Accelerated Continuous-Time Approximate Dynamic Programming via Data-Assisted Hybrid Control*

Daniel E. Ochoa*, Jorge I. Poveda

[a]*Department of Electrical, Energy and Computer Engineering. University of Colorado Boulder, Boulder, 80305, Colorado, USA*

Abstract

We introduce a new closed-loop architecture for the online solution of approximate optimal control problems in the context of continuous-time systems. Specifically, we introduce the first algorithm that incorporates dynamic momentum in actor-critic structures to control continuous-time dynamic plants with an affine structure in the input. By incorporating dynamic momentum in our algorithm, we are able to accelerate the convergence properties of the closed-loop system, achieving superior transient performance compared to traditional gradient-descent based techniques. In addition, by leveraging the existence of past recorded data with sufficiently rich information properties, we dispense with the persistence of excitation condition traditionally imposed on the regressors of the critic and the actor. Given that our continuous-time momentum-based dynamics also incorporate periodic discrete-time resets that emulate restarting techniques used in the machine learning literature, we leverage tools from hybrid dynamical systems theory to establish asymptotic stability properties for the closed-loop system. We illustrate our results with a numerical example.

Keywords: Approximate dynamic programming, concurrent learning, hybrid systems, Lyapunov theory.

1. Introduction

Recent technological advances in computation and sensing have incentivized the development and implementation of data-assisted feedback control techniques previously deemed intractable due to their computational complexity. Among these techniques, reinforcement learning (RL) has emerged as a practically viable tool with remarkable degrees of success in robotics [1], autonomous driving [2], water-distribution systems [3], among other cyber-physical applications, see [4]. These types of algorithms, are part of a large landscape of adaptive systems that aim to control a plant while simultaneously optimizing a performance index in a model-free way, with closed-loop stability guarantees.

In this paper, we focus on a particular class of infinite horizon RL problems from the perspective of approximate optimal control and approximate adaptive dynamic programming (AADP). Specifically, we study the optimal control problem for nonlinear continuous-time and control-affine deterministic plants, interconnected with approximate adaptive optimal controllers [5] in an actor-critic configuration. These types of adaptive controllers aim to find, in real time, the solution to the Hamilton-Jacobi-Bellman (HJB) equation by measuring the output of the nonlinear dynamical system while making use of two approximation structures:

- a critic, used to estimate the optimal value function of the optimal control problem, and
- an actor, used to estimate the optimal feedback controller.

Our goal is to design online adaptive dynamics for the real-time tuning of the aforementioned structures, while simultaneously achieving closed-loop stability and high transient performance. To achieve this, and motivated by

*Research supported in part by NSF grant number CNS-1947613.

*Corresponding Author.

Email addresses: daniel.ochoa@colorado.edu (Daniel E. Ochoa), jorge.poveda@colorado.edu (Jorge I. Poveda)

the widespread usage of momentum-based gradient dynamics in practical RL settings [6], we study continuous-time actor-critic dynamics inspired by a class of ordinary differential equations (ODEs) that can be seen as continuous-time counterparts of Nesterov's accelerated optimization algorithm [7]. Such types of algorithms have gained popularity in optimization and related fields due to the fact that they can minimize smooth convex functions at a rate of order $\mathcal{O}(1/t^2)$ [8]. The main source for the acceleration property in these ODEs comes from the addition of momentum to gradient-based dynamics, in conjunction with a vanishing dynamic damping coefficient. However, as recently shown in [9] and [10], the non-uniform convergence properties that emerge in these types of dynamics complicates their use in feedback systems with plant dynamics in the loop. In this paper, we overcome these challenges by incorporating resets into the proposed momentum-based algorithms, similar to restarting heuristics studied in the machine learning literature, see [11] and [7]. Our resulting actor-critic controller is naturally modeled by a hybrid dynamical system that incorporates continuous-time and discrete-time dynamics, which we analyze using tools from [12].

A traditional assumption in the literature of continuous-time actor-critic RL is that the regressors used in the parameterizations satisfy a persistence of excitation condition along the trajectories of the plant. However, in practice, this condition can be difficult to verify a priori. To circumvent this issue, in this paper we consider a data-assisted approach, where a finite amount of past "sufficiently rich" recorded data is used to guarantee asymptotic learning in the closed-loop system. As a consequence, the resulting data-assisted hybrid control algorithm concurrently uses real-time and recorded data, similar in spirit to concurrent-learning (CL) techniques [13]. By using Lyapunov-based tools for hybrid dynamical systems, we analyze the interconnection of an actor-critic neural-network (NN) controller and the nonlinear plant, establishing that the trajectories of the closed-loop system remain ultimately bounded around the origin of the plant and the optimal actor and critic NN parameters. Since the resulting closed-loop system has suitable regularity properties in terms of continuity of the dynamics, our stability results are in fact robust with respect to arbitrarily small additive disturbances that can be adversarial in nature, or that can arise due to numerical implementations. To the best knowledge of the authors, these are the first theoretical stability guarantees of continuous-time accelerated actor-critic algorithms for neural network-based adaptive dynamic programming controllers in nonlinear deterministic settings.

The rest of this paper is organized as follows: Section 2 presents the notation and some concepts on hybrid dynamical systems, Section 3 presents the problem statement and some preliminaries on optimal control. Section 4 introduces the hybrid momentum-based dynamics for the update of the critic NN, Section 5 presents the update dynamics for the actor NN, and Section 6 studies the properties of closed-loop system. In Section 7 we study a numerical example illustrating our theoretical results.

2. Preliminaries

Notation: We denote the real numbers by $\mathbb{R}$, and we use $\mathbb{R}_{\geq 0} \subset \mathbb{R}$ to denote the non-negative real line. We use $\mathbb{R}^n$ to represent the n-dimensional Euclidean space and $|\cdot|$ to denote its usual vector norm. Given $A \in \mathbb{R}^{n\times n}$, we use $|A|$ to denote the induced 2-norm for matrices, and we infer its distinction with the vector norm depending on the context. We use $\mathrm{Tr}\,(A)$ to denote the trace operator on matrices. Given a compact set $\mathcal{A} \subset \mathbb{R}^n$ and a vector $z \in \mathbb{R}^n$, we use $|z|_{\mathcal{A}} \coloneqq \min_{s\in\mathcal{A}} |z - s|$ to represent the minimum distance of z to $\mathcal{A}$. We also use $r\mathbb{B}$ to denote a closed ball in the Euclidean space, of radius $r > 0$, and centered at the origin. We use $I_n \in \mathbb{R}^{n\times n}$ to denote the identity matrix, and (x, y) for the concatenation of the vectors x and y, i.e., $(x, y) \coloneqq [x^\top, y^\top]^\top$. A function $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to be of class-$\mathcal{K}$ ($\gamma \in \mathcal{K}$), if it is continuous, zero at zero, and nondecreasing. A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to be of class-$\mathcal{KL}$ ($\beta \in \mathcal{KL}$) if $\beta(\cdot, s) \in \mathcal{K}$ for each $s \in \mathbb{R}_{\geq 0}$, it is non-increasing in its second argument, and $\lim_{s\to\infty} \beta(r, s) = 0$ for each $r \in \mathbb{R}_{\geq 0}$. The gradient of a real valued function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as a column vector and denoted by ∇f. For a vector valued function $g : \mathbb{R}^n \to \mathbb{R}^m$, we use $\frac{\partial g(x)}{\partial x} \in \mathbb{R}^{m\times n}$ to denote its Jacobian matrix.

Hybrid Dynamical Systems: To study our algorithms, we will use tools from hybrid dynamical systems (HDS) theory [12]. A HDS with state $x \in \mathbb{R}^n$, has dynamics

$$x \in C,\ \dot{x} = F(x), \quad \text{and} \quad x \in D,\ \ x^+ = G(x), \tag{1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is called the *flow map*, $G : \mathbb{R}^n \to \mathbb{R}^n$ is called the *jump map*, and $C \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^n$ are closed sets, called the *flow set* and the *jump set*, respectively. We use $\mathcal{H} = (C, F, D, G)$ to denote the elements

of the HDS $\mathcal{H}$. Solutions $x : \text{dom}(x) \to \mathbb{R}^n$ to system (1) are indexed by a continuous-time parameter t, which increases continuously during flows, and a discrete-time index j, which increases by one during jumps. Thus, the notation $\dot{x}$ in (1) represents the derivative $\frac{dx(t,j)}{dt}$; and x^+ in (1) represents the value of x after an instantaneous jump, i.e., $x(t, j+1)$. Therefore, solutions $x : \text{dom}(x) \to \mathbb{R}^n$ to system (1) are defined on *hybrid time domains*. For a precise definition of hybrid time domains and solutions to HDS of the form (1), we refer the reader to [12, Ch.2]. The following definitions will be instrumental to study the stability and convergence properties of systems of the form (1).

Definition 1. *The compact set $\mathcal{A} \subset C \cup D$ is said to be* uniformly asymptotically stable *(UAS) for system* (1) *if* $\exists$ $\beta \in \mathcal{KL}$ *and* $r > 0$ *such that every solution* x *with* $x(0,0) \in r\mathbb{B} \cap (C \cup D)$ *satisfies:*

$$|x(t,j)|_{\mathcal{A}} \leq \beta(|x(0,0)|_{\mathcal{A}}, t+j), \ \forall\ (t,j) \in \textit{dom}(x). \tag{2}$$

When $\beta(r,s) = c_1 r e^{-c_2 s}$ for some $c_1, c_2 > 0$, the set $\mathcal{A}$ is said to be uniformly exponentially stable *(UES).* □

3. Problem Statement

Consider a control-affine nonlinear dynamical plant

$$\dot{x} = f(x) + g(x)u, \tag{3}$$

where $x \in \mathbb{R}^n$ is the state of the system, $u \in U \subset \mathbb{R}^m$ is the input, and $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ are locally Lipschitz functions. Our goal is to design a stable algorithm able to find –in real time– a control law u^* that minimizes the cost functional $V : \mathbb{R}^n \times \mathcal{U}_V \to \mathbb{R}$ given by:

$$V(x_0, u) \coloneqq \int_0^\infty r\Big(x\big(\tau\big), u\left(x(\tau)\right)\Big) d\tau, \tag{4}$$

where $x\big(t\big)$ represents a solution to (3) from the initial condition $x(0) = x_0$, that results from implementing a feedback law u, belonging to a class of admissible control laws $\mathcal{U}_V$ characterized as follows:

Definition 2. *[14, Definition 1] Given the dynamical system in* (3)*, a feedback control $u : \mathbb{R}^n \to \mathbb{R}^m$ is* admissible *with respect to the cost functional V in* (4) *if*

- *u is continuous,*
- *u renders system* (3) *UAS,*
- *$V(x_0, u) < \infty$ for all $x_0 \in \mathbb{R}^n$.* □

We denote the set of admissible feedback laws *as $\mathcal{U}_V$.*

In (4), we consider cost functions $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ of the form $r(x,u) \coloneqq Q(x) + R(u)$, where the state-cost is given by $Q(x) \coloneqq x^\top \Pi_x x$ with $\Pi_x \succ 0$, and the control-cost is given by $R(u) \coloneqq u^\top \Pi_u u$ with $\Pi_u \succ 0$. To find the optimal control law that minimizes (4), we study the *Hamiltonian function* $H : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ related to (3) and (4), given by

$$H(x, u, \nabla V) \coloneqq \nabla V^\top (f(x) + g(x)u) + Q(x) + R(u). \tag{5}$$

Using (5), a necessary optimality condition for u^* is given by Pontryagin's maximum principle [15]:

$$u^*(x) = \underset{u \in \mathcal{U}_V}{\arg\min}\, H(x, u, \nabla V^*) \implies u^*(x) = -\frac{1}{2}\Pi_u^{-1} g(x)^\top \nabla V^*(x), \tag{6}$$

where V^* represents the *optimal value function*:

$$V^*(x) \coloneqq \inf_{u \in \mathcal{U}_V} V(x, u(\cdot))$$

On the other hand, under the assumption that V^* is continuously differentiable, the optimal value function can be shown to satisfy the Hamilton-Jacobi-Bellman equation [5, Ch. 1.4]:

$$\frac{\partial V^*}{\partial t} = -H(x, u^*, \nabla V^*) \quad \forall x \in \mathbb{R}^n.$$

Since the functional in (4) does not have an explicit dependence on t, it follows that $\frac{\partial V^*}{\partial t} = 0$, and hence $H(x, u^*, \nabla V^*) = 0$, meaning that for all $x \in \mathbb{R}^n$, the following holds:

$$\nabla {V^*}^\top \big(f(x) + g(x)u^*(x)\big) + Q(x) + R\Big(u^*(x)\Big) = 0. \tag{7}$$

The time-invariant Hamilton-Jacobi-Bellman equation in (7), allows for a state-dependent characterization of optimality. Therefore, by using the optimal control law in (6), and assuming that the system dynamics (3) are known, the form (7) could be leveraged to find V^*. Unfortunately, finding an explicit closed-form expression for V^*, and thus for the optimal control law, is, in general, an intractable problem. However, the utility of (7) is not completely lost. As we shall show in the following sections, online and historical "measurements" of (7) can be leveraged in real time to estimate the optimal control law u^* while concurrently rendering a neighborhood of the origin of system (3) asymptotically stable.

4. Data-Assisted Critic Dynamics

To leverage the form of (7), we consider the following parameterization of the optimal value function $V^*(x)$:

$$V^*(x) = {\theta_c^*}^\top \phi_c(x) + \epsilon_c(x) \quad \forall x \in K, \tag{8}$$

where $K \subset \mathbb{R}^n$ is a compact set, $\theta_c^* \in \mathbb{R}^{l_c}$, $\phi_c : \mathbb{R}^n \to \mathbb{R}^{l_c}$ is a vector of continuously differentiable basis functions, and $\epsilon_c : \mathbb{R}^n \to \mathbb{R}$ is the approximation error. The parameterization (8) is always possible on compact sets due to the continuity properties of V and the universal approximation theorem [16]. This parametrization results in an optimal Hamiltonian of the form $H_p^* \coloneqq H(x, u^*, \frac{\partial \phi_c}{\partial x}^\top \theta_c^* + \nabla \epsilon_c)$ given by:

$$H_p^*(x) = {\theta_c^*}^\top \psi(x, u^*(x)) + Q(x) + R\left(u^*(x)\right) + \nabla \epsilon_c(x)^\top \left(f(x) + g(x)u^*(x)\right), \tag{9}$$

where we defined $\psi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{l_c}$ as:

$$\psi(x, u) \coloneqq \frac{\partial \phi_c(x)}{\partial x} \left(f(x) + g(x)u\right). \tag{10}$$

We note that the explicit dependence of $\psi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{l_c}$ on the control action u, defined in (10), is a fundamental departure from the previous approaches studied in the context of concurrent learning (CL) NN actor-critic controllers, such as those considered in [17] and [18]. In particular, we note that in the context of CL the data used to estimate the optimal value function V^* is generated from measurements of the optimal Hamiltonian which, by definition, incorporates the optimal control law u^*. Hence, the need to include u as part of the regressor vectors ψ becomes crucial; this dependence characterizes how far our recorded measurements of a Hamiltonian are from the optimal Hamiltonian H_p^*. Indeed, this distance will explicitly emerge in our convergence and stability analysis. Naturally, the dependence of (10) on u will impose stronger conditions on the recorded data needed to estimate V^*.
Assuming we have access to ϕ_c, we can define a *critic* neural network as:

$$\hat{V}(x) \coloneqq \theta_c^\top \phi_c(x), \ \ \forall x \in K, \tag{11}$$

which will serve as an approximation of the optimal value function V^* in (8). This critic NN results in an estimated Hamiltonian:

$$H\left(x, u, \nabla \hat{V}\right) \coloneqq \theta_c^\top \psi\left(x, u\right) + Q(x) + R(u), \tag{12}$$

which we will use to design the update dynamics of the critic parameters θ_c. In particular, our goal is to use previously recorded data from trajectories of the plant to ensure asymptotic stability of the set of optimal critic parameters $\{\theta_c^*\}$, while simultaneously enabling the incorporation of instantaneous measurements from the plant. Towards this end, we will assume enough "richness" properties in the recorded data, a notion that is captured by a relaxed (and finite-time) version of *persistence of excitation* (PE); see [13] and [19].

Assumption 1. *Let $\{\psi(x_k, u^*(x_k))\}_{k=1}^N$ be a sequence of recorded data, and define:*

$$\Lambda \coloneqq \sum_{k=1}^{N} \Psi(x_k, u^*(x_k))\Psi(x_k, u^*(x_k))^\top, \quad \Psi(x,u) \coloneqq \frac{\psi(x,u)}{1+\psi(x,u)^\top\psi(x,u)}. \tag{13}$$

There exists $\underline{\lambda} \in \mathbb{R}_{>0}$ such that $\Lambda \succeq \underline{\lambda} I_n$, i.e., the data is $\underline{\lambda}$-sufficiently-rich ($\underline{\lambda}$-SR). □

Remark 1. *In this paper, we study reinforcement learning dynamics that do not make explicit usage of exploration signals with standard PE properties, which can be difficult to guarantee in practice. Instead, we assume access to samples obtained by observing the action of optimal values $u^*(x_k)$ acting on the plant. Note however that this does not imply knowledge of the optimal control policy as a whole, but only of a finite number of* demonstrations *from an "expert" policy. Similar requirements commonly arise in the literature of imitation learning, or inverse reinforcement learning, and have been recently shown in practice to reduce the exploratory requirements of online reinforcement learning algorithms, with mild assumptions in the sampling of the demonstrations. For recent discussions on these topics in the discrete-time stochastic reinforcement learning setting we refer the reader to [20] and [21].*

Now, we consider the instantaneous and data-dependent errors of the estimated Hamiltonian with respect to the optimal one:

$$\begin{aligned} e^i(\theta_c, x, u) &\coloneqq H\left(x, u, \nabla\hat{V}\right) - H\left(x, u^*(x), \nabla V^*\right) \\ &= \theta_c^\top \psi(x,u) + Q(x) + R(u), \\ e_k^d(\theta_c) &\coloneqq H\left(x_k, u^*(x_k), \nabla\hat{V}\right) - H\left(x_k, u^*(x_k), \nabla V^*\right) \\ &= \theta_c^\top \psi(x_k, u^*(x_k)) + Q(x_k) + R(u^*(x_k)), \end{aligned}$$

where we used the fact that $H(x, u^*(x), \nabla V^*) = 0$. Moreover, we define the *joint instantaneous and data-dependent* error as:

$$e(\theta_c, x, u) \coloneqq \frac{1}{2}\left(\rho_i \frac{e^i(x, \theta_c, u)^2}{\left(1+|\psi(x,u)|^2\right)^2} + \rho_d \sum_{k=1}^{N} \frac{e_k^d(\theta_c)^2}{\left(1+|\psi(x_k, u^*(x_k))|^2\right)^2}\right), \tag{14}$$

where $\rho_i \in \mathbb{R}_{\geq 0}$ and $\rho_d \in \mathbb{R}_{>0}$ are tunable gains. Since we are interested in designing real-time training dynamics for the estimation of the optimal parameters θ_c^*, we compute the the gradient of (14) with respect to θ_c as follows:

$$\begin{aligned} \nabla_{\theta_c} e(\theta_c, x, u) = \rho_i &\left(\Psi(x,u)\Psi(x,u)^\top\theta_c + \frac{\psi(x,u)\left[Q(x)+R(u)\right]}{\left(1+\psi(x,u)^\top\psi(x,u)\right)^2} \right) \\ &+ \rho_d \left(\Lambda\theta_c + \sum_{k=1}^{N} \frac{\psi(x_k, u^*(x_k))\left[Q(x_k)+R(u^*(x_k))\right]}{\left(1+\psi(x_k, u^*(x_k))^\top \psi(x_k, u^*(x_k))\right)^2} \right), \end{aligned} \tag{15}$$

where Λ and Ψ are defined in Assumption 1.
The "propagated" error to the HJB equation that results from the approximate parametrization of V^* in (8), is given by:

$$\epsilon_{\text{HJB}}(x) \coloneqq H(x, u^*(x), \nabla V^*) - H\left(x, u^*, \frac{\partial \phi_c(x)}{\partial x}^\top \theta_c^*\right)$$

$$= -\nabla \epsilon_c^\top(x)\Big(f(x) + g(x)u^*(x)\Big). \tag{16}$$

The following assumption is standard, and it is satisfied when the involved functions are continuous and K is compact.

Assumption 2. *There exist* $\overline{\phi_c},\ \overline{d\phi_c},\ \overline{\epsilon_c},\ \overline{d\epsilon_c},\ \overline{\epsilon_{HJB}}, \overline{g} \in \mathbb{R}_{>0}$ *such that*

$$|\phi_c(x)| \le \overline{\phi_c}, \quad \left|\frac{\partial \phi_c(x)}{\partial x}\right| \le \overline{d\phi_c}, \quad |\epsilon_c(x)| \le \overline{\epsilon_c},$$
$$|\nabla \epsilon_c(x)| \le \overline{d\epsilon_c}, \ |\epsilon_{HJB}(x)| \le \overline{\epsilon_{HJB}}, \ |g(x)| \le \overline{g} \quad \forall x \in K,$$

where K *is the same set considered in* (8). □

4.1. Critic Dynamics via Data-Driven Hybrid Momentum-Based Control

To design fast asymptotically stable dynamics for the estimate θ_c, we propose a new class of momentum-based critic dynamics inspired by accelerated gradient flows with restarting mechanisms, such as those studied in [7] and [11]. Specifically, we consider the following hybrid dynamics of the form (1), with state $y := (\theta_c, p, \tau)$ and elements:

$$C_c := \left\{y \in \mathbb{R}^{2l_c+1} \ : \ \tau \in [T_0, T]\right\}, \qquad F_c(y, x, u) := \begin{pmatrix} \frac{2}{\tau}(p - \theta_c) \\ -2k_c \nabla_{\theta_c} e(\theta_c, x, u) \\ \frac{1}{2} \end{pmatrix}, \tag{17a}$$

$$D_c := \left\{y \in \mathbb{R}^{2l_c+1} \ : \ \tau = T\right\}, \qquad G_c(y) := \begin{pmatrix} \theta_c \\ \theta_c \\ T_0 \end{pmatrix}, \tag{17b}$$

where $k_c \in \mathbb{R}_{>0}$ is a tunable gain, and (p, τ) are auxiliary states that are periodically reset every time $\tau = T$ via the jump map (17b), with $\infty > T > T_0 > 0$. The dynamical system in (17) flows in continuous time according to (17a) whenever the timer variable τ is in $[T_0, T]$. As soon as τ hits T, the algorithm (17) resets the timer variable to T_0, as well as the momentum variable p to θ_c, while leaving θ_c unaffected. Accordingly, after the first reset, the system exhibits periodic resets every $\Delta T = 2(T - T_0)$ intervals of time. The following assumption provides data-dependent tuning guidelines for the resetting frequency of the timer variable τ, which will be leveraged in our stability results.

Assumption 3. *The tunable parameters* $(T_0, T, k_c, \rho_i, \rho_d)$ *satisfy* $2\rho_d \underline{\lambda} > \rho_i$ *and*

$$T_0^2 + \frac{1}{2k_c \underline{\lambda} \rho_d} < T^2 < \frac{8\rho_d \underline{\lambda}}{k_c \rho_i^2}, \tag{18}$$

where $\underline{\lambda}$ *is the level of richness of the recorderd data defined in Assumption 1.* □

For system (17), we study stability properties with respect to the compact set:

$$\mathcal{A}_c := \mathcal{A}_{\theta_c, p} \times [T_0, T], \tag{19a}$$
$$\mathcal{A}_{\theta_c, p} := \left\{(\theta_c, p) \in \mathbb{R}^{2l_c} : p_c = \theta_c, \ \theta_c = \theta_c^*\right\}. \tag{19b}$$

The following theorem is the first main result of this paper. All the proofs are presented in the Appendices.

Theorem 1. *Given a number* l_c *of basis functions* ϕ_c *parametrizing the critic NN, and a compact set* $K \subset \mathbb{R}^n$*, suppose that Assumptions 1, 2 and 3 are satisfied. Then, there exists* $(\kappa, c) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ *and class-*$\mathcal{K}_\infty$ *functions* γ_1 *and* γ_2*, such that for every solution* $y = (\theta_c, p, \tau)$ *to* (17) *with initial condition* $y(0,0) = (\theta_c(0,0), p(0,0), \tau(0,0))$*, and using the control policy* $u(\cdot) \in \mathcal{U}_V$ *on the plant, the critic parameters* θ_c *satisfy*

$$|\theta_c(t,j) - \theta_c^*| \le \kappa e^{-c(t+j)} \, |y(0,0)|_{\mathcal{A}_c} + \gamma_2\left(|\tilde{u}(x(t,j))|\right) + \gamma_1(\overline{\epsilon_{HJB}}), \tag{20}$$

where $\tilde{u}(x(t,j)) := u(x(t,j)) - u^*(x(t,j))$*, for all* $(t,j) \in$ *dom* (y) □

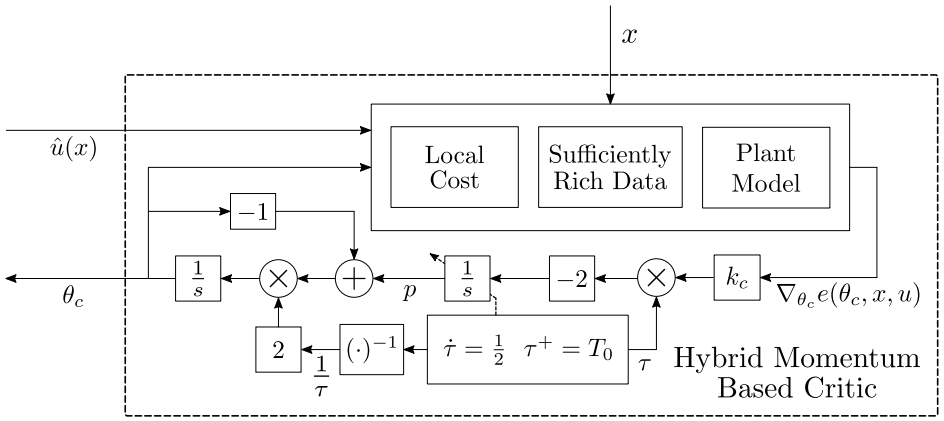


Figure 1: Proposed Hybrid Momentum Based Dynamics for the training of the Critic subsystem

The presence of a residual optimal-control mismatch term in (20) of the form $\gamma_2(|u(x) - u^*(x)|)$, represents a crucial difference with respect to previous CL adaptive dynamic approaches, such as those studied in [17] and [5, Ch. 4]. This term is a direct byproduct of our definition of ψ in (10), its dependence on the control action u, and its appearance in the error gradient (15). In principle, the emergence of this term in Theorem 1 is agnostic to the particular gradient-based update dynamics for the critic NN, regardless of the inclusion or not of momentum. Since $\gamma_2 \in \mathcal{K}$, the larger the difference between the nominal input u and the optimal feedback law u^*, the greater the residual error in the convergence of θ_c. In particular, the bound (20) describes a semi-global practical input-to-state stability property that, to the best knowledge of the authors, is novel in the context of CL-based RL. In the next section we will show that the residual error $\gamma_2(|\tilde{u}|)$ can be removed by incorporating an additional actor NN in the system.

Remark 2. *In contrast to standard data-driven gradient-descent dynamics for the estimation of the optimal value function V^*, which can achieve exponential rates of convergence proportional to $\underline{\lambda}$ (cf. [18, 13]), under the assumptions of Theorem 1 the critic update dynamics* (17) *can achieve exponential convergence with rates proportional to $\sqrt{\underline{\lambda}}$. As shown in [9], momentum-based dynamics of this form can achieve these rates using the restarting parameter*

$$T = T^* := e\sqrt{\frac{1}{2k_c\rho_d\underline{\lambda}} + T_0^2}. \tag{21}$$

This property is particularly useful in settings where the level of richness of the data-set is limited, i.e., when $\underline{\lambda} \ll 1$, which is common in practical applications.

Theorem 1 guarantees exponential convergence to a neighborhood of the optimal parameters $\{\theta_c^*\}$ that define the optimal value function V^*. Consequently, by continuity, and on compact sets, $\hat{V}$ would converge to an ϵ-approximation of V^*, which can be leveraged by the control law (6) to stabilize system (3). However, as noted in [22], implementing only critic structures for the control of nonlinear dynamical systems of the form (3) can lead to poor closed-loop transient performance. To tackle this issue, we consider an auxiliary dynamical system, called the *actor*, which will serve as an estimator of the optimal controller that acts on the plant.

5. Actor Dynamics

Using the optimal value parametrization described in Section 4 the optimal control law can written as:

$$u^*(x) = -\frac{1}{2}\Pi_u^{-1} g(x)^\top \left[\frac{\partial \phi_c(x)}{\partial x}^\top \theta_c^* + \nabla \epsilon_c(x)\right], \quad \forall x \in K. \tag{22}$$

Therefore, using $\frac{\partial \phi_c(x)}{\partial x}$ and $g(x)$ we can implement an actor neural-network given by:

$$\hat{u}(x) = \omega(x)^\top \theta_u, \tag{23}$$

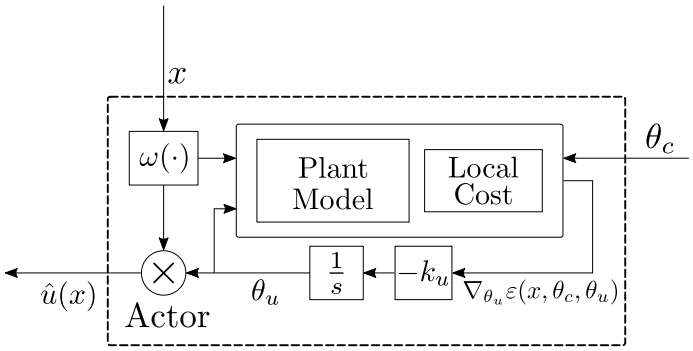


Figure 2: Actor Subsystem

where $\omega : \mathbb{R}^n \to \mathbb{R}^{l_c \times m}$ is defined as:

$$\omega(x) \coloneqq -\frac{1}{2}\frac{\partial \phi_c(x)}{\partial x} g(x) \Pi_u^{-1}. \tag{24}$$

To guarantee convergence of $\hat{u}$ to u^*, we design update dynamics for $\theta_u \in \mathbb{R}^{l_c}$ based on the minimization of the error:

$$\begin{aligned} \varepsilon(x, \theta_c, \theta_u) &\coloneqq \frac{1}{2}\left[\alpha_1 \frac{\varepsilon_a(x,\theta_c,\theta_u)^\top \varepsilon_a(x,\theta_c,\theta_u)}{1 + \operatorname{Tr}\left(\omega(x)^\top \omega(x)\right)} + \alpha_2 \varepsilon_b(\theta_c,\theta_u)^\top \varepsilon_b(\theta_c,\theta_u)\right], \\ \varepsilon_a(x,\theta_c,\theta_u) &\coloneqq \hat{u}(x) - \omega(x)^\top \theta_c = \omega(x)^\top \left(\theta_u - \theta_c\right), \\ \varepsilon_b(\theta_c,\theta_u) &\coloneqq \theta_u - \theta_c, \end{aligned} \tag{25}$$

which satisfies:

$$\nabla_{\theta_u}\varepsilon(x,\theta_c,\theta_u) = \Omega(x)(\theta_u - \theta_c),$$

where

$$\Omega(x) \coloneqq \alpha_1 \frac{\omega(x)\omega(x)^\top}{1 + \operatorname{Tr}\left(\omega(x)^\top \omega(x)\right)} + \alpha_2 I \in \mathbb{R}^{l_c \times l_c} \quad \forall x \in \mathbb{R}^n. \tag{26}$$

Based on these definitions, we consider the following gradient-descent dynamics for the actor neural-network:

$$\dot{\theta}_u = F_u(\theta_u, x, \theta_c) \coloneqq -k_u \nabla_{\theta_u}\varepsilon(x,\theta_c,\theta_u), \tag{27}$$

where $k_u \in \mathbb{R}_{>0}$ is a tunable gain. A scheme representing these update dynamics is shown in Figure 2.

6. Momentum-Based Actor-Critic Feedback System

Consider the closed-loop resulting from the interconnection between the plant (3), the critic update dynamics (17), the actor update dynamics (27) and the feedback law in (23) shown in Figure 3a, and given by:

$$\dot{x} = f(x) + g(x)\hat{u}(x), \qquad x^+ = x, \tag{28a}$$

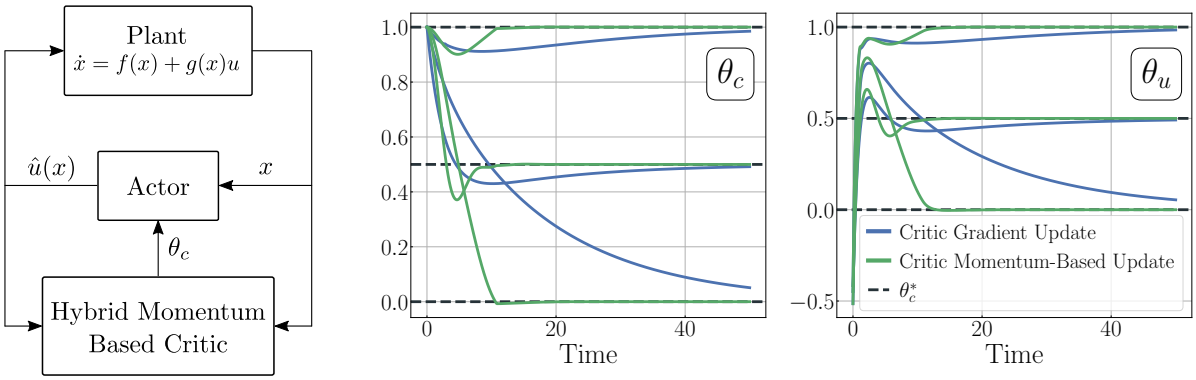
$$\dot{y} = F_c(y, x, \hat{u}(x)), \qquad y^+ = G_c(y), \tag{28b}$$

$$\dot{\theta}_u = F_u(\theta_u, x, \theta_c), \qquad \theta_u^+ = \theta_u, \tag{28c}$$

and with flow set and jump set given by $C = \mathbb{R}^n \times C_c \times \mathbb{R}^{l_c}$ and $D = \mathbb{R}^n \times D_c \times \mathbb{R}^{l_c}$ respectively, where C_c and D_c are as defined in (17). Let $z \coloneqq (x, y, \theta_u)$ be the overall state of the closed-loop system, and define:

$$\mathcal{A} \coloneqq \{0\} \times \mathcal{A}_c \times \{\theta_c^*\}.$$

The following is the main result of this paper.



(a) Closed-Loop System

(b) Convergence of the critic (left) and actor (right) neural networks' weights to the optimal values.

Figure 3: Closed-Loop System Diagram and Numerical Example

Theorem 2. *Given the vector of basis functions $\phi_c : \mathbb{R}^n \to \mathbb{R}^{l_c}$ parametrizing the critic NN and a compact set $K_z \coloneqq K \times K_y \times K_\theta \subset \mathbb{R}^n \times \mathbb{R}^{2l_c+1} \times \mathbb{R}^{l_c}$, where K is given as in* (8)*, suppose that Assumption 1-3 are satisfied. Then, there exists $\beta \in \mathcal{KL}$, $\gamma \in \mathcal{K}$ and tunable parameters $(\rho_i, \rho_d, k_c, k_u, \alpha_1, \alpha_2)$, such that for every solution $z = (x, y, \theta_u)$ to the closed-loop system* (28)*, with initial condition $z(0,0) = (x(0,0), y(0,0), \theta_u(0,0)) \in K_z$, there exists $\tilde{T} > 0$ such that for all $(t, j) \in dom(z)$:*

$$|z(t,j)|_{\mathcal{A}} \leq \beta(|z(0,0)|_{\mathcal{A}}, t+j) + \gamma(\left|\left(\overline{\epsilon_{HJB}}, \overline{d\epsilon_c}\right)\right|) + \nu,$$

for all $0 \leq t + j \leq \tilde{T}$, and

$$|z(t,j)|_{\mathcal{A}} \leq \gamma(\left|\left(\overline{\epsilon_{HJB}}, \overline{d\epsilon_c}\right)\right|) + \nu, \quad \forall\, \tilde{T} \leq t + j,$$

for some $\nu > 0$ constant. □

Theorem 2 establishes asymptotic convergence to a neighborhood of the compact set $\mathcal{A}$ as $\left(\overline{\epsilon_{\text{HJB}}}, \overline{d\epsilon_c}\right) \to 0$ from any compact set K_z modulo some error ν, under a suitable choice of tunable parameters. To the best knowledge of the authors this is the first result providing stability certificates for continuous-time actor-critic reinforcement learning using recorded data and accelerated value-function estimation dynamics with momentum. In addition, since the resulting closed-loop system in (28) is given by a well-posed hybrid system, the stability results are robust with respect to arbitrarily small additive disturbances on the states and dynamics [12, Ch. 7].

7. Numerical Example

In this section, we present a numerical experiment that illustrates our theoretical results. In particular, we study the following nonlinear control-affine plant:

$$\dot{x} = f(x) + g(x)u, \tag{29a}$$

$$f(x) = \begin{bmatrix} -x_1 + x_2 \\ -\frac{1}{2}\left(x_1 - x_2\left(1 - \cos(2x_1 + 2)^2\right)\right) \end{bmatrix}, \tag{29b}$$

$$g(x) \coloneqq \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}, \tag{29c}$$

with local state and control costs given by $Q(x) = x^\top x$ and $R(u) = u^2$ [18]. The optimal value function for this setting is given by $V^*(x) = \frac{1}{2}x_1^2 + x_2^2$ with optimal control law given by $u^*(x) = -(\cos(2x_1) + 2)x_2$. Using this information, we choose $\phi_c(x) = (x_1^2, x_1x_2, x_2^2)$, and we implement the prescribed hybrid momentum-based dynamics

in (17) for the update of the critic neural network, and the update dynamics for the actor described in (27). We obtain the results shown in Figure 3b with $x(0,0) = (-10, 10)$, $\theta_c(0,0) = (1,1,1)$ and $\theta_u \in [0,1]^3$. We compare the results with the case in which the critic neural-network is updated with the gradient-descent dynamics of [17], and where the sufficiently rich data is a set of 16 data points obtained by sampling the dynamics (29) in a grid around the origin of size 4×4. In our simulations we use $T_0 = 0.1, T = 5.5$ for the momentum-based dynamics in (17). These particular values are obtained by using the level of richness $\underline{\lambda}$ of the data-set, and the inequalities in (18) in order to ensure compliance with Assumption 3. For both reinforcement learning dynamics we use $k_c = 1, k_u = 1, \rho_d = 1$ and $\rho_i = 1$. As shown in the figure both update dynamics are able to converge to $\{\theta_c^*\}$, with $\theta_c^* = (1/2, 0, 1)$ describing the optimal value function V^*. However, the hybrid-based dynamics are able to significantly improve the transient performance of the learning mechanism.[1]

8. Conclusions

In this paper, we introduced the first stability guarantees for deterministic continuous-time actor-critic reinforcement learning with accelerated training of neural network structures. To do so, we studied a novel hybrid momentum-based estimation dynamical system for the critic NN, which estimates, in real time, the optimal value function. Our stability analysis leveraged the existence of rich recorded data taken from a finite number of samples along optimal trajectories and inputs of the system. We showed that this finite sequence of samples can be used to train the controller to achieve online optimal performance with fast transient performance. Closed-loop stability was established using tools from hybrid dynamical systems theory. Potential extensions include the study of similar accelerated training dynamics for the actor subsystem, as well as considering reinforcement learning problems in hybrid plants.

References

[1] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: lessons we have learned," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 698–721, 2021.

[2] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[3] J. Martinez-Piazuelo, D. E. Ochoa, N. Quijano, and L. F. Giraldo, "A multi-critic reinforcement learning method: An application to multi-tank water systems," *IEEE Access*, vol. 8, pp. 173227–173238, 2020.

[4] K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, "Handbook of reinforcement learning and control," 2021.

[5] R. Kamalapurkar, P. Walters, J. Rosenfeld, and W. E. Dixon, *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*. Springer, 2018.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[7] W. Su, S. Boyd, and E. Candes, "A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights," *J. of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.

[8] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *Proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. E7351–E7358, 2016.

[9] J. I. Poveda and N. Li, "Robust hybrid zero-order optimization algorithms with acceleration via averaging in continuous time," *Automatica*, vol. 123, 2021.

[10] J. I. Poveda and A. R. Teel, "The heavy-ball ode with time-varying damping: Persistence of excitation and uniform asymptotic stability," in *2020 American Control Conference (ACC)*, pp. 773–778, IEEE, 2020.

[11] O'Donoghue and E. J. Candès, "Adaptive restart for accelerated gradient schemes," *Foundations of Computational Mathematics*, vol. 15, no. 3, pp. 715–732, 2013.

[12] R. Goebel, R. G. Sanfelice, and A. R. Teel, "Hybrid dynamical systems: modeling stability, and robustness," 2012.

[13] G. Chowdhary and E. Johnson, "Concurrent learning for convergence in adaptive control without persistency of excitation," in *49th IEEE Conference on Decision and Control (CDC)*, pp. 3674–3679, IEEE, 2010.

[14] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized hamilton-jacobi-bellman equation," *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[15] D. Liberzon, *Calculus of variations and optimal control theory*. Princeton university press, 2011.

[16] 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.

[17] 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, 2015.

[1] The code used to implement this simulation can be found in the following repository: https://github.com/deot95/Accelerated-Continuous-Time-Approximate-Dynamic-Programming-through-Data-Assisted-Hybrid-Control

[18] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, 2016.
[19] K. J. Astrom and B. Wittenmark, *Adaptive Control*. Addison-Wesley Publishing Company, 1989.
[20] K. Ciosek, "Imitation learning by reinforcement learning," in *International Conference on Learning Representations*, 2022.
[21] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell, "Bridging offline reinforcement learning and imitation learning: A tale of pessimism," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
[22] K. Doya, "Reinforcement learning in continuous time and space," *Neural computation*, vol. 12, no. 1, pp. 219–245, 2000.
[23] C. Cai and A. R. Teel, "Characterizations of input-to-state stability for hybrid systems," *Systems & Control Letters*, vol. 58, no. 1, pp. 47–53, 2009.
[24] H. K. Khalil, *Nonlinear Systems*. Upper Saddle River, NJ: Prentice Hall, 2002.
[25] D. E. Ochoa, J. I. Poveda, A. Subbaraman, G. S. Schmidt, and F. R. Pour-Safaei, "Accelerated concurrent learning algorithms via data-driven hybrid dynamics and nonsmooth odes," in *Learning for Dynamics and Control*, pp. 866–878, PMLR, 2021.

Appendix A. Proof Theorem 1

Appendix A.1. Gradient of Critic Error-Function in Deviation Variables

First, using (16) together with $H(x, u^*(x), \nabla V^*) = 0$ for all x, we obtain:

$$\psi(x, u^*(x))^\top \theta_c^* + Q(x) + R\left(u^*(x)\right) = \epsilon_{\text{HJB}}(x). \tag{A.1}$$

Thus, using (15) and (A.1), we can rewrite the gradient of $e(\theta_c, x, u)$ as follows:

$$\nabla_{\theta_c} e(\theta_c, x, u) = \Theta(x, u)\left(\theta_c - \theta_c^*\right) + \upsilon_\epsilon(x, u) + \chi(x, u), \tag{A.2}$$

where

$$\Theta(x, u) \coloneqq \rho_i \Psi(x, u)\Psi(x, u)^\top + \rho_d \Lambda, \tag{A.3}$$

and

$$\upsilon_\epsilon(x, u) \coloneqq \rho_i \frac{\psi(x, u)\epsilon_{\text{HJB}}(x)}{\left(1 + |\psi(x, u)|^2\right)^2} + \rho_d \sum_{k=1}^{N} \frac{\psi(x_k, u^*(x_k))\epsilon_{\text{HJB}}(x_k)}{\left(1 + |\psi(x_k, u^*(x_k))|^2\right)^2} \in \mathbb{R}^{l_c}, \tag{A.4}$$

$$\chi(x, u) \coloneqq \frac{\rho_i \psi(x, u)\left[\frac{\partial \phi_c(x)}{\partial x} g(x)\left(u - u^*(x)\right)\right]^\top \theta_c^*}{\left(1 + |\psi(x, u)|^2\right)^2} + \frac{\rho_i \psi(x, u)\left[R(u) - R(u^*(x))\right]}{\left(1 + |\psi(x, u)|^2\right)^2} \in \mathbb{R}^{l_c}, \tag{A.5}$$

which, by using the fact that $\frac{r}{(1+r^2)^2} \leq \frac{3\sqrt{3}}{16}, \forall r \in \mathbb{R}_{\geq 0}$, satisfy:

$$|\upsilon_\epsilon(x, u)| \leq \frac{3\sqrt{3}}{16} \overline{\epsilon_{\text{HJB}}}\left(\rho_i + N\rho_d\right), \tag{A.6a}$$

$$|\chi(x, u)| \leq \rho_i \frac{3\sqrt{3}}{16}\left(\overline{g}\left(\overline{d\phi_c}\left[1 + |\theta_c^*|\right] + \overline{d\epsilon_c}\right)|u - u^*(x)| + \lambda_{\max}\left(\Pi_u\right)|u - u^*(x)|^2\right). \tag{A.6b}$$

The following Lemma will be instrumental for our results.

Lemma 1. *If the data is $\underline{\lambda}$-sufficiently-rich, then there exist $\overline{\Theta}, \underline{\Theta} \in \mathbb{R}_{>0}$ such that*

$$\underline{\Theta} I_n \preceq \Theta(x, u) \preceq \overline{\Theta} I_n \qquad \forall x \in \mathbb{R}^n,\ \forall u \in \mathbb{R}^m.$$

Proof. Let $\theta \in \mathbb{R}^{l_c}$ be arbitrary. Since, by assumption, the data is $\underline{\lambda}$-SR it follows that:

$$\begin{aligned}\theta^\top \Theta(x, u)\theta &= \theta^\top \rho_i \Psi(x, u)\Psi(x, u)^\top \theta + \theta^\top \rho_d \Lambda \theta \\ &\geq \rho_d \underline{\lambda} |\theta|^2\end{aligned}$$

$$\implies \Theta(x,u) \succeq \underline{\Theta} I_{l_c},\ \forall (x,u) \in \mathbb{R}^n \times \mathbb{R}^m, \tag{A.7}$$

where $\underline{\Theta} \coloneqq \rho_d \underline{\lambda}$. On the other hand, using the fact that $\left|aa^\top\right| = |a|^2,\ \forall a \in \mathbb{R}^n$, we obtain that:

$$\left|\Psi(x,u)\Psi(x,u)^\top\right| = |\Psi(x,u)|^2 \leq 1,\ \forall (x,u) \in \mathbb{R}^n \times \mathbb{R}^m,$$

we obtain:

$$\begin{aligned}\theta^\top \Theta(x,u)\theta &= \theta^\top \rho_i \psi(x,u)\psi(x,u)^\top \theta + \theta^\top \rho_d \Lambda \theta \\ &\leq \left(\rho_i + \rho_d \lambda_{\max}(\Lambda)\right)|\theta|^2 \\ &\implies \Theta(x,u) \preceq \overline{\Theta} I_{l_c}, \quad \forall (x,u) \in \mathbb{R}^n \times \mathbb{R}^m,\end{aligned}$$

where $\overline{\Theta} \coloneqq \rho_i + \rho_d \lambda_{\max}(\Lambda)$. ■

Appendix A.2. Lyapunov-Based Analysis

Recall from Section 4 that $y = (\theta_c, p, \tau)$, suppose that the assumptions of Theorem 1 hold and consider the Lyapunov candidate function $V_c : \mathbb{R}^{l_c} \times \mathbb{R}^{l_c} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ given by:

$$V_c(y) \coloneqq \frac{|p-\theta_c|^2}{4} + \frac{|p-\theta_c^*|^2}{4} + k_c \rho_d \tau^2 \frac{(\theta_c - \theta_c^*)\top \Lambda (\theta_c - \theta_c^*)}{2}, \tag{A.8}$$

where Λ was defined in Assumption 1 and which satisfies:

$$\begin{aligned}\underline{c}\,|y|_{\mathcal{A}_c}^2 \leq V_c(y) \leq \overline{c}\,|y|_{\mathcal{A}_c}^2, \\ \underline{c} \coloneqq \min\left\{\frac{1}{4}, \frac{k_c \rho_d T_0^2 \underline{\lambda}}{2}\right\}, \quad \overline{c} \coloneqq \left\{\frac{3}{4},\ \frac{1}{2}\left(1 + k_c \rho_d T^2 \overline{\lambda}\right)\right\},\end{aligned} \tag{A.9}$$

where $\overline{\lambda} \coloneqq \lambda_{\max}(\Lambda)$. Now, let $u \in \mathcal{U}_V$, and consider the time derivative of V_c along the continuous-time evolution of the critic subsystem, i.e., $\dot{V}_c = \nabla_y V_c(y)^\top \dot{y}$. Then, by using (A.2) and Lemma 1, and some algebraic manipulation, $\dot{V}_c$ can be shown to satisfy

$$\dot{V}_c \leq -\begin{pmatrix}|p-\theta_c| & |\theta_c - \theta_c^*|\end{pmatrix} M(\tau) \begin{pmatrix}|p-\theta_c| \\ |\theta_c - \theta_c^*|\end{pmatrix} + 2\sqrt{2} k_c y_{\mathcal{A}_c}\left(|\upsilon_\epsilon(x)| + |\chi(x,u(x))|\right), \tag{A.10}$$

where

$$M(\tau) \coloneqq \begin{pmatrix}\frac{2}{k_c \tau^2} & -\frac{\rho_i}{2} \\ -\frac{\rho_i}{2} & \underline{\Theta}\end{pmatrix}, \tag{A.11}$$

and $\mathcal{A}_c$ was defined in Section 4. Since $2\rho_d \underline{\lambda} > \rho_i$ and $T^2 < \frac{8\rho_d \underline{\lambda}}{k_c \rho_i^2}$ by means of Asssumption 2, and $\tau(t,j) \in [T_0, T],\ \forall (t,j) \in \operatorname{dom}(y)$ by construction of the critic update dynamics (17), it follows that $M(\tau) \succeq \underline{r}$ with $\underline{r} \coloneqq \underline{\Theta} - \frac{\rho_i}{2}$. Hence, from (A.10) and using (A.6), we obtain that:

$$\dot{V}_c \leq -\underline{r}\,|y|_{\mathcal{A}_c}^2 + |y|_{\mathcal{A}_c}\left(\gamma_\nu\left(\overline{\epsilon_{\text{HJB}}}\right) + \gamma_\chi\left(|u(x) - u^*(x)|\right)\right), \tag{A.12}$$

where $\gamma_\nu, \gamma_\chi \in \mathcal{K}_\infty$ are given by:

$$\begin{aligned}\gamma_\nu(r) &\coloneqq \frac{3\sqrt{6}}{8}\left(\rho_i + N\rho_d\right) r, \quad \gamma_\chi(r) \coloneqq c_\chi(r + r^2), \\ c_\chi &\coloneqq \frac{3\sqrt{6}}{8}\rho_i \max\left\{\overline{g}\left(\overline{d\phi_c}\left[1 + |\theta_c^*|\right] + \overline{d\epsilon_c}\right),\ \lambda_{\max}(\Pi_u)\right\}.\end{aligned}$$

Thus, letting $d_c \in (0,1)$, and using (A.9), (A.12):

$$\dot{V}_c \leq -\frac{\underline{r}(1-d_c)}{\overline{c}} V_c(y), \quad \forall\, |y|_{\mathcal{A}_c} \geq \frac{1}{d_c}\left(\gamma_\nu\left(\overline{\epsilon_{\text{HJB}}}\right) + \gamma_\chi\left(|u(x) - u^*(x)|\right)\right). \tag{A.13a}$$

On the other hand, the change of V_c during the jumps in the update dynamics for the critic (17), satisfies:

$$V_c\left(y^+\right) - V_c(y) \leq -\eta V_c(y), \tag{A.14}$$

with $\eta \coloneqq 1 - \frac{T_0^2}{T^2} - \frac{1}{2k_c\rho_d\underline{\lambda}T^2}$ which satisfies $\eta \in (0,1)$ by means of Assumption 2. Together, (A.13) and (A.14), in conjuction with the quadratic bounds of (A.9), imply the results of Theorem 1 via [23, Prop 2.7] and the fact that $|\theta_c(t,j) - \theta_c^*| \leq |y(t,j)|_{\mathcal{A}_c} \leq |(\theta_c(t,j), p(t,j))|_{\mathcal{A}_{\theta_c,p}}$ for all $(t,j) \in \mathrm{dom}\,(y)$. ■

Appendix B. Proof of Theorem 2

Appendix B.1. Gradient of Actor Error-Function in Deviation Variables

First, note that we we can write (25) as:

$$\nabla_{\theta_u}\varepsilon_a(x,\theta_c,\theta_u) = \Omega(x)\left(\theta_u - \theta_c^* - (\theta_c - \theta_c^*)\right),$$

and consider the following Lemma, instrumental for our results.

Lemma 2. *There exists $\overline{\Omega}, \underline{\Omega} \in \mathbb{R}_{>0}$ such that*

$$\underline{\Omega} I_{l_c} \preceq \Omega(x) \preceq \overline{\Omega} I_{l_c}.$$

Proof. Let $\theta \in \mathbb{R}^{l_c}$ be arbitrary. Then, by the definition of $\Omega : \mathbb{R}^n \to \mathbb{R}^{l_c \times l_c}$ in (26), it follows that:

$$\theta^\top \Omega(x)\theta = \alpha_1 \frac{\left|\omega(x)^\top\theta\right|^2}{1 + \mathrm{Tr}\left(\omega(x)^\top\omega(x)\right)} + \alpha_2 |\theta|^2 \geq \alpha_2 |\theta|^2 \implies \Omega(x) \succeq \underline{\Omega} I_{l_c}, \quad \forall x \in \mathbb{R}^n,$$

where $\underline{\Omega} \coloneqq \alpha_2$. On the other hand, we obtain:

$$\theta^\top \Omega(x)\theta = \left(\alpha_1 \frac{|\omega(x)|^2}{1 + |\omega(x)|_F^2} + \alpha_2\right)|\theta|^2 \leq \overline{\Omega}\,|\theta|^2 \implies \Omega(x) \preceq \overline{\Omega} I_{l_c}, \quad \forall x \in \mathbb{R}^n,$$

where $\overline{\Omega} \coloneqq \alpha_1 + \alpha_2$, $|A|_F$ represents the Frobenius norm and where we used $|A| \leq |A|_F$, $\forall A \in \mathbb{R}^{l_c \times l_c}$ and $\frac{r^2}{1+r^2} \leq 1\ \forall r \in \mathbb{R}$. ■

Now, consider the Lyapunov function

$$\mathcal{V}(z) \coloneqq V_o(x) + V_c(y) + V_a(\theta_u), \tag{B.1a}$$

$$V_o(x) \coloneqq V^*(x), \quad V_a(\theta_u) \coloneqq \frac{1}{2}|\theta_u - \theta_u^*|^2, \tag{B.1b}$$

where V_c was defined in (A.8) and where we recall that $z = (x, y, \theta_u)$. By [24, Lemma 4.3], and since $V_o = V^*$ is a continuous and positive definite function in $\mathbb{R}^n$, there exist $\underline{\gamma}_o, \overline{\gamma}_o \in \mathcal{K}$ such that $\underline{\gamma}_o(|x|) \leq V_o(x) \leq \overline{\gamma}_o(|x|)$. Hence, using (A.9), and the fact that sum of class $\mathcal{K}$ is in turn of class $\mathcal{K}$, there exist $\underline{\gamma}_{\mathcal{V}}, \overline{\gamma}_{\mathcal{V}} \in \mathcal{K}$ such that:

$$\underline{\gamma}_{\mathcal{V}}\left(|z|_{\mathcal{A}}\right) \leq \mathcal{V}(z) \leq \overline{\gamma}_{\mathcal{V}}\left(|z|_{\mathcal{A}}\right) \tag{B.2}$$

Now, the time derivative of $\dot{V}_o = \nabla V_o(x)^\top \dot{x}$ along the trajectories of (28) satisfies:

$$\dot{V}_o \leq -Q(x) + \frac{\overline{g}^2 \lambda_{\max}\left(\Pi_u^{-1}\right)}{2}\left(\overline{d\phi_c}\,|\theta_c^*| + \overline{d\epsilon_c}\right)\left(\overline{d\phi_c}\,|\theta_u - \theta_c^*| + \overline{d\epsilon_c}\right). \tag{B.3}$$

On the other hand, making use of Lemma 2, for the time derivative of $\dot{V}_a = \nabla_{\theta_u} V_a(\theta_u)^\top \theta_u$ we obtain:

$$\dot{V}_a \leq -k_u\alpha_2 |\theta_u - \theta_c^*|^2 + k_u\overline{\Omega}\,|\theta_u - \theta_c^*|\,|\theta_c - \theta_c^*|. \tag{B.4}$$

Hence, using (A.10), (B.3), and (B.4), together with the upper bounds in (A.6), we obtain that the time derivative of $\mathcal{V}$ along the trajectories of the closed-loop system satisfies:

$$\begin{aligned}\dot{\mathcal{V}} \leq & -Q(x) - \underline{r}\,|y|^2_{\mathcal{A}_c} - k_u\alpha_2\,|\theta_u - \theta_c^*|^2 \\ & + c_y\,|y|_{\mathcal{A}_c} + c_u\,|\theta_u - \theta_c^*| + c_{yu}\,|\theta_u - \theta_c^*|\,|y|_{\mathcal{A}_c} \\ & + c_{yu^2}\,|y|_{\mathcal{A}_c}\,|\theta_u - \theta_c^*|^2 + c_0,\end{aligned} \tag{B.5}$$

where

$$\begin{aligned}c_y &\coloneqq \frac{3\sqrt{6}}{8}k_c\left(\overline{\epsilon_{\text{HJB}}}\,(\rho_i + N\rho_d) + \frac{1}{2}\overline{g}^2\rho_i\left[\lambda_{\max}\left(\Pi_u^{-1}\right)\left(\overline{d\phi_c}\,[1 + |\theta_c^*|] + \overline{d\epsilon_c}\right)\overline{d\epsilon_c} + \lambda_{\max}\left(\Pi_u\right)\lambda_{\max}\left(\Pi_u^{-1}\right)^2\overline{d\epsilon_c}^2\right]\right),\\ c_u &\coloneqq \frac{1}{2}\left(\overline{d\phi_c}\,|\theta_c^*| + \overline{d\epsilon_c}\right)\overline{g}^2\lambda_{\max}\left(\Pi_u^{-1}\right)\overline{d\phi_c},\\ c_{yu} &\coloneqq \frac{3\sqrt{6}k_c}{16}\left(2k_u\overline{\Omega} + \overline{g}^2\rho_i\lambda_{\max}\left(\Pi_u^{-1}\right)\left(\overline{d\phi_c}\,[1 + |\theta_c^*|] + \overline{d\epsilon_c}\right)\overline{d\phi_c}\right),\\ c_{yu^2} &\coloneqq \frac{3\sqrt{6}}{16}k_c\overline{g}^2\rho_i\lambda_{\max}\left(\Pi_u\right)\lambda_{\max}\left(\Pi_u^{-1}\right)^2\overline{d\phi_c}^2,\\ c_0 &\coloneqq \frac{1}{2}\left(\overline{d\phi_c}\,|\theta_c^*| + \overline{d\epsilon_c}\right)\overline{g}^2\lambda_{\max}\left(\Pi_u^{-1}\right)\overline{d\epsilon_c}.\end{aligned}$$

Then, for all $|\theta_u - \theta_c^*| \leq \frac{c_{yu}}{c_{yu^2}}$, by using $Q(x) = x^\top\Pi_x x$ and letting $d_1 \in (0,1)$, from (B.5), $\dot{\mathcal{V}}$ can be further upper bounded as:

$$\begin{aligned}\dot{\mathcal{V}} \leq & -\lambda_{\min}\left(\Pi_x\right)|x|^2 - (1-d_1)\left(\underline{r}\,|y|^2_{\mathcal{A}_c} + k_u\alpha_2\,|\theta_u - \theta_c^*|^2\right)\\ & + c_y\,|y|_{\mathcal{A}_c} + c_u\,|\theta_u - \theta_c^*| + c_0\\ & - \begin{pmatrix}|y|_{\mathcal{A}_c} & |\theta_u - \theta_c^*|\end{pmatrix}\begin{pmatrix}d_1\underline{r} & -c_{yu}\\ -c_{yu} & d_1k_u\alpha_2\end{pmatrix}\begin{pmatrix}|y|_{\mathcal{A}_c}\\ |\theta_u - \theta_c^*|\end{pmatrix}.\end{aligned} \tag{B.6}$$

Now, pick a set of tunable parameters $(\rho_i, \rho_d, k_c, k_u)$ such that $\underline{r} \geq \frac{c_{yu}^2}{d_1^2 k_u\alpha_2}$ so that from (B.6), we obtain:

$$\dot{\mathcal{V}} \leq -(1-d_2)d_z\,|z|^2_{\mathcal{A}}, \quad \forall\,|z|_{\mathcal{A}} \geq \max\left\{\frac{c_0}{2d_{yu}},\ \frac{2d_{yu}}{d_2d_z}\right\},\ |\theta_u - \theta_c^*| \leq \frac{c_{yu}}{c_{yu^2}}, \tag{B.7a}$$

with

$$\begin{aligned}d_z &\coloneqq \min\left\{\lambda_{\min}\left(\Pi_x\right),\ (1-d_1)\underline{r},\ k_u\alpha_2\right\},\\ d_{yu} &\coloneqq \max\left\{2c_{yu},\ c_0\right\}, \quad d_2 \in (0,1).\end{aligned}$$

Notice that for every compact set K_θ of initial conditions for θ_u we can pick suitable $\rho_i, \rho_d, \alpha_1, \alpha_2, k_c, k_u$ to satisfy $K_\theta \subset \frac{c_{yu}}{cyu^2}\mathbb{B}$ such that (B.7) holds for every trajectory with $\theta_u(0,0) \in K_\theta$. Now, during jumps x and θ_u do not change, and hence $\mathcal{V}$ satisfies:

$$\mathcal{V}(z^+) - \mathcal{V}(z) = V_c(y^+) - V_c(y) \leq -\eta V_c(y). \tag{B.8}$$

The result of the theorem follows by using the strong-decrease of $\mathcal{V}$ during flows outside a neighborhood of $\mathcal{A}$ described in (B.7), the non-increase of $\mathcal{V}$ during jumps given in (B.8), by noting that, by design, the closed-loop dynamics are a well-posed HDS which experiences periodic jumps followed by intervals of flow of length $T - T_0 > 0$ (c.f. [25]), and by following the same arguments of [12, Prop 3.27] and [23, Prop. 2.7]. ■