and the Lipschitz continuity of \( \phi_x \). By Corollary 1,

\[
\|\lambda(t) - \tilde{\lambda}(t)\| \leq \|\lambda(0) - \tilde{\lambda}(0)\|_* + \kappa \sup_{t \in [0,T]} \|\phi_x(t, x(t), u(t)) - \phi_x(t, \tilde{x}(t), \tilde{u}(t))\|_*
\]

\[
+ \left( \ell_{fx} k + \ell_{fx} \ell_{fu} k^2 \right) \sup_{t \in [0,T]} \|u(t) - \tilde{u}(t)\|_U,
\]
where
\[
\|\lambda(T) - \hat{\lambda}(T)\| = \|\psi_x(x(T)) - \psi_x(\hat{x}(T))\|, \\
\leq \ell_{\psi_x,x}\|x(T) - \hat{x}(T)\| \leq \ell_{\psi_x,x} \sup_{t \in [0,T]} \|x(t) - \hat{x}(t)\| \leq \ell_{\psi_x,x} \sup_{t \in [0,T]} \|u(t) - \bar{u}(t)\|_U
\]
and
\[
\|\hat{\phi}_x(t, x(t), u(t)) - \phi_x(t, \hat{x}(t), \bar{u}(t))\| \leq \ell_{\hat{\phi}_x,u} \sup_{t \in [0,T]} \|x(t) - \hat{x}(t)\| + \ell_{\hat{\phi}_x,u} \sup_{t \in [0,T]} \|u(t) - \bar{u}(t)\|_U \leq \ell_{\hat{\phi}_x,u} \|u(t) - \bar{u}(t)\|_U \leq \ell_{\hat{\phi}_x,u} \sup_{t \in [0,T]} \|u(t) - \bar{u}(t)\|_U.
\]
As a consequence of Lemma 1, \(\min_{T \in [0,T]} \|x(T) - \hat{x}(T)\| \leq \ell_{\hat{\phi}_x,u} \sup_{t \in [0,T]} \|u(t) - \bar{u}(t)\|_U\), so we simplify
\[
\|\lambda(t) - \hat{\lambda}(t)\| \leq \ell_{\hat{\phi}_x,u} \ell_{\hat{f}_x,u}\kappa \sup_{t \in [0,T]} \|u(t) - \bar{u}(t)\|_U + \kappa \ell_{\hat{\phi}_x,u} \sup_{t \in [0,T]} \|u(t) - \bar{u}(t)\|_U + \ell_{\hat{\phi}_x,u} \sup_{t \in [0,T]} \|u(t) - \bar{u}(t)\|_U.
\]
Substituting the state and costate difference bounds into (17) completes the proof of statement 4. Then statement 4 is a standard consequence of the Banach fixed point theorem.

**F. Proof of Corollary 2**

We first establish that the fixed points of the MSA operator are precisely the controls that satisfy PMP. One direction is obvious: \(u^* = \text{MSA}(u^*)\) implies that \(u^*\) satisfies PMP. Now suppose that \(u^*\) satisfies PMP, and let \(x^*, \lambda^*\) be the corresponding state and costate trajectories. Then \(u^*(t) \in \arg\min_{u \in U} H(t, x^*(t), \lambda^*(t), u)\) for all \(t \in [0, T]\), so the assumption that the Hamiltonian has a unique minimizer implies that \(u^*(t) = h(t, x^*(t), \lambda^*(t))\) for all \(t \in [0, T]\), and thus \(u^* = \text{MSA}(u^*)\).

We then establish that the MSA iteration converges to a unique fixed point \(\hat{u}\). For \(T\) sufficiently small or \(c\) sufficiently large, \(\kappa\) is sufficiently small that \(Lip(\text{MSA}) \leq b_1\kappa + b_2\kappa^2 < 1\), by Theorem 4. Then the Banach fixed point theorem establishes that a unique fixed point \(\hat{u}\) exists, and that the iteration from any initial guess converges to \(\hat{u}\).

Since an optimal control \(u^*\) exists, it is a fixed point of MSA, and the fixed point of MSA is unique. Furthermore, if a control \(u^*\) satisfies PMP, then it is a fixed point of MSA, and hence is equal to the optimal control.

**REFERENCES**

[1] V. V. Aleksandrov, “On the accumulation of perturbations in the linear systems with two coordinates,” Vestnik MGU, vol. 3, pp. 67–76, Jan. 1968.

[2] M. Athans and P. L. Falb, *Optimal Control: An Introduction to the Theory and its Applications*. New York, NY, USA: McGraw-Hill, 1966.

[3] J. T. Betts, *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*. Philadelphia, PA, USA: SIAM, 2010.

[4] L. Böttcher, N. Antulov-Fantulin, and T. Asikis, “AI Pontryagin or how artificial neural networks learn to control dynamical systems,” *Nat. Commun.*, vol. 13, no. 1, pp. 1–9, 2022, doi: 10.1038/s41467-021-27590-0.

[5] A. Bressan and B. Piccoli, *Introduction to the Mathematical Theory of Control*, Springfield, MO, USA: Amer. Inst. Math. Sci., 2007.

[6] F. Bullo, *Control System Theory for Dynamical Systems*, 1st ed. Kindle Direct Publ., 2022. [Online]. Available: http://motion.me.ucsb.edu/book-cds

[7] F. L. Chernousko and A. A. Lyubushin, “Method of successive approximations for solution of optimal control problems,” *Optimal Control Appl. Methods*, vol. 5, no. 2, pp. 101–114, 1982, doi: 10.1002/oca.4660030201.

[8] B. A. Conway, “A survey of methods available for the numerical optimization of continuous dynamic systems,” *J. Optim. Theory Appl.*, vol. 152, no. 2, pp. 271–306, 2012, doi: 10.1007/s10957-011-9918-z.

[9] P. E. Crouch, F. Lannabbi-Lagarrigue, and A. J. van der Schaft, “Adjoint and Hamiltonian input-output differential equations,” *IEEE Trans. Autom. Control*, vol. 40, no. 4, pp. 603–615, Apr. 1995, doi: 10.1109/9.376115.

[10] W. E, “A proposal on machine learning via dynamical systems,” *Commun. Math. Stat.*, vol. 1, no. 5, pp. 1–11, 2017, doi: 10.1007/s40304-017-0103-z.

[11] M. Green and D. J. N. Limebeer, *Linear Robust Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1995.

[12] H. B. Keller, *Numerical Methods for Two-Point Boundary-Value Problems*. New York, NY, USA: Dover Publ., 2018.

[13] H. J. Kelley, R. E. Kopp, and H. G. Moyer, “Successive approximation techniques for trajectory optimization,” Grumman Aircraft Eng. Corp., Bethpage, NY, USA, Rep., 1961. [Online]. Available: https://apps.dtic.mil/Annex/MIT/AD0268321.pdf

[14] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[15] O. Kouba and D. S. Bernstein, “What is the adjoint of a linear system? [lecture notes],” *IEEE Control Syst. Mag.*, vol. 40, no. 3, pp. 62–70, Jun. 2020, doi: 10.1109/MCS.2020.2976389.

[16] I. A. Krylov and F. L. Chernous’ko, “On a method of successive approximations for the solution of problems of optimal control,” *USSR Comput. Math. Math. Phys.*, vol. 2, no. 6, pp. 1371–1382, 1963, doi: 10.1016/0041-5553(63)90353-7.

[17] S. Lenhart and J. T. Workman, *Optimal Control Applied to Biological Models*. Boca Raton, FL, USA: Chapman Hall, 2007.

[18] Q. Li, L. Chen, C. Tai, and W. E, “Maximum principle based algorithms for deep learning,” *J. Mach. Learn. Res.*, vol. 18, no. 165, pp. 1–29, 2018. [Online]. Available: http://jmlr.org/papers/v18/17-653.html

[19] Q. Li and S. Hao, “An optimal control approach to deep learning and applications to discrete-weight neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2985–2994. [Online]. Available: https://openreview.net/pdf?id=SB17-653.html

[20] M. McAsley, L. Mou, and W. Han, “Convergence of the forward-backward sweep method in optimal control,” *Comput. Optim. Appl.*, vol. 53, no. 1, pp. 207–226, 2012, doi: 10.1007/s10589-011-9454-7.

[21] S. K. Mitter, “Successive approximation methods for the solution of optimal control problems,” *Automatica*, vol. 3, nos. 3–4, pp. 135–149, 1966, doi: 10.1016/0003-6813(66)90009-4.

[22] A. V. Rao, “A survey of numerical methods for optimal control,” *Adv. Astronaut. Sci.*, vol. 135, no. 1, pp. 497–528, 2009.

[23] O. V. Stryk, “Numerical solution of optimal control problems by direct collocation,” in *Optimal Control*, Basel, Switzerland: Springer, 1993, pp. 129–143, doi: 10.1007/978-3-0348-7539-4_10.

[24] P. Tabuada and B. Gharesifard, “Universal approximation power of deep residual neural networks through the lens of control,” *IEEE Trans. Autom. Control*, early access, Jul. 12, 2022, doi: 10.1109/TAC.2022.3190051.

[25] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong, “You only propagate once: Accelerating adversarial training via maximal principle,” in *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY, USA: Curran, 2019, doi: 10.48550/arXiv.1905.00877.