2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Safe Intermittent Reinforcement Learning for Nonlinear Systems

Yongliang Yang[1], Kyriakos G. Vamvoudakis[2], Hamidreza Modares[3],Wei He[1],
Yixin Yin[1], and Donald C. Wunsch[4],

Abstract— **In this paper, an online intermittent actor-critic reinforcement learning method is used to stabilize nonlinear systems optimally while also guaranteeing safety. A barrier function-based transformation is introduced to ensure that the system does not violate the user-defined safety constraints. It is shown that the safety constraints of the original system can be guaranteed by assuring the stability of the equilibrium point of an appropriately transformed system. Then, an online intermittent actor-critic learning framework is developed to learn the optimal safe intermittent controller. Also, Zeno behavior is guaranteed to be excluded. Finally, numerical examples are conducted to verify the efficacy of the learning algorithm.**

Index Terms— **Safety control, intermittent feedback, reinforcement learning.**

I. Introduction

One of the fundamental issues in safety-critical applications of controller design is the ability of the controlled system to achieve not only stability and safety but also a user-defined performance [1], [2]. In addition, the decision making designs are usually implemented on digital platforms where sensors and controllers communicate through shared resources [3], [4]. Traditional digital controller implementations depend on periodical execution that may result in unnecessary sampling and communication between the controller and the plant. Efficient utilization of the shared resources is critical for large-scale cyber-physical and complex systems [5].

Related work

In optimal control theory [6], enforcing safety and state constraints is challenging. This is even more critical when reinforcement learning [7]–[9] is used to find the optimal control solution in real time. The work of [1], [10] presented a barrier Lyapunov function to restrict the system evolution within a given region while dealing with explicit constraints. This method is later extended to more comprehensive constraints [11]–[13]. In [2], [14], the authors provide user-defined performance in the presence of exogenous disturbances and system uncertainties using a set-theoretic model reference adaptive control framework, where the performance denotes the difference between the state of the controlled system and the given reference model. However, the aforementioned existing methods are based on continuous feedback. It is desired to develop an intermittent feedback control design to guarantee safety and optimal performance.

In an intermittent feedback control design, the system runs in an open-loop fashion until the feedback-loop is closed when a user-designed triggering condition is satisfied [15]–[18]. In the event-triggering design, the goal is to minimize the communication burden while guaranteeing closed-loop stability of the equilibrium point based on the input-to-state stability (ISS) theory. Note that the ISS Lyapunov function needs to be known a priori for the event triggering condition design, which is difficult to obtain for general nonlinear systems. This paper deals with this issue by using adaptive/approximate dynamic programming [19], [20], which can efficiently learn the optimal value function in an online manner [21] and has been widely used in control applications [22]. The work of [23], [24] developed online actor-critic learning algorithms with continuous feedback to deal with input saturation while the work of [25], [26] developed model-free intermittent actor-critic learning algorithms for linear systems.

Contributions: The contributions of the present paper are twofold. First, a novel barrier function based system transformation method is used to transform the constrained system to an equivalent system without state constraints. It is guaranteed that if the initial state is within the prescribed bound, the constraints of the original system will not be violated if the transformed system remains stable. Then, a novel intermittent actor-critic-barrier learning algorithm is used to solve the constrained regulation problem in an online fashion while excluding Zeno behavior.

Structure: The remainder of this paper is structured as follows. Section II gives the preliminaries and problem formulation. A barrier function based system transformation is presented to consider the full-state constraints. In Section III, an optimal control policy based intermittent feedback design

This work was supported in part by the National Natural Science Foundation of China under Grant 61903028 and Grant 61333002, in part by the Fundamental Research Funds for the China Central Universities of USTB under grant No. FRF-TP-18-031A1 and No. FRF-BD-17-002A, in part by the China Post-Doctoral Science Foundation under Grant 2018M641197, in part by the National Science Foundation under grant NSF CAREER CPS-1851588, in part by ONR Minerva under grant No. N00014-18-1-2160, in part by NATO under grant No. SPS G5176, in part by the Mary K. Finley Endowment, in part by the Missouri S&T Intelligent Systems Center, and in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-18-2-0260.

[1]Y. Yang, Y. Yin and W. He are with the School of Automation and Electrical Engineering, university of Science and Technology Beijing, Beijing 100083, China email: yangyongliang@ieee.org; weihe@ieee.org; yyx@ies.ustb.edu.cn.

[2]K. G. Vamvoudakis is with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA email: kyriakos@gatech.edu.

[3]H. Modares is with the Department of Mechanical Engineering, Michigan State University, East Lansing, MI 48824, USA email: modaresh@msu.edu.

[4]D. C. Wunsch II is with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA email: dwunsch@mst.edu.

is proposed. In Section IV, the online intermittent actor-critic learning algorithm is developed. In order to validate the effectiveness of the presented algorithm, a simulation example is conducted in Section V. Finally Section VI concludes and talks about future work.

II. Problem Formulation

Consider the following continuous-time affine nonlinear dynamical system $\forall t \geqslant 0$,

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\vdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= f(x) + g(x)\,u,\ x(t_0) = x_0 \end{aligned} \tag{1}$$

where $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^n$ with $x_i \in \mathbb{R}$ is the system state, $u \in \mathbb{R}$ is the control input, and $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are Lipschitz continuous functions.

In order to reduce the communication burden, an intermittent feedback control design is used. In such a design, the system state is sampled sporadically at the instants characterized by a monotone increasing sequence $\{t_k\}_{k=0}^{\infty}$ with $\lim_{k\to\infty} t_k = \infty$. That is, the sampled state remains constant between two successive events. i.e.,

$$\hat{x}(t) = \begin{cases} x(t_k), & t \in [t_k, t_{k+1}) \\ x(t), & t = t_{k+1}. \end{cases} \tag{2}$$

Problem 1. Consider the system (1), find an intermittent feedback control input $u(t) = u(\hat{x}(t))$ with the triggering instants $\{t_k\}_{k=0}^{\infty}$ such that the closed-loop system has an asymptotic stable equilibrium while the state $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^{\mathrm{T}}$ satisfies the constraints defined as

$$\begin{aligned} x_1 &\in (a_1, A_1) \\ &\vdots \\ x_n &\in (a_n, A_n), \end{aligned} \tag{3}$$

where $a_i < 0$, $A_i > 0$, $i = 1, 2, ..., n$. □

A. Barrier Function

We now define the barrier function as follows.

Definition 1. The function $b(\cdot) : \mathbb{R} \to \mathbb{R}$ defined on (a, A) is referred to as barrier function if it satisfies

1) The barrier function takes a finite value when its arguments satisfy the constraints.
2) The barrier function approaches infinity as the state approaches the boundary of the constraints, i.e.,

$$\begin{aligned} \lim_{z\to a^+} b(z; a, A) &= -\infty \\ \lim_{z\to A^-} b(z; a, A) &= +\infty. \end{aligned} \tag{4}$$

3) The barrier function defined in (6) vanishes at the equilibrium of the system (1), i.e.,

$$b(0; a, A) = 0 \in (a, A). \tag{5}$$

□

We select the barrier function as

$$b(z; a, A) = \log\left(\frac{A}{a}\frac{a - z}{A - z}\right), \forall z \in (a, A) \tag{6}$$

where a and A are two constants satisfying $a < A$. Moreover, the barrier function is invertible on the interval (a, A), i.e.,

$$b^{-1}(y; a, A) = aA\frac{e^{\frac{y}{2}} - e^{-\frac{y}{2}}}{ae^{\frac{y}{2}} - Ae^{-\frac{y}{2}}}, \forall y \in \mathbb{R}, \tag{7}$$

with a time derivative given as

$$\frac{\mathrm{d}b^{-1}(y; a, A)}{\mathrm{d}y} = \frac{Aa^2 - aA^2}{a^2e^y - 2aA + A^2e^{-y}}. \tag{8}$$

B. System Transformation

Consider the barrier function based state transformation as

$$\begin{aligned} s_i &= b(x_i; a_i, A_i), \\ x_i &= b^{-1}(s_i; a_i, A_i), \end{aligned} \quad i = 1, ..., n \tag{9}$$

then,

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \frac{\mathrm{d}x_i}{\mathrm{d}s_i}\frac{\mathrm{d}s_i}{\mathrm{d}t}, \tag{10}$$

which can yield,

$$\begin{aligned} \dot{s}_i &= \frac{x_{i+1}(s_{i+1})}{\left.\frac{\mathrm{d}b^{-1}(y;a_i,A_i)}{\mathrm{d}y}\right|_{y=s_i}} \\ &= \frac{a_{i+1}A_{i+1}\left(e^{\frac{s_{i+1}}{2}} - e^{-\frac{s_{i+1}}{2}}\right)}{a_{i+1}e^{\frac{s_{i+1}}{2}} - A_{i+1}e^{-\frac{s_{i+1}}{2}}}\frac{A_i^2e^{-s_i} - 2a_iA_i + a_i^2e^{s_i}}{A_ia_i^2 - a_iA_i^2}, \\ &= F_i(s_i, s_{i+1}), \qquad i = 1, ..., n-1 \end{aligned} \tag{11}$$

$$\begin{aligned} \dot{s}_n &= \frac{f(x) + g(x)\,u}{\left.\frac{\mathrm{d}b^{-1}(y;a_n,A_n)}{\mathrm{d}y}\right|_{y=s_n}} \\ &= [f(x) + g(x)\,u]\frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2} \\ &= F_n(s) + g_n(s)\,u, \end{aligned} \tag{12}$$

where

$$\begin{aligned} F_n(s) &= \frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2} \times \\ &\quad f\left(\begin{bmatrix} b_1^{-1}(s_1) & \cdots & b_n^{-1}(s_n) \end{bmatrix}\right) \\ g_n(s) &= \frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2} \times \\ &\quad g\left(\begin{bmatrix} b_1^{-1}(s_1) & \cdots & b_n^{-1}(s_n) \end{bmatrix}\right). \end{aligned} \tag{13}$$

The dynamics of of $s = \begin{bmatrix} s_1 & \cdots & s_n \end{bmatrix}^{\mathrm{T}}$ can be expressed in a compact form as

$$\dot{s} = F(s) + G(s)\,u, \tag{14}$$

with

$$F(s) = \begin{bmatrix} F_1(s_1, s_2) \\ \vdots \\ F_n(s) \end{bmatrix}, G(s) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_n(s) \end{bmatrix}. \tag{15}$$

Assumption 1. The system dynamics satisfy:

1) $F(s)$ is Lipschitz with $F(0) = 0$, and there exists a constant b_f such that, for $x \in \Omega$, $\|F(s)\| \leqslant b_f \|s\|$ where Ω is a compact set containing the origin.
2) $G(s)$ is bounded on Ω, i.e., there exists a constant b_g such that $\|G(s)\| \leqslant b_g$.
3) The system (1) is controllable over the compact set Ω.
4) The performance functional (7) satisfies zero-state observability. □

The barrier transformation (9) can be written in a compact form as

$$\begin{aligned} s &= b(x; a, A) \\ x &= b^{-1}(s; a, A). \end{aligned} \tag{16}$$

Therefore, with the sporadic sampling in (2), one has

$$\hat{s} = b(\hat{x}; a, A) \tag{17}$$

with

$$\hat{s} = \begin{bmatrix} \hat{s}_1 & \cdots & \hat{s}_n \end{bmatrix}^{\mathrm{T}},\ \hat{s}_i = b(\hat{x}_i; a_i, A_i).$$

In the intermittent feedback design, the controller depends on the sampled state, i.e.,

$$u(t) = u(\hat{s}(t)). \tag{18}$$

Let $e(t)$ denote the gap between the sampled state $\hat{s}(t)$ and the current state $s(t)$, i.e.,

$$e(t) = \hat{s}(t) - s(t), \tag{19}$$

with dynamics given as,

$$\dot{e}(t) = -\dot{x}(t),\ t \in [t_k, t_{k+1}). \tag{20}$$

Then, the closed-loop dynamics of the system (14) is

$$\dot{s}(t) = F(s) + G(s)\, u(s + e),\ t \in [t_k, t_{k+1}). \tag{21}$$

To solve Problem 1 with safety constraints using the intermittent feedback (18), an equivalent problem without constraints is considered as follows.

Problem 2. Consider the system (14). Find an intermittent feedback control input u in (18) with the triggering instants $\{t_k\}_{k=0}^{\infty}$ such that the closed-loop system (21) has an asymptotically stable equilibrium point in the ISS sense. □

The following lemma guarantees that Problems 1 and 2 are equivalent in the sense that the intermittent feedback controller of Problem 2 with state s satisfies of Problem 1 with state x and the full-state constraints (3).

Lemma 1. *Suppose that $u^\star(\cdot)$ solves Problem 2 for the system give by* (14)*. Then, $u^\star(\cdot)$ also solves Problem 1 provided that the initial state x_0 of system* (1) *satisfies the constraints in* (3)*.*

Proof. This proof follows from [1, Lemma 1]. □

III. Intermittent Feedback Using Optimal Policy

In this section, we investigate intermittent optimal feedback to solve Problem 2. It is shown that the optimal continuous feedback with a user-defined event-triggering instants solves Problem 1.

A. Optimal Control Policy

We consider the value function $\forall t \geqslant 0$ as

$$V(s(t)) = \int_t^{\infty} U(s, u)\mathrm{d}\tau,\ \forall s,\ t \geqslant 0, \tag{22}$$

where $U(s, u) := Q(s) + u^{\mathrm{T}} R u$ with $Q(s) > 0$ and $R > 0$.

Definition 2. *(Admissible Policy)* A control policy $\mu(s)$ is said to be admissible with respect to (26) on $\Omega \in \mathbb{R}^n$, denoted by $\mu(s) \in \pi(\Omega)$, if

- $\mu(s)$ is continuous on Ω,
- $\mu(0) = 0$,
- $u(s) = \mu(s)$ stabilizes (14) on Ω,
- $V(s)$ is finite $\forall s \in \Omega$. □

Given an admissible policy u, the Hamiltonian is defined as

$$\mathcal{H}\left(s, u, \frac{\partial V}{\partial s}\right) = \left(\frac{\partial V}{\partial s}\right)^{\mathrm{T}} [F(s) + G(s)\, u] + U(s, u), \tag{23}$$

The corresponding value function $V(x)$ (22) for a given admissible policy $u(\cdot)$ satisfies the Bellman equation

$$\begin{aligned} 0 &= \mathcal{H}\left(s, u, \frac{\partial V}{\partial x}\right) \\ &= \left(\frac{\partial V}{\partial s}\right)^{\mathrm{T}} [F(s) + G(s)\, u] + U(s, u). \end{aligned} \tag{24}$$

Based on the optimal control theory [6], the stationary condition in the Hamiltonian (23) yields the optimal control policy

$$\begin{aligned} u^\star(s) &= \arg\min_{u \in \pi(\Omega)} \mathcal{H}\left(s, u, \frac{\partial V^\star}{\partial x}\right) \\ &= -\frac{1}{2} R^{-1} G^{\mathrm{T}}(s) \frac{\partial V^\star(s)}{\partial s} \end{aligned} \tag{25}$$

where $V^\star(x)$ is the optimal value function defined as

$$V^\star(s(t)) = \min_{u(\cdot) \in \pi(\Omega)} \int_t^{\infty} U(s, u)\mathrm{d}\tau. \tag{26}$$

Inserting the optimal control policy (25) into the Bellman equation (24) yields the Hamilton-Jacobi-Bellman equation

$$\begin{aligned} 0 = &\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} F(s) + \lambda^2 \bar{R}^{\mathrm{T}} \ln\left[\mathbf{1_m} - \tanh^2(D^\star)\right] \\ &+ Q(s), \end{aligned} \tag{27}$$

where $\mathbf{1_m}$ is a vector of m ones.

B. Intermittent Design

The following are classical assumptions [15]–[17].

Assumption 2. Lipschitz continuity of the controller with respect to the gap $e(t)$

$$\|u^{\star}(x)-u^{\star}(y)\| \leqslant L\|x-y\|.$$

where $L \in \mathbb{R}^{+}$ is the Lipschitz constant. □

Assumption 3. Lipschitz continuity of the closed-loop system with respect to the state and to the gap $e(t)$. □

Lemma 2. *Suppose that Assumptions 2 and 3 hold, with an intermittent feedback given by*

$$u=-\frac{1}{2}R^{-1}g^{\mathrm{T}}(\hat{s})\frac{\partial V^{\star}(\hat{s})}{\partial \hat{s}}, \tag{28}$$

and the triggering instants are determined by the triggering condition designed as

$$\|e\|^{2} \geqslant \frac{\underline{\lambda}(Q)\left(1-\sigma^{2}\right)}{L}\|x\|^{2}+\frac{1}{L}\|u(\hat{s})\|^{2},$$

where $\underline{\lambda}(Q)$ is the minimum eigenvalue of Q and for some user defined parameter $\sigma \in (0,1)$, then the closed-loop system (21) *has an asymptotically stable equilibrium point and the Problem 1 is solved. In addition, Zeno behavior is guaranteed to be excluded.*

Proof. The proof is an extension of the proof provided in [27]. ■

As shown in (25), the optimal constrained control solution $u^{\star}(s)$ depends on solving the HJB equation (27) for the optimal value function $V^{\star}(s)$. However, the HJB equation (27) is a nonlinear partial differential and extremely difficult to solve. In the following, an online algorithm is presented to find an approximate solution to the HJB equation (27).

IV. Intermittent Feedback Using Online Actor-Critic-Barrier Learning

In this section, we employ the actor-critic online RL algorithm to learn the optimal feedback policy in an adaptive fashion while using an intermittent actor.

A. Value Function Approximation

Definition 3. *(Persistency of Excitation)* A vector signal $y(t) \in \mathbb{R}^{p}$ is exciting over the interval $[t, t+T]$ with $T \in \mathbb{R}^{+}$ if there exist $\beta_1 \in \mathbb{R}^{+}$ and $\beta_2 \in \mathbb{R}^{+}$ such that $\forall t$,

$$\beta_1 I_{p\times p} \leqslant \int_{t}^{t+T} y(\tau)\, y^{\mathrm{T}}(\tau)\,\mathrm{d}\tau \leqslant \beta_2 I_{p\times p},$$

where $I_{p\times p}$ is an identity matrix of order p. □

Assumption 4. There exists a basis function such that the optimal value function $V^{*}(s)$ and its gradient $\nabla V^{*}(s) := \frac{\partial V^{*}(s)}{\partial s}$ can be uniformly approximated with a critic network, within a set $\Omega \subseteq \mathbb{R}^{n}$ that contains the origin, as

$$V^{\star}(s) = W^{\mathrm{T}}\phi(s)+\varepsilon(s) \tag{29}$$

$$\nabla V^{\star}(s) = [\nabla\phi(s)]^{\mathrm{T}}W+\nabla\varepsilon(s) \tag{30}$$

where $W \in \mathbb{R}^{N}$ is the critic weight, $\phi(\cdot): \mathbb{R}^{n} \to \mathbb{R}^{N}$ is the critic basis, $\varepsilon(s)$ and $\nabla\varepsilon(s)$ are the bounded approximation errors satisfying $\|\varepsilon(s)\| \leqslant b_{\varepsilon}$ and $\|\nabla\varepsilon(s)\| \leqslant b_{d\varepsilon}$, $\phi(s)$ and $\nabla\phi(s)$ satisfying $\|\phi(s)\| \leqslant b_{\phi}$ and $\|\nabla\phi(s)\| \leqslant b_{d\phi}$ $\forall s \in \Omega$. □

The ideal weight, W in (29), provides the best approximate to the optimal value function $V^{\star}(s)$ on the compact set Ω and is unknown. Therefore, the estimation of W is implemented by the critic network with the approximations of the value function and value gradient

$$\hat{V}(s) = W_{c}^{\mathrm{T}}\phi(s) \tag{31}$$

$$\nabla\hat{V}(s) = [\nabla\phi(s)]^{\mathrm{T}}W_{c}. \tag{32}$$

Then, for a given policy $u(\cdot)$, the residual of Bellman equation approximation can be determined as

$$e_{c}(t) := U(s(t),u(t))+W_{c}^{\mathrm{T}}(t)\sigma(t) \tag{33}$$

where σ is a N-dimensional vector signal defined as

$$\sigma := \nabla\phi(s)\left[F(s)+G(s)\,u^{\star}\right]. \tag{34}$$

For the optimal control policy $u^{\star}(s)$, the Bellman equation (24) approximation error using the value function approximation (29) can be expressed as

$$\varepsilon_{B} = U(s,u^{\star})+W^{\mathrm{T}}\sigma. \tag{35}$$

Define the critic weight approximation error as $\tilde{W}_{c} = W - W_{c}$. Then, from (35), the relation between Bellman residual e_{c} and the Bellman equation approximation error ε_{B} can be written in terms of the critic weight error $\tilde{W}_{c}$ as

$$e_{c} = \varepsilon_{B}-\tilde{W}_{c}^{\mathrm{T}}\sigma. \tag{36}$$

The policy evaluation for an admissible control policy $u(\cdot)$ can be formulated as adapting the critic weight $\hat{W}_{c}$ to minimize the objective function [21]

$$E_{c} = \frac{1}{2}\frac{[e_{c}(t)]^{2}}{\left(1+\sigma^{\mathrm{T}}\sigma\right)^{2}}. \tag{37}$$

Then $e_{c} \to \varepsilon_{B}$ as $W_{c} \to W$. Using the chain rule yields the gradient descent algorithm for minimizing E_{c} given by [21]

$$\dot{W}_{c} = -\alpha_{c}\frac{\partial E_{c}}{\partial W_{c}} = -\alpha_{c}\frac{\sigma}{\left(1+\sigma^{\mathrm{T}}\sigma\right)^{2}}\left[\sigma^{\mathrm{T}}W_{c}+U(s,u)\right]. \tag{38}$$

Theorem 1. *Let u be any admissible control policy. Let the critic network* (31) *with the adaptive tuning law* (38) *be used to evaluate the given control policy. Suppose that the signal $\bar{\sigma} = \frac{\sigma(t)}{1+\sigma^{\mathrm{T}}(t)\sigma(t)}$ satisfies the PE condition. Then, $\tilde{W}_{c}$ is uniformly ultimately bounded (UUB).*

Proof. The proof follows from [21]. ■

B. Intermittent Actor Learning

The intermittent feedback given by (28) using the optimal policy $u^\star(\cdot)$ can be re-written in terms of the value function approximation (30) as

$$u^\star = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(\hat{s})\left[(\nabla\phi(\hat{s}))^{\mathrm{T}}W + \nabla\varepsilon(\hat{s})\right], \quad t\in[t_k, t_{k+1}), \tag{39}$$

which can written as,

$$u(\hat{s}) = (W_a^\star)^{\mathrm{T}}\phi_a(\hat{s}) + \varepsilon_a(\hat{s}), \tag{40}$$

where $\|\varepsilon_a\| \leqslant b_{\varepsilon a}$, $\|\phi_a\| \leqslant b_{\phi a}$ and $\|\phi_a(s)\| \leqslant \|k_a\|\|s\|$ for $s\in\Omega_s$ with $W_a^\star$ an unknown weight to be approximated by using a actor network as

$$u_a(\hat{s}) = (W_a)^{\mathrm{T}}\phi_a(\hat{s}). \tag{41}$$

To find the tuning for the actor approximator, we need to define the error e_u in the following form

$$e_a = W_a^{\mathrm{T}}\phi_a(\hat{s}) + \frac{1}{2}R^{-1}g^{\mathrm{T}}(\hat{s})[\nabla\phi(\hat{s})]^{\mathrm{T}}W_c. \tag{42}$$

The objective for the actor is to minimize the following squared error performance,

$$E_a = \frac{1}{2}e_a^{\mathrm{T}}(t)\,e_a(t). \tag{43}$$

The actor is updated only at the event instants t_k, for $k = 0,1,2,....$ Therefore, the actor weight is constant during the inter-event interval, i.e., $\dot{W}_a = 0,\ t\in(t_k, t_{k+1})$. At the event instant, the jump equation to compute W_a is determined by using a gradient descent law

$$\begin{aligned} W_a^+ &= W_a - \alpha_a\frac{\partial E_u}{\partial W_a} \\ &= W_a - \alpha_a\phi_a(s(t))[\phi_a(s(t))]^{\mathrm{T}}W_a \\ &\quad -\frac{1}{2}\alpha_a\phi_a(s(t))W_c^{\mathrm{T}}\nabla\phi(s(t))g(s(t))R^{-1}, \end{aligned} \tag{44}$$

for $t = t_k,\ k = 0,1,2,....$ Define the actor weight approximation error $\tilde{W}_a = W_a^\star - W_a$. Then, the dynamics of the actor weight error can be expressed as

$$\begin{aligned} \tilde{W}_a^+ &= \tilde{W}_a(t) - \alpha_a\phi_a(s(t))[\phi_a(s(t))]^{\mathrm{T}}\tilde{W}_a(t) \\ &\quad -\alpha_a\phi_a(s(t))[\phi_a(s(t))]^{\mathrm{T}}\varepsilon_a \\ &\quad -\frac{1}{2}\alpha_a\phi_a(s(t))\tilde{W}_c^{\mathrm{T}}\nabla\phi(s(t))g(s(t))R^{-1} \\ &\quad -\frac{1}{2}\alpha_a\phi_a(s(t))\nabla\varepsilon g(s(t))R^{-1},\ t = t_k \\ \dot{\tilde{W}}_a &= 0,\ t\in(t_k, t_{k+1}). \end{aligned} \tag{45}$$

Also, one can obtain the dynamics of the critic weight error $\tilde{W}_c(t)$ based on the critic learning (38) as

$$\begin{aligned} \dot{\tilde{W}}_c &= -\alpha_c\frac{\sigma_a\sigma_a^{\mathrm{T}}}{(1+\sigma_a^{\mathrm{T}}\sigma_a)^2}\tilde{W}_c \\ &\quad +\alpha_c\frac{\sigma_a}{(1+\sigma_a^{\mathrm{T}}\sigma_a)^2}\underbrace{\left[-\frac{\partial\varepsilon}{\partial s}(F+Gu_a)\right]}_{\varepsilon_c}, t\in(t_k, t_{k+1}) \\ \tilde{W}_c^+ &= 0,\ t = t_k, \end{aligned} \tag{46}$$

with $\|\varepsilon_c\| \leqslant b_{\varepsilon c}$ and $\|\nabla\varepsilon_c\| \leqslant b_{d\varepsilon c}$. By defining the augmented state $\psi := \begin{bmatrix} s^{\mathrm{T}} & \hat{s}^{\mathrm{T}} & \tilde{W}_c^{\mathrm{T}} & \tilde{W}_a^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$ with the flow dynamics on $t\in(r_j, t_{k+1}],\ j\in\mathbb{N}$ as

$$\dot{\psi} = \begin{bmatrix} F(s) + G(s)\left[\left(W_a^\star - \tilde{W}_a\right)^{\mathrm{T}}\phi_a(\hat{s}) + \eta\right] \\ 0 \\ -\alpha_c\frac{\sigma_a\sigma_a^{\mathrm{T}}}{(1+\sigma_a^{\mathrm{T}}\sigma_a)^2}\tilde{W}_c + \alpha_c\frac{\sigma_a}{(1+\sigma_a^{\mathrm{T}}\sigma_a)^2}\varepsilon_c, \\ 0 \end{bmatrix} \tag{47}$$

and the jump dynamics at event instant $t = t_k$ as

$$\psi^+ \quad = \psi(t) + \begin{bmatrix} 0 \\ s(t) \\ 0 \\ \mathrm{vec}(\Phi(t)) \end{bmatrix}, \tag{48}$$

with $\mathrm{vec}(\cdot)$ being the vectorization operator obtained by stacking the columns of a matrix on top of one another

$$\begin{aligned} \Phi(t) := -\alpha_a\phi_a(s(t))\Big[&\phi_a{}^{\mathrm{T}}(s(t))\tilde{W}_a(t) + \phi_a(s(t))^{\mathrm{T}}\varepsilon_u \\ &+ \tilde{W}_c^{\mathrm{T}}\frac{\partial\phi(s(t))}{\partial x}G(s(t))R^{-1} + \frac{\partial\varepsilon_c}{\partial x}G(s(t))R^{-1}\Big]. \end{aligned}$$

Theorem 2. *Consider system given by* (14) *with the actor* (41) *and the critic* (31) *with online learning* (38) *and* (45)*, respectively. Then, there exists a compact set* $\Omega = (\Omega_s \times \Omega_{\hat{s}} \times \Omega_{\tilde{W}_c} \times \Omega_{\tilde{W}_a})$ *such that the augmented state* ψ *converges exponentially to the set* Ω *given tht the event instants* $\{t_k\}_{k=0}^{\infty}$ *are determined by the following triggering condition,*

$$\|e\|^2 \leqslant \frac{(1-\beta^2)\underline{\lambda}(Q)}{L^2\bar{\lambda}(R)}\|s\|^2 + \frac{\underline{\lambda}(R)}{L^2\bar{\lambda}(R)}\left\|W_a^T\phi_a(\hat{s})\right\|^2 \tag{49}$$

where $\underline{\lambda}(\cdot)$, $\bar{\lambda}(\cdot)$*, define the minimum and maximum eigenvalues respectively,* $\beta\in(0,1)$, $\alpha_c > \frac{1}{4}\sqrt{\frac{1}{\underline{\lambda}(\bar{\sigma})}}$ *for the critic and* $0 < \alpha_a < 2\frac{\left(b_{\phi a}^2 - \frac{3}{2}\right)}{b_{\phi a}^4}$; $b_{\phi a} > \sqrt{\frac{3}{2}}$ *for the actor. In addition, Zeno behavior is excluded.*

Proof. We prove closed-loop stability of the system with flow dynamics (47) and jump dynamics (48) using an impulsive system approach. Consider the Lyapunov candidate $\mathcal{V}: \mathbb{R}^{2n}\times\mathbb{R}^{2n}\times\mathbb{R}^h\times\mathbb{R}^{h_2} \to \mathbb{R}$ for the continuous part (47) of the impulsive model,

$$\mathcal{V}(\psi) \quad = V^\star(s) + V^\star(\hat{s}) + V_c + V_a, \tag{50}$$

where $V^\star(s)$ and $V^\star(\hat{s})$ are the state $s(t)$ and its sampled version $\hat{s}(t)$. The functions $V_c := \left\|\tilde{W}_c\right\|^2$ and $V_a := \frac{\alpha_a^{-1}}{2}\mathrm{tr}\{\tilde{W}_a^{\mathrm{T}}\tilde{W}_a\}$ are the Lyapunov functions for the critic and actor error dynamics given by (46) and (45).

1) For the flow dynamics, note that $\dot{V}_a = \dot{V}^\star(\hat{s}) = 0$. Therefore,

$$\dot{\mathcal{V}} = \underbrace{\left(\frac{\partial V^\star(s)}{\partial s}\right)^{\mathrm{T}}\left\{F(s) + G(s)\left(W_a^\star - \tilde{W}_a\right)^{\mathrm{T}}\phi_a(\hat{s})\right\}}_{\dot{V}^\star(s)}$$

$$\underbrace{-\alpha_c\left(\frac{\partial V_c}{\partial \tilde{W}_c}\right)^{\mathrm{T}}\frac{\sigma_a\,{\sigma_a}^{\mathrm{T}}}{{\sigma_a}^{\mathrm{T}}\sigma_a+1^2}\tilde{W}_c+\alpha_c\left(\frac{\partial V_c}{\partial \tilde{W}_c}\right)^{\mathrm{T}}\frac{\sigma_a}{\left({\sigma_a}^{\mathrm{T}}\sigma_a+1\right)^2}\varepsilon_c}_{\dot{V}_c}$$

From the HJB equation (27), one has,

$$\frac{\partial V^\star(s)}{\partial s}^{\mathrm{T}}F(s)=-\frac{\partial V^\star(s)}{\partial s}^{\mathrm{T}}G(s)\,u^\star(s)-U(s,u^\star(s)) \quad (51)$$

Since $\frac{\partial V^\star}{\partial s}^{\mathrm{T}}G(s)=-Ru^\star(s)^{\mathrm{T}}$ we have,

$$\dot{V}^\star(s)=-\frac{1}{2}s^{\mathrm{T}}Qs+\frac{1}{2}u^\star(s)^{\mathrm{T}}Ru^\star(s)-u^\star(s)^{\mathrm{T}}RW_a^{\mathrm{T}}\phi_a(\hat{s}). \quad (52)$$

From Assumption 2, one has,

$$\begin{aligned}&-u^\star(s)^{\mathrm{T}}RW_a^{\mathrm{T}}\phi_a(\hat{s})+\frac{1}{2}u^\star(s)^{\mathrm{T}}Ru^\star(s)\\ =\;&\frac{1}{2}\bar{\lambda}(R)\left\|u^\star(s)-W_a^{\mathrm{T}}\phi_a(\hat{s})\right\|^2-\frac{1}{2}\underline{\lambda}(R)\left\|W_a^{\mathrm{T}}\phi_a(\hat{s})\right\|^2\\ \leqslant\;&\frac{1}{2}L^2\bar{\lambda}(R)\|s\|^2-\frac{1}{2}\underline{\lambda}(R)\left\|W_a^{\mathrm{T}}\phi_a(\hat{s})\right\|^2. \quad (53)\end{aligned}$$

By substituting (53) in (52) we have

$$\begin{aligned}\dot{V}^\star(s)\;\leqslant\;&-\frac{1}{2}\beta^2\underline{\lambda}(Q)\|s\|^2-\frac{1}{2}(1-\beta^2)\underline{\lambda}(Q)\|s\|^2\\ &+\frac{1}{2}L^2\bar{\lambda}(R)\|s\|^2-\frac{1}{2}\underline{\lambda}(R)\left\|W_a^{\mathrm{T}}\phi_a(\hat{s})\right\|^2\\ \leqslant\;&-\frac{1}{2}\beta^2\underline{\lambda}(Q)\|s\|^2,\end{aligned}$$

after using (49). Now for the term $\dot{V}_c$ we have,

$$\begin{aligned}\dot{V}_c\;\leqslant\;&-2\alpha\underline{\lambda}(\bar{\sigma})\left\|\tilde{W}_c\right\|^2+\frac{1}{4\alpha}\left\|\tilde{W}_c\right\|b_{\varepsilon c},\\ \leqslant\;&-\left(2\alpha\underline{\lambda}(\bar{\sigma})-\frac{1}{8\alpha}\right)\left\|\tilde{W}_c\right\|^2+\frac{1}{8\alpha}b_{\varepsilon c}^2. \quad (54)\end{aligned}$$

One has $\dot{V}_c<0$ whenever $\tilde{W}_c$ lies outside the compact set $\Omega_{\tilde{W}_c}=\left\{\tilde{W}_c:\left\|\tilde{W}_c\right\|\leqslant b_{\varepsilon c}\sqrt{\frac{1}{16\alpha^2\underline{\lambda}(\bar{\sigma})-1}}\right\}$, which guarantees that $\tilde{W}_c$ is UUB.

2) For the jump dynamics given by (48), we consider the following difference Lyapunov function candidate,

$$\begin{aligned}\Delta\mathcal{V}(\psi)\;=\;&\underbrace{V^\star\left(s\left(t_k^+\right)\right)-V^\star\left(s\left(t_k\right)\right)}_{\Delta V^\star(s)}+\underbrace{V^\star\left(\hat{s}\left(t_k^+\right)\right)-V^\star\left(\hat{s}\left(t_k\right)\right)}_{\Delta V^\star(\hat{s})}\\ &+\underbrace{\frac{\alpha_a^{-1}}{2}\mathrm{tr}\left\{\tilde{W}_a^{\mathrm{T}}\left(t_k^+\right)\tilde{W}_a\left(t_k^+\right)\right\}-\frac{\alpha_a^{-1}}{2}\left\{\tilde{W}_a^{\mathrm{T}}\left(t_k\right)\tilde{W}_a\left(t_k\right)\right\}}_{\Delta V_a}\\ &+\underbrace{V_c\left(\tilde{W}_c\left(t_k^+\right)\right)-V_c\left(\tilde{W}_c\left(t_k\right)\right)}_{\Delta V_c}\end{aligned}$$

From the previous analysis, the state s converges asymptotically to zero, and the critic estimation error $\tilde{W}_c$ is UUB, thus we have that $V^\star(s^+)\leqslant V^\star(s(t_k))$ if s lies outside the compact set Ω_s, and $V_c(\tilde{W}_c^+)\leqslant V_c(\tilde{W}_c(t_k))$ if $\tilde{W}_c$ lies outside the compact set $\Omega_{\tilde{W}_c}$. It follows that during jumps one has $s^+=\hat{s}^+$, we have that $V^\star(\hat{s}^+)\leqslant V^\star(\hat{s})$. Then one can write $\Delta\mathcal{V}(\hat{s}):=V^\star(\hat{s}^+)-V^\star(\hat{s}(t_k))\leqslant -k(\|\hat{s}\|)$ where k is a class-$\mathcal{K}$ function [28].

For V_a, one has

$$\begin{aligned}\Delta V_a&=\frac{\alpha_a^{-1}}{2}\mathrm{tr}\left\{\tilde{W}_a^{+T}\tilde{W}_a^+\right\}-\frac{\alpha_a^{-1}}{2}\mathrm{tr}\left\{\tilde{W}_a(t_k)^{\mathrm{T}}\tilde{W}_a\left(t_k\right)\right\}\\ &=\frac{\alpha_a^{-1}}{2}\mathrm{tr}\left\{\left[\tilde{W}_a\left(t_k\right)+\Phi\left(t_k\right)\right]^{\mathrm{T}}\left[\tilde{W}_a\left(t_k\right)+\Phi\left(t_k\right)\right]\right\} \quad (55)\end{aligned}$$

After multiplying each terms in the bracket, one has

$$\begin{aligned}\Delta V_a=&-\mathrm{tr}\left\{\tilde{W}_a^{\mathrm{T}}\left(t_k\right)\phi_a\left(s\left(t_k\right)\right)\phi_a(s\left(t_k\right))^{\mathrm{T}}\tilde{W}_a\left(t_k\right)\right\}\\ &-\mathrm{tr}\left\{\tilde{W}_a^{\mathrm{T}}\left(t_k\right)\phi_a\left(s\left(t_k\right)\right)\phi_a(s\left(t_k\right))^{\mathrm{T}}\varepsilon_u\right\}\\ &-\mathrm{tr}\left\{\tilde{W}_a^{\mathrm{T}}\left(t_k\right)\phi_a\left(s\left(t_k\right)\right)\tilde{W}_c^{\mathrm{T}}\frac{\partial\phi\left(s\left(t_k\right)\right)}{\partial x}G\left(s\left(t_k\right)\right)R^{-1}\right\}\\ &-\mathrm{tr}\left\{\tilde{W}_a^{\mathrm{T}}\left(t_k\right)\phi_a\left(s\left(t_k\right)\right)\frac{\partial\varepsilon_c}{\partial s}G\left(s\left(t_k\right)\right)R^{-1}\right\}\\ &+\frac{\alpha_a}{2}tr\left\{\left\|\phi_a\left(s\left(t_k\right)\right)\left[\phi_a(s\left(t_k\right))^{\mathrm{T}}\tilde{W}_a\left(t_k\right)+\phi_a(s\left(t_k\right))^{\mathrm{T}}\varepsilon_u\right.\right.\\ &+\left.\left.\tilde{W}_c^{\mathrm{T}}\frac{\partial\phi\left(s\left(t_k\right)\right)}{\partial s}G\left(s\left(t_k\right)\right)R^{-1}+\frac{\partial\varepsilon_c}{\partial s}G\left(s\left(t_k\right)\right)R^{-1}\right]\right\|^2\right\} \quad (56)\end{aligned}$$

Based on the actor error jump dynamics (45), applying Young's inequality to (56) yields the results in (57) (see top of next page), where ρ is the known bound since we have proved that the critic estimation error is UUB. Therefore, it is guaranteed that $\Delta V(\tilde{W}_a)<0$ when $\tilde{W}_a$ lies outside the compact set $\Omega_{\tilde{W}_a}=\left\{\tilde{W}_a:\left\|\tilde{W}_a\right\|\leqslant\sqrt{\frac{\rho}{\left(b_{\phi a}^2-\frac{3}{2}-\frac{\alpha_a}{2}b_{\phi a}^4\right)}}\right\}$.

To guarantee that the tracking error $s(t)$ remains in the compact set $\Omega_s\subseteq\mathbb{R}^n$, define the following compact set

$$\Omega_{\mathcal{V}_{\max}}:=\left\{s\in\mathbb{R}^n|\mathcal{V}(t)\leqslant\mathcal{V}_{\max})\right\}\subset\mathbb{R}^n$$

where $\mathcal{V}_{\max}$ is chosen as the largest constant so that $\Omega_{\mathcal{V}_{\max}}\subseteq\Omega$. Since by assumption $s(0)\in\Omega_s$, and $\Omega_s\subset\Omega$ then we can conclude that $s(0)\in\Omega$. While $s(t)$ remains inside Ω, we have seen that $\dot{\mathcal{V}}\leqslant 0$ and therefore $s(t)$ must remain inside $\Omega_{\mathcal{V}_{\max}}\subset\Omega$. The fact that $s(t)$ remains inside a compact set also excludes the possibility of finite escape time and therefore one has global existence of solution.

3) We shall now prove the absence of the Zeno behavior by showing that the inter-event time $\mathrm{T}_j:=t_{k+1}-t_k,\ \forall j\in\mathbb{N}$ is strictly positive.

At the triggering instant t_k, one has,

$$\begin{aligned}\|e\|^2\;\geqslant\;&\frac{(1-\beta^2)\underline{\lambda}(Q)}{L^2\bar{\lambda}(R)}\|s\|^2+\frac{\underline{\lambda}(R)}{L^2\bar{\lambda}(R)}\left\|W_a^{\mathrm{T}}\phi_a(\hat{s})\right\|^2\\ \geqslant\;&\frac{(1-\beta^2)\underline{\lambda}(Q)}{L^2\bar{\lambda}(R)}\|s\|^2. \quad (58)\end{aligned}$$

Therefore, $y:=\frac{\|e\|}{\|s\|}$ evolves from 0 to $K_y:=\sqrt{\frac{(1-\beta^2)\underline{\lambda}(Q)}{L^2\bar{\lambda}(R)}}$ during each inter-event interval. The dynamics of $s(t)$ using the intermittent actor-critic learning satisfies

$$\|\dot{s}\|\leqslant\underbrace{\left(b_f+b_g\left\|W_a^T\right\|\|k_a\|\right)}_{:=K_s}(\|s\|+\|e\|)$$

Based on [3], the evolution of y satisfies

$$\dot{y}\leqslant K_s(1+y)^2, t\in[t_k,t_{k+1})$$

by using the comparison lemma [28], one has

$$y\leqslant\frac{(t-t_k)K_s}{1-(t-t_k)K_s}, t\in[t_k,t_{k+1}).$$

$$\Delta V(\tilde{W}_a) \leqslant -(b_{\phi a}^2 - \frac{3}{2} - \frac{\alpha_a}{2} b_{\phi a}^4)\left\|\tilde{W}_a(t_k)\right\|^2 + \frac{1}{2}\left(b_{\phi a}^2 b_{\varepsilon a}\right)^2 + \frac{1}{8}\left[\left\|\tilde{W}_c\right\| b_{\phi a} b_{d\phi} \underline{\lambda}\left(R^{-1}\right)\right]^2 + \frac{1}{8}\left[b_{\phi a} b_{d\varepsilon} \underline{\lambda}\left(R^{-1}\right)\right]^2 + \frac{\alpha_a}{2} b_{\phi a}^4 b_{\varepsilon a}^2 + \frac{\alpha_a}{32} b_{\phi a}^2 \left\|\tilde{W}_c\right\|^2 b_{d\phi}^2 \underline{\lambda}(R^{-1})^2 + \frac{\alpha_a}{32} b_{\phi a}^2 b_{d\varepsilon c}^2 \underline{\lambda}(R^{-1})^2 + \frac{1}{2}\left(\alpha_a b_{\phi a}^3 b_{\varepsilon a}^2\right)^2 + \frac{1}{4}\alpha_a b_{\phi a} b_{d\phi} \underline{\lambda}\left(R^{-1}\right)\left\|\tilde{W}_c\right\| + \frac{1}{4}\alpha_a b_{\phi a} b_{d\varepsilon} \underline{\lambda}\left(R^{-1}\right) + \frac{1}{4}\alpha_a b_{\phi a} b_{\varepsilon a} b_{d\phi} \underline{\lambda}\left(R^{-1}\right)\left\|\tilde{W}_c\right\| + \frac{1}{8}\alpha_a b_{d\varepsilon} b_{d\phi} \underline{\lambda}\left(R^{-1}\right)\left\|\tilde{W}_c\right\| \Bigg\} = \rho \tag{57}$$

Evaluate the above inequality at event instant t_{k+1} yields

$$K_y \leqslant \frac{\|e(t_{k+1})\|}{\|s(t_{k+1})\|} \leqslant \frac{(t_{k+1} - t_k) K_s}{1 - (t_{k+1} - t_k) K_s}, \forall k$$

which is equivalent to the fact that

$$(t_{k+1} - t_k) \geqslant \frac{K_y}{K_s (1 + K_y)}, \forall k \in \mathbb{N}.$$

Therefore, the inter-event interval is strictly positive. This completes the proof. ■

V. Simulation Study

Consider the controlled Van der Pol oscillator with dynamics given by,

$$\dot{x} = \begin{bmatrix} x_2 \\ -x_1 + 0.5\left(1 - x_2^2\right) x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ x_1 \end{bmatrix} u. \tag{59}$$

In this scenario, it is desired that the state $x = [x_1\ x_2]^{\mathrm{T}}$ satisfies the following constraints,

$$x_1 \in (-0.6, 0.2), x_2 \in (-0.2, 0.2) \tag{60}$$

According to the converse HJB method [29], when the performance parameters are selected as $Q = I_{2\times 2}$ and $R = 1$, the optimal controller would be $u^\star(x) = -x_1 x_2$. When applying this optimal control policy to the system (59), the state evolution of $x(t)$ is shown in Figure 1, where the solid lines represent the state evolution and the dashed lines denote the asymmetric bounds for the states. The system state trajectories of $x(t)$ in the two-dimensional space are shown in the left of Figure 2, where the black box denotes the safety region. It is desired to drive the system states to the origin without violating the safety constraints. Even the states are regulated to the origin asymptotically, the full state constraints can not be guaranteed.

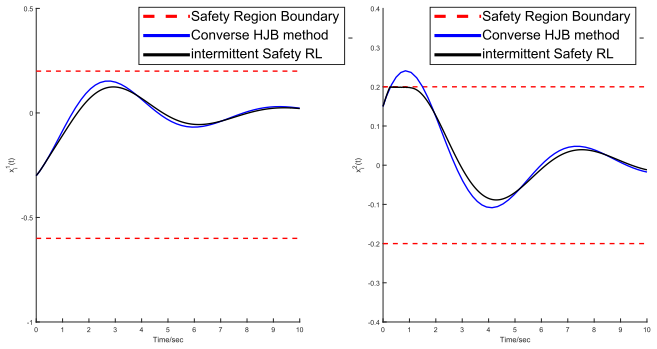

Fig. 1. The evolution of the states of the system with and without our proposed framework.

Next, we apply the safe reinforcement learning algorithm with the actor-critic-barrier structure developed in Section IV, where the corresponding state evolution of $x(t)$ is given in the right of Figure 1. In this case, we start the system from the same initial condition. The system state trajectories in two-dimensional space are shown in the right of Figure 2. In contrast to the previous case, one can observe that the state approach to the origin without violating the full state constraints. Both the full-state constraints and the closed-loop stability can be guaranteed when using the actor-critic-barrier learning algorithm. Finally, the state evolution of $s(t)$ as well as its sampled version $\hat{s}(t)$ is shown in Figure 3.

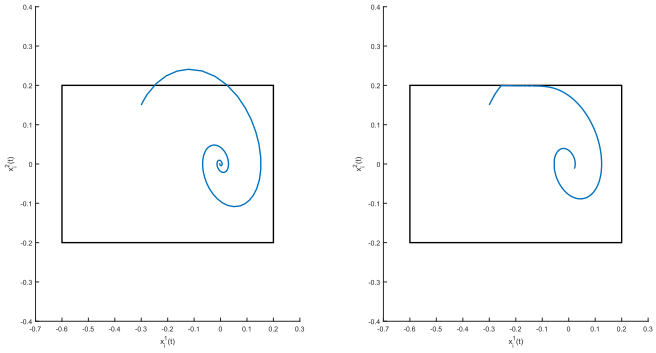

Fig. 2. Two-dimensional phase plot of state trajectories using the converse HJB approach [29] and the actor-critic-barrier algorithm. The black box denotes the safety region.

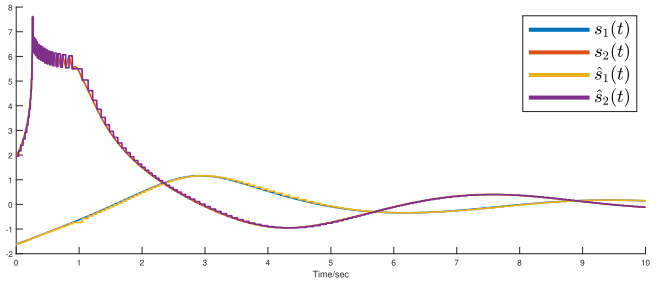

Fig. 3. Evolution of the state trajectories when applying the actor-critic-barrier algorithm.

VI. Conclusion

This paper presents a barrier-function-based system transformation to capture the full-state constraints in the regulation problem. With the presented system transformation, the full-state-constrained regulation problem is equivalent to an unconstrained problem. Then, a novel actor-critic-barrier

algorithm with intermittent feedback is presented to solve the constrained regulation problem in an online fashion. It is shown that the boundedness and convergence of the actor-critic weights to the optimal ones is guaranteed, and Zeno behavior is excluded. Future work will focus on using different kinds of intermittent protocols.

REFERENCES

[1] K. P. Tee, S. S. Ge, and E. H. Tay, "Barrier lyapunov functions for the control of output-constrained nonlinear systems," *Automatica*, vol. 45, no. 4, pp. 918 – 927, 2009.

[2] E. Arabi and T. Yucelen, "Set-theoretic model reference adaptive control with time-varying performance bounds," *International Journal of Control*, vol. 0, no. 0, pp. 1–12, 2018.

[3] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1680–1685, Sep. 2007.

[4] X. Ge, F. Yang, and Q.-L. Han, "Distributed networked control systems: A brief overview," *Information Sciences*, vol. 380, pp. 117 – 131, 2017.

[5] K. Vamvoudakis and S. Jagannathan, *Control of Complex Systems: Theory and Applications*. Butterworth-Heinemann, 2016.

[6] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.

[7] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal adaptive control and differential games by reinforcement learning principles*. IET, 2013, vol. 2.

[8] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Leader-follower output synchronization of linear heterogeneous systems with active leader using reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2139–2153, June 2018.

[9] Y. Yang, Z. Guo, H. Xiong, D. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2019. [Online]. Available: doi:10.1109/TNNLS.2019.2897814

[10] B. Ren, S. S. Ge, K. P. Tee, and T. H. Lee, "Adaptive neural control for output feedback nonlinear systems using a barrier lyapunov function," *IEEE Transactions on Neural Networks*, vol. 21, no. 8, pp. 1339–1345, Aug 2010.

[11] Y.-J. Liu and S. Tong, "Barrier lyapunov functions-based adaptive control for a class of nonlinear pure-feedback systems with full state constraints," *Automatica*, vol. 64, pp. 70 – 75, 2016.

[12] ——, "Barrier lyapunov functions for nussbaum gain adaptive control of full state constrained nonlinear systems," *Automatica*, vol. 76, pp. 143 – 152, 2017.

[13] W. He, Z. Li, and C. L. P. Chen, "A survey of human-centered intelligent robots: issues and challenges," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 602–609, 2017.

[14] E. Arabi, B. C. Gruenwald, T. Yucelen, and N. T. Nguyen, "A set-theoretic model reference adaptive control architecture for disturbance rejection and uncertainty suppression with strict performance guarantees," *International Journal of Control*, vol. 91, no. 5, pp. 1195–1208, 2018.

[15] R. Postoyan, P. Tabuada, D. Nei, and A. Anta, "A framework for the event-triggered stabilization of nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 982–996, April 2015.

[16] W. P. M. H. Heemels, K. H. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, Dec 2012, pp. 3270–3285.

[17] A. Molin and S. Hirche, "On the optimality of certainty equivalence for event-triggered control systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 470–474, Feb 2013.

[18] Y. Yang, H. Modares, K. G. Vamvoudakis, Y. Yin, and D. C. Wunsch, "Dynamic intermittent feedback design for H∞ containment control on a directed graph," *IEEE Transactions on Cybernetics*, 2019. [Online]. Available: doi:10.1109/TCYB.2019.2933736

[19] 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, June 2018.

[20] Y. Yang, D. Wunsch, and Y. Yin, "Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 8, pp. 1929–1940, Aug 2017.

[21] K. G. Vamvoudakis and F. L. Lewis, "Online actorcritic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878 – 888, 2010.

[22] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Optimal containment control of unknown heterogeneous systems with active leaders," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 3, pp. 1228–1236, May 2019.

[23] H. Modares, F. L. Lewis, and M. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1513–1525, Oct 2013.

[24] K. G. Vamvoudakis, M. F. Miranda, and J. P. Hespanha, "Asymptotically stable adaptiveoptimal 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, Nov 2016.

[25] K. G. Vamvoudakis and H. Ferraz, "Model-free event-triggered control algorithm for continuous-time linear systems with optimal performance," *Automatica*, vol. 87, pp. 412 – 420, 2018.

[26] Y. Yang, K. G. Vamvoudakis, H. Ferraz, and H. Modares, "Dynamic intermittent Q-learning-based model-free suboptimal co-design of L_2-stabilization," *International Journal of Robust and Nonlinear Control*, vol. 29, no. 9, pp. 2673–2694, 2019.

[27] K. G. Vamvoudakis, A. Mojoodi, and H. Ferraz, "Event-triggered optimal tracking control of nonlinear systems," *International Journal of Robust and Nonlinear Control*, vol. 27, no. 4, pp. 598–619, 2017.

[28] H. K. Khalil, *Nonlinear control*. Pearson New York, 2015.

[29] V. Nevistic and J. A. Primbs, "Constrained nonlinear optimal control: a converse hjb approach," *California Institute of Technology*, 1996.