2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Online Learning-based Optimal Control of Nonlinear Systems with Finite-Time Convergence Guarantees

Nick-Marios T. Kokolakis and Kyriakos G. Vamvoudakis

***Abstract*— This paper develops a critic-only reinforcement learning-based algorithm for learning the solution to the Hamilton-Jacobi-Bellman equation in finite time. In particular, a non-Lipschitz experience replay-based learning law utilizing recorded and current data is introduced for updating the critic weights to learn the value function. The non-Lipschitz property of the dynamics gives rise to finite-time convergence and stability, while the experience replay-based approach eliminates the need to satisfy the persistence of excitation condition if the recorded data is sufficiently rich. Simulation results demonstrate the efficacy of the proposed approach.**

***Index Terms*— Adaptive learning, finite-time stability, optimal control, reinforcement learning, autonomy.**

I. Introduction

Exploiting the benefits of *reinforcement learning* (RL) [1], the control systems community has conducted a considerable effort towards enabling *cognitive autonomy* by designing control mechanisms that run in real-time and adapt to changes in the environment. This gives rise to *intelligent autonomous systems* (IAS) exhibiting features such as the strong ability to learn new tasks, adaptivity under uncertainty, real-time optimality, and tolerance to unpredictable failures [2]. Nevertheless, to ensure the effective operation of IAS without human intervention, it is necessary for the decision-making mechanism to generate optimal policies in finite-time rather than in an infinite time.

Optimal control theory deals with finding a control law for a given dynamical system so that a user-prescribed cost functional is optimized [3]. In the infinite horizon optimal control problem, the notions of optimality and asymptotic stability are intertwined [4]. In particular, the optimal control strategy is a state feedback law establishing asymptotic stability while minimizing the performance measure. In fact, to derive the optimal control policy, one needs first to determine the optimal cost function (value function) by solving a nonlinear partial differential equation, the so-termed Hamilton–Jacobi–Bellman (HJB) equation [5]. Nonetheless, analytically solving the HJB equation is a challenging task while being usually computationally intractable, thereby giving rise to the development of *adaptive dynamic programming* (ADP) techniques [6]–[9].

N-M. T. Kokolakis and K. G. Vamvoudakis are with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA e-mail: nmkokolakis@gatech.edu, kyriakos@gatech.edu.

This work was supported in part by ARO grant No. W911NF-19 − 1 − 0270, ONR Minerva grant No. N00014 − 18 − 1 − 2160, NSF grant Nos. CAREER CPS-1851588 and SATC-1801611, and the Onassis Scholarship [Scholarship ID: F ZR 025/1-2021/2022].

Related work

ADP unifies *optimal* [5] and *adaptive* [10] control towards developing adaptive learning mechanisms enabling the learning of solutions to optimal control problems by employing measured data along the system trajectories [11]–[19]. The ADP algorithms are developed through an *actor-critic structure* involving two approximators. Specifically, a *critic network* that evaluates the performance of a control policy and an *actor network* that computes this policy. It is evident that the vast majority of the existing adaptive learning algorithms for solving optimal control problems [2], [20] converge to a near-optimal control law provided that a *persistence of excitation* (PE) [10] condition is satisfied. On the other hand, *concurrent learning/experience replay*-based ADP algorithms [21], [22] allow the learning of the solution to the optimal control problem by requiring a weaker form of a PE condition to be satisfied [23], [24]. In fact, these algorithms are data-driven and leverage recorded and instantaneous data concurrently for the adaptation of the critic weights.

The aforementioned approaches and the references therein concern the design of adaptive learning-based mechanisms for solving the HJB equation associated with the *infinite horizon optimal control problem* by means of an actor-critic structure. However, as we already mentioned, the solution to the infinite horizon optimal control problem renders the equilibrium point of the closed-loop system asymptotically stable [4]. The concept of *asymptotic stability* in dynamical systems allows the convergence of system trajectories to a Lyapunov stable equilibrium point over the infinite horizon [25]. On the contrary, the notion of *finite-time stability* enables the convergence of the system solutions to a Lyapunov stable equilibrium state in finite time [26]. In real-world applications, it is imperative to design decision-making mechanisms guaranteeing optimality as well as finite-time stability. In the context of optimal control, this necessity is captured by the *finite-time optimal control* problem first stated in [27], that is, the problem of finding state-feedback control laws that optimize a given performance functional while guaranteeing finite-time stability of the closed-loop system. To the best of our knowledge, an ADP approach enabling learning of the solution to the HJB equation in finite-time is absent from the literature.

Contributions: The contributions of the present paper are threefold. First, a RL-based framework is developed for learning online and in finite time the optimal value function and the optimal control policy. Then, a non-Lipschitz expe-

rience replay-based adaptive learning law for updating the critic weights is introduced while ensuring finite-time stability properties provided that the recorded data is sufficiently rich. Finally, the proposed scheme relies on the use of only a critic network, allowing the simultaneous learning of the value function and the optimal strategy, thus leading to a less computationally expensive learning structure.

Structure: The remainder of the paper is structured as follows. Section II states the finite-time optimal control problem. In Section III, a critic-only learning framework is developed for learning online and in finite time the solution to the optimal control problem. Section IV provides simulation results. Finally, Section V concludes and provides future work directions.

Notation: The notation used in this paper is standard. Specifically, $\|\cdot\|_p \triangleq \left[\sum_{i=1}^{n} |x_i|^p\right]^{1/p},\ 1 \leqslant p < \infty$, denotes the Hölder p-norm of a vector. The induced 2-norm for the matrix $Q \in \mathbb{R}^{m\times n}$ is defined as $\|Q\| \triangleq \sqrt{\lambda_{\max}\left(Q^T Q\right)} = \sigma_{\max}(Q)$, with $\lambda_{\max}$ (resp., $\lambda_{\min}$) denoting the maximum (resp., minimum) eigenvalue and $\sigma_{\max}$ (resp., $\sigma_{\min}$) denoting the maximum (resp., minimum) singular value. The gradient of a scalar-valued function V with respect to a vector-valued variable x is defined as a row vector and is denoted by $V'(x)$. We define the open ball $\mathcal{B}_\varepsilon\left(x_{\mathrm{e}}\right) \triangleq \{x \in \mathbb{R}^n : \|x - x_{\mathrm{e}}\| < \varepsilon\}$ centered at x_{e} with radius ε in the Euclidean norm, while the corresponding closed ball is denoted as $\mathcal{B}_\varepsilon[x_{\mathrm{e}}]$. Let $\lceil\cdot\rfloor^\eta \triangleq |\cdot|^\eta \operatorname{sign}(\cdot)$, where $|\cdot|$ and $\operatorname{sign}(\cdot)$ operate componentwise and $\eta > 0$. The distance of a point $x_0 \in \mathbb{R}^n$ to a closed set $C \subseteq \mathbb{R}^n$ in the norm $\|\cdot\|$ is defined as $\operatorname{dist}\left(x_0, C\right) \triangleq \inf_{x\in C}\{\|x_0 - x\|\}$. The $\mathcal{X} \times \mathcal{Y}$ is the Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$. Finally, $\partial\mathcal{S}$ and $\mathcal{S}^{\mathrm{c}}$ denote the boundary and the complement of the set $\mathcal{S}$, respectively.

II. Finite-Time Optimal Control Problem

Consider the following continuous-time dynamical system

$$\begin{aligned}\dot{x}(t) =& F(x(t), u(t))\\ =& f(x(t)) + G(x(t))u(t), \quad x(0) = x_0, \quad t \geqslant 0, \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the control input, $f : \mathbb{R}^n \to \mathbb{R}^n$ and $G : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are continuous on $\mathbb{R}^n$ with $f(0) = 0$.

Define the cost functional associated with (1) as

$$J\left(x_0, u(\cdot)\right) \triangleq \int_0^\infty L(x(t), u(t))\mathrm{d}t \tag{2}$$

with

$$L(x, u) \triangleq L_1(x) + L_2(x)u + u^{\mathrm{T}} R(x)u, \tag{3}$$

where $L_1 : \mathbb{R}^n \to \mathbb{R},\ L_2 : \mathbb{R}^n \to \mathbb{R}^{1\times m}$, and $R : \mathbb{R}^n \to \mathbb{R}^{m\times m}$ are continuous on $\mathbb{R}^n$, and $R(x) > 0,\ x \in \mathbb{R}^n$.

We now introduce the notion of *finite-time stability*.

Definition 1. [26]. The zero solution $x(t) \equiv 0$ to (1) with $u(t) \equiv 0$ is *finite-time stable* if it is Lyapunov stable and finite-time convergent, i.e., for all $x(0) \in \mathcal{N}\backslash\{0\}$, where $\mathcal{N} \subseteq \mathbb{R}^n$ is some open neighborhood of the origin, $\lim_{t\to T(x(0))} x(t) = 0$, where $T(\cdot)$ is the settling-time function such that $T(x(0)) < \infty,\ x(0) \in \mathcal{N}$. The zero solution $x(t) \equiv 0$ to (1) with $u(t) \equiv 0$ is *globally finite-time stable* if it is finite-time stable with $\mathcal{N} = \mathbb{R}^n$. □

Definition 2. [28]. Let $\mathcal{N} \subseteq \mathbb{R}^n$ be an open neighborhood of the origin. The compact set $M \subset \mathcal{N}$ is *finite-time attractive* with respect to (1) with $u(t) \equiv 0$ if for every $x(0) \in \mathcal{N}$, the solution $x(t),\ t \geqslant 0$, satisfies $\operatorname{dist}\left(x(t), M\right) = 0,\ t \geqslant T(x(0))$, where $T(\cdot)$ is the settling-time function such that $T(x(0)) < \infty,\ x(0) \in \mathcal{N}$. Furthermore, the compact set M is *globally finite-time attractive* if it is finite-time attractive with $\mathcal{N} = \mathbb{R}^n$. □

Next, we state the finite-time optimal stabilization problem by following the formulation of [27].

Problem 1. For each initial condition $x_0 \in \mathbb{R}^n$, define the set of *globally finite-time stabilizing controllers* $\mathcal{S}\left(x_0\right) \triangleq \{u(\cdot) : x(\cdot)$ given by (1) satisfying $x(t) \to 0$ as $t \to T(x(0))\}$. The objective amounts to finding a *globally finite-time stabilizing optimal control law* $u^\star(\cdot) \in \mathcal{S}\left(x_0\right),\ x_0 \in \mathbb{R}^n$, rendering the equilibrium point of the closed-loop system (1) with $u = u^\star(x)$ globally finite-time stable while minimizing the performance index (2). □

The finite-time optimal stabilization problem involves the minimization

$$V(x_0) \triangleq \min_{u(\cdot)\in\mathcal{S}(x_0)} J\left(x_0, u(\cdot)\right), \quad x_0 \in \mathbb{R}^n,$$

subject to (1). Note that the function $V(\cdot)$ is the value function and can be thought of as the optimal cost (cost-to-go) from x_0.

Define the Hamiltonian function

$$\begin{aligned} H(x, u, V'^{\mathrm{T}}(x)) \triangleq L(x, u) + V'(x)F(x, u),\\ (x, u) \in \mathbb{R}^n \times \mathbb{R}^m. \end{aligned} \tag{4}$$

Applying the stationary condition to the Hamiltonian function (4), one obtains the feedback control law $u^\star(x)$, which is the global minimizer of the Hamiltonian function for all $x \in \mathbb{R}^n$ since $H(x, u, V'^{\mathrm{T}}(x))$ is convex in u. Namely,

$$\begin{aligned} u^\star(x) \triangleq& \arg\min_{u\in\mathbb{R}^m} H\left(x, u, V'^{\mathrm{T}}(x)\right)\\ =& -\frac{1}{2}R^{-1}(x)\left[L_2(x) + V'(x)G(x)\right]^{\mathrm{T}}. \end{aligned} \tag{5}$$

Plugging (5) into (4), one can derive the HJB equation

$$\begin{aligned} 0 =& L_1(x) + V'(x)f(x) - \frac{1}{4}\left[V'(x)G(x) + L_2(x)\right]\\ &\cdot R^{-1}(x)\left[V'(x)G(x) + L_2(x)\right]^{\mathrm{T}}. \end{aligned} \tag{6}$$

Alternatively, the HJB equation (6) can be written in the compact form

$$H(x, u^\star(x), V'^{\mathrm{T}}(x)) = \min_{u\in\mathbb{R}^m} H\left(x, u, V'^{\mathrm{T}}(x)\right) = 0. \tag{7}$$

The next theorem provides sufficient conditions allowing us to characterize an optimal feedback controller attaining stabilization of the closed-loop system in finite time.

Theorem 1. *Consider the controlled nonlinear dynamical system* (1) *with performance index* (2)*. Suppose that there exist a radially unbounded continuously differentiable function* $V : \mathbb{R}^n \to \mathbb{R}$*, real numbers* $c > 0$ *and* $\beta \in (0, 1)$*, and a continuous control law* $u^\star : \mathbb{R}^n \to \mathbb{R}^m$ *such that*

$$\begin{aligned} u^\star(0) &= 0, \\ V(0) &= 0, \\ V(x) &> 0, \quad x \in \mathbb{R}^n \backslash \{0\}, \\ V'(x)F(x, u^\star(x)) &\leqslant -c(V(x))^\beta, \quad x \in \mathbb{R}^n, \\ H(x, u^\star(x), V'^{\mathrm{T}}(x)) &= 0, \quad x \in \mathbb{R}^n, \\ H(x, u, V'^{\mathrm{T}}(x)) &\geqslant 0, \quad (x, u) \in \mathbb{R}^n \times \mathbb{R}^m. \end{aligned}$$

Then, with the feedback control $u(\cdot) = u^\star(x(\cdot))$*, the zero solution* $x(t) \equiv 0$ *to* (1) *is globally finite-time stable with a settling-time function* $T : \mathbb{R}^n \to [0, \infty)$ *such that*

$$T(x_0) \leqslant \frac{1}{c(1-\beta)} \left(V(x_0)\right)^{1-\beta}, \quad x_0 \in \mathbb{R}^n, \tag{8}$$

and

$$J(x_0, u^\star(x(\cdot))) = V(x_0), \quad x_0 \in \mathbb{R}^n.$$

Furthermore, the feedback control $u(\cdot) = u^\star(x(\cdot))$ *minimizes* $J(x_0, u(\cdot))$ *in the sense that*

$$J(x_0, u^\star(x(\cdot))) = \min_{u(\cdot) \in \mathcal{S}(x_0)} J(x_0, u(\cdot)).$$

Proof. The proof follows from [27]. ∎

Remark 1. Although we deal with the *finite-time optimal control problem*, we assess the system performance over the infinite horizon. In particular, the settling-time function depends on the system's initial conditions and satisfies the inequality (8). Thus, in view of radial unboundedness, it follows that the time of convergence to the equilibrium point may increase (possibly unboundedly) as the vector norm of the initial condition increases. □

The problem of the *finite-time optimal control* amounts to solving the HJB equation (7), which is in general intractable aside from special cases. Thus, the next section will devise *learning-based techniques* for approximating the solution of the HJB equation.

III. Finite-Time Stable Online Learning

In this section, we develop a *learning-based* algorithm for learning *online* and in *finite time* the solution of the HJB equation (7) by utilizing data gathered along the system trajectories. Towards this, we will employ a critic structure, i.e., an approximator allowing us to *simultaneously* approximate the value function and the optimal controller.

A. Finite-Time Stable Tuning

According to the Weierstrass higher-order approximation theorem [29], we can locally approximate the value function $V(x)$ and its gradient on a compact set $\mathcal{X} \subset \mathbb{R}^n$ that includes the origin with a neural network approximator as

$$V(x) = W^{\star\mathrm{T}}\phi(x) + \epsilon(x), \quad x \in \mathcal{X},$$

$$V'^{\mathrm{T}}(x) = \phi'^{\mathrm{T}}(x)W^\star + \epsilon'^{\mathrm{T}}(x), \quad x \in \mathcal{X}, \tag{9}$$

where $W^\star \in \mathbb{R}^N$ is an ideal constant weight vector satisfying $\|W^\star\|_2 \leqslant W_{\mathrm{m}}$ for some $W_{\mathrm{m}} > 0$, $\phi : \mathcal{X} \to \mathbb{R}^N$ is a vector of basis functions such that $\varphi_i(0) = 0$ and $\varphi_i'(0) = 0$, $i = 1, \ldots, N$, N is the number of neurons in the hidden layer of the neural network, and $\epsilon(x)$ is an approximation error. Note that one has to select the basis functions $\varphi_i(x)$, $i = 1, \ldots, N$, properly in order that they form a complete independent basis set [30].

The optimal control law can be approximated as

$$u^\star = -\frac{1}{2}R^{-1}\left(L_2(x) + \left(\phi'^{\mathrm{T}}(x)W^\star + \epsilon'^{\mathrm{T}}(x)\right)^{\mathrm{T}} G(x)\right)^{\mathrm{T}}, \quad x \in \mathcal{X}.$$

Substituting (9) into (7), we obtain the approximate HJB equation

$$\begin{aligned} H\left(x, u^\star(x), \phi'^{\mathrm{T}}(x)W^\star\right) =& L(x, u^\star(x)) \\ &+ W^{\star\mathrm{T}}\phi'(x)F(x, u^\star(x)) \\ =& \epsilon_{\mathrm{HJB}}, \quad x \in \mathcal{X}, \end{aligned}$$

where $\epsilon_{\mathrm{HJB}} \triangleq -\epsilon'(x)F(x, u^\star(x))$ is the residual error coming from the value function approximation error.

However, the ideal weights $W^\star$ are unknown, and thus we consider a critic with estimates $\hat{W} \in \mathbb{R}^N$ of the form

$$\hat{V}(x) \triangleq \hat{W}^{\mathrm{T}}\phi(x), \quad x \in \mathcal{X}, \tag{10}$$

and an approximate optimal controller given by

$$\hat{u} \triangleq -\frac{1}{2}R^{-1}\left(L_2(x) + \left(\phi'^{\mathrm{T}}(x)\hat{W}\right)^{\mathrm{T}} G(x)\right)^{\mathrm{T}}, \quad x \in \mathcal{X}. \tag{11}$$

Plugging the approximate value function (10) and the approximate optimal control law (11) into (4), one obtains the approximate HJB equation

$$\hat{H}\left(x, \hat{u}, \phi'^{\mathrm{T}}(x)\hat{W}\right) \triangleq \hat{W}^{\mathrm{T}}\phi'(x)F(x, \hat{u}) + L(x, \hat{u}), \quad x \in \mathcal{X}, \tag{12}$$

which is available for measurement, unlike the parameter error $\tilde{W} \coloneqq \hat{W} - W^\star$, which is not since $W^\star$ is unknown.

Define the Hamiltonian estimation error corresponding to the data collected at the current time $t \geqslant 0$ as

$$\begin{aligned} e(t) \triangleq& \hat{H}\left(x(t), \hat{u}(t), \phi'^{\mathrm{T}}(x(t))\hat{W}(t)\right) \\ &- H\left(x(t), u^\star(x(t)), V'^{\mathrm{T}}(x(t))\right) \\ =& \hat{H}\left(x(t), \hat{u}(t), \phi'^{\mathrm{T}}(x(t))\hat{W}(t)\right), \quad t \geqslant 0, \end{aligned}$$

and the Hamiltonian error associated with the recorded data at the time instants $0 \leqslant t_1, \ldots, t_k < t$ as

$$\begin{aligned} e(t_i, t) \triangleq& \hat{H}\left(x(t_i), \hat{u}(t_i), \phi'(x(t_i))^{\mathrm{T}}\hat{W}(t)\right) \\ \triangleq& \hat{W}^{\mathrm{T}}(t)\omega(t_i) + L(x(t_i), \hat{u}(t_i)), \end{aligned}$$

where $\omega(t) \triangleq \phi'(x(t))F(x(t), \hat{u}(t))$, $t \geqslant 0$.

Next, define the cost function of the current and past Hamiltonian estimation errors for $\gamma \in (0,1)$ as

$$E(\hat{W}(t)) \triangleq \frac{1}{\gamma+1}\left(\left|\frac{e(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\right|^{\gamma+1} + \sum_{i=1}^{k}\left|\frac{e\left(t_i,t\right)}{\omega^{\mathrm{T}}\left(t_i\right)\omega\left(t_i\right)+1}\right|^{\gamma+1}\right), \quad t \geqslant 0.$$

The *finite-time convergent data-driven* learning law for updating the critic weights is derived using a gradient descent algorithm as

$$\begin{aligned}\dot{\hat{W}} =& -\alpha\frac{\partial E(\hat{W}(t))}{\partial \hat{W}}\\ =& -\alpha\frac{\omega(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\left[\frac{e(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\right]^{\gamma}\\ &-\alpha\sum_{i=1}^{k}\frac{\omega(t_i)}{\omega^{\mathrm{T}}(t_i)\omega(t_i)+1}\left[\frac{e\left(t_i,t\right)}{\omega^{\mathrm{T}}\left(t_i\right)\omega\left(t_i\right)+1}\right]^{\gamma},\\ &\hat{W}\left(0\right)=\hat{W}_0, \quad t\geqslant 0, \end{aligned} \tag{13}$$

where $\alpha > 0$ is a constant gain that dictates the learning rate.

In the sequel, by taking into account (12) and (13), one can derive the parameter error dynamics as

$$\begin{aligned}\dot{\tilde{W}}(t) =& -\alpha\frac{\omega(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\left[\frac{\omega^{\mathrm{T}}(t)\tilde{W}(t)+\epsilon_{\mathrm{H}}(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\right]^{\gamma}\\ &-\alpha\sum_{i=1}^{k}\frac{\omega(t_i)}{\omega^{\mathrm{T}}(t_i)\omega(t_i)+1}\\ &\cdot\left[\frac{\omega^{\mathrm{T}}(t_i)\tilde{W}(t)+\epsilon_{\mathrm{H}}(t_i)}{\omega^{\mathrm{T}}\left(t_i\right)\omega\left(t_i\right)+1}\right]^{\gamma},\\ &\tilde{W}\left(0\right)=\tilde{W}_0, \quad t\geqslant 0, \end{aligned} \tag{14}$$

where $\epsilon_{\mathrm{H}} \triangleq H\left(x,\hat{u},\phi'^{\mathrm{T}}(x)W^{\star}\right)$ behaves as a disturbance stemming from the value function approximation error, and thus it is imperative to investigate its boundedness relative to the number of neurons N in the hidden layer of the critic neural network. The next proposition examines this issue.

Proposition 1. *For every $\epsilon_m > 0$ there exist $L(\epsilon_m) > 0$ and $N_0(\epsilon_m) > 0$ such that $\sup_{x\in\mathcal{X}}|\epsilon_H| < L(\epsilon_m),\ N \geqslant N_0(\epsilon_m)$. Furthermore, If $N \to \infty$, then $\epsilon_H \equiv 0$.*

Proof. It has been omitted due to space limitations and will be presented in the journal version of this work. ■

Before proceeding to our main theorem establishing the *finite-time stability properties* of our concurrent learning law, the following definition introducing the concept of a *sufficiently rich* data set is needed.

Definition 3. The recorded data set $\{\omega\left(t_i\right)\}_{i=1}^{k}$ is *k-sufficiently rich* if the matrix $\Omega \triangleq [\omega\left(t_1\right)\ldots\omega\left(t_k\right)]$ has $\mathrm{rank}(\Omega) = N$. □

It follows from Definition 3 that a recorded data set is *k-sufficiently rich* if and only if the set $\{\omega\left(t_i\right)\}_{i=1}^{k}$ contains N linearly independent vectors.

Theorem 2. *Consider the weight parameter error dynamics given by* (14)*. Define $\bar{\omega}(\cdot) \triangleq \frac{\omega(\cdot)}{\omega^{T}(\cdot)\omega(\cdot)+1}$, let $\bar{\epsilon}_{\mathrm{Hm}},\ \bar{\omega}_m > 0$, and $\theta \in (0,1)$, and assume that the recorded data set $\{\bar{\omega}\left(t_i\right)\}_{i=1}^{k}$ is k-sufficiently rich. Then the following statements hold:*

i) If $\epsilon_H \equiv 0$, then the zero solution $\tilde{W}(t) \equiv 0$ to (14) *is globally uniformly finite-time stable with a settling-time function $T : \mathbb{R}^N \to [0,\infty)$ such that*

$$T\left(\tilde{W}\left(0\right)\right) \leqslant \frac{\|\tilde{W}(0)\|_2^{1-\gamma}}{\sigma_{\min}^{\gamma+1}(\bar{\Omega})\alpha(1-\gamma)}, \quad \tilde{W}\left(0\right)\in\mathbb{R}^N,$$

where $\gamma \in (0,1)$.

ii) Define

$$\nu \triangleq \sigma_{\min}^{\gamma+1}(\bar{\Omega})(1-\theta)(1-\gamma).$$

If $\epsilon_H \not\equiv 0$, then the solutions to (14) *are globally uniformly ultimately bounded with the ultimate bound*

$$\mu \triangleq \left(\frac{(k+1)\bar{\epsilon}_{\mathrm{Hm}}^{\gamma}\bar{\omega}_m}{\theta\sigma_{\min}^{\gamma+1}(\bar{\Omega})}\right)^{\frac{1}{\gamma}}$$

and a settling-time function $T : \mathbb{R}^N \to [0,\infty)$ such that

$$T(\tilde{W}(0)) \leqslant \frac{\|\tilde{W}(0)\|_2^{1-\gamma}-\mu^{1-\gamma}}{\alpha\nu}, \quad \tilde{W}\left(0\right)\in\mathbb{R}^N.$$

In particular, for every initial condition $\tilde{W}(0) \in \mathbb{R}^N$, the solution $\tilde{W}(t),\ t \geqslant 0$, to (14) *satisfies*

$$\begin{aligned}\|\tilde{W}(t)\|_2 &\leqslant \sqrt{2\alpha}\left(\frac{\|\tilde{W}(0)\|_2^{1-\gamma}}{(2\alpha)^{\frac{1-\gamma}{2}}}-\frac{(2\alpha)^{\frac{\gamma+1}{2}}\nu}{2}t\right)^{\frac{1}{1-\gamma}}, && t < T\left(\tilde{W}(0)\right),\\ \|\tilde{W}(t)\|_2 &\leqslant \mu, && t \geqslant T\left(\tilde{W}(0)\right).\end{aligned}$$

Proof. It has been omitted due to space limitations and will be presented in the journal version of this work. ■

Remark 2. In the absence of the value function approximation error, the critic weights will converge to the optimal weights $W^{\star}$ in *finite time*. However, even though the settling-time function is unknown, it is upper bounded by a strictly increasing function of $\|\tilde{W}(0)\|_2$ that depends on the parameter α. Thus, the larger the learning rate α is, the faster the convergence of the parameter error to the origin will be. On the other hand, in the presence of the value function approximation error, it turns out that for every initial condition $\tilde{W}(0) \in \mathbb{R}^N$, the solution $\tilde{W}(t),\ t \geqslant 0$, to (14) reaches the compact set $\mathcal{B}_{\mu}[0]$ in *finite time*, that is, at most in time $\frac{\|\tilde{W}(0)\|_2^{1-\gamma}-\mu^{1-\gamma}}{\alpha\sigma_{\min}^{\gamma+1}(\bar{\Omega})(1-\theta)(1-\gamma)}$, and remains therein for all future time. Note that $T(\tilde{W}(0)) = 0$ if and only if $\tilde{W}(0) \in \mathcal{B}_{\mu}[0]$. Finally, one can reduce the parameter error along with the settling time by choosing the parameters γ and k properly since they determine the size of μ, that is, the size of the ball $\mathcal{B}_{\mu}[0]$, as well as the upper bound of the settling time $\frac{\|\tilde{W}(0)\|_2^{1-\gamma}-\mu^{1-\gamma}}{\alpha\sigma_{\min}^{\gamma+1}(\bar{\Omega})(1-\theta)(1-\gamma)}$ for every initial condition $\tilde{W}\left(0\right)\in\mathbb{R}^N$. □

The next theorem investigates the stability properties of the augmented dynamics composed of the adaptive law (13) for updating the critic and the resulting closed-loop dynamics after substituting in (1) the approximate optimal controller (11), that is,

$$\dot{x}(t) = f(x(t)) + G(x(t))\hat{u}(t), \quad x(0) = x_0, \quad t \geqslant 0. \tag{15}$$

Theorem 3. *Consider the weight parameter error dynamics* (14) *with the closed-loop dynamics* (15)*, and let* $\tilde{Z} \triangleq [\tilde{W}^{\mathrm{T}},\ x^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^N \times \mathcal{X}$ *be the state vector of the augmented dynamics* (14) *and* (15)*. Define* $\bar{\omega}(\cdot) \triangleq \frac{\omega(\cdot)}{\omega^{\mathrm{T}}(\cdot)\omega(\cdot)+1}$*, let* $\bar{\epsilon}_{\mathrm{Hm}},\ \bar{\omega}_m,\ d > 0$*, and* $\gamma,\ \theta \in (0,1)$*. Suppose that the recorded data set* $\{\bar{\omega}(t_i)\}_{i=1}^{k}$ *is* k*-sufficiently rich, let* $\epsilon_H \not\equiv 0$*, and define*

$$\bar{\mu} \triangleq \left(\frac{d + (k+1)\bar{\epsilon}_{\mathrm{Hm}}^{\gamma}\bar{\omega}_m}{\theta\sigma_{\min}^{\gamma+1}(\bar{\Omega})}\right)^{\frac{1}{\gamma}}.$$

Then, the compact set $\mathcal{Z} \triangleq \mathcal{B}_{\bar{\mu}}[0] \times \mathcal{X} \subset \mathbb{R}^N \times \mathcal{X}$ *is finite-time attractive with a settling-time function* $\bar{T} : \mathbb{R}^N \times \mathcal{X} \to [0,\infty)$ *such that*

$$\bar{T}\left(\tilde{Z}(0)\right) \leqslant \frac{\|\tilde{W}(0)\|_2^{1-\gamma} - \bar{\mu}^{1-\gamma}}{\alpha\sigma_{\min}^{\gamma+1}(\bar{\Omega})(1-\theta)(1-\gamma)}, \quad \tilde{Z}(0) \in \mathbb{R}^N \times \mathcal{X}.$$

In particular, for every initial condition $\tilde{Z}(0) \in \mathbb{R}^N \times \mathcal{X}$*, the solutions* $\tilde{Z}(t),\ t \geqslant 0$*, to* (14) *and* (15) *satisfy* $\operatorname{dist}\left(\tilde{Z}(t), \mathcal{Z}\right) = 0,\ t \geqslant \bar{T}\left(\tilde{Z}(0)\right)$.

Proof. It has been omitted due to space limitations and will be presented in the journal version of this work. ■

Remark 3. According to Theorem 3, it follows that for every initial condition $\tilde{Z}(0) \in \mathbb{R}^N \times \mathcal{X}$, there exists $\bar{T}\left(\tilde{Z}(0)\right) \geqslant 0$ such that the solution $\tilde{Z}(t),\ t \geqslant 0$, to (14) and (15) converges to $\mathcal{Z}$ in *finite time*. However, note that $\bar{T}\left(\tilde{Z}(0)\right) = 0$ if and only if $\tilde{Z}(0) \in \mathcal{Z}$. Finally, note that $\mathcal{Z}$ is a compact set and not necessarily a ball associated with a particular norm. □

IV. Simulation Results

Consider a spacecraft with one axis of symmetry given by

$$\dot{\omega}_1(t) = I_{23}\omega_3\omega_2(t) + u_1(t),\ \omega_1(0) = \omega_{10}, \quad t \geqslant 0, \tag{16}$$

$$\dot{\omega}_2(t) = -I_{23}\omega_3\omega_1(t) + u_2(t),\ \omega_2(0) = \omega_{20}, \tag{17}$$

where $I_{23} \triangleq (I_2 - I_3)/I_1$, I_1, I_2, and I_3 denote the spacecraft principal moments of inertia such that $0 < I_1 = I_2 < I_3$, $\omega_1 : [0,\infty) \to \mathbb{R}$, $\omega_2 : [0,\infty) \to \mathbb{R}$, and $\omega_3 \in \mathbb{R}$ are the components of the angular velocity vector with respect to a given inertial reference frame expressed in a central body reference frame, and u_1 and u_2 denote the spacecraft control moments. Note that the dynamical system (16) and (17) can be cast in the form of (1) with $n = 2$, $m = 2$, $x = [\omega_1,\ \omega_2]^{\mathrm{T}}$, $f(x) = [I_{23}\omega_3\omega_2,\ -I_{23}\omega_3\omega_1]^{\mathrm{T}}$, $G(x) = I_2$, and $u = [u_1,\ u_2]^{\mathrm{T}}$.

Next, towards structuring the finite-time optimal control problem, the terms composing the performance integrand (3) are given by $L_1(x) = \left(\frac{2}{3}\omega_1\|x\|_2^{-\frac{2}{3}} + I_{23}\omega_3\omega_2\right)^2 + \left(-\frac{2}{3}\omega_2\|x\|_2^{-\frac{2}{3}} + I_{23}\omega_3\omega_1\right)^2$, $L_2(x) = 2\left[-I_{23}\omega_3\omega_2,\ I_{23}\omega_3\omega_1\right]$, and $R(x) = I_2$.

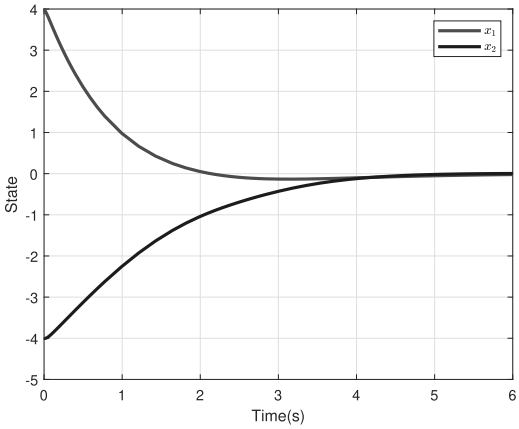

Fig. 1. The time evolution of the state trajectories $x(t),\ t \geqslant 0$. Once the learning procedure terminates, the state trajectories converge to the origin in finite time.

However, by employing inverse optimal control arguments, the authors of [27] have unveiled that the value function together with the optimal control law are given by $V(x) = \left(x^{\mathrm{T}}x\right)^{\frac{2}{3}}$ and $u^\star(x) = \left[\ -\frac{2}{3}\omega_1\|x\|_2^{-\frac{2}{3}} - I_{23}\omega_3\omega_2,\quad -\frac{2}{3}\omega_2\|x\|_2^{-\frac{2}{3}} + I_{23}\omega_3\omega_1\ \right]^{\mathrm{T}}$, respectively, whereas the settling-time function $T : \mathbb{R}^2 \to [0,\infty)$ is such that $T(x_0) \leqslant \frac{9}{4}\left(\omega_{10}^2 + \omega_{20}^2\right)^{\frac{1}{3}},\ x_0 \in \mathbb{R}^2$.

Let $I_1 = I_2 = 0.4\ \mathrm{kg}\cdot\mathrm{m}^2$, $I_3 = 0.2\ \mathrm{kg}\cdot\mathrm{m}^2$, $\omega_{10} = 4$ Hz, $\omega_{20} = -4$ Hz, $\omega_3 = 1$ Hz, $\alpha = 10$, and $\gamma = 0.9$. Concerning critic, the initial weights are randomly initialized within the interval $[0,1]$ and the basis functions are selected as $\phi(x) = \left[\ x_1^2,\quad x_1x_2,\quad x_2^2\ \right]^{\mathrm{T}}$. To enable the collection of sufficiently rich data along the closed-loop system trajectories, we inject a dithering excitation to the control input (11) in the form of $\rho(t) = 0.01\left(\sin(1.3\pi t) + \cos(1.3\pi t)\right),\ 0 \leqslant t \leqslant 0.5$.

Fig. 1 depicts the evolution of the controlled state trajectories over time. Note that $x(t) = 0$ for $t \geqslant 4.787$ sec, which confirms that $T(x_0) \leqslant 7.1433$ sec. Fig. 2 illustrates the finite-time convergence of the critic weights, while Fig. 3 shows the evolution of the approximate optimal control law (11). It is evident that the learning is attained in a *finite time* of $t = 4.452$ sec.

V. Conclusion and Future Work

This paper developed a critic-only RL-based algorithm for learning in finite time the value function alongside the optimal control policy associated with the finite-time optimal control problem for nonlinear systems. Under the assumption of sufficiently rich data, which is an easier condition to satisfy compared to the traditional PE condition, we devised a non-Lipschitz data-driven learning law for updating the critic weights while establishing finite-time stability via Lyapunov analysis. We also highlighted that the proposed learning

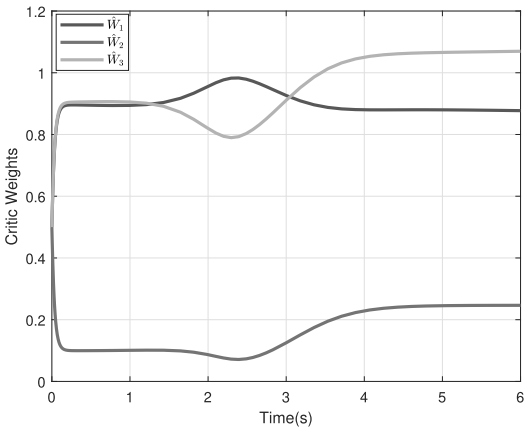

Fig. 2. The time evolution of the critic weights $\hat{W}(t),\ t \geqslant 0$. Note that the learning is achieved in a finite time of $t = 4.452$ sec.

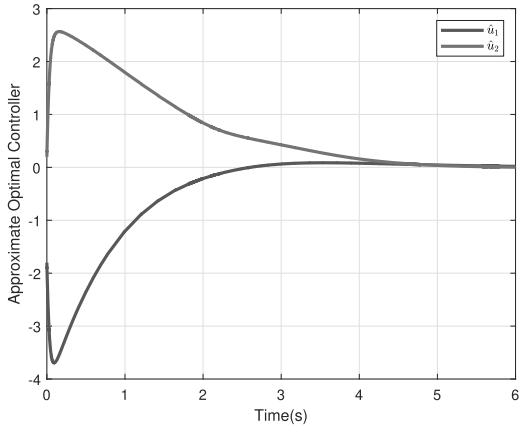

Fig. 3. The time evolution of the approximate optimal controller $\hat{u}(t),\ t \geqslant 0$.

mechanism is composed only of a critic and thus exhibits a lower complexity than other architectures in the literature whose structure additionally requires an actor. Simulation results validated the feasibility of the proposed learning algorithm. Future research endeavors will extend this framework to address the finite-time optimal trajectory tracking control problem.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press, 2018.

[2] K. G. Vamvoudakis and N.-M. T. Kokolakis, *Synchronous Reinforcement Learning-Based Control for Cognitive Autonomy*. Boston - Delft: Now Publishers, 2020, vol. 8, no. 1–2.

[3] D. Liberzon, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton, NJ: Princeton University Press, 2011.

[4] D. S. Bernstein, "Nonquadratic cost and nonlinear feedback control," *International Journal of Robust and Nonlinear Control*, vol. 3, no. 3, pp. 211–229, 1993.

[5] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. Hoboken, NJ: John Wiley & Sons, 2012.

[6] J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, *Handbook of Learning and Approximate Dynamic Programming*. Hoboken, NJ: John Wiley & Sons, 2004, vol. 2.

[7] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Hoboken, NJ: John Wiley & Sons, 2007, vol. 703.

[8] F. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 39–47, 2009.

[9] D. P. Bertsekas, *Reinforcement Learning and Optimal Control*. Belmont, MA: Athena Scientific, 2019.

[10] P. Ioannou and B. Fidan, *Adaptive Control Tutorial*. Philadelphia, PA: SIAM, 2006, vol. 11.

[11] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[12] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems*, vol. 32, no. 6, pp. 76–105, 2012.

[13] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, UK: Springer Science & Business Media, 2012.

[14] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ: John Wiley & Sons, 2013, vol. 17.

[15] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming with Applications in Optimal Control*. Cham, Switzerland: Springer, 2017.

[16] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042–2062, 2017.

[17] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. Hoboken, NJ: John Wiley & Sons, 2017.

[18] R. Kamalapurkar, P. Walters, J. Rosenfeld, and W. Dixon, *Reinforcement Learning for Optimal Feedback Control*. Cham, Switzerland: Springer, 2018.

[19] Z.-P. Jiang, T. Bian, and W. Gao, *Learning-Based Control: A Tutorial and Some Recent Results*. Boston - Delft: Now Publishers, 2020, vol. 8, no. 3.

[20] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London, UK: IET, 2013, vol. 2.

[21] G. Chowdhary and E. Johnson, "Concurrent learning for convergence in adaptive control without persistency of excitation," in *Proc. IEEE Conference on Decision and Control*, 2010, pp. 3674–3679.

[22] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine Learning*, vol. 8, no. 3-4, pp. 293–321, 1992.

[23] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[24] K. G. Vamvoudakis, M. F. Miranda, and J. P. Hespanha, "Asymptotically stable adaptive–optimal control algorithm with saturating actuators and relaxed persistence of excitation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2386–2398, 2016.

[25] W. M. Haddad and V. Chellaboina, *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach*. Princeton, NJ: Princeton University Press, 2011.

[26] S. P. Bhat and D. S. Bernstein, "Finite-time stability of continuous autonomous systems," *SIAM Journal on Control and Optimization*, vol. 38, no. 3, pp. 751–766, 2000.

[27] W. M. Haddad and A. L'Afflitto, "Finite-time stabilization and optimal feedback control," *IEEE Transactions on Automatic Control*, vol. 61, no. 4, pp. 1069–1074, 2015.

[28] A. Polyakov, "Nonlinear feedback design for fixed-time stabilization of linear control systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 8, pp. 2106–2110, 2011.

[29] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, no. 5, pp. 551–560, 1990.

[30] F. Lewis, S. Jagannathan, and A. Yesildirak, *Neural Network Control of Robot Manipulators and Non-Linear Systems*. Philadelphia , PA: CRC press, 2020.