Safe Intermittent Reinforcement Learning With Static and Dynamic Event Generators

Yongliang Yang, *Member, IEEE*, Kyriakos G. Vamvoudakis, *Senior Member, IEEE*, Hamidreza Modares, *Member, IEEE*, Yixin Yin, *Member, IEEE*, Donald C. Wunsch II, *Fellow, IEEE*

***Abstract*— In this article, we present an intermittent framework for safe reinforcement learning (RL) algorithms. First, we develop a barrier function-based system transformation to impose state constraints while converting the original problem to an unconstrained optimization problem. Second, based on optimal derived policies, two types of intermittent feedback RL algorithms are presented, namely, a static and a dynamic one. We finally leverage an actor/critic structure to solve the problem online while guaranteeing optimality, stability, and safety. Simulation results show the efficacy of the proposed approach.**

***Index Terms*— Actor/critic structures, asymptotic stability, barrier functions, reinforcement learning (RL), safety-critical systems.**

I. Introduction

CONSTRAINTS are inevitability present in engineering systems as demonstrated in variety of engineering applications, including flexible joint robots [1] and flight control [2]. Such constraints can be categorized as input based [3], [4], output based [5], and state based [6].

Conventional Lyapunov analysis guarantees closed-loop stability of the equilibrium point, but without any conclusions about the transient behavior of the input and output signals. To impose constraints on the behavior of the system, nonquadratic Lyapunov functions, referred to as barrier Lyapunov functions [7], [8], have been used in time-delay systems [9] and in uncertain systems with unknown control directions [10]. In order to satisfy a user-defined transient performance, a prescribed performance adaptive control method is proposed in [11].

A. Related Work

Coping with input constraints is a challenging task [12], especially in optimization-based approaches, such as [13], for which the solution requires the solution of Hamilton–Jacobi–Bellman (HJB) equations [14], [15]. Due to the "curse of dimensionality," a closed-form solution to the HJB equation is difficult to obtain even for systems with simple dynamics and without any constraints.

Reinforcement learning (RL) [16]–[19] is a machine learning method that provides an efficient way to solve the HJB equation online in real time [20]. Variants of RL have been applied widely in control, including regulation [21]–[23], cooperative control [24], [25], robust control [26], [27], differential games [28]–[30], and constrainted control [31]–[33]. However, guaranteeing that the state and input constraints are not violated, as in model predictive control [34]–[36], is challenging especially for mechanisms with intermittent feedback. The aforementioned results may lead to unnecessary communication and computation loads due to continuous feedback. To address this concern, several works with intermittent feedback have been used [37]–[44].

There are primarily two types of approaches to safe RL. Such techniques include modification of the optimization criterion with a safety component, such as barrier functions, by transforming the operational constraints into soft constraints; and modifying the exploration process through the incorporation of external system knowledge or historical data. This article is toward the latter direction and combines advantages from both approaches.

B. Contributions

The contributions of this article are threefold. First, full-state and input constraints are considered simultaneously for designing safe policies, with the use of barrier Lyapunov functions and proper performance design. Second, two types of intermittent policies are presented to reduce the communication burden, namely, static and dynamic. Finally, a safe RL algorithm

Manuscript received July 2, 2019; revised November 9, 2019; accepted January 12, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61903028, in part by the China Post-Doctoral Science Foundation under Grant 2018M641197, in part by the Fundamental Research Funds for the Central Universities under Grant FRF-TP-18-031A1 and Grant FRF-BD-17-002A, in part by the National Science Foundation under Grant S&AS-1849264 and Grant CPS-1851588, in part by ONR Minverva under Grant N00014-18-1-2874, in part by the DARPA/Microsystems Technology Office, and in part by the Army Research Laboratory under Grant W911NF-18-2-0260. *(Corresponding author: Yongliang Yang.)*

Yongliang Yang and Yixin Yin are with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China, and also with the Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, University of Science and Technology Beijing, Beijing 100083, China (e-mail: yangyongliang@ieee.org; yyx@ies.ustb.edu.cn).

Kyriakos G. Vamvoudakis is with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: kyriakos@gatech.edu).

Hamidreza Modares is with the Department of Mechanical Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: modaresh@msu.edu).

Donald C. Wunsch II is with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401 USA (e-mail: wunsch@ieee.org).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2967871

is developed to learn the solution to the constrained problem by using the past and current data concurrently.

C. Structure

The remainder of this article is structured as follows. Section I formulates the problem. A novel barrier function is developed to deal with the constraints on the state and the inputs in Section II. Section III develops the basis for the intermittent design with static and dynamic triggers. An actor-critic-barrier structure is introduced in Section IV. Section V shows the efficacy with simulations, and finally, Section VI concludes and talks about future work.

II. Problem Formulation

Consider the continuous-time nonlinear dynamical system

$$\begin{aligned}\dot{x}_i &= x_{i+1}, \quad i = 1, \ldots, n-1\\ \dot{x}_n &= f(x) + g(x)u, \quad t \geq 0\end{aligned} \tag{1}$$

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

The constrained control problem of system (1) with full state constraints can be formulated as follows.

Problem 1 (Safety Control Problem With Full-State Constraints and Input Saturation): Consider the system (1), find a policy $u(\cdot) : \mathbb{R}^n \to \mathbb{R}$ such that the following performance is minimized for every x_0 and $t_0 \geq 0$:

$$\bar{V}(\cdot) = \int_{t_0}^{\infty} r(x, u)dt \tag{2}$$

given the equality constraints (1). Note that the state $x = [\, x_1 \ \ldots \ x_n \,]^{\mathrm{T}}$ and the control input u satisfy the following constraints:

$$\|u\| \leq \gamma \tag{3}$$

$$x_j \in (a_j, A_j) \quad \forall j = 1, \ldots, n \tag{4}$$

where $\gamma > 0$, $a_i < 0$, $A_i > 0$, and $r(x, u) := H(x) + \Theta(u)$ with $H(x)$ and $\Theta(u)$ positive definite functions. □

A. Barrier Function Design for Full-State Constraints

In this section, a barrier function-based system transformation is designed to deal with asymmetric full-state constraints.

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

$$b(z; a, A) = \log\left(\frac{A}{a}\frac{a - z}{A - z}\right) \quad \forall z \in (a, A) \tag{5}$$

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

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

with a derivative given as

$$\frac{db^{-1}(y; a, A)}{dy} = \frac{Aa^2 - aA^2}{a^2e^y - 2aA + A^2e^{-y}}. \qquad \square$$

Based on Definition 1, we consider the barrier function-based state transformation as follows:

$$\begin{aligned}s_i &:= b(x_i; a_i, A_i)\\ x_i &:= b^{-1}(s_i; a_i, A_i), \quad i = 1, \ldots, n.\end{aligned} \tag{6}$$

The dynamics are given by

$$\frac{dx_i}{dt} = \frac{dx_i}{ds_i}\frac{ds_i}{dt}, \quad t \geq 0$$

and after using Definition 1 we get

$$\begin{aligned}\dot{s}_i &= \frac{a_{i+1}A_{i+1}\big(e^{\frac{s_{i+1}}{2}} - e^{-\frac{s_{i+1}}{2}}\big)}{a_{i+1}e^{\frac{s_{i+1}}{2}} - A_{i+1}e^{-\frac{s_{i+1}}{2}}}\frac{A_i^2e^{-s_i} - 2a_iA_i + a_i^2e^{s_i}}{A_ia_i^2 - a_iA_i^2}\\ &:= F_i(s_i, s_{i+1}), \quad i = 1, \ldots, n-1\\ \dot{s}_n &= [f(x) + g(x)u]\frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2}\\ &:= F_n(s) + g_n(s)u\end{aligned}$$

where

$$\begin{aligned}F_n(s) &= \frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2}\\ &\quad \times f\big(\big[\, b_1^{-1}(s_1) \quad \ldots \quad b_n^{-1}(s_n) \,\big]\big)\\ g_n(s) &= \frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2}\\ &\quad \times g\big(\big[\, b_1^{-1}(s_1) \quad \ldots \quad b_n^{-1}(s_n) \,\big]\big).\end{aligned}$$

The dynamics of the system with the augmented state $s := [\, s_1 \ \ldots \ s_n \,]^{\mathrm{T}}$ can be expressed in a compact form as

$$\dot{s} = F(s) + G(s)u, \quad t \geq 0. \tag{7}$$

with

$$\begin{aligned}F(s) &:= [\, F_1(s_1, s_2) \quad \ldots \quad F_n(s) \,]^{\mathrm{T}}\\ G(s) &:= [\, 0 \quad \ldots \quad 0 \quad g_n(s) \,]^{\mathrm{T}}.\end{aligned}$$

Assumption 1: The dynamics given by (7) satisfy the following.

1) $F(s)$ is Lipschitz with $F(0) = 0$, and there exists a constant b_f such that for $s \in \Omega$, $\|F(s)\| \leq b_f\|s\|$, where $\Omega \subseteq \mathbb{R}^n$ is a compact set containing the origin.
2) $G(s)$ is bounded on Ω, i.e., there exists a constant b_g such that $\|G(s)\| \leq b_g$.
3) The system stabilizable for every $s \in \Omega$. □

Property 1: A barrier function has the following properties.

1) Given that the state s of the system (7) is bounded, then the constraints (4) on the state x of system is satisfied, that is,
$$|b(z; a, A)| < +\infty \quad \forall z \in (a, A).$$
2) If the state x of the system (1) approaches the boundary of the safe region (a_i, A_i), the state s will approach infinity, that is,
$$\lim_{z \to a^+} b(z; a, A) = -\infty; \quad \lim_{z \to A^-} b(z; a, A) = +\infty.$$
3) The state s of the system (7) is regulated, i.e., $s \equiv 0$ if and only if the state x of the system (1) is regulated, that is,
$$b(0; a, A) = 0 \quad \forall a < A.$$

4) The barrier function is a monotonic mapping, and hence, the inverse exists. □

Remark 1: The barrier function-based system transformation guarantees the following.

1) The stabilization of s is equivalent to the stabilization of x.
2) The boundedness of s is equivalent to the satisfaction of a proper condition on x. □

B. Penalty Function Design for Input Saturation

In this section, in order to deal with input saturations, Problem 1 is converted to an unconstrained optimization problem with a nonquadratic control input penalty function.

Problem 2 (Optimal Control Problem With Input Saturation): Find a policy $u(\cdot):\mathbb{R}^n \to \mathbb{R}^m$ such that the performance

$$V(\cdot) = \int_{t_0}^{\infty} U(s,u)dt \quad \forall s_0,\ t_0 \geq 0 \tag{8}$$

is minimized given the equality constraints (7), and $U(s,u) := Q(s) + \Theta(u)$, with $Q(s) := s^{\mathrm{T}}Qs,\ Q \succeq 0$, and the penalty function on the control in out is selected as [3], [31]

$$\Theta(u) := 2\int_0^u r\theta^{-1}(v)\,dv \tag{9}$$

where $r \in \mathbb{R}^+$. □

Definition 2 (Zero-State Observability [13]): A nonlinear system $\dot{x} = f(x,t)$ with a measured output $y = h(x)$ is zero-state observable if $y(t) \equiv 0\ \forall t \geq 0$ implies that $x(t) \equiv 0$ $\forall t \geq 0$. □

Assumption 2: The performance function defined by (8) satisfies the zero-state observability property. □

Definition 3 (Admissible Policy [3]): A control policy $\mu(s)$ is said to be admissible with respect to (8) on $\Omega \subseteq \mathbb{R}^n$, denoted by $\mu(s) \in \pi(\Omega)$, if the following is satisfied.

1) $\mu(s)$ is continuous on Ω.
2) $\mu(0) = 0$.
3) $u(s) := \mu(s)$ stabilizes (7) on Ω.
4) $V(s)$ is finite $\forall s \in \Omega$. □

The penalty function $\theta_i(\cdot)$ on the control input is selected as

$$\theta(v) = \gamma \tanh\left(\frac{v}{\gamma}\right). \tag{10}$$

We can now rewrite $\Theta(u)$ in (9) as [3], [31]

$$\begin{aligned}\Theta(u) &= 2r\int_0^u \gamma\left[\tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv \\ &= 2r\gamma u\tanh^{-1}\left(\frac{u}{\gamma}\right) + r\gamma^2\log\left(1-\left(\frac{u}{\gamma}\right)^2\right).\end{aligned} \tag{11}$$

The penalty function design given by (10) and (11) has the following properties.

Property 2: To deal with input saturation (3), the function $\theta_i(\cdot)$ satisfies [3] the following.

1) It is a one-to-one real-analytic integrable function of class $\mathrm{L}_2(\Omega)$ and C^p with $p \geq 1$.
2) It maps $\mathbb{R}$ onto the interval $(-\gamma, \gamma)$.
3) $\theta_i(0) = 0$.
4) It is a monotonic odd function with a first derivative bounded by a constant M. □

Remark 2: According to Property 2, we have the following.

1) $\Theta(u) = 0$ if and only if $u = 0$.
2) $\Theta(u)$ is a positive definite function of its argument.
3) $\Theta(u)$ is bounded if and only if the condition is satisfied.
4) $\Theta(u)$ approaches to infinity as $\|u_i\| \to \gamma$. □

C. Equivalence Between Problems 1 and 2

Given the barrier function-based system transformation, one can convert the Problem 1 with full-state constraints and input saturation to Problem 2, which is an unconstrained optimization problem.

In this section, we provide a formal results of equivalence.

Lemma 1: Suppose that Assumptions 1 and 2 hold and that $u^\star(\cdot)$ solves Problem 2 for (7) with (3) and (4). Then, the following holds.

1) The closed-loop system satisfies (4) provided that the initial state x_0 of the system (1) is within the region described by the constants a_i, A_i, and $\forall i$.
2) The performance described by (8) is equivalent to that of (2) given that the penalty functions on x and s are the same.

Proof: From (8), one can obtain that $\dot{V}^\star(t) \leq 0$, that is,

$$V^\star(s(t)) \leq V^\star(s(0)), \quad t \geq 0.$$

Then, $V^\star(s(t))$ remains bounded given that $V^\star(s(0))$ is bounded, which is guaranteed by the condition that the initial condition $x(0)$ of the system (1) satisfies (4). Finally, one can infer that

$$x_\ell(t) \in (a_\ell, A_\ell), \quad \ell = 1,2,\ldots,n,\ t \geq 0.$$

Therefore, given $u^\star$, the constraints of Problem 1 are satisfied.

Now consider the barrier-function-based state transformation described by (6). Each element of the state $s = [\,b_1(x_1)\ \ldots\ b_n(x_n)\,]^{\mathrm{T}}$ is finite given that x satisfies the constraints given in (4).

Next, comparing the two performance functions (2) and (8) provides an equivalent results given that the penalty functions of x and s are the same. This completes the proof. ■

III. Intermittent Feedback Design

Two types of suboptimal intermittent feedback designs with static and dynamic triggering conditions that guarantee the input saturation and full-state constraints are now developed.

A. Continuous Feedback

Define the Hamiltonian as

$$\mathcal{H}\left(s,u,\frac{\partial V}{\partial s}\right) = \left(\frac{\partial V}{\partial s}\right)^{\mathrm{T}}[f(s)+g(s)u] + U(s,u). \tag{12}$$

Then, by differentiating (8) along the system trajectories yields the Bellman equation as

$$0 = \mathcal{H}\left(s,u,\frac{\partial V}{\partial x}\right) = \left(\frac{\partial V}{\partial s}\right)^{\mathrm{T}}[F(s)+G(s)u] + U(s,u). \tag{13}$$

Define the optimal value function as

$$V^\star(s(t)) = \min_{u(\cdot)\in\pi(\Omega)} \int_t^\infty [Q(s(\tau)) + \Theta(u(\tau))]d\tau.$$

The necessary condition for optimality is

$$0 = \min_{u\in\pi(\Omega)} \mathcal{H}\left(s, u, \frac{\partial V^\star}{\partial s}\right).$$

The optimal control $u^\star$ can be found by applying the stationary condition to the Hamiltonian

$$u^\star(s) = \operatorname*{argmin}_{u\in\pi(\Omega)} \mathcal{H}\left(s, u, \frac{\partial V^\star}{\partial s}\right) = -\gamma\tanh(D^\star(s))$$
$$D^\star(s) = \frac{1}{2\gamma r}G^{\mathrm{T}}(s)\frac{\partial V^\star(s)}{\partial s}. \tag{14}$$

Inserting the optimal control policy (14) into (9) yields

$$\Theta(u^\star) = \gamma\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)\tanh(D^\star(s)) + \gamma^2 r\ln[1-\tanh^2(D^\star(s))] \tag{15}$$

and then, finally, one has the HJB as [13], [31]

$$0 = Q(s) + \left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} F(s) + \gamma^2 r\ln[1-\tanh^2(D^\star)]. \tag{16}$$

B. Intermittent Feedback

In order to reduce the computation and communication burden, an intermittent feedback is developed by introducing an aperiodic sampling mechanism, that is,

$$\hat{s}(t) := \begin{cases} s(t_k), & \forall t\in[t_k, t_{k+1}) \\ s(t), & t = t_{k+1} \end{cases}$$

where $\{t_k\}_{k=0}^\infty$ is a sequence of event instants.

The gap is denoted as

$$e(t) := \hat{s}(t) - s(t). \tag{17}$$

In the sequel, an intermittent control law with aperiodic sampling is introduced as[1]

$$u_e(t) := u^\star(\hat{s}) = -\gamma\tanh\left(\frac{1}{2\gamma r}G^{\mathrm{T}}(\hat{s})\frac{\partial V^\star(s)}{\partial s}\bigg|_{s=\hat{s}}\right). \tag{18}$$

Given u_e from (18), the closed-loop dynamics of (7) can be written as

$$\dot{s}(t) = F(s(t)) + G(s(t))u^\star(s(t)+e(t)), \quad t\geq 0. \tag{19}$$

The following assumptions are adopted from [37] and [38].

Assumption 3: The function $D^\star(s)$ defined in (14) is Lipschitz continuous on Ω satisfying

$$\|D^\star(a) - D^\star(b)\| \leq L_D\|a-b\| \quad \forall a, b\in\Omega. \qquad \square$$

[1]In the intermittent feedback design, the event-triggered control consists of the feedback control mapping and the event triggering condition. Here, the notation u_e denotes the event-triggered control design with the optimal policy $u^\star(\cdot)$. Combined with different event-triggering conditions, one can obtain different intermittent feedback designs. This will be illustrated later as u_s and u_d.

Assumption 4: At the intermittent instant, t_k $\forall k\in\mathbb{Z}^+$, finite-time stabilization is not achieved, i.e., $s(t_k)\neq 0$. $\square$

For the intermittent feedback control policy (18), the following lemma holds.

Theorem 1 (Static Intermittent Feedback Design): Suppose that Assumptions 1–4 hold. Consider the constrained-input nonlinear system given by (7), and let $V^\star(s)$ be the optimal solution to the HJB equation (16). Then, the following holds.

1) The closed-loop system has an asymptotically stable equilibrium point with
$$u_s(t) := -\gamma\tanh\left(\frac{1}{2\gamma r}G^{\mathrm{T}}(\hat{s})\frac{\partial V^\star(s)}{\partial s}\bigg|_{s=\hat{s}}\right) \tag{20}$$
and an intermittent condition
$$\|e\|^2 \leq \frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}\|s\|^2 + \frac{1}{L_e}\Theta(u^\star(\hat{s})) \tag{21}$$
where $L_e := \gamma^2 L_D^2 r$, and $\beta\in(0,1)$ is a design parameter.
2) The interevent time defined by $\bar{\delta}_k := t_{k+1} - t_k$ $\forall k\in\mathbb{Z}^+$ is strictly positive and has a lower bound. That is, the Zeno behavior is excluded.

Proof: See the Appendix. ■

Remark 3: As a result of Theorem 1, the sampling instants determined by (21) can be expressed as

$$t_0 = 0$$
$$t_{k+1} = \inf_{t\in\mathbb{R}^+}\{t > t_k \wedge p\leq 0\} \tag{22}$$

where

$$p := (1-\beta^2)\lambda_{\min}(Q)\|s\|^2 + \Theta(u^\star(\hat{s})) - L_e\|e\|^2. \tag{23}$$

Note that the parameters of the intermittent condition (21) are time invariant, and $p\geq 0$ has to be always satisfied. Moreover, the intermittent feedback $u_s(t)$ depends on the solution of the HJB equation (16), which is a nonlinear partial differential and extremely difficult to be solved analytically. $\square$

To formulate the dynamic triggering condition, we shall introduce the following internal dynamics

$$\dot{\eta} = -\mu\eta + p \quad \forall\eta(t_0) = \eta_0, \quad t\geq 0 \tag{24}$$

where p is defined in (23) and $\mu\in\mathbb{R}^+$ is a parameter to be designed later.

Consider the dynamic triggering condition given by

$$\eta(t) + \vartheta p(t) \leq 0 \tag{25}$$

where $\vartheta\in\mathbb{R}^+$ is a parameter to be designed later. The intermittent instants are then determined as

$$t_0 = 0$$
$$t_{k+1} = \inf_{t\in\mathbb{R}^+}\{(t > t_k)\wedge(\eta(t) + \vartheta p(t)\leq 0)\}. \tag{26}$$

The properties of the dynamic intermittent condition (25) are presented in the following lemma.

Lemma 2: Let μ be a positive constant, $\eta_0, \vartheta\in\mathbb{R}_0^+$, and p is defined as in (23). Then, the following holds.

1) $\eta(t) + \vartheta p(t)\geq 0 \quad \forall t\geq 0$.
2) $\eta\geq 0 \quad \forall t\geq 0$.

Proof: Using (26), the following condition holds:

$$\eta(t) + \vartheta p(t)\geq 0.$$

If $\vartheta = 0$, then the first proposition $\eta(t) + \vartheta p(t) \geq 0$ implies that $\eta \geq 0\ \forall t \geq 0$. Now, if $\vartheta \neq 0$, it follows from the first proposition that $p(t) \geq -(1/\vartheta)\eta(t)$. Considering (24), one can obtain $\dot{\eta}(t) \geq -(\mu + (1/\vartheta))\eta(t)$ for $\forall \eta_0$ and $t \geq 0$. Let $y(\eta_0, t)$ be the solution of the differential equation

$$\dot{y}(t) = -\left(\mu + \frac{1}{\vartheta}\right) y(t), \quad y(0) = \eta_0,\ t \geq 0.$$

Then, from the comparison principle [45], we have that $\eta(t) \geq y(t) \geq 0$. This completes the proof. ■

The closed-loop stability of the equilibrium point is provided in the next theorem.

Theorem 2 (Dynamic Intermittent Feedback Design): Suppose that Assumptions 1–3 hold. Consider the intermittent feedback control given by[2]

$$u_d(t) := -\gamma \tanh\left(\frac{1}{2\gamma r} G^{\mathrm{T}}(\hat{s}) \frac{\partial V^\star(s)}{\partial s}\bigg|_{s=\hat{s}}\right) \tag{27}$$

with the intermittent instants defined in (26). Then, the following holds.

1) The closed-loop system of (19) has an asymptotically stable equilibrium point.
2) Let $\{t_k^s\}_{k=0}^\infty$ and $\{t_k^d\}_{k=0}^\infty$ be the triggering time sequences determined by the static and the dynamic intermittent feedback laws, respectively. Assume also that $t_i^s = t_j^d$ and $x(t_i^s) = x(t_j^d)$. Then, by denoting the next triggering instants by the static and the dynamic intermittent feedback laws as t_{i+1}^s and t_{j+1}^d, respectively, one has $t_{k+1}^d \geq t_{k+1}^s$. That is, Zeno behavior is excluded.

Proof: Consider a Lyapunov candidate as $W = V^\star + \eta$ for the augmented system of (24) and the system (19) with (27). Then, according to Theorem 1, $\dot{W}$ satisfies $\dot{W} = \dot{V}^\star(x) + \dot{\eta} \leq -p + (-\mu\eta + p)$. According to Lemma 2, $\eta \geq 0$. Then, $\dot{W}$ is negative definite, which implies that the closed-loop system has an asymptotically stable equilibrium point.

Assume now that $t_{j+1}^d < t_{i+1}^s$. Then, based on (26), one has that $p(t_{j+1}^d) > 0$, i.e., $(1-\beta^2)\lambda_{\min}(Q)\|s\|^2 + \Theta(u^\star(\hat{s})) \geq L_e\|e\|^2$. Based on (26) and Lemma 2, one has $\eta(t_{j+1}^d) + \vartheta p(t_{j+1}^d) \leq 0$, that is,

$$\begin{aligned} 0 &\geq \eta\big(t_{\mathrm{j+1}}^{\mathrm{d}}\big) \\ &\quad + \vartheta[(1-\beta^2)\lambda_{\min}(Q)\|s\|^2 + \Theta(u^\star(\hat{s})) - L_e\|e\|^2] \\ &\geq \vartheta[(1-\beta^2)\lambda_{\min}(Q)\|s\|^2 + \Theta(u^\star(\hat{s})) - L_e\|e\|^2] \\ &= \vartheta p\big(t_{\mathrm{j+1}}^{\mathrm{d}}\big). \end{aligned}$$

Thus, $p(t_{j+1}^d) \leq 0$, which contradicts the assumption that $p(t_{j+1}^d) > 0$. Therefore, $t_{j+1}^d \geq t_{i+1}^s$. Note that during the instant $t_i^s = t_j^d$, then the interevent time of the dynamic intermittent feedback is no smaller than the static intermittent feedback. Based on Theorem 1, we can conclude that the dynamic intermittent excludes the Zeno behavior. This completes the proof. ■

[2]From (20) and (27), one can observe that the static and dynamic intermittent feedback designs share the same feedback control mapping. However, as given in (22) and (26), u_s and u_d are not the same due to the different event instants.

IV. Actor/Critic Learning

In Section III, both the continuous and intermittent feedback designs depend on the optimal policy mapping $u^\star(\cdot)$, which is obtained by solving the HJB equation (16). In this section, a novel online RL algorithm is presented to obtain the optimal control policy in an online fashion.

A. Actor-Critic Network

To find the solution to HJB equation (16) in an online fashion, we employ an actor-critic RL algorithm.

According to the Weierstrass high-order approximation theorem [46], we shall use a function approximator for the value function, in a compact set $\Omega \subseteq \mathbb{R}^n$. There exists a critic approximator such that

$$\begin{aligned} V^\star(s) &= (W^\star)^{\mathrm{T}}\phi_c(s) + \varepsilon_c(s) \\ \nabla V^\star(s) &= [\nabla\phi_c(s)]^{\mathrm{T}} W^\star + \nabla\varepsilon_c(s) \end{aligned} \tag{28}$$

where $W^\star \in \mathbb{R}^N$ are the critic weights, $\phi_c(\cdot)$:$\mathbb{R}^n \to \mathbb{R}^N$ is the basis, $\varepsilon(s)$ and $\nabla\varepsilon(s)$ are the approximation errors. The ideal weights $W^\star$ that provide the best N-dimensional approximation to the value function $V^\star(s)$ on the compact set Ω_s are unknown. Therefore, we shall estimate $W^\star$ by actual weights W_c as follows:

$$\begin{aligned} V(s) &= W_c^{\mathrm{T}}\phi_c(s) \\ \nabla V(s) &= [\nabla\phi_c(s)]^{\mathrm{T}} W_c. \end{aligned} \tag{29}$$

As shown in (14), the optimal control policy depends on $(\partial V^\star(s)/\partial s)$, and hence, the policy can be determined as

$$\begin{aligned} u_c(\hat{s}) &= -\gamma \tanh(D_c(\hat{s})) \\ D_c(\hat{s}) &= \frac{1}{2\gamma r} G^{\mathrm{T}}(\hat{s})[\nabla\phi_c(\hat{s})]^{\mathrm{T}} W_c. \end{aligned} \tag{30}$$

In order to ensure stability of the equilibrium point, the policy that is going to be applied to the system will be implemented by an actor network as follows:

$$\begin{aligned} u_a(\hat{s}) &= -\gamma \tanh(D_a(\hat{s})) \\ D_a(\hat{s}) &= \frac{1}{2\gamma r} G^{\mathrm{T}}(\hat{s})[\nabla\phi_c(\hat{s})]^{\mathrm{T}} W_a. \end{aligned} \tag{31}$$

B. Critic and Previous Data

In this section, the parameter update of the critic network is presented. To obviate the requirement of persistency of excitation (PE) condition while guaranteeing the parameter convergence, history data are efficiently used for the critic learning.

Using (28) in the Bellman equation (13), yields:

$$\begin{aligned} \varepsilon_B &= U(s, u^\star) + (W^\star)^{\mathrm{T}}\nabla\phi_c(s)[F(s) + G(s)u^\star] \\ &= U(s, u^\star) + (W^\star)^{\mathrm{T}}\sigma \\ \sigma &= \nabla\phi_c(s)[F(s) + G(s)u^\star]. \end{aligned} \tag{32}$$

From (28), one can observe that the residual ε_B is due to the value gradient approximation error $\nabla\varepsilon_c(s)$, that is,

$$\varepsilon_B = -[\nabla\varepsilon_c(s)]^{\mathrm{T}}[F(s) + G(s)u^\star]. \tag{33}$$

Accordingly, the HJB equation (16) can be represented using the value function approximation (28) with a residual as

$$\begin{aligned}\varepsilon_{hjb}(s) &= (W^\star)^{\mathrm{T}}\nabla\phi(s)F(s) + \gamma^2 r\ln[1-\tanh^2(D_W)] \\ &\quad + Q(s) \\ D_W &= \frac{1}{2\gamma r}G^{\mathrm{T}}(s)(\nabla\phi_c)^{\mathrm{T}}W^\star.\end{aligned}$$

In order to derive the update law for the critic network we need to define the following error:

$$e_c(t) := U(s(t), u^\star(t)) + W_c^{\mathrm{T}}(t)\sigma(t). \tag{34}$$

By denoting $\tilde{W}_c := W^\star - W_c$ and from (32) and (34), the relationship between e_c and the Bellman residual ε_B can be expressed in terms of the error $\tilde{W}_c$ as

$$e_c = \varepsilon_B - \tilde{W}_c^{\mathrm{T}}\sigma. \tag{35}$$

The policy evaluation procedure can be formulated by defining the following performance:

$$E_c(t) = \frac{1}{2}\frac{[e_c(t)]^2}{\left[1+\sigma^{\mathrm{T}}(t)\sigma(t)\right]^2}.$$

In order to achieve that $e_c \to \varepsilon_B$ as $W_c \to W^\star$, we need to apply the gradient descent algorithm as follows:

$$\dot{W}_c = -\alpha_c\frac{\partial E_c}{\partial W_c} = -\alpha_c\frac{\sigma}{\left(1+\sigma^{\mathrm{T}}\sigma\right)^2}[\sigma^{\mathrm{T}}W_c + U(s,u)].$$

Assumption 5: For the critic network, the following holds on a compact set Ω_s:

1) $W^\star$ is bounded, i.e., $\|W^\star\| \le W_{\max}$.
2) $\|\varepsilon_c(s)\| \le \varepsilon_{\mathrm{cmax}}$ and $\|\nabla\varepsilon_c(s)\| \le \varepsilon_{\mathrm{cdmax}}$.
3) The basis as well as its gradient are bounded, i.e., $\|\phi_c(s)\| \le \phi_{\mathrm{cmax}}$, $\|\nabla\phi_c(s)\| \le \phi_{\mathrm{cdmax}}$.
4) The HJB and the Bellman residuals are bounded, i.e., $\|\varepsilon_{\mathrm{hjb}}\| \le \varepsilon_h$ and $\|\varepsilon_B\| \le \varepsilon_{\mathrm{Bmax}}$.
5) The basis gradient is Lipschitz continuous in the sense that there exists a constant L_ϕ such that $\|\nabla\phi_c(s_1) - \nabla\phi_c(s_2)\| \le L_\phi\|s_1 - s_2\|$. □

As shown in [47], in order to guarantee convergence of the weights W_c to the ideal ones $W^\star$, the signal σ is required to be persistently excited in the following sense.

Definition 4 (Persistency of Excitation [47]): A vector signal $y(t) \in \mathbb{R}^p$ is exciting over the interval $[t, t+T_{\mathrm{PE}}]$ with $T_{\mathrm{PE}} \in \mathbb{R}^+$ if there exists $\beta_1 \in \mathbb{R}^+$ and $\beta_2 \in \mathbb{R}^+$ such that $\forall t$

$$\beta_1 I_{p\times p} \le \int_t^{t+T_{\mathrm{PE}}} y(\tau)y^{\mathrm{T}}(\tau)d\tau \le \beta_2 I_{p\times p}. \qquad \square$$

Remark 4: The PE condition on the signal $\sigma(t)$ is equivalent to the requirement of positive definiteness of the matrix $\bar{M}$ on an arbitrary finite interval $[t, t+T]$, where

$$\bar{M}(t) = \int_t^{t+T_{\mathrm{PE}}} M(\tau)d\tau; \quad M(\tau) = \frac{\sigma(\tau)\sigma^{\mathrm{T}}(\tau)}{\left[1+\sigma^{\mathrm{T}}(\tau)\sigma(\tau)\right]^2}.$$

Note that the PE condition requires future information and cannot be checked during the learning phase. □

To relax the dependence on future information and rely only on past data, the following modified objective will be used:

$$\bar{E}_c = E_c(t) + E_c(t_k); \quad E_c(t_k) := \sum_{k=1}^{N}\frac{e_{ck}^2}{\left(1+\sigma_k^{\mathrm{T}}\sigma_k\right)^2}$$

where $e_{ck} := U_k + W_c^{\mathrm{T}}(t)\sigma_k$, $U_k := U(s(t_k), u(t_k))$, $\sigma_k := \sigma(t_k)$ and t_k denotes the intermittent instant. Then, the tuning law can be obtained as

$$\begin{aligned}\dot{W}_c &= -\alpha_c\frac{\partial\bar{E}_c}{\partial W_c} \\ &= -\alpha_c\frac{\sigma(t)e_c(t)}{\left[1+\sigma^{\mathrm{T}}(t)\sigma(t)\right]^2} - \alpha_c\sum_{k=1}^{N}\frac{\sigma_k e_{ck}}{\left[1+\sigma_k^{\mathrm{T}}\sigma_k\right]^2}.\end{aligned} \tag{36}$$

Condition 1: Let $Z = \left[\sigma_1 \ \dots \ \sigma_p\right]$ be the history stack. Then, Z contains as many linearly independent elements as the number of basis in (28). That is $\mathrm{rank}(Z) = N$. □

Fact 1: For an arbitrary vector ω, one has

$$\left\|\frac{\omega}{1+\omega^{\mathrm{T}}\omega}\right\| \le \frac{1}{2}, \quad \left\|\frac{1}{1+\omega^{\mathrm{T}}\omega}\right\| \le 1$$
$$\left\|\frac{\omega\omega^{\mathrm{T}}}{1+\omega^{\mathrm{T}}\omega}\right\| \le 1, \quad \left\|\frac{\omega\omega^{\mathrm{T}}}{(1+\omega^{\mathrm{T}}\omega)^2}\right\| \le \frac{1}{4}. \qquad \square$$

Theorem 3: Let u be any admissible control policy. Let the critic network (29) with an experience replay tuning law given by (36) be used to evaluate the given control policy. Suppose that the history stack satisfies Condition 1. Then, the critic weight estimation error $\tilde{W}_c$ converges exponentially to the residual set $R_s = \{\tilde{W}_c | \|\tilde{W}_c\| \le c\varepsilon_{\mathrm{Bmax}}\}$, where $\varepsilon_{\mathrm{Bmax}}$ is a bound for $\varepsilon_B(t)$ and c is a positive constant.

Proof: Given (35), the dynamics of $\tilde{W}_c$ can be expressed as

$$\begin{aligned}\dot{\tilde{W}}_c &= -\alpha_c\left[\frac{\sigma(t)\sigma^{\mathrm{T}}(t)}{\left[1+\sigma^{\mathrm{T}}(t)\sigma(t)\right]^2} + \sum_{k=1}^{N}\frac{\sigma_k\sigma_k^{\mathrm{T}}}{\left[1+\sigma_k^{\mathrm{T}}\sigma_k\right]^2}\right]\tilde{W}_c \\ &\quad + \alpha_c\left[\frac{\sigma(t)\varepsilon_B(t)}{\left[1+\sigma^{\mathrm{T}}(t)\sigma(t)\right]^2} + \sum_{k=1}^{N}\frac{\sigma_k\varepsilon_B(t_k)}{\left[1+\sigma_k^{\mathrm{T}}\sigma_k\right]^2}\right].\end{aligned} \tag{37}$$

Consider the Lyapunov function

$$V_c = \frac{1}{2}\tilde{W}_c^{\mathrm{T}}\alpha_1^{-1}\tilde{W}_c.$$

Differentiating V_c along the trajectories (37) yields

$$\dot{V}_c = -\tilde{W}_c^{\mathrm{T}}[\Gamma(t) + \Gamma_k]\tilde{W}_c + \tilde{W}_c^{\mathrm{T}}[\Psi(t) + \Psi_k]$$

where

$$\Gamma(t) = \frac{\sigma(t)\sigma^{\mathrm{T}}(t)}{\left[1+\sigma^{\mathrm{T}}(t)\sigma(t)\right]^2}, \quad \Gamma_k = \sum_{k=1}^{N}\frac{\sigma_k\sigma_k^{\mathrm{T}}}{\left[1+\sigma_k^{\mathrm{T}}\sigma_k\right]^2}$$
$$\Psi(t) = \frac{\sigma(t)\varepsilon_B(t)}{\left[1+\sigma^{\mathrm{T}}(t)\sigma(t)\right]^2}, \quad \Psi_k = \sum_{k=1}^{N}\frac{\sigma_k\varepsilon_B(t_k)}{\left[1+\sigma_k^{\mathrm{T}}\sigma_k\right]^2}.$$

Under Condition 1, then $\Gamma_k \succ 0$ can be guaranteed, and in addition, based on Fact 1, one has

$$\|\Psi(t) + \Psi_k\| \le \frac{N+1}{2}\varepsilon_{\mathrm{Bmax}}.$$

Therefore

$$\dot{V}_c \le -\lambda_{\min}(\Gamma_k)\|\tilde{W}_c\|^2 + \frac{N+1}{2}\varepsilon_{\mathrm{Bmax}}\|\tilde{W}_c\|.$$

Then, $\dot{V}_c$ is negative definite provided that

$$\|\tilde{W}_c\| \ge c\varepsilon_{\mathrm{Bmax}}$$

where $c = (N + 1/2\lambda_{\min}(\Gamma_k)) > 0$. This completes the proof. ■

C. Actor Intermittent Learning

In this section, two types of intermittent conditions for the actor learning using intermittent feedback are developed. We shall use an impulsive system formulation to analyze the closed-loop stability of the equilibrium point with the intermittent actor-critic learning.

Definition 5 (Impulsive System): The impulsive system can be described as

$$\begin{cases} \dot{\chi} = h_{\text{flow}}(\chi) & \forall t \in (t_k, t_{k+1}] \\ \chi^+ = h_{\text{jump}}(\chi), & t = t_k. \end{cases}$$

where $\chi \in \mathbb{R}^{n_\chi}$ is the state, $\chi^+ := \lim_{s \searrow t} \chi(s)$, $\{t_k\}_{k=0}^{\infty}$ is a monotonically increasing sequence of sampling instants with t_k being the kth consecutive sampling instant satisfying $\lim_{k\to\infty} t_k = \infty$. The functions h_{flow} and h_{jump} are the flow and the jump dynamics from $\mathbb{R}^{n_\chi}$ to $\mathbb{R}^{n_\chi}$, respectively. □

Given (30) and (31), we need to define the objective function for the actor as

$$\begin{aligned} E_a &= \frac{1}{2} e_a^{\mathrm{T}} R e_a \\ e_a &= u_c - u_a = \gamma[\tanh(D_a) - \tanh(D_c)] \end{aligned}$$

where e_a denotes the difference between u_a in (31) and u_c in (30). The actor learns in an intermittent fashion, i.e., the actor weight updates only at the triggering instants, and held constant otherwise.

Using a gradient descent algorithm for minimizing E_a yields

$$\begin{aligned} \Delta W_a(t_k) &= W_a^+ - W_a(t_k) \\ &= -\alpha_a[\nabla\phi_c G e_a + \nabla\phi_c G \tanh^2(D_a) e_a + Y W_a]. \end{aligned}$$

Considering the intermittent actor tuning law, one can write the update as an impulsive system

$$\begin{aligned} \dot{W}_a(t) &= 0 \quad \forall t \in \mathbb{R}^+ \Big\backslash \bigcup_{k\in\mathbb{N}} t_k \\ W_a^+ &= W_a - \alpha_a[\nabla\phi_c G e_a + \nabla\phi_c G \tanh^2(D_a) e_a + Y W_a] \\ &\quad \forall t = t_k \end{aligned} \tag{38}$$

with $W_a^+ = \lim_{\tau \searrow t_k} W_a(\tau)$. Accordingly, the dynamics of the actor weight error $\tilde{W}_a := W^\star - W_a$ is

$$\begin{aligned} \dot{\tilde{W}}_a(t) &= 0 \quad \forall t \in \mathbb{R}^+ \Big\backslash \bigcup_{k\in\mathbb{N}} t_k \\ \tilde{W}_a^+ &= \tilde{W}_a + \alpha_a\{\gamma\nabla\phi_c G[\tanh(D_a) - \tanh(D_c)] \\ &\quad + \tanh^2(D_a)\gamma\nabla\phi_c G e_a + Y W^\star - Y\tilde{W}_a\} \quad \forall t = t_k. \end{aligned} \tag{39}$$

Given that (31) is applied to the system (7) yields

$$\begin{aligned} \dot{W}_c &= -\alpha_c \left\{ \frac{\sigma_a(t) e_c^a(t)}{\left[1 + \sigma_a^{\mathrm{T}}(t)\sigma_a(t)\right]^2} + \sum_{k=1}^{N} \frac{\sigma_{ak} e_{ck}^a}{\left[1 + \sigma_{ak}^{\mathrm{T}}\sigma_{ak}\right]^2} \right\} \\ e_c^a(t) &= U_a(t) + W_c^{\mathrm{T}}(t)\sigma_a(t) \\ e_{ck}^a &= U_{ak} + W_c^{\mathrm{T}}(t)\sigma_{ak} \end{aligned} \tag{40}$$

where

$$\begin{aligned} \sigma_a(t) &= \nabla\phi_c[F(s(t)) + G(s(t))u_a(t)], \quad \sigma_{ak} = \sigma_a(t_k) \\ U_a(t) &= U(s(t), \quad u_a(t)), \quad U_{ak} = U_a(t_k). \end{aligned}$$

Based on (33) and (35), one has

$$\begin{aligned} e_c^a &= \varepsilon_B^a - \tilde{W}_c^{\mathrm{T}}\sigma_a, \\ \varepsilon_B^a &= -[\nabla\varepsilon(s)]^{\mathrm{T}}[F(s) + G(s)u_a]. \end{aligned} \tag{41}$$

From (41) and (40), the dynamics of the critic error $\tilde{W}_c$ can be expressed as

$$\begin{aligned} \tilde{W}_c^+ &= 0 \quad \forall t = t_k, \\ \dot{\tilde{W}}_c &= -\alpha_c[\Gamma_a(t) + \Gamma_{ak}]\tilde{W}_c + \alpha_c[\Psi_a(t) + \Psi_{ak}] \\ &\quad \forall t \in \mathbb{R}^+ \Big\backslash \bigcup_{k\in\mathbb{N}} t_k \end{aligned} \tag{42}$$

where

$$\begin{aligned} \Gamma_a(t) &= \frac{\sigma_a\sigma_a^{\mathrm{T}}}{\left[1 + \sigma_a^{\mathrm{T}}\sigma_a\right]^2}, \quad \Gamma_{ak} = \sum_{k=1}^{N} \frac{\sigma_{ak}\sigma_{ak}^{\mathrm{T}}}{\left[1 + \sigma_{ak}^{\mathrm{T}}\sigma_{ak}\right]^2} \\ \Psi_a(t) &= \frac{\sigma_a\varepsilon_B}{\left[1 + \sigma_a^{\mathrm{T}}\sigma_a\right]^2}, \quad \Psi_{ak} = \sum_{k=1}^{N} \frac{\sigma_{ak}\varepsilon_B(t_k)}{\left[1 + \sigma_{ak}^{\mathrm{T}}\sigma_{ak}\right]^2}. \end{aligned}$$

Due to the unmatched parameterization of the value function, most existing RL-based adaptive optimal learning algorithms only guarantee the uniformly ultimately boundedness of the state and the actor-critic weights [31], [47]. In order to guarantee asymptotical stability of the equilibrium point of the closed-loop system, an additional robustifying term is added into the actor network to improve control performance, that is,

$$\bar{u}_a = u_a + \delta, \quad \delta = -B\|s\|^2 \frac{1}{(A + s^{\mathrm{T}}s)} \tag{43}$$

where u_a is defined in (31), A and B are positive design parameters.

The closed-loop system dynamics can now be written as

$$\begin{aligned} \dot{s} &= F(s) - G(s)\left[\gamma \tanh\left(\frac{1}{2\gamma r} G^{\mathrm{T}}(\hat{s})[\nabla\phi_c(\hat{s})]^{\mathrm{T}} W_a\right) + \delta\right] \\ &\quad \forall t \in \mathbb{R}^+ \Big\backslash \bigcup_{k\in\mathbb{N}} t_k \\ \hat{s}^+ &= \hat{s} + e, \quad t = t_{\mathrm{k}}. \end{aligned} \tag{44}$$

Finally, by combining (39), (42), and (44), the dynamics of the system with an augmented state as $\chi := \left[\, s^{\mathrm{T}} \;\; \hat{s}^{\mathrm{T}} \;\; \tilde{W}_c^{\mathrm{T}} \;\; \tilde{W}_a^{\mathrm{T}} \,\right]^{\mathrm{T}}$ can be expressed in terms of the impulsive system as shown in (45), as shown at the bottom of the next page.

The following theorem provides the stability analysis of the augmented system (45) for the intermittent safe RL algorithm with the actor-critic-barrier structure. Before moving on, we define notations in (46), as shown at the bottom of the next page, for theoretical discussions.

Theorem 4 (Safe RL With Static Intermittent Feedback): Suppose that Assumptions 1–5 and Condition 1 hold. Consider the dynamical system (1) with the control input (43). Let the control input applied to the system be represented as u_a in (31) with the gradient-descent-based actor-critic learning given by (38) and (40). Denote $\bar{\Omega} := \{\Omega_s \times \Omega_{\hat{s}} \times \Omega_{\tilde{W}_c} \times \Omega_{\tilde{W}_a}\}$. Then, the following hold.

1) The equilibrium point of the closed-loop system with augmented state χ is asymptotically stable for $\chi_0 \in \bar{\Omega}$ provided that the event instants $\{t_k\}_{k=0}^{\infty}$ are determined by

$$\|e\|^2 \le \frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}\|s\|^2 + \frac{1}{L_e}\Theta(u_a(\hat{s})) \tag{47}$$

with the parameters α_a, Y, r_a, and r_c satisfying

$$q_a - \frac{r_a}{2} > 0, \quad \lambda_{\min}(\Gamma_{ak}) - \frac{r_c}{2} > 0 \tag{48}$$

and the robustfying term given in (43) satisfying

$$B > \frac{(\rho_s + \rho_a + \rho_c)(A + \|s\|^2)}{\bar{V}_d b_G \|s\|^2}. \tag{49}$$

2) Zeno behavior is excluded.

Proof: See the Appendix. ■

Remark 5: The intermittent instants can be expressed as

$$t_0 = 0, \ t_{k+1} = \inf_{t \in \mathbb{R}^+} \{t > t_k \wedge q \leq 0\} \tag{50}$$

with

$$q := \frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}\|s\|^2 + \frac{1}{L_e}\Theta(u_a(\hat{s})) - \|e\|^2. \tag{51}$$

□

To introduce a dynamic intermittent feedback, an additional internal dynamical system is provided

$$\dot{\varsigma} = -\Xi\varsigma + q, \quad \varsigma(t_0) = \varsigma_0, \ t \in \mathbb{R}_0^+ \tag{52}$$

where q is defined in (51) and $\Xi \in \mathbb{R}^+$ is a parameter to be designed later. An event is triggered when the following condition is satisfied:

$$\varsigma(t) + \phi q(t) \leq 0, \quad t \geq 0 \tag{53}$$

where $\phi \in \mathbb{R}^+$ is a parameter to be designed later. The intermittent instants are determined by the following rule:

$$\begin{aligned} t_0 &= 0 \\ t_{k+1} &= \inf_{t \in \mathbb{R}^+} \{(t > t_k) \wedge (\varsigma(t) + \phi q(t) \leq 0)\}. \end{aligned} \tag{54}$$

Lemma 3 [37]: There exists constants A_e and B_e such that the dynamics of $s(t)$ satisfies

$$\|\dot{s}\| \leq A_e\|s\| + B_e\|e\|. \quad ■$$

To this end, the co-design of the dynamic triggering condition and feedback gain based on intermittent RL can be formulated in the next theorem.

Theorem 5 (Safe RL With Dynamic Intermittent Feedback): Under Assumptions 1–5 and Condition 1, consider the dynamical system (1) with the control input (43). Let the control input applied to the system be represented as u_a in (31) with the gradient-descent-based actor-critic learning as given in (38) and (40). Then, the following hold.

1) The equilibrium point of the closed-loop system with the augmented state χ is asymptotically stable for $\chi_0 \in \bar{\Omega}$ provided that the event instants $\{t_k\}_{k=0}^{\infty}$ is determined as (54).
2) Zeno behavior is excluded.

Proof: See the Appendix. ■

Corollary 1: Let $\{\bar{t}_k^s\}_{k=0}^{\infty}$ and $\{\bar{t}_k^d\}_{k=0}^{\infty}$ be the triggering time sequences determined by the static and dynamic intermittent RL as designed in Theorems 4 and 5, respectively. Assume also that $\bar{t}_i^s = \bar{t}_j^d$. Then, by denoting the next triggering instants by the static and the dynamic intermittent RL as $\bar{t}_{i+1}^s$ and $\bar{t}_{j+1}^d$, respectively, one has $\bar{t}_{j+1}^d \geq \bar{t}_{i+1}^s$.

Proof: The proof follows from Theorem 2. ■

Remark 6: The static and dynamic event-triggering conditions (50) and (54) contain the parameters β, Ξ and ϕ. As $\phi \to \infty$ the dynamic intermittent feedback becomes the static case. The parameter Ξ act as the time constant of the filter (52), which can not be too fast compared with the time constant of the signal q. Also, one can reduce the event-triggering frequency by selecting β close to 0. For more details of the effect of parameters β, Ξ and ϕ on the static and dynamic intermittent feedback designs, readers are referred to [38], [41], and [48]. □

The online safe RL algorithm is shown in Fig. 1. The learning framework consists of a barrier-function-based system transformation and an actor-critic online learning structure. In contrast to the offline iterative algorithms, both the critic and the actor updates are performed simultaneously in real-time. The barrier function based system transformation

$$\dot{\chi} = \begin{bmatrix} F(s) - G(s)\left[\gamma \tanh\left(\frac{1}{2\gamma r}G^{\mathrm{T}}(\hat{s})[\nabla\phi_c(\hat{s})]^{\mathrm{T}}(W^\star - \tilde{W}_a)\right) + \delta\right] \\ 0 \\ -\alpha_c[\Gamma_a(t) + \Gamma_{ak}]\tilde{W}_c + \alpha_c[\Psi_a(t) + \Psi_{ak}] \\ 0 \end{bmatrix}$$

$$\chi^+ = \chi(t) + \begin{bmatrix} 0 \\ e \\ 0 \\ \alpha_a\{\gamma\nabla\phi_c G[\tanh(D_a) - \tanh(D_c)] + \tanh^2(D_a)\gamma\nabla\phi_c G e_a + YW^\star - Y\tilde{W}_a\} \end{bmatrix} \tag{45}$$

$$\rho_a = \frac{1}{2r_a}m_a^2 + n_a, \quad \rho_s = \gamma^2\bar{D}, \quad \rho_c = \frac{(N+1)^2}{8r_c}\varepsilon_{\mathrm{Bmax}}^2, \ q_a = \lambda_{\min}(2Y - \alpha_a Y^{\mathrm{T}}Y)$$

$$m_a = 2[2\gamma\phi_{\mathrm{cdmax}}b_G + 2\gamma^2\phi_{\mathrm{cdmax}}b_G + 2\alpha_a\|Y\|\gamma^2\phi_{\mathrm{cdmax}}b_G + \|Y\|W_{\max} + \alpha_a\|Y^{\mathrm{T}}Y\|W_{\max} + 2\gamma\alpha_a\|Y\|\phi_{\mathrm{cdmax}}b_G]$$

$$n_a = \alpha_a(2\gamma\phi_{\mathrm{cdmax}}b_G)^2 + \alpha_a(2\gamma^2\phi_{\mathrm{cdmax}}b_G + \|Y\|W_{\max})^2 + 2\alpha_a(2\gamma\phi_{\mathrm{cdmax}}b_G)(2\gamma^2\phi_{\mathrm{cdmax}}b_G + \|Y\|W_{\max})$$

$$\bar{D} = \frac{1}{4\gamma^2 r}\left(2\tilde{W}_{\mathrm{admax}}\phi_{\mathrm{cdmax}}b_G^2\varepsilon_{\mathrm{cdmax}} + b_G^2\phi_{\mathrm{cdmax}}^2\tilde{W}_{\mathrm{admax}}^2 + \varepsilon_{\mathrm{cdmax}}^2 b_G^2\right), \quad \bar{V}_d = (W_{\max}\phi_{\mathrm{cdmax}} + \varepsilon_{\mathrm{cdmax}})$$

$$\Omega_s = \left\{s \,\middle|\, \|s\| \leq \sqrt{\frac{\rho_s}{\lambda_{\min}(Q)\beta^2}}\right\}, \quad \Omega_{\tilde{W}_c} = \left\{\tilde{W}_c \,\middle|\, \|\tilde{W}_c\| \leq \sqrt{\frac{\rho_c}{\lambda_{\min}(\Gamma_{ak}) - \frac{r_c}{2}}}\right\}, \ \Omega_{\tilde{W}_a} = \left\{\tilde{W}_a \,\middle|\, \|\tilde{W}_a\| \leq \sqrt{\frac{\rho_a}{q_a - \frac{r_a}{2}}}\right\} \tag{46}$$

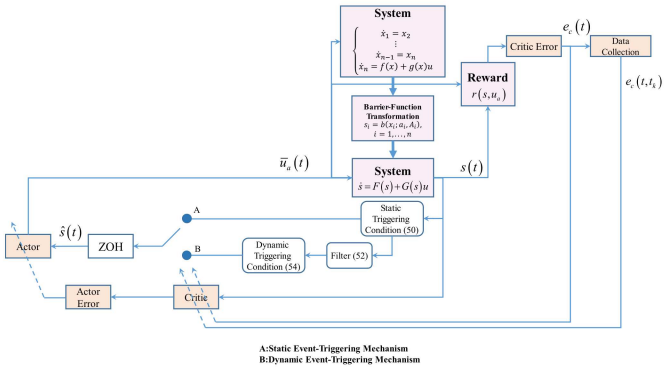


Fig. 1. Intermittent online actor-critic-barrier safe RL framework.

Algorithm 1 Actor-Critic-Barrier Online Learning Algorithm

Require: Begin with initial state $x(0)$ and initial actor-critic weights $W_c(0)$ and $W_a(0)$. Set $i = 0$ and $t_i = 0$ then propagate the time using ordinary differential equation solver such as the Runge Kutta method with the time step increment h;

1: (*Barrier-function-based system transformation*): Transform the system (1) with state x to the equivalent system (7) with state s;
2: (*Penalty function design*): Design the control input penalty function $\Theta(\cdot)$ for the performance index(8);
3: (*Robustifying the actor*): Apply the actor with the robust term δ (43) to system (7);
4: (*System evolution*): Update the impulsive system (44);
5: (*Actor-critic learning*): Update the actor-critic networks according to (40) and (38);
6: (*Event instant determination*): Determine the instant for sampling update using static or dynamic event-triggering condition ((50) or (54));
7: (*Online data collection*): Collect the online data to store the term σ_k and e_{ck} for the critic learning (40) until Condition 1 is satisfied;
8: Set $t_{i+1} = t_i + h$ and go to Step 3.

is used to tackle the full-state constraints. To obviate the PE condition, the online data is collected until Condition 1 is satisfied. Options A and B in Fig. 1 represent the static and dynamic event-triggering conditions. It can be seen that the dynamic event-triggering condition can be viewed as a filtered version of the static event-triggering condition. Finally, the online actor-critic-barrier RL algorithm can be summarized in Algorithm 1.

V. Simulations

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

$$\dot{x} = \begin{bmatrix} x_2 \\ -x_1 + 0.5\big(1 - x_2^2\big)x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ x_1 \end{bmatrix} u, \quad t \geq 0.$$

The constraints on the input control and the state are

$$x_1 \in (-0.6, 0.2), \quad x_2 \in (-0.2, 0.2); \; u \in (0, 1).$$

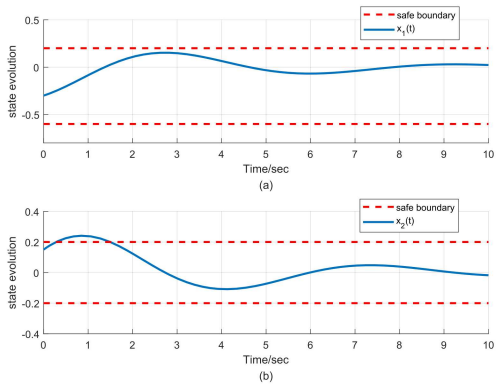


Fig. 2. Evolution of the state trajectories by using a converse HJB approach without constraints. The evolution of the state (a) $x_1(t)$ and (b) $x_2(t)$.

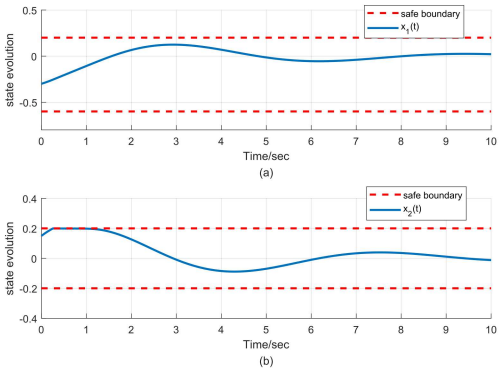


Fig. 3. Evolution of the state trajectories given our safe static intermittent RL algorithm. The evolution of the state (a) $x_1(t)$ and (b) $x_2(t)$.

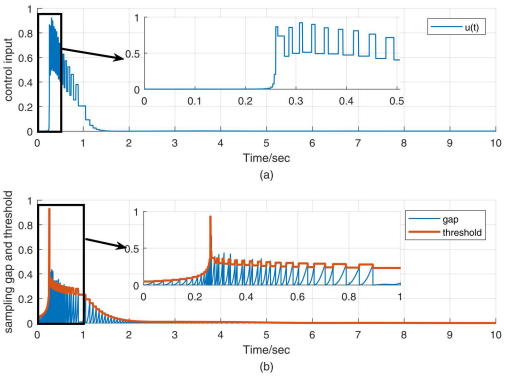


Fig. 4. Evolution of the control input signal, the threshold and the gap signal given our safe static intermittent RL algorithm. (a) Control input signal. (b) Threshold and the gap signal.

A. Case 1: Converse HJB Method Without Safety Constraints

According to the converse HJB method [49], given that the performance parameters are selected as $Q = I_{2\times 2}$ and $R = 1$, the optimal controller is $u^\star(x) = -x_1 x_2$. Fig. 2 shows with solid lines the state evolutions and with dashed lines the bounds where one can see that the constraints are violated.

B. Case 2: Proposed Solution

Next, we apply the intermittent static feedback to the safe-critical system to the safe-critical system. The state evolution is shown in Fig. 3. The control input signal is shown in Fig. 4(a) and the evolution of the gap signal is shown

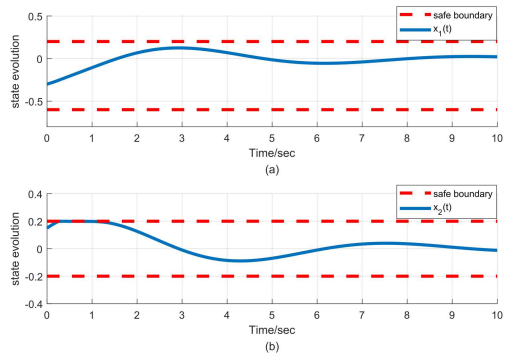


Fig. 5. Evolution of the state trajectories with our safe dynamic intermittent RL algorithm. The evolution of the state (a) $x_1(t)$ and (b) $x_2(t)$.

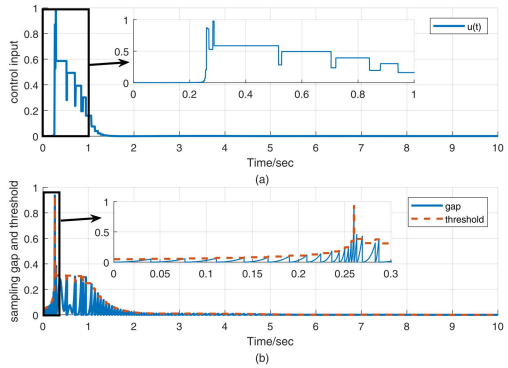


Fig. 6. Evolution of the control input signal and the gap for our safe dynamic intermittent RL. (a) Control input signal. (b) Threshold and the gap signal.

in Fig. 4(b). In contrast to the previous case, one can observe that the state approaches the origin without violating the input and state constraints.

We now apply the intermittent dynamic feedback to the safe-critical system, where the corresponding results are given in Figs. 5 and 6. The state constraints, the input saturation and the closed-loop stability of the equilibrium point are guaranteed. In addition, the comparison between Figs. 4 and 5 can tell that the dynamic intermittent feedback design could further reduce the sampling update.

VI. Conclusion

This article presents a novel safe RL algorithm using intermittent static and dynamic feedback. A barrier function transformation is used to transform the system to deal with the full-state and input constraints. Then, based on an actor-critic structure, a novel, safe RL algorithm is developed to find the optimal safe controller in an online fashion for both cases. In contrast to the persistent excitation condition, experience replay technique is used in a way that recorded history data is utilized together with the current data. Finally, simulation results show the efficacy of the proposed framework.

Future work will focus on extending the results to multi-agent systems with cooperative and noncooperative agents.

Appendix
Proof of Theorem 1

Consider the optimal value function $V^\star(s)$ as a Lyapunov function candidate. First, differentiating $V^\star(s)$ along the trajectories yields

$$\dot{V}^\star(s) = \left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} F(s) + \left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)u^\star(\hat{s}).$$

By inserting the HJB in $\dot{V}^\star(s)$, one has

$$\dot{V}^\star(s) = -Q(s) - \gamma^2 r \ln[1-\tanh^2(D^\star)] + \left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)u^\star(\hat{s}). \quad (55)$$

From (15), the term $-\gamma^2 r \ln[1-\tanh^2(D^\star)]$ in (55) can be rewritten as

$$-\gamma^2 r \ln[1-\tanh^2(D^\star)] = -\Theta(u^\star(s)) + \gamma\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)\tanh(D^\star). \quad (56)$$

By using (11), we can rewrite (55) as

$$-\gamma^2 r \ln[1-\tanh^2(D^\star(s))] = -\Theta(u^\star(\hat{s})) - \int_{u^\star(\hat{s})}^{u^\star(s)} 2r\gamma\left[\tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv + \gamma\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)\tanh(D^\star(s)). \quad (57)$$

By considering (14), one has

$$r\int_{u^\star(s)}^{u^\star(\hat{s})} [D^\star(s)]dv = rD^\star(s)[u^\star(\hat{s}) - u^\star(s)]. \quad (58)$$

By considering (58), the term $[(\partial V^\star(s)/\partial s)]^{\mathrm{T}} G(s)u^\star(\hat{s})$ in (55) can be equivalently expressed as

$$\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)u^\star(\hat{s}) = -\gamma\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)\tanh(D^\star) + 2r\gamma\int_{u^\star(s)}^{u^\star(\hat{s})} (D^\star)dv. \quad (59)$$

Collecting the results in (57) and (59), one can obtain

$$\dot{V}^\star = -Q(s) - \Theta(u^\star(\hat{s})) + 2r\gamma\int_{u^\star(s)}^{u^\star(\hat{s})}\left[D^\star + \tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv. \quad (60)$$

Denote $w := -\tanh^{-1}((v/\gamma))$, then one has

$$v = -\gamma\tanh(w), \quad dv = -\gamma[1-\tanh^2(w)]dw.$$

The integral in (60) satisfies

$$2r\gamma\int_{u^\star(s)}^{u^\star(\hat{s})}\left[D^\star + \tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv \le 2r\gamma^2\int_{D^\star(s)}^{D^\star(\hat{s})}[w - D^\star]dw \le r\gamma^2 L_D^2\|e\|^2. \quad (61)$$

Inserting (61) into (60) yields

$$\dot{V}^\star(s) \le -s^{\mathrm{T}}Qs - \Theta(u^\star(\hat{s})) + r\gamma^2 L_D^2 e^{\mathrm{T}}e \le -\beta^2 s^{\mathrm{T}}Qs$$

given that the following condition is satisfied

$$r\gamma^2 L_D^2\|e\|^2 \le (1-\beta^2)\lambda_{\min}(Q)\|s\|^2 + \Theta(u^\star(\hat{s})).$$

According to Assumption 4, one has $s(t) \neq 0$ at the event instants. Thus, $e(t) = 0$ and $(\|e(t)\|/\|s(t)\|) = 0$ when $t = t_k$. At the event instant $t = t_k$, the triggering condition (21) is violated, that is,

$$\begin{aligned}\|e\|^2 &\geq \frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}\|s\|^2 + \frac{1}{L_e}\Theta(u^\star(\hat{s})) \\ &\geq \frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}\|s\|^2\end{aligned}$$

which is equivalent to

$$\|e\| \geq \sqrt{\frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}}\|s\|. \tag{62}$$

That is, $(\|e(t)\|/\|s(t)\|)$ evolves from 0 to $(((1-\beta^2)\lambda_{\min}(Q)/L_e))^{1/2}$ between two successive event instants. Then, the time that $(\|e(t)\|/\|s(t)\|)$ evolves from 0 to $(((1-\beta^2)\lambda_{\min}(Q)/L_e))^{1/2}$ provides a lower bound on the interevent time.

Next, we analyze the dynamics of $(\|e(t)\|/\|s(t)\|)$. Note that $\tanh(z) = 1 - z^2$. Then, the Lipschitz constant of the function $\tanh(\cdot)$, denoted as $L_{\tanh}$, is not greater than one. Therefore, one has

$$\begin{aligned}&\|u^\star(s+e)\| - \|u^\star(s)\| \\ &\quad\leq \gamma\|\tanh(D^\star(s+e) - \tanh(D^\star(s)))\| \\ &\quad\leq \gamma L_{\tanh} L_D\|e\| \leq \gamma L_D\|e\|.\end{aligned}$$

The control input $u^\star(s+e)$ satisfies

$$\|u^\star(s+e)\| \leq \gamma L_D(\|s\| + \|e\|).$$

By applying the intermittent feedback (20), in (19) one has

$$\|\dot{s}\| \leq (b_F + b_G\gamma L_D)(\|s\| + \|e\|).$$

Then, $(\|e\|/\|x\|)$ satisfies, for $\forall t \in (t_k, t_{k+1}]$ and $\forall k \in \mathbb{Z}^+$

$$\frac{d}{dt}\left(\frac{\|e\|}{\|s\|}\right) \leq (b_F + b_G\gamma L_D)\left(1 + \frac{\|e\|}{\|s\|}\right)^2.$$

Based on the comparison lemma from [45], one has

$$\frac{\|e\|}{\|x\|} \leq \frac{(t-t_{\rm k})\,(b_F + b_G\gamma L_D)}{1-(t-t_{\rm k})\,(b_F + b_G\gamma L_D)} \qquad \forall t \in (t_k, t_{k+1}],\quad k \in \mathbb{Z}^+.$$

Evaluating (63) at $t = t_{k+1}$ and combining with (62) yields

$$\begin{aligned}\sqrt{\frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}} &\leq \frac{\|e(t)\|}{\|s(t)\|} \\ &\leq \frac{(t_{\rm k+1}-t_{\rm k})(b_F + b_G\gamma L_D)}{1-(t_{\rm k+1}-t_{\rm k})(b_F + b_G\gamma L_D)} \\ &\qquad \forall k \in \mathbb{Z}^+\end{aligned}$$

which further results in, for $\forall k \in \mathbb{Z}^+$

$$t_{k+1} - t_k \geq \frac{\frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}}{(b_F + b_G\gamma L_D)\left(1 + \sqrt{\frac{(1-\beta^2)\lambda_{\min}(Q)}{L_e}}\right)}. \tag{63}$$

Therefore, Zeno behavior is guaranteed to be excluded. This completes the proof. ■

PROOF OF THEOREM 4

Given that a policy u_a is applied to the system (7), we shall use the following Lyapunov equation:

$$\mathcal{V} = V^\star(s) + V^\star(\hat{s}) + \underbrace{\frac{1}{2}\tilde{W}_c^{\rm T}\alpha_1^{-1}\tilde{W}_c}_{V_c} + \underbrace{\frac{1}{2}\tilde{W}_a^{\rm T}\alpha_2^{-1}\tilde{W}_a}_{V_a}$$

where $V^\star(s)$ is the optimal value function, $\tilde{W}_c$ and $\tilde{W}_a$ are the critic and actor weight errors, respectively. In the following, we shall analyze the stability of the augmented system (45).

First, based on the augment system dynamics (45), one has $\Delta V^\star(s) = \Delta V_c(\tilde{W}_c) = 0$. As shown later, since the state s is asymptotically stable, then, $V^\star(\hat{s}^+) \leq V^\star(\hat{s})$ and there exists a class-$\mathcal{K}$ function $\kappa(\cdot)$ such that $\Delta V^\star(\hat{s}) = V^\star(\hat{s}^+) - V^\star(\hat{s}) \leq -\kappa(\hat{s})$.

Consider the actor weight error dynamics in (39), the difference of the Lyapunov function $V_a(\tilde{W}_a)$ can be written as in (64), as shown at the bottom of this page. Based on Assumption 5, $\Delta V_a(\tilde{W}_a)$ can be upper bounded as

$$\Delta V_a(\tilde{W}_a) \leq -q_a\|\tilde{W}_a\|^2 + m_a\|\tilde{W}_a\| + n_a \tag{65}$$

where are q_a, m_a, and n_a are defined in (46). Based on the parameter design (48), $q_a > 0$. Note that from (65), $\Delta V_a(\tilde{W}_a)$ can be further upper bounded as

$$\begin{aligned}\Delta V_a(\tilde{W}_a) &\leq -q_a\|\tilde{W}_a\|^2 + \frac{r_a}{2}\|\tilde{W}_a\|^2 + \frac{1}{2r_a}m_a^2 + n_a \\ &= -\left(q_a - \frac{r_a}{2}\right)\|\tilde{W}_a\|^2 + \rho_a\end{aligned} \tag{66}$$

where $\rho_a = (1/2r_a)m_a^2 + n_a$ and r_a is selected such that the condition (48) is satisfied. Hence, $\Delta V_a(\tilde{W}_a) \leq 0$ if the actor network estimation error satisfies $\|\tilde{W}_a\| \geq ((\rho_a/q_a - (r_a/2)))^{1/2}$. Thus, $\tilde{W}_a$ converges to the residual set $\Omega_{\tilde{W}_a}$, which is defined in (46).

Differentiating $\mathcal{V}$ yields

$$\dot{\mathcal{V}} = \dot{V}^\star(s) + \dot{V}^\star(\hat{s}) + \dot{V}_c(\tilde{W}_c) + \dot{V}_a(\tilde{W}_a).$$

From (45), one has

$$\dot{V}^\star(\hat{s}) = \dot{V}_a(\tilde{W}_a) = 0. \tag{67}$$

$$\begin{aligned}\Delta V_a(\tilde{W}_a) = {}& 2\tilde{W}_a^{\rm T}\{\gamma\nabla\phi G[\tanh(D_a) - \tanh(D_c)] + \tanh^2(D_a)\gamma\nabla\phi G e_a + YW^\star \\ &+ \alpha_a Y^{\rm T}[\tanh^2(D_a)\gamma\nabla\phi G e_a + YW^\star] + \alpha_a Y^{\rm T}\gamma\nabla\phi G[\tanh(D_a) - \tanh(D_c)]\} \\ &+ \alpha_a\|\gamma\nabla\phi G[\tanh(D_a) - \tanh(D_c)]\|^2 + \alpha_a\|\tanh^2(D_a)\gamma\nabla\phi G e_a + YW^\star\|^2 \\ &+ 2\alpha_a\{\gamma\nabla\phi G[\tanh(D_a) - \tanh(D_c)]\}^{\rm T}\{\tanh^2(D_a)\gamma\nabla\phi G e_a + YW^\star\} - 2\tilde{W}_a^{\rm T}Y\tilde{W}_a + \alpha_a\tilde{W}_a^{\rm T}Y^{\rm T}Y\tilde{W}_a.\end{aligned} \tag{64}$$

First, we analyze the stability of the flow dynamics $s(t)$. Based on (56) and (57), one has

$$\begin{aligned}&\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} F(s)\\&\quad=\gamma\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)\tanh(D^\star(s))\\&\qquad-\int_{u^\star(\hat{s})}^{u^\star(s)} 2r\gamma\left[\tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv-Q(s)-\Theta(u^\star(\hat{s})).\end{aligned}\tag{68}$$

One can then obtain

$$\begin{aligned}&\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)\,u_a\\&\quad=-\gamma\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}} G(s)\tanh(D^\star)+2r\gamma\int_{u^\star(\hat{s})}^{u_a(\hat{s})}(D^\star)dv\\&\qquad+2r\gamma\int_{u^\star(s)}^{u^\star(\hat{s})}(D^\star)dv.\end{aligned}\tag{69}$$

Therefore, $\dot{V}^\star(s)$ can be expressed as in (70), as shown at the bottom of this page, where the last inequality results from (43) and Assumption 5. From (61), one can obtain

$$\begin{aligned}2r\gamma\int_{u^\star(\hat{s})}^{u_a(\hat{s})}&\left[D^\star(s)+\tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv\\&\le r\gamma^2[D_a(\hat{s})-D^\star(\hat{s})]^2.\end{aligned}\tag{71}$$

From the previous stability analysis for the jump dynamics, it can be inferred that the actor weight error is bounded, i.e., there exists a constant $\tilde{W}_{\mathrm{admax}}$ such that $\|\tilde{W}_a\|\le\tilde{W}_{\mathrm{admax}}$. Then, from Assumption 5, one can obtain

$$[D_a(\hat{s})-D^\star(\hat{s})]^{\mathrm{T}}R[D_a(\hat{s})-D^\star(\hat{s})]\le\bar{D}.\tag{72}$$

From (61) and (70)–(72), $V^\star(s)$ can be upper bounded as

$$\begin{aligned}\dot{V}^\star(s)&\le-Q(s)-\Theta(u_a(\hat{s}))+r\gamma^2L_D^2\|e\|^2\\&\quad+\gamma^2\bar{D}-\bar{V}_d b_G B\|s\|^2\frac{1}{A+s^{\mathrm{T}}s}\\&=-Q(s)-\Theta(u_a(\hat{s}))+r\gamma^2L_D^2\|e\|^2+\rho_s\\&\quad-\bar{V}_d b_G B\|s\|^2\frac{1}{A+s^{\mathrm{T}}s}\\&\le-\beta^2 s^{\mathrm{T}}Qs+\rho_s-\bar{V}_d b_G B\|s\|^2\frac{1}{A+s^{\mathrm{T}}s}\end{aligned}\tag{73}$$

where $\rho_s=\gamma^2\bar{D}$ and the last inequality is guaranteed by the intermittent condition (47). Then, without the robustifying term δ, $\dot{V}^\star(s)<0$ if $\|s\|>((\rho_s/\lambda_{\min}(Q)\beta^2))^{1/2}$, i.e., s converges to the residual set Ω_s, which is defined in (46). When s is outside the set Ω_s, as shown later, the robustifying term δ will be used to guarantee the closed-loop asymptotic stability.

Second, for the critic weight error flow dynamics (42), differentiating $V_c(\tilde{W}_c)$ yields

$$\begin{aligned}\dot{V}_c(\tilde{W}_c)&=-\tilde{W}_c^{\mathrm{T}}[\Gamma_a(t)+\Gamma_{ak}]\tilde{W}_c+\tilde{W}_c^{\mathrm{T}}[\Psi_a(t)+\Psi_{ak}]\\&\le-\lambda_{\min}(\Gamma_{ak})\|\tilde{W}_c\|^2+\frac{N+1}{2}\varepsilon_{\mathrm{Bmax}}\|\tilde{W}_c\|\\&\le-\left[\lambda_{\min}(\Gamma_{ak})-\frac{r_c}{2}\right]\|\tilde{W}_c\|^2+\rho_c\end{aligned}\tag{74}$$

where $\rho_c=((N+1)^2/8r_c)\varepsilon_{\mathrm{Bmax}}^2$ and r_c is selected such that the condition (48) is satisfied. Then, from (74), one can infer that $\dot{V}_c<0$ if $\|\tilde{W}_c\|\ge((\rho_c/\lambda_{\min}(\Gamma_{ak})-\frac{r_c}{2}))^{1/2}$. Thus, $\tilde{W}_c$ converges to the residual set $\Omega_{\tilde{W}_c}$, which is defined in (46).

Hence, from (67), (73), and (74), the derivative of $\mathcal{V}$ can be expressed as

$$\begin{aligned}\dot{\mathcal{V}}&=\dot{V}^\star(s)+\dot{V}^\star(\hat{s})+\dot{V}_a(\tilde{W}_a)+\dot{V}_c(\tilde{W}_c)\\&\le-\beta^2 s^{\mathrm{T}}Qs-\left[\lambda_{\min}(\Gamma_{ak})-\frac{r_c}{2}\right]\|\tilde{W}_c\|^2\\&\quad+\rho_c+\rho_s-\bar{V}_d b_G B\|s\|^2\frac{1}{A+s^{\mathrm{T}}s}.\end{aligned}\tag{75}$$

Finally, consider the facts in (66) and (75), the closed-loop augmented system (45) is asymptotically stable provided the robustifying term δ is designed to satisfy the condition (49). This completes the proof.

The interevent time can be shown to be lower bounded by a positive constant following the proof of Theorem 1. This completes the proof. ■

Proof of Theorem 5

Consider the augmented system with dynamics (45), and let $\mathcal{W}:=\mathcal{V}+L_e\varsigma$ be a Lyapunov candidate. Then, based on (67), (73), and (74), the time derivative of $\mathcal{W}(t)$ can be bounded as

$$\begin{aligned}\dot{\mathcal{W}}&\le-Q(s)-\left[\lambda_{\min}(\Gamma_{ak})-\frac{r_c}{2}\right]\|\tilde{W}_c\|^2-\Theta(u_a(\hat{s}))\\&\quad+L_e\|e\|^2+L_e(-\Xi\varsigma+q)\\&=-Q(s)-\left[\lambda_{\min}(\Gamma_{ak})-\frac{r_c}{2}\right]\|\tilde{W}_c\|^2-L_e\Xi\varsigma\\&\quad+(1-\beta^2)\lambda_{\min}(Q)\|s\|^2\\&\le-\beta^2\lambda_{\min}(Q)\|s\|^2-\left[\lambda_{\min}(\Gamma_{ak})-\frac{r_c}{2}\right]\|\tilde{W}_c\|^2-L_e\Xi\varsigma.\end{aligned}$$

$$\begin{aligned}\dot{V}^\star(s)&=-Q(s)-\Theta(u_a(\hat{s}))+2r\gamma\int_{u^\star(s)}^{u^\star(\hat{s})}[D^\star(s)+\tanh^{-1}\left(\frac{v}{\gamma}\right)]dv+2r\gamma\int_{u^\star(\hat{s})}^{u_a(\hat{s})}[D^\star(s)+\tanh^{-1}(\frac{v}{\gamma})]dv\\&\quad-\left[\frac{\partial V^\star(s)}{\partial s}\right]^{\mathrm{T}}G(s)B\|s\|^2\frac{1}{(A+s^{\mathrm{T}}s)}\\&\le-Q(s)-\Theta(u_a(\hat{s}))+2r\gamma\int_{u^\star(s)}^{u^\star(\hat{s})}\left[D^\star(s)+\tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv+2r\gamma\int_{u^\star(\hat{s})}^{u_a(\hat{s})}\left[D^\star(s)+\tanh^{-1}\left(\frac{v}{\gamma}\right)\right]dv\\&\quad-\bar{V}_d b_G B\|s\|^2\frac{1}{A+s^{\mathrm{T}}s}\end{aligned}\tag{70}$$

$$\begin{aligned}\dot{\xi} &\le \frac{\sqrt{a\phi}}{\sqrt{\varpi+b\phi\|s\|^2}}(A_e\|s\|+B_e\|e\|)+\frac{\sqrt{a\phi}\|e\|}{2(\varpi+b\phi\|s\|^2)^{\frac{3}{2}}}[\Xi\varpi-b\|s\|^2+a\|e\|^2+2b\phi A_e\|s\|^2+2b\phi B_e\|e\|\|s\|]\\ &\le A_{\rm e}\sqrt{\frac{a}{b}}+B_{\rm e}\xi+B_{\rm e}\sqrt{\frac{b}{a}}\xi^2+\frac{1}{2\phi}\xi^3+\frac{\sqrt{a\phi}\|e\|}{2(\varpi+b\phi\|x\|^2)^{\frac{3}{2}}}(\Xi\varpi-b\|s\|^2+2b\phi A_{\rm e}\|s\|^2)\\ &\le A_{\rm e}\sqrt{\frac{a}{b}}+B_{\rm e}\xi+B_{\rm e}\sqrt{\frac{b}{a}}\xi^2+\frac{1}{2\phi}\xi^3+\frac{\Xi}{2}\xi+\frac{b\phi\|s\|^2}{2(\varpi+b\phi\|s\|^2)}\left(-\Xi-\frac{1}{\phi}+2A_{\rm e}\right)\xi\\ &\le A_{\rm e}\sqrt{\frac{a}{b}}+\left(B_{\rm e}+\frac{\Xi}{2}\right)\xi+B_{\rm e}\sqrt{\frac{b}{a}}\xi^2+\frac{1}{2\phi}\xi^3 \end{aligned} \tag{79}$$

From Theorem 4, one can obtain that $\dot{\mathcal{W}}\le 0$, which shows that the equilibrium point is asymptotically stable during the flow.

For the jump dynamics, note that the variable ς in (52) only evolves during flows and is kept constant otherwise. Hence, $\Delta\mathcal{W}(t)=\Delta\mathcal{V}(t)$. As shown in Theorem 4, the asymptotic stability of the equilibrium point can be proved.

Denote $b=(1-\beta^2)\lambda_{\min}(Q)$ and $a=L_e$. Then, from (53), the following holds:

$$a(\varsigma+\phi q)=\varpi+\phi(b\|s\|^2+\Theta(u_a(\hat{s}))-a\|e\|^2)\le 0 \tag{76}$$

where $\varpi=a\varsigma$ satisfies

$$\dot{\varpi}=-a\varpi+b\|s\|^2+\Theta(u_a(\hat{s}))-a\|e\|^2. \tag{77}$$

From (76), one can obtain

$$a\phi\|e\|^2\ge\varpi+b\phi\|s\|^2+\phi\Theta(u_a(\hat{s}))\ge\varpi+b\phi\|s\|^2.$$

Therefore, in the interval of $[t_k,t_{k+1})$, the variable $\xi(t):=(\sqrt{a\phi}\|e(t)\|/(\varpi(t)+b\phi\|s(t)\|^2)^{1/2})$ evolves from 0 to 1. We can now bound the interevent interval by writing the dynamics of $\xi(t)$ as

$$\dot{\xi}=\frac{\sqrt{a\phi}e^{\rm T}\dot{e}}{\sqrt{\varpi+b\phi\|s\|^2}\|e\|}-\frac{\sqrt{a\phi}\|e\|}{2(\varpi+b\phi\|s\|^2)^{\frac{3}{2}}}(\dot{\varpi}+2b\phi s^{\rm T}\dot{s}) \tag{78}$$

where $\xi(0)=0$. From Lemma 3, (17), and (77), one has

$$\begin{aligned}\dot{e}&=-\dot{s}\\ \|\dot{s}\|&\le A_{\rm e}\|s\|+B_{\rm e}\|e\|\\ \dot{\varpi}&\ge-\Xi\varpi+(b\|s\|^2-a\|e\|^2).\end{aligned}$$

Then, from (78), the dynamics of ξ satisfies (79), as shown at the top of this page, where the last inequality holds if $\Xi\in(0,2A_e)$ and $\phi\in(0,(1/2A_e-\Xi)]$. Denote $\zeta(t,\zeta_0)$ as the solution of the following differential equation:

$$\dot{\zeta}=A_e\sqrt{\frac{a}{b}}+\left(B_e+\frac{\gamma}{2}\right)\zeta+B_e\sqrt{\frac{b}{a}}\zeta^2+\frac{1}{2\phi}\zeta^3,\quad \zeta_0=\xi_0.$$

Based on the comparison principle [45] and (79), $\xi(t)$ satisfies $\xi(t)\le\zeta(\zeta_0,t)$. Then, the time needed by $\xi(t)$ to evolve from 0 to 1 is lower bounded by a positive constant τ_d given as

$$\tau_d=\int_0^1\frac{1}{A_e\sqrt{\frac{a}{b}}+\left(B_e+\frac{\mu}{2}\right)v+B_e\sqrt{\frac{b}{a}}v^2+\frac{1}{2\phi}v^3}dv$$

Therefore, condition (53) is Zeno-free. ■

REFERENCES

[1] S. Ling, H. Wang, and P. X. Liu, "Adaptive fuzzy tracking control of flexible-joint robots based on command filtering," *IEEE Trans. Ind. Electron.*, to be published, doi: 10.1109/TIE.2019.2920599.

[2] L. Sonneveldt, Q. P. Chu, and J. A. Mulder, "Nonlinear flight control design using constrained adaptive backstepping," *J. Guid., Control, Dyn.*, vol. 30, no. 2, pp. 322–336, Mar. 2007.

[3] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, May 2005.

[4] M. Chen, S. S. Ge, and B. Ren, "Adaptive tracking control of uncertain MIMO nonlinear systems with input constraints," *Automatica*, vol. 47, no. 3, pp. 452–465, Mar. 2011.

[5] 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, Apr. 2009.

[6] Y.-J. Liu, S. Lu, S. Tong, X. Chen, C. P. Chen, and D.-J. Li, "Adaptive control-based Barrier Lyapunov functions for a class of stochastic nonlinear systems with full state constraints," *Automatica*, vol. 87, pp. 83–93, Jan. 2018.

[7] W. He, B. Huang, Y. Dong, Z. Li, and C.-Y. Su, "Adaptive neural network control for robotic manipulators with unknown deadzone," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2670–2682, Sep. 2018.

[8] Y.-J. Liu, M. Gong, S. Tong, C. L. P. Chen, and D.-J. Li, "Adaptive fuzzy output feedback control for a class of nonlinear systems with full state constraints," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2607–2617, Oct. 2018.

[9] D. Li and D. Li, "Adaptive tracking control for nonlinear time-varying delay systems with full state constraints and unknown control coefficients," *Automatica*, vol. 93, pp. 444–453, Jul. 2018.

[10] Y.-J. Liu and S. Tong, "Barrier Lyapunov functions for Nussbaum gain adaptive control of full state constrained nonlinear systems," *Automatica*, vol. 76, pp. 143–152, Feb. 2017.

[11] C. P. Bechlioulis and G. A. Rovithakis, "Robust adaptive control of feedback linearizable MIMO nonlinear systems with prescribed performance," *IEEE Trans. Autom. Control*, vol. 53, no. 9, pp. 2090–2099, Oct. 2008.

[12] C. Wen, J. Zhou, Z. Liu, and H. Su, "Robust adaptive control of uncertain nonlinear systems in the presence of input saturation and external disturbance," *IEEE Trans. Autom. Control*, vol. 56, no. 7, pp. 1672–1678, Jul. 2011.

[13] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.

[14] Y. Yang, D. Wunsch, and Y. Yin, "Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1929–1940, Aug. 2017.

[15] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[16] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[17] Z. Ni, H. He, J. Wen, and X. Xu, "Goal representation heuristic dynamic programming on maze navigation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2038–2050, Dec. 2013.

[18] Z. Ni, H. He, D. Zhao, X. Xu, and D. V. Prokhorov, "GrDHP: A general utility function representation for dual heuristic dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 614–627, Mar. 2015.

[19] Z. Ni, H. He, X. Zhong, and D. V. Prokhorov, "Model-free dual heuristic dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1834–1839, Aug. 2015.

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

[21] L. Dong, Y. Tang, H. He, and C. Sun, "An event-triggered approach for load frequency control with supplementary ADP," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 581–589, Jan. 2017.

[22] D. Wang, M. Ha, and J. Qiao, "Self-learning optimal regulation for discrete-time nonlinear systems under event-driven formulation," *IEEE Trans. Autom. Control*, to be published, doi: 10.1109/TAC.2019.2926167.

[23] L. Dong, X. Zhong, C. Sun, and H. He, "Event-triggered adaptive dynamic programming for continuous-time systems with control constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1941–1952, Aug. 2017.

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

[25] 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 Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2139–2153, Mar. 2018.

[26] Y. Yang, Z. Guo, H. Xiong, D.-W. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3735–3747, Dec. 2019.

[27] D. Wang, H. He, and D. Liu, "Adaptive critic nonlinear robust control: A survey," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3429–3451, Oct. 2017.

[28] Q. Zhang and D. Zhao, "Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 2874–2885, Aug. 2019.

[29] Y. Tang, H. He, J. Wen, and J. Liu, "Power system stability control for a wind farm based on adaptive dynamic programming," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 166–177, Jan. 2015.

[30] J. Yan, H. He, X. Zhong, and Y. Tang, "Q-learning-based vulnerability analysis of smart grid against sequential topology attacks," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 200–210, Jan. 2017.

[31] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, Jul. 2014.

[32] 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 Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2386–2398, Nov. 2016.

[33] B. Fan, Q. Yang, X. Tang, and Y. Sun, "Robust ADP design for continuous-time nonlinear systems with output constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2127–2138, Jun. 2018.

[34] D. A. Copp, K. G. Vamvoudakis, and J. P. Hespanha, "Distributed output-feedback model predictive control for multi-agent consensus," *Syst. Control Lett.*, vol. 127, pp. 52–59, May 2019.

[35] J. B. Rawlings and D. Q. Mayne, *Model Predictive Control Theory and Design*. Madison, WI, USA: Nob Hill, 2009.

[36] P. D. Christofides, R. Scattolini, D. M. De La Peña, and J. Liu, "Distributed model predictive control: A tutorial review and future research directions," *Comput. Chem. Eng.*, vol. 51, pp. 21–41, Apr. 2013.

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

[38] A. Girard, "Dynamic triggering mechanisms for event-triggered control," *IEEE Trans. Autom. Control*, vol. 60, no. 7, pp. 1992–1997, Jul. 2015.

[39] D. Wang and D. Liu, "Neural robust stabilization via event-triggering mechanism and adaptive learning technique," *Neural Netw.*, vol. 102, pp. 27–35, Jun. 2018.

[40] D. Wang, C. Mu, H. He, and D. Liu, "Event-driven adaptive robust control of nonlinear systems with uncertainties through NDP strategy," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 7, pp. 1358–1370, Jul. 2017.

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

[42] Q. Zhang, D. Zhao, and D. Wang, "Event-based robust control for uncertain nonlinear systems using adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 37–50, Jan. 2018.

[43] Y. Zhu, D. Zhao, H. He, and J. Ji, "Event-triggered optimal control for partially unknown constrained-input systems via adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4101–4109, May 2017.

[44] X. Yang and H. He, "Adaptive critic learning and experience replay for decentralized event-triggered control of nonlinear interconnected systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2019.2898370.

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

[46] B. A. Finlayson, *The Method of Weighted Residuals and Variational Principles*. Philadelphia, PA, USA: SIAM, 2013, vol. 73.

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

[48] 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 Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2019.2933736.

[49] J. A. Primbs and V. Nevistić, "Optimality of nonlinear design techniques: A converse HJB approach," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CIT-CDS 96-021, 1996.

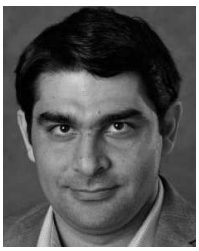
Yongliang Yang (Member, IEEE) received the B.S. degree in electrical engineering from Hebei University, Baoding, China, in 2011, and the Ph.D. degree in electrical engineering from the University of Science and Technology Beijing (USTB), Beijing, China, in 2017.

He was a Visiting Scholar with the Missouri University of Science and Technology, Rolla, MO, USA, from 2015 to 2017, supported by the China Scholarship Council. He is currently an Assistant Professor with the School of Automation and Electrical Engineering, USTB. His research interests include adaptive optimal control, distributed optimization and control, and cyber-physical systems (CPSs).

Dr. Yang was a recipient of the Best Ph.D. Dissertation of China Association of Artificial Intelligence, the Best Ph.D. Dissertation of USTB, the Chancellors Scholarship in USTB, and the Excellent Graduates Awards in Beijing. He also serves as a Reviewer of several international journals and conferences, including *Automatica*, the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, the IEEE TRANSACTIONS ON CYBERNETICS, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Kyriakos G. Vamvoudakis (Senior Member, IEEE) was born in Athens, Greece. He received the Diploma degree (Hons.) (a five-year degree, equivalent to M.Sc.) in electronic and computer engineering from the Technical University of Crete, Chania, Greece, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from The University of Texas at Arlington, Arlington, TX, USA, in 2008 and 2011, respectively, under the supervision of Dr. F. L. Lewis. From 2011 to 2012, he was an Adjunct Professor and a Faculty Research Associate with The University of Texas at Arlington and the Automation and Robotics Research Institute, Fort Worth, TX, USA. From 2012 to 2016, he was a Project Research Scientist with the Center for Control, Dynamical Systems and Computation, University of California at Santa Barbara, Santa Barbara, CA, USA. He was an Assistant Professor with the Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA, USA, until 2018. He currently serves as an Assistant Professor with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His research interests include approximate dynamic programming, game theory, cyber-physical security, networked control, smart grid, and safe autonomy.

Dr. Vamvoudakis is a Senior Member of AIAA. He is a member of the Technical Chamber of Greece. He is a member of Tau Beta Pi, Eta Kappa Nu, and Golden Key Honor societies. He currently is a member of the Technical Committee on Intelligent Control of the IEEE Control Systems Society, the Technical Committee on Adaptive Dynamic Programming and Reinforcement Learning of the IEEE Computational Intelligence Society, and the IEEE Control Systems Society Conference Editorial Board. He was a recipient of the 2018 NSF CAREER Award, the 2019 ARO YIP Award, and several international awards, including the 2016 International Neural Network Society Young Investigator Award, the Best Paper Award for Autonomous/Unmanned Vehicles at the 27th Army Science Conference in 2010, the Best Presentation Award at the World Congress of Computational Intelligence in 2010, and the Best Researcher Award from the Automation and Robotics Research Institute in 2011. He has also served on various international program committees and has organized special sessions, workshops, and tutorials for several international conferences. He is currently an Associate Editor of *Automatica*, the *IEEE Computational Intelligence Magazine*, the *Journal of Optimization Theory and Applications*, and the IEEE CONTROL SYSTEMS LETTERS. He is a Registered Electrical/Computer Engineer (PE). He is listed in Who's Who in the World, Who's Who in Science and Engineering, and Who's Who in America.



Hamidreza Modares (Member, IEEE) received the B.S. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2004, the M.S. degree in electrical engineering from the Shahrood University of Technology, Shahroud, Iran, in 2006, and the Ph.D. degree in electrical engineering from The University of Texas at Arlington, Arlington, TX, USA, in 2015.

He was a Faculty Research Associate with The University of Texas at Arlington from 2015 to 2016, and an Assistant Professor with the Missouri University of Science and Technology, Rolla, MO, USA, from 2016 to 2018. He is currently an Assistant Professor with the Department of Mechanical Engineering, Michigan State University, East Lansing, MI, USA. His current research interests include cyber-physical systems, reinforcement learning, distributed control, robotics, and machine learning.

Dr. Modares was a recipient of the Best Paper Award from the 2015 IEEE International Symposium on Resilient Control Systems. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Yixin Yin (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Science and Technology Beijing (USTB), Beijing, China, in 1982, 1984, and 2002, respectively.

He was a Visiting Scholar with several universities in Japan, including The University of Tokyo, Tokyo, Japan, the Kyushu Institute of Technology, Kitakyushu, Japan, Kanagawa University, Yokohama, Japan, Chiba University, Chiba, Japan, and the Muroran Institute of Technology, Muroran, Japan. He served as the Dean for the School of the Information Engineering, USTB, from 2000 to 2011, where he was the Dean of the School of Automation and Electrical Engineering, from 2011 to 2017. He is currently a Professor with the School of Automation and Electrical Engineering, USTB. His major research interests include modeling and control of complex industrial processes, computer-aided design of control system, intelligent control, and artificial life.

Prof. Yin is a fellow of the Chinese Society for Artificial Intelligence and a member of the Chinese Society for Metals and the Chinese Association of Automation. He was a recipient of several national awards, including the Outstanding Young Educator in 1993, the Award of Science and Technology Progress in Education in 1994, the Award of National Science and Technology Progress in 1995, the Special Allowance of the State Council in 1994, the Best Paper of Japanese Acoustical Society in 1999, and the Award of Metallurgical Science and Technology in 2014.

Donald C. Wunsch II (Fellow, IEEE) received the B.S. degree in applied mathematics from The University of New Mexico, Albuquerque, NM, USA, in 1984, the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 1987 and 1991, respectively, and the M.B.A. degree (executive) in business administration from Washington University in St. Louis, St. Louis, MO, USA, in 2006.

He was with Texas Tech University, Lubbock, TX, USA, Boeing, Seattle, WA, USA, Rockwell International, Albuquerque, NM, USA, and International Laser Systems, Albuquerque, NM, USA. He is currently the Mary K. Finley Missouri Distinguished Professor with the Missouri University of Science and Technology (Missouri S&T), Rolla, MO, USA, where he is also the Director of the Applied Computational Intelligence Laboratory, a multidisciplinary research group. His current research interests include clustering/unsupervised learning, biclustering, adaptive resonance, and adaptive dynamic programming architectures, hardware, and applications, neurofuzzy regression, autonomous agents, games, and bioinformatics.

Dr. Wunsch is an INNS fellow. He was the INNS President. He was a recipient of the NSF CAREER Award, the 2015 International Neural Networks Society (INNS) Gabor Award, and the 2019 Ada Lovelace Service Award. He has produced 22 Ph.D. recipients in computer engineering, electrical engineering, systems engineering, and computer science. He served as the International Joint Conference on Neural Networks (IJCNN) General Chair, and on several boards, including the St. Patricks School Board, the IEEE Neural Networks Council, the INNS, and the University of Missouri Bioinformatics Consortium, and the Chair of the Missouri S&T Information Technology and Computing Committee and the Student Design and Experiential Learning Center Board.