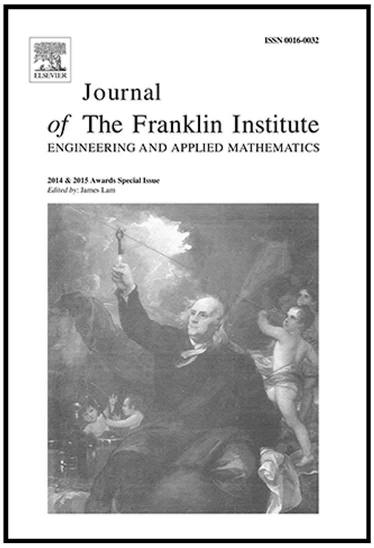
Journal Pre-proof



Online Barrier-Actor-Critic Learning for H$_\infty$ Control with Full-State Constraints and Input Saturation

Yongliang Yang, Da-Wei Ding, Haoyi Xiong, Yixin Yin, Donald C. Wunsch

PII: S0016-0032(19)30904-4
DOI: https://doi.org/10.1016/j.jfranklin.2019.12.017
Reference: FI 4334

To appear in: *Journal of the Franklin Institute*

Received date: 17 June 2019
Revised date: 11 October 2019
Accepted date: 9 December 2019

Please cite this article as: Yongliang Yang, Da-Wei Ding, Haoyi Xiong, Yixin Yin, Donald C. Wunsch, Online Barrier-Actor-Critic Learning for H$_\infty$ Control with Full-State Constraints and Input Saturation, *Journal of the Franklin Institute* (2019), doi: https://doi.org/10.1016/j.jfranklin.2019.12.017

Online Barrier-Actor-Critic Learning for H_∞ Control with Full-State Constraints and Input Saturation[☆]

Yongliang Yang[a,b,*], Da-Wei Ding[a,b], Haoyi Xiong[c,d], Yixin Yin[a,b], Donald C. Wunsch[e]

[a] *School of Automation & Electrical Engineering, University of Science and Technology Beijing, Beijing 10083, China*
[b] *Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, University of Science and Technology Beijing, Beijing 10083, China*
[c] *Big Data Laboratory, Baidu Research, Beijing 100193, China*
[d] *National Engineering Laboratory of Deep Learning Technology and Application, Beijing 100193, China*
[e] *Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA*

Abstract

This paper develops a novel adaptive optimal control design method with full-state constraints and input saturation in the presence of external disturbance. First, to consider the full-state constraints, a barrier function is developed for system transformation. Moreover, it is shown that, with the barrier-function-based system transformation, the stabilization of the transformed system is equivalent to the original constrained control problem. Second, the disturbance attenuation problem is formulated within the zero-sum differential games framework. To determine the optimal control and the worst-case disturbance, a novel barrier-actor-critic algorithm is presented for adaptive optimal learning while guaranteeing the full-state constraints and input saturation. It is proven that the closed-loop signals remain bounded during the online learning phase. Finally, simulation studies are conducted to demonstrate the effectiveness of the presented barrier-actor-critic learning algorithm.

Keywords: **full-state constraints**, **input saturation**, **disturbance attenuation**, **adaptive dynamic programming**, **barrier-actor-critic learning**

[☆]This work was supported in part by the National Natural Science Foundation of China under Grant No. 61903028, No. 61873028 and No. 61333002, 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 FRF-BD-19-002A, in part by the Lifelong Learning Machines program from DARPA/Microsystems Technology Office and in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-18-2-0260. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

*Corresponding author: Da-Wei Ding (ddaweiauto@163.com)
Email addresses: yangyongliang@ieee.org (Yongliang Yang), xionghaoyi@baidu.com (Haoyi Xiong), yyx@ies.ustb.edu.cn (Yixin Yin), wunsch@ieee.org (Donald C. Wunsch)

1. Introduction

Nonlinear dynamics commonly exists in engineering applications, such as input saturation [1, 2, 3] and dead-zone [4, 5], output constraints [6, 7], friction dynamics [8, 9], backlash-like hysteresis [10, 11, 12], unmodeled dynamics [13], etc. Modern control theory, such as the H_∞ control method [14, 15] and adaptive control method [16, 17], has received considerable attention to compensate for the system uncertainty and attenuate the effect of external disturbance for nonlinear systems. In addition to the closed-loop stability, practical constraints captured by user-defined performance is desired to be guaranteed. However, classical H_∞ control and adaptive control methods cannot guarantee the user-defined performance. In this paper, a novel adaptive optimal controller design is developed to stabilize the nonlinear systems while considering both the prescribed performance on full-state and input saturation simultaneously.

For the nonlinear systems with imperfect dynamical behavior, such as exogenous disturbance and system uncertainties, the adaptive control method is widely used for feedback design to compensate the system uncertainty and attenuate exogenous disturbances [16, 17]. However, classical adaptive control design methods only consider the closed-loop stability. In addition to the closed-loop stability, practical constraints are important for controller design. For example, in the control of Euler-Lagrange systems, the link and joint velocity cannot be arbitrarily large and has to be remained in the bounded region due to limitation imposed by mechanical characteristics. In many applications, the constraints are usually captured by the user-defined performance. Many efforts have been made to address this issue. Compared to classical quadratic Lyapunov function design, Lyapunov analysis is combined with barrier function design [18] to consider the constraints on output, which is essentially partial-state constraints [19, 20]. Since then, the barrier Lyapunov function design is extended to consider full-state constraints for stochastic nonlinear systems [21], pure-feedback systems [22], Euler-Lagrange systems [23], time-delay systems [24], to name a few. Another type of constrained controller design adopts a prescribed transient performance to develop a system transformation [25]. In the prescribed performance adaptive control, the prescribed transient performance is captured by a user-defined performance bound, which specifies the safety region for the tracking error. Recently, the prescribed performance adaptive control method is extended to deal with output feedback control problem [26], consensus problem of multi-agent systems [27], nonlinear systems with input dead-zone [28], controller design for flexible joint robots [29], synchronization problem of teleoperation robotics [30], and so on. To relax the requirement that both the reference signal along with its derivatives and every element of the state variable are available for feedback design, Arabia and Yucelen developed a set-theoretic model reference adaptive control framework [31]. In the set-theoretic model reference adaptive control framework, the norm of the gap between the system state and the reference signal is guaranteed to be within a user-defined constant bound. However, in the existing adaptive controller design methods, only closed-loop stability and the prescribed user-defined performance constraints is considered without optimality discussion. In this paper, a novel adaptive constrained controller is presented with optimality discussions.

The centerpiece of optimal control theory is the Hamilton-Jacobi-Bellman/ Hamilton-Jacobi-Isaacs (HJB/HJI) equations for nonlinear systems, which is necessary and sufficient for the optimality condition [32]. However, the HJ equations are difficult to solve due to the inherent nonlinearity. Therefore, adaptive dynamic programming (ADP) has been developed to approximate the nonlinear HJ equations in an online fashion, where an intelligent agent seeks optimal decisions to maximize the lone-term cumulative reward [33]. Variants of ADP has been applied widely in control applications to solve the optimal control problems, including iterative ADP algorithms in discrete-time [34] and continuous-time [35] for optimal regulation problems, model-free learning algorithm for H_∞ control problem [36], online actor-critic learning algorithm [37] for optimal tracking control problems [38, 39], robust stabilization problem [40], guaranteed cost control problem [41, 42], consensus control problem of multi-agent systems [43, 44], event-triggered control [45], to name a few. Besides, ADP has been successfully applied to differential games [46]. In addition, ADP extensions have been made to deal with constraints of input saturation in [47] and constraints on the state in [48]. However, these existing results do not consider the case with external disturbance, input saturation, and full-state constraints. In this paper, all these issues are considered in a comprehensive framework.

The contributions of this paper are threefold. First, in this paper, both the full-state constraints and input saturation are considered simultaneously for the controller design problem. This is achieved by introducing a barrier function based system transformation. It is also discussed theoretically that the transform equivalence can be guaranteed in the sense that the stabilization of the transformed system ensures the constraints of the original system. Second, the disturbance attenuation is achieved within the framework of zero-sum differential games. A novel barrier-actor-critic algorithm is developed for adaptive optimal learning with the full-state constraints and input saturation. Finally, to obviate the requirement of persistent excitation condition, the experience replay technique is employed to utilize the history and current date concurrently.

The remainder of this paper is organized as follows. In Section 2, the problem of constrained control design with full-state constraints and input saturation is given. Section 3 presents the barrier-function-based system transformation to deal with full-state constraints. In Section 4, a novel actor-critic-barrier algorithm is developed for the online learning of the adaptive optimal constrained controller.

2. Preliminaries

2.1. Notations and Definitions

The following standard notation will be adopted.

$\mathbb{R}^+$	$\triangleq$	the set of positive real numbers.
$\mathbb{R}^n$	$\triangleq$	n-dimensional vector space.
I	$\triangleq$	Identity matrix with proper dimension.
$\mathbf{1}$	$\triangleq$	vector with all entries being 1.
$\Vert\mathcal{M}\Vert$	$\triangleq$	$\sqrt{tr\left(\mathcal{M}\mathcal{M}^H\right)}$, the matrix Frobenius norm of matrix $\mathcal{M}$.
$\Vert v\Vert$	$\triangleq$	the euclidean norm of vector v.
$\mathbb{Z}$	$\triangleq$	the set of integers.
$\lambda_{\min}(A)$	$\triangleq$	the minimum eigenvalue of matrix A.

Definition 1. (Zero-State Observality) [15] The system (1) with the measured output $y = h(x)$ is zero-state observable if $y(t) \equiv 0$ for $\forall t \geqslant 0$ implies that $x(t) \equiv 0$ for $\forall t \geqslant 0$.

Definition 2. (Persistent Excitation Condition) [16] The vector signal $z(\cdot) \in \mathbb{R}^n$ is said to be persistently excited (PE) on the interval $[T_1, T_2]$ if there exists positive constants $\gamma_1 > 0$ and $\gamma_2 > 0$ such that, for all $t \in [T_1, T_2]$,

$$\gamma_1 I \leqslant \int_t^{t+T_1} z(\tau)\, z^{\mathrm{T}}(\tau)\, d\tau \leqslant \gamma_2 I$$

Definition 3. (Uniformly Ultimately Bounded Stability) [16] Consider the nonlinear system

$$\dot{x} = F(x,t), \quad \forall t \in \mathbb{R}^+, \quad x(t_0) = x_0 \tag{1}$$

with $x(t) \in \mathbb{R}^n$ is the system state and x_0 is the initial condition. The equilibrium point x_e of system (1) is said to be uniformly ultimately bounded (UUB) if there existes a compact set $\Omega \subset \mathbb{R}^n$ so that for all $x_0 \subset \Omega$, there exists a bound B and a time $T(B, x_0)$ such that $\Vert x(t) - x_0\Vert \leqslant B$ for all $t \geqslant t_0 + T$.

Lemma 1. *[49] For $\forall w \in \mathbb{R}$, there exists a bounded $\tilde{w}$ satisfying $\Vert\tilde{w}\Vert \leqslant \ln 4$, such that*

$$-2\ln\left(1 + e^{-2w}\right) = 2w - 2w\mathrm{sgn}(w) + \tilde{w},$$

Lemma 2. *[50] The following inequality holds for any $a > 0$ and $y \in \mathbb{R}$*

$$0 \leqslant |y| - y \tanh\left(\frac{y}{a}\right) \leqslant \kappa a \tag{2}$$

where $\kappa = 0.2785$.

2.2. Problem Statement

In this paper, we consider the following continuous-time affine nonlinear dynamical systems

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\vdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= f(x) + g(x)u + k(x)d \end{aligned} \tag{3}$$

where $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^{\mathrm{T}} \in R^n$ is the system state, $u(\cdot) : R^n \to R^{m_1}$ is the control policy, $d(\cdot) : R^n \to R^{m_2}$ is the external disturbance, $f(\cdot), g(\cdot), k(\cdot) : R^n \to R$ are Lipschitz continuous nonlinear functions. The constrained H_∞ control problem for system (3) with full-state constraints and input saturation can be formulated as follows.

Problem 1. Design the proper performance output $L(x, u)$, where $L(\cdot, \cdot)$ is a positive definite function of its argument, and the optimal policy u^* such that, with the saturation constraints on the control input as

$$\|u_i\| \leqslant \lambda, \forall i = 1, ..., m_1 \tag{4}$$

where $u = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}^{\mathrm{T}}$, and the full-state constraints as

$$\begin{aligned} x_1 &\in (a_1, A_1) \\ &\vdots \\ x_n &\in (a_n, A_n) \end{aligned} \tag{5}$$

for $\forall d \in \mathcal{L}_2$, system (3) have L_2-gain less than or equal to γ, i.e.,

$$\frac{\int_t^\infty L(x(\tau), u(\tau))d\tau}{\int_t^\infty \|d(\tau)\|^2 d\tau} \leqslant \gamma^2, \tag{6}$$

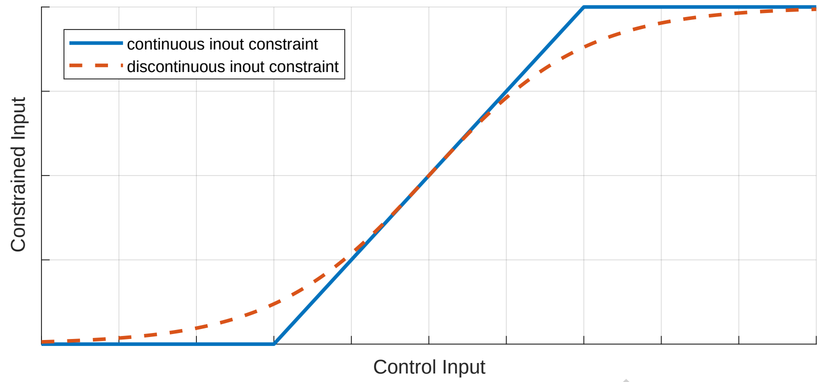


Figure 1: Evolution of the two-dimensional phase plot of the state trajectories $[x_1(t) \quad x_2(t)]$. The black box denotes the safe region.

Remark 1. To deal with the input constraints (4), the saturation function can be applied, which is defined as [51, 52, 53, 54]

$$\Gamma(u_i) = \begin{cases} u_i, & if \quad u_i \leqslant \lambda \\ sign(u_i), & if \quad u_i > \lambda \end{cases}$$

Then, the system dynamics can be denoted as

$$\begin{aligned} \dot{x}_i &= x_{i+1}, i = 1, 2, ..., n-1 \\ \dot{x}_n &= f(x) + g(x)\Gamma(u) + k(x)d \end{aligned}$$

Note that the saturation function $\Gamma(\cdot)$ is a discontinuous function, which leads to discontinuity in the system dynamics. In this paper, we consider continuous constraints on the input signal, which is shown in Figure 1 and widely used in the literature, such as [33, 47, 55]. As shown later, the nonquadratic penalty function on the control input signal (17) is presented, which guarantees the boundedness of the optimal control input (23).

As shown by (4) – (6), the objective of Problem 1 can be divided into three parts, i.e., disturbance attenuation, input saturation and full-state constraints. For the full-state constraints, we introduce the following barrier function.

Definition 4. (Barrier Function) The function $B(\cdot) : \mathbb{R} \to \mathbb{R}$ defined on (a, A) is referred to as barrier function if

$$B(z; a, A) = \ln\left(\frac{A}{a}\frac{a-z}{A-z}\right), \forall z \in (a, A) \tag{7}$$

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

$$B^{-1}(y; a, A) = aA\frac{e^{\frac{y}{2}} - e^{-\frac{y}{2}}}{ae^{\frac{y}{2}} - Ae^{-\frac{y}{2}}}, \forall y \in \mathbb{R} \tag{8}$$

with the derivative

$$\frac{dB^{-1}(y; a, A)}{dy} = \frac{Aa^2 - aA^2}{a^2e^y - 2aA + A^2e^{-y}} \tag{9}$$

Remark 2. To guarantee that the full-state constraints is not violated for Problem 1, the barrier function in Definition 4 has the following desired properties

1) The barrier function $B(\cdot)$ takes finite value when the its arguments are within the user-defined region (a, A).
2) The barrier function $B(\cdot)$ approach to infinity as the state approach the boundary of the prescribed region (a, A), i.e.,

$$\lim_{z \to a^+} B(z; a, A) = -\infty$$
$$\lim_{z \to A^-} B(z; a, A) = +\infty$$

3) The barrier function $B(\cdot)$ vanishes at the equilibrium of the system (3), i.e.,

$$B(0; a, A) = 0, \forall a < A$$

3. Barrier-Function-Based Zero-Sum Game

In this section, the system (3) with full-state constraints is transformed into another system without state constraints by using the barrier function in Definition 4. Consider the barrier-function-based state transformation as

$$\begin{aligned} s_i &= B(x_i; a_i, A_i), \\ x_i &= B^{-1}(s_i; a_i, A_i), i = 1, \cdots, n \end{aligned} \tag{10}$$

Then, by using the chain rule, one has

$$\frac{dx_i}{dt} = \frac{dx_i}{ds_i}\frac{ds_i}{dt} \tag{11}$$

From (11), the dynamics of the transformed state s can be written as

$$\begin{aligned}
\dot{s}_i &= \frac{x_{i+1}\left(s_{i+1}\right)}{\left.\frac{dB^{-1}(y;a_i,A_i)}{dy}\right|_{y=s_i}} \\
&= \frac{a_{i+1}A_{i+1}\left(e^{\frac{s_{i+1}}{2}}-e^{-\frac{s_{i+1}}{2}}\right)}{a_{i+1}e^{\frac{s_{i+1}}{2}}-A_{i+1}e^{-\frac{s_{i+1}}{2}}}\frac{A_i^2e^{-s_i}-2a_iA_i+a_i^2e^{s_i}}{A_ia_i^2-a_iA_i^2} \\
&= F_i\left(s_i,s_{i+1}\right), \qquad i=1,\dots,n-1 \\
\dot{s}_n &= \frac{f\left(x\right)+g\left(x\right)u+k(x)d}{\left.\frac{dB^{-1}(y;a_n,A_n)}{dy}\right|_{y=s_n}} \\
&= \left[f\left(x\right)+g\left(x\right)u+k(x)d\right]\frac{A_n^2e^{-s_n}-2a_nA_n+a_n^2e^{s_n}}{A_na_n^2-a_nA_n^2} \\
&= F_n\left(s\right)+g_n\left(s\right)u+k_n(s)d
\end{aligned} \tag{12}$$

with

$$\begin{aligned}
F_n\left(s\right) &= \frac{A_n^2e^{-s_n}-2a_nA_n+a_n^2e^{s_n}}{A_na_n^2-a_nA_n^2}f\left(\begin{bmatrix} B_1^{-1}\left(s_1\right) & \dots & B_n^{-1}\left(s_n\right)\end{bmatrix}\right) \\
g_n\left(s\right) &= \frac{A_n^2e^{-s_n}-2a_nA_n+a_n^2e^{s_n}}{A_na_n^2-a_nA_n^2}g\left(\begin{bmatrix} B_1^{-1}\left(s_1\right) & \dots & B_n^{-1}\left(s_n\right)\end{bmatrix}\right) \\
k_n\left(s\right) &= \frac{A_n^2e^{-s_n}-2a_nA_n+a_n^2e^{s_n}}{A_na_n^2-a_nA_n^2}k\left(\begin{bmatrix} B_1^{-1}\left(s_1\right) & \dots & B_n^{-1}\left(s_n\right)\end{bmatrix}\right)
\end{aligned} \tag{13}$$

Note that system (12) with the state $s=\begin{bmatrix} s_1 & \cdots & s_n\end{bmatrix}^{\mathrm{T}}$ can be expressed in a compact form as

$$\dot{s}=F\left(s\right)+G\left(s\right)u+K\left(s\right)d \tag{14}$$

with $F\left(s\right)=\begin{bmatrix} F_1\left(s_1,s_2\right) \\ \vdots \\ F_n\left(s\right)\end{bmatrix}$, $G\left(s\right)=\begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_n\left(s\right)\end{bmatrix}$, $K\left(s\right)=\begin{bmatrix} 0 \\ \vdots \\ 0 \\ k_n\left(s\right)\end{bmatrix}$.

The following assumptions are imposed on system (14), which is commonly used for nonlinear systems controller design [47, 55].

Assumption 1. The system dynamics (14) is assumed to have the following properties.

1) $F(s)$ is Lipschitz with $F(0)=0$, and there exists a constant b_f such that, for $s\in\Omega$, $\|F(s)\|\leqslant b_f\|s\|$ where Ω is a compact set containing the origin.
2) $G(s)$ and $K(s)$ are bounded on Ω, i.e., there exists a constant b_g and b_k such that $\|G(s)\|\leqslant b_g$ and $\|K(s)\|\leqslant b_k$, respectively.

3) The system (3) is controllable over the compact set Ω.

In the following, to consider the input saturation and disturbance attenuation in Problem 1, the framework of the zero-sum differential game is introduced. For system (14) with the control input $u(t)$ and the disturbance policy $d(t)$, consider the following cost function

$$V(s_0; u, d) = \int_{t_0}^{\infty} U(s, u, d) dt \tag{15}$$

where $U(s, u, d)$ is the reward function with

$$\begin{aligned} U(s, u, d) &= L(x, u) - \gamma^2 \|d\|^2, \\ L(x, u) &= Q(s) + \Theta(u) - \gamma^2 \|d\|^2 \end{aligned} \tag{16}$$

where $Q(s)$ being a positive definite monotonically increasing function and $\Theta(u)$ being a positive definite integrand function. To deal with input saturation, the nonquadratic penalty function is used,

$$\begin{aligned} \Theta(u) &= 2\int_0^u \left[\lambda \tanh^{-1}\left(\frac{v}{\lambda}\right)\right] R dv \\ &= 2\lambda\left(\tanh^{-1}(u/\lambda)\right)^{\mathrm{T}} R u + \lambda^2 \bar{R} \ln\left(\mathbf{1} - u^2/\lambda^2\right) \end{aligned} \tag{17}$$

where $\lambda > 0$ is the saturation limit for the control input, $R = \mathrm{diag}(r_1, ..., r_m)$ and $\bar{R} = [r_1, ..., r_m] \in \mathbb{R}^{1\times m}$ with $r_i > 0$ for $i = 1, ..., m$ is the weight on control effort for each input.

Problem 2. For system (14) with the control policy u and disturbance policy d, find the Nash equilibrium (u^*, d^*) of the zero-sum game with the constraints of input saturation (4).

Define the Hamiltonian for the cost (15) with the control policy u and disturbance policy d as

$$H(u, d, V) = \left(\frac{\partial V}{\partial s}\right)^{\mathrm{T}} [F(s) + G(s) u + K(s) d] + U(s, u, d) \tag{18}$$

Then, differential equivalent of the cost (15) can be expressed in terms of the Hamiltonian (18) as

$$H\left(s, u, d, \frac{\partial V}{\partial s}\right) = 0 \tag{19}$$

which is referred to as the Bellman equation.

Based on the game theory [56], the disturbance attenuation problem is equivalent to solving the following two-player zero-sum game,

$$V^*(s) = \min_u \max_d V(s; u, d) \tag{20}$$

This two-player zero-sum game has a unique solution if the Nash condition holds

$$V^*(s) = \min_u \max_d V(s; u, d) = \max_d \min_u V(s; u, d) \tag{21}$$

According to [32], the stationary condition for optimality is

$$\frac{\partial H(u, d, V^*)}{\partial u} = 0, \frac{\partial H(u, d, V^*)}{\partial d} = 0 \tag{22}$$

Then, one can obtain the optimal control input u^* and the worst-case disturbance d^*, respectively, as

$$u^*(s) = -\lambda \tanh\left(\frac{1}{2\lambda} R^{-1} G^{\mathrm{T}}(s) \frac{\partial V^*(s)}{\partial s}\right) \tag{23}$$

$$d^*(s) = \frac{1}{2\gamma^2} K^{\mathrm{T}}(s) \frac{\partial V^*(s)}{\partial s} \tag{24}$$

where (u^*, d^*) is termed as Nash equilibrium for zero-sum game. Inserting the optimal control policy and disturbance term (23) in (17) results in [55]

$$\Theta(u^*) = \lambda \left[\frac{\partial V^*(s)}{\partial s}\right]^{\mathrm{T}} G(s) \tanh(D^*) + \lambda^2 \bar{R} \ln\left[1 - \tanh^2(D^*)\right] \tag{25}$$

where $D^* = (1/2\lambda) R^{-1} G(s)^{\mathrm{T}} \frac{\partial V^*(s)}{\partial s}$. Inserting the Nash equilibrium (u^*, d^*) into (19) and using (25), the Bellman equation becomes the Hamilton-Jacobi-Isaacs (HJI) equation

$$\begin{aligned} 0 = {} & Q(s) + \left[\frac{\partial V^*(s)}{\partial s}\right]^{\mathrm{T}} F(s) + \lambda^2 \bar{R} \ln\left[1 - \tanh^2(D^*)\right] \\ & + \frac{1}{4\gamma^2} \left[\frac{\partial V^*(s)}{\partial s}\right]^{\mathrm{T}} K(s) K(s)^{\mathrm{T}} \frac{\partial V^*(s)}{\partial s} \end{aligned} \tag{26}$$

The following assumption on the cost function (15), which has been widely used in [14, 15], is employed in this paper.

Assumption 2. The performance functional (15) satisfies zero-state observability.

The following lemma discusses the equivalence between Problems 1 and 2

Lemma 3. *Suppose that the pair of policy $\{u^*(\cdot), d^*(\cdot)\}$ solve Problem 2 for system* (14)*. Then, the optimal control policy $\{u^*(\cdot)\}$ also solves Problem 1 provided that the initial state x_0 of system* (3) *satisfies the constraints in* (5)*.*

Under Assumptions 1 and 2, suppose that $\mu^* = \{u^*, d^*\}$ solves Problem 2 for system (14) *with performance* (15) *and reward* (16)*, then the following hold:*

1) *The closed-loop system satisfies the constraints* (5) *provided that the initial state x_0 of system* (3) *is within the region (a_i, A_i), $\forall i = 1, ..., n$.*

2) The disturbance attenuation condition (6) *can be guaranteed if the performance output* $L(x,u)$ *is designed as*

$$L(x,u) = Q(s) + \Theta(u).$$

Proof. 1) Based on Assumptions 1 and 2, the existence of a positive definite and continuously differentiable optimal value function $V^*(s)$ can be guaranteed. From (18), one can obtain that $\dot{V}^*(t) \leqslant 0$, i.e.,

$$V^*(s(t)) \leqslant V^*(s(0)), \forall t \geqslant 0.$$

Then, $V^*(s(t))$ remains bounded if $V^*(s(0))$ is bounded, which is guaranteed by the condition that the initial condition $x(0)$ of system (3) satisfies the constraints in (5). Finally, from the discussions in Remark 2, one can infer that

$$x_i(t) \in (a_i, A_i), \ i = 1, 2, \cdots, n.$$

Therefore, given $\mu^* = \{u_1^*, u_2^*\}$, the constraints of Problem 1 are satisfied.

2) Now consider the barrier-function-based state transformation described by (10). Then, each element of the state $s = \begin{bmatrix} b_1(x_1) & \cdots & b_n(x_n) \end{bmatrix}^{\mathrm{T}}$ is finite given that x satisfies the constraints given in (5). Note that the Nash equilibrium (u^*, d^*) and the optimal value function V^* satisfies the Bellman equation (19), i.e., $H\left(s, u^*, d^*, \frac{\partial V^*}{\partial s}\right) = 0$. Then, considering (16) and the performance output $L(x,u)$, one has,

$$H\left(s, u^*, d^*, \frac{\partial V^*}{\partial s}\right) = 0 \Rightarrow \frac{\int_t^\infty \|z(\tau)\|^2 d\tau}{\int_t^\infty \|d(\tau)\|^2 d\tau} \leqslant \gamma^2$$

provided that $L(x,u) = Q(s) + \Theta(u)$. This completes the proof. ∎

Remark 3. As shown in (25), the optimal constrained control and disturbance solution $u^*(s)$ and $d^*(s)$ depend on solving the HJI equation (26) for the optimal value function $V^*(s)$. However, the HJI equation (26) 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 HJI equation (26).

4. Online Actor-Critic-Barrier Learning

As shown in Lemma 3, with the barrier-function-based system transformation (10), the equivalence between Problems 1 and 2 can be guaranteed. In this section, we present a novel barrier-actor-critic online algorithm to learn the optimal control policy and the worst disturbance with

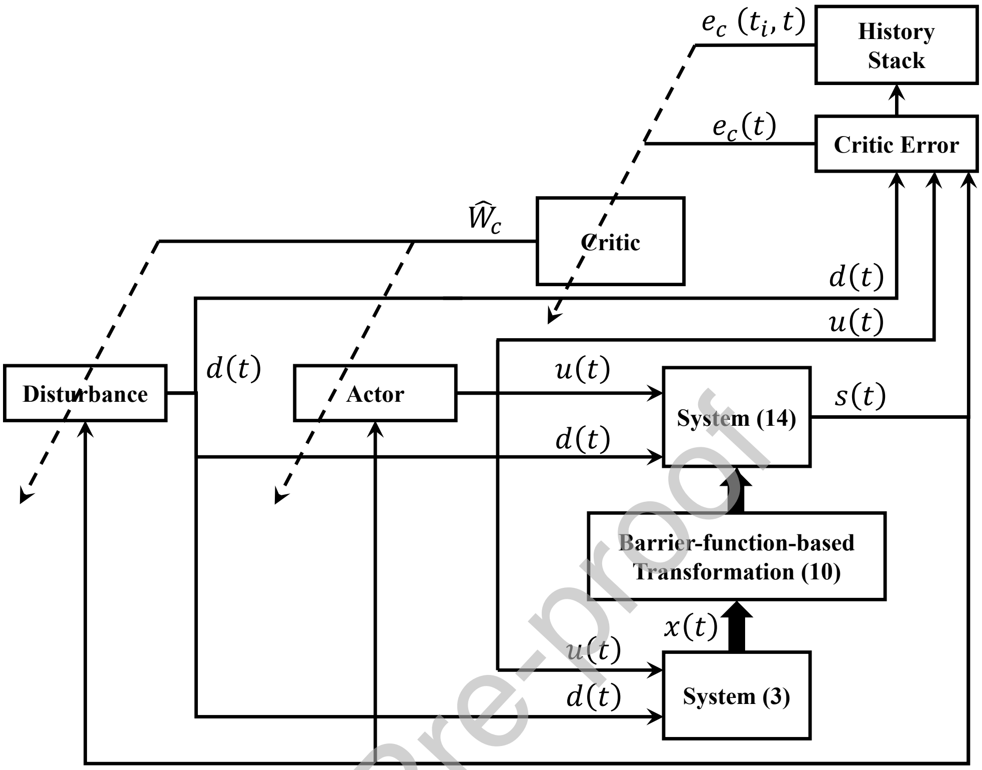


Figure 2: The overall barrier-actor-critic algorithm for disturbance attenuation with input saturation and full-state constraints. 1) Based on the barrier function defined in Definition 4, a novel system transformation is applied to original system (3) to obtain the transformed system (14). 2) The barrier-function-based system transformation is then combined with the actor-critic online algorithm to learn the optimal control policy $u*$ and worst-case disturbance $d*$. 3) To obviate the requirement of PE condition for online critic learning, the experience replay technique is employed to concurrently utilize the online and history data.

respect to the performance of Problem 2. First, value function approximation for the critic learning is represented by using neural networks. Online critic learning is designed to approximate the HJI equation (26). In addition, two actor NNs are designed to learn the optimal control policy (23) and the worst-case disturbance (24), respectively.

4.1. Value Function Approximation

Using the NN approximation theorem, there exists a single-layer NN such that the value function $V(s)$ and its gradient $\nabla V\,(s)$ can be uniformly approximated with a critic NN as the number of basis sets increases, within a compact set $\Omega \subseteq \mathbb{R}^n$ that contains the origin, as

$$V^*(s) = (W^*)^{\mathrm{T}} \phi(s) + \varepsilon(s) \tag{27}$$

$$\nabla V^*(s) = [\nabla \phi(s)]^{\mathrm{T}} W^* + \nabla \varepsilon(s) \tag{28}$$

where $W^* \in \mathbb{R}^N$ is an ideal weight vector for the best N-dimensional value function approximation, $\phi(\cdot) : \mathbb{R}^n \to \mathbb{R}^N$ is the NN basis function, $\nabla = \partial/\partial s$, $\varepsilon(s)$ and $\nabla\varepsilon(s)$ are the NN approximation residual. For the value function approximation (27) and (28), the following standard assumption is adopted in this paper.

Assumption 3. The value function approximation as shown in (27) and (28) are assumed to have the following properties.

1) The ideal weight W is bounded by a constant, i.e., $\|W^*\| \leqslant b_*$;
2) The value function approximation residual ε and $\nabla\varepsilon$ satisfies $\|\varepsilon(s)\| \leqslant b_\varepsilon$ and $\|\nabla\varepsilon(s)\| \leqslant b_{\mathrm{d}\varepsilon}$;
3) The NN basis function $\phi(s)$ and its gradient $\nabla\phi(s)$ satisfies $\|\phi(s)\| \leqslant b_\phi$ and $\|\nabla\phi(s)\| \leqslant b_{\mathrm{d}\phi}$ for $\forall s \in \Omega$.

For the optimal control policy $u^*(s)$ and the optimal disturbance inputs $d^*(s)$, the Bellman equation (19) approximation error using the value function approximation (27) can be expressed as

$$\xi = U(s, u^*, d^*) + (W^*)^{\mathrm{T}} \sigma \tag{29}$$

where σ is a N-dimensional vector signal defined as

$$\sigma = \nabla\phi(s)\left[F(s) + G(s)u^* + K(s)d^*\right] \tag{30}$$

Considering the value gradient approximation (28), one can obtain that the Bellman residual results from the value gradient approximation error $\nabla\epsilon(s)$, i.e.,

$$\xi = -[\nabla\varepsilon(s)]^{\mathrm{T}}\left[F(s) + G(s)u^* + K(s)d^*\right] \tag{31}$$

Similarly, with the value function approximation (27), the HJI equation (26) can be approximated with a residual expressed as

$$\begin{aligned} \zeta &= Q(s) + (W^*)^{\mathrm{T}}\sigma + \Theta(-\lambda\tanh(D_u)) - \frac{1}{4\gamma^2}(W^*)^{\mathrm{T}} D_d W^* \\ &= Q(s) + (W^*)^{\mathrm{T}}\nabla\phi(s)F(s) + \lambda^2\bar{R}\ln\left(1 - \tanh^2(D_u^*)\right) + \frac{1}{4\gamma^2}(W^*)^{\mathrm{T}} D_d W^* \end{aligned} \tag{32}$$

with

$$\begin{aligned} D_u^* &= \frac{1}{2\lambda} R^{-1} G^{\mathrm{T}}(s) [\nabla\phi(s)]^{\mathrm{T}} W^* \\ D_d &= \nabla\phi(s) K(s) K^{\mathrm{T}}(s) [\nabla\phi(s)]^{\mathrm{T}} \end{aligned} \tag{33}$$

Remark 4. From Assumptions 1 and 3, the policy representations in (23) and (24), the Bellman equation approximation residual ξ is bounded in the sense that there exists a constant b_ξ such that $\|\xi\| \leqslant b_\xi$. Similarly, the HJI approximation residual ζ using the ideal N-dimensional value function approximation (27) and (28) is bounded as $\zeta \leqslant b_\zeta$.

4.2. Critic Learning

The ideal weight, W in (27), provides the best approximate to the optimal value function $V^*(s)$ 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

$$\hat{V}(s) = \hat{W}_c^{\mathrm{T}} \phi_c(s) \tag{34}$$

$$\nabla\hat{V}(s) = [\nabla\phi_c(s)]^{\mathrm{T}} \hat{W}_c \tag{35}$$

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

$$\begin{aligned} e_c(t) &= U(s(t), u(t), d(t)) + \hat{W}_c^{\mathrm{T}} \sigma(t) \\ &= -(\nabla\varepsilon)^{\mathrm{T}} [F(s(t)) + G(s(t)) u(t) + K(s(t)) d(t)] \end{aligned} \tag{36}$$

Define the critic weight approximation error as

$$\tilde{W}_c = W^* - \hat{W}_c \tag{37}$$

Then, from (29), the relation between Bellman residual e_c and the Bellman equation approximation error ζ can be written in terms of the critic weight error $\tilde{W}_c$ as

$$e_c(t) = \xi(t) - \tilde{W}_c^{\mathrm{T}}(t)\sigma(t) \tag{38}$$

$$e_c(t_i, t) = \xi(t_i) - \tilde{W}_c^{\mathrm{T}}(t)\sigma(t_i) \tag{39}$$

Then $e_c \to \xi$ as $\hat{W}_c \to W^*$. 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

$$E_c = \frac{1}{2}\left(\frac{[e_c(t)]^2}{(1+\sigma^{\mathrm{T}}(t)\sigma(t))^2} + \sum_{i=1}^{k} \frac{[e_c^2(t_i, t)]^2}{(1+\sigma^{\mathrm{T}}(t_i)\sigma(t_i))^2} \right) \tag{40}$$

Using the chain rule yields adaptive critic online learning as

$$\begin{aligned}\dot{\hat{W}}_c &= -\alpha_c \frac{\partial E_c}{\partial \hat{W}_c} \\ &= -\alpha_c \frac{\sigma(t)\, e_c(t)}{\left[1+\sigma^{\mathrm{T}}(t)\,\sigma(t)\right]^2} - \alpha_c \sum_{i=1}^{k} \frac{\sigma(t_i)\, e_c(t_i,t)}{\left[1+\sigma^{\mathrm{T}}(t_i)\,\sigma(t_i)\right]^2} \\ &= -\alpha_c \frac{\sigma(t)\left[\xi(t)-\sigma(t)^{\mathrm{T}}\tilde{W}_c(t)\right]}{\left[1+\sigma^{\mathrm{T}}(t)\,\sigma(t)\right]^2} - \alpha_c \sum_{i=1}^{k} \frac{\sigma(t_i)\left[\xi(t_i)-\sigma(t_i)^{\mathrm{T}}\tilde{W}_c(t)\right]}{\left[1+\sigma^{\mathrm{T}}(t_i)\,\sigma(t_i)\right]^2} \qquad (41)\end{aligned}$$

where $\alpha_c > 0$ is the critic learning rate.

Condition 1. The recorded data matrix $\begin{bmatrix} \sigma(t_1) & \cdots & \sigma(t_k) \end{bmatrix}$ is full column rank.

Theorem 1. *Let u be any given admissible control policy. Then, under Condition 1, the critic weight approximation error $\tilde{W}_c$ in* (37) *is UUB with the adaptive critic learning* (41).

Proof. Based on (37) and (41), the dynamics of $\tilde{W}_c$ can be expressed as

$$\dot{\tilde{W}}_c(t) = -N_1 \tilde{W}_c(t) + N_2 \qquad (42)$$

where

$$N_1 = \alpha_c \left(\frac{\sigma(t)\,\sigma(t)^{\mathrm{T}}}{\left[1+\sigma^{\mathrm{T}}(t)\,\sigma(t)\right]^2} + \sum_{i=1}^{k} \frac{\sigma(t_i)\,\sigma(t_i)^{\mathrm{T}}}{\left[1+\sigma^{\mathrm{T}}(t_i)\,\sigma(t_i)\right]^2} \right) \qquad (43)$$

$$N_2 = \alpha_c \left(\frac{\sigma(t)\,\xi(t)}{\left[1+\sigma^{\mathrm{T}}(t)\,\sigma(t)\right]^2} + \sum_{i=1}^{k} \frac{\sigma(t_i)\,\xi(t_i)}{\left[1+\sigma^{\mathrm{T}}(t_i)\,\sigma(t_i)\right]^2} \right) \qquad (44)$$

Note the fact that $\left\| \frac{y}{1+y^{\mathrm{T}}y} \right\| \leqslant \frac{1}{2}$ and $\left\| \frac{1}{1+y^{\mathrm{T}}y} \right\| \leqslant 1$ for arbitrary vector signal y. Then, from Remark 4, N_2 in (42) satisfies $\|N_2\| \leqslant \frac{\alpha_c}{2}(k+1)\, b_\xi$. Consider the following Lyapunov function:

$$V_c = \frac{1}{2\alpha_c} \tilde{W}_c^{\mathrm{T}}(t) \tilde{W}_c(t) \qquad (45)$$

By differentiating (45) along the critic weight error dynamics (42), one has

$$\dot{V}_c = -\tilde{W}_c^{\mathrm{T}}(t) \left(\frac{\sigma(t)\,\sigma^{\mathrm{T}}(t)}{\left[1+\sigma^{\mathrm{T}}(t)\,\sigma(t)\right]^2} + \Lambda \right) \tilde{W}_c(t) + \tilde{W}_c^{\mathrm{T}}(t) N_2 \qquad (46)$$

with

$$\Lambda = \sum_{i=1}^{k} \frac{\sigma(t_i)\,\sigma^{\mathrm{T}}(t_i)}{\left(1+\sigma^{\mathrm{T}}(t_i)\,\sigma(t_i)\right)^2} > 0 \qquad (47)$$

which is guaranteed by Condition 1. Therefore, $\dot{V}_c$ is negative if

$$\left\| \tilde{W}_c(t) \right\| > \frac{\alpha_c (k+1)\, b_\xi}{2\lambda_{\min}(\Lambda)}$$

Then, the critic weight error $\tilde{W}_c$ converges to the residual set

$$\Omega_c = \left\{ \tilde{W}_c \middle| \left\| \tilde{W}_c(t) \right\| > \frac{\alpha_c (k+1) b_\xi}{2\lambda_{\min}(\Lambda)} \right\}$$

Therefore, the critic weight error $\tilde{W}_c$ is UUB. This completes the proof. ■

Remark 5. In contrast to the stability analysis as in [37, 55] where the PE condition on the signal is required for the signal $\sigma(t)$, in this paper, only Condition 1 is required to be satisfied for the signal $\sigma(t_i)$ in the history stack. Note that Condition 1 is weaker than the traditional PE condition and is easier to be checked for online implementation.

4.3. Actor and Disturbance Learning

As shown in (23) and (24), the optimal control policy and disturbance depend on the optimal value gradient $\frac{\partial V^*(s)}{\partial s}$. Therefore, consider the value gradient approximation with the critic weight $\hat{W}_c$ in (35), the control and disturbance policies can be determined using the critic weight as

$$\begin{aligned} u_c(s) &= -\lambda \tanh\left(\hat{D}_c\right) && (48) \\ \hat{D}_c &= \frac{1}{2\lambda} R^{-1} G^{\mathrm{T}} (\nabla\phi)^{\mathrm{T}} \hat{W}_c && (49) \\ d_c(s) &= \frac{1}{2\gamma^2} K^{\mathrm{T}} (\nabla\phi)^{\mathrm{T}} \hat{W}_c && (50) \end{aligned}$$

However, this policy improvement does not guarantee the stability of the closed-loop system [36, 37, 47, 55]. Therefore, to ensure the closed-loop stability, the policy applied to the system is implemented by alternative approximators using actor and disturbance network as

$$\begin{aligned} u_a(s) &= -\lambda \tanh\left(\hat{D}_u\right) && (51) \\ \hat{D}_u &= \frac{1}{2\lambda} R^{-1} G^{\mathrm{T}} (\nabla\phi)^{\mathrm{T}} \hat{W}_u && (52) \\ d_a(s) &= \frac{1}{2\gamma^2} K^{\mathrm{T}} (\nabla\phi)^{\mathrm{T}} \hat{W}_d && (53) \end{aligned}$$

where $\hat{W}_u$ is the actor network weight and $\hat{W}_d$ is the disturbance network weight. Define the weight estimation errors for the actor and the disturbance as,

$$\tilde{W}_u = W^* - \hat{W}_u, \quad \tilde{W}_d = W^* - \hat{W}_d \tag{54}$$

The actor network is designed to minimize the objective function

$$E_u = \frac{1}{2} e_u^{\mathrm{T}} R e_u \tag{55}$$

where

$$e_u = u_a - u_c = \lambda \left[\tanh\left(D_c\right) - \tanh\left(D_a\right)\right] \tag{56}$$

denotes the difference between the actor u_a (51) applied to the system and the control input u_c (48). Applying the actor (51) and disturbance (53) to the system (14) yields the closed-loop dynamics

$$\begin{aligned} \dot{s}\left(t\right) &= \sigma_a\left(t\right) \\ &= F\left(s\right) - G\left(s\right)\lambda \tanh\left(\hat{D}_u\right) + \frac{1}{2\gamma^2}K\left(s\right)K\left(s\right)^{\mathrm{T}}\left[\nabla\phi\left(s\right)\right]^{\mathrm{T}}\hat{W}_d \end{aligned} \tag{57}$$

Define

$$\begin{aligned} \xi_1 &= \left[\frac{\sigma_a\sigma_a^{\mathrm{T}}}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}\sigma_{ai}^{\mathrm{T}}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2}\right] \\ \xi_2 &= \left[\frac{\sigma_a}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2}\right] \\ \xi_3 &= \frac{\sigma_a\pi}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}\pi_i}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2} \\ \xi_4 &= -\frac{\alpha_c}{4\gamma^2}\left[\frac{\sigma_a}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2}\right] \\ \psi\left(t\right) &= \nabla\phi G\lambda\left[\tanh\left(\frac{\hat{D}_u}{\rho}\right) - \tanh\left(\hat{D}_u\right)\right] \\ \pi\left(t\right) &= W^{\mathrm{T}}\nabla\phi G\lambda\left[\tanh\left(\frac{D_u^*}{\rho}\right) - \tanh\left(\frac{\hat{D}_u}{\rho}\right)\right] + \varepsilon_J \end{aligned} \tag{58}$$

where $\sigma_{ai} = \sigma_a(t_i)$ and $\pi_i = \pi(t_i)$. Then, the stability and convergence of all the signals in the closed-loop system with the barrier-actor-disturbance learning algorithm is discussed in the following theorem.

Theorem 2. *Consider the dynamical system* (14) *with the critic* (34)*, the actor* (51)*, the disturbance input* (53) *with the design parameters in* (58) *and the following adaptive learning rules for the critic weight $\hat{W}_c$, actor weight $\hat{W}_u$ and disturbance $\hat{W}_d$, respectively,*

$$\begin{aligned} \dot{\hat{W}}_c &= -\alpha_c\frac{\sigma_a\left(t\right)\left[U\left(s\left(t\right),u_a,d_a\right) + \hat{W}_c^{\mathrm{T}}\sigma_a\left(t\right)\right]}{\left(1+\sigma_a^{\mathrm{T}}\left(t\right)\sigma_a\left(t\right)\right)^2} \\ &\quad -\alpha_c\sum_{i=1}^{k}\frac{\sigma_a\left(t_i\right)\left[U\left(s\left(t_i\right),u_a\left(t_i\right),d_a\left(t_i\right)\right) + \hat{W}_c^{\mathrm{T}}\left(t\right)\sigma_a\left(t_i\right)\right]}{\left(1+\sigma_a^{\mathrm{T}}\left(t_i\right)\sigma_a\left(t_i\right)\right)^2} \end{aligned} \tag{59}$$

$$\dot{\hat{W}}_u = -\alpha_u\left[Y_u\hat{W}_u + \nabla\phi G e_u + \nabla\phi G\tanh^2\left(\hat{D}_u\right)e_u\right], \tag{60}$$

$$\dot{\hat{W}}_d = -\alpha_d\left(Y_{d1}\hat{W}_d - Y_{d2}\hat{W}_c + D_d\hat{W}_d\xi_4^{\mathrm{T}}\hat{W}_c\right) \tag{61}$$

where $\alpha_c \in \mathbb{R}$*,* $\alpha_u \in \mathbb{R}$ *and* $\alpha_d \in \mathbb{R}$ *are the learning rate for the critic, actor and disturbance networks,* $Y_u \in \mathbb{R}^{N\times N}$*,* $Y_{d1} \in \mathbb{R}^{N\times N}$ *and* $Y_{d2} \in \mathbb{R}^{N\times N}$ *are the feedback gains for the actor and disturbance networks. Then, the augmented state* $X = \begin{bmatrix} s^{\mathrm{T}} & \tilde{W}_c^{\mathrm{T}} & \tilde{W}_u^{\mathrm{T}} & \tilde{W}_d^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ *is UUB provided that the design parameters are selected such that*

$$\begin{aligned}
&q > 0 \\
&-\xi_1 + \frac{r_c}{2}\xi_2\xi_2^{\mathrm{T}} + \frac{1}{2r_{d1}}I + \frac{r_{d2}}{2}\xi_4\xi_4^{\mathrm{T}} < 0 \\
&\frac{1}{2r_c}\psi\psi^{\mathrm{T}} - Y_u < 0 \\
&Y_{d1} + D_d\xi_4^{\mathrm{T}}W^* + \frac{r_{d1}}{2}Y_{d2}Y_{d2}^{\mathrm{T}} + \frac{1}{2r_{d2}}D_dW^*(W^*)^{\mathrm{T}}D_d^{\mathrm{T}} < 0
\end{aligned} \tag{62}$$

where r_c*,* r_{d1} *and* r_{d2} *are positive constants to be determined.*

Proof. Consider the following Lyapunov candidate function:

$$J(X) = V^*(s) + V_c\left(\tilde{W}_c\right) + V_u\left(\tilde{W}_u\right) + V_d\left(\tilde{W}_d\right) \tag{63}$$

where $V^*(\cdot)$ is the optimal value function satisfying the HJI equation and

$$V_c(s) = \frac{1}{2}\tilde{W}_c^{\mathrm{T}}\alpha_c^{-1}\tilde{W}_c, \quad V_u(s) = \frac{1}{2}\tilde{W}_u^{\mathrm{T}}\alpha_u^{-1}\tilde{W}_u, \quad V_d(s) = \frac{1}{2}\tilde{W}_d^{\mathrm{T}}\alpha_d^{-1}\tilde{W}_d$$

The derivative of the Lyapunov function (63) is given by

$$\dot{J} = \dot{V}^* + \dot{V}_c + \dot{V}_u + \dot{V}_d \tag{64}$$

For the first term of (64), one has

$$\begin{aligned}
\dot{V}^* &= \left[(W^*)^{\mathrm{T}}\nabla\phi + (\nabla\varepsilon)^{\mathrm{T}}\right]\left[F(s) + G(s)u_a + K(s)d_a\right] \\
&= (W^*)^{\mathrm{T}}\nabla\phi F - (W^*)^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right) + \frac{1}{2\gamma^2}(W^*)^{\mathrm{T}}D_d\hat{W}_d + \varepsilon_0
\end{aligned} \tag{65}$$

with D_d is defined in (33) and

$$\begin{aligned}
\varepsilon_0 &= (\nabla\varepsilon)^{\mathrm{T}}\sigma_a \\
\sigma_a &= F(s) - G(s)\lambda\tanh\left(\hat{D}_u\right) + \frac{1}{2\gamma^2}K(s)K(s)^{\mathrm{T}}[\nabla\phi(s)]^{\mathrm{T}}\hat{W}_d
\end{aligned} \tag{66}$$

Based on Assumptions 1 and 3 and Remark 4, ε_0 can be upper bounded as

$$\varepsilon_0 \leqslant b_{d\varepsilon}b_f\|s\| + b_{d\varepsilon}b_g\lambda + \frac{1}{2\gamma^2}b_{d\varepsilon}b_k^2b_{d\phi}b_* - \frac{1}{2\gamma^2}(\nabla\varepsilon)^{\mathrm{T}}K(s)K(s)^{\mathrm{T}}[\nabla\phi(s)]^{\mathrm{T}}\tilde{W}_d \tag{67}$$

From (25) and (32), one has

$$\begin{aligned}
(W^*)^{\mathrm{T}}\nabla\phi F &= -Q(s) - \Theta(-\lambda\tanh(D_u^*)) + (W^*)^{\mathrm{T}}\nabla\phi G\lambda\tanh(D_u^*) \\
&\quad -\frac{1}{4\gamma^2}(W^*)^{\mathrm{T}}D_dW^* + \zeta
\end{aligned}$$

Journal Pre-proof

with

$$\Theta\left(-\lambda\tanh\left(D_u^*\right)\right) = \left(W^*\right)^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_u^*\right)+\lambda^2\bar{R}\ln\left(1-\tanh^2\left(D_u^*\right)\right) \tag{68}$$

where D_u has been defined as in (33). Inserting $\left(W^*\right)^{\mathrm{T}}\nabla\phi F$ and (68) into (65) yields

$$\begin{aligned}\dot{V}^* = & -Q\left(s\right)-\Theta\left(-\lambda\tanh\left(D_u^*\right)\right)+\left(W^*\right)^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_u^*\right)\\ & -\frac{1}{4\gamma^2}\left(W^*\right)^{\mathrm{T}}D_dW^*-\left(W^*\right)^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)+\frac{1}{2\gamma^2}\left(W^*\right)^{\mathrm{T}}D_dW^*\\ & -\frac{1}{2\gamma^2}\left(W^*\right)^{\mathrm{T}}D_d\tilde{W}_d+\zeta+\varepsilon_0\end{aligned} \tag{69}$$

Since $Q(\cdot)$ and $\Theta(\cdot)$ are positive definite functions, then, there exists a positive constant $q>0$ such that

$$s^{\mathrm{T}}qs \leqslant Q(s) \leqslant Q(s)+\Theta\left(-\lambda\tanh\left(D_u^*\right)\right) \tag{70}$$

The third term in (69) can be upper bounded by

$$\left(W^*\right)^{\mathrm{T}}\nabla\phi\left(x\right)G\lambda\tanh\left(D_u^*\right) \leqslant \lambda b_g b_{d\phi} b_* \tag{71}$$

Considering $W^*=W_u+\tilde{W}_u$, then, the forth term in (69) can be rewritten as

$$\begin{aligned} & -\left(W^*\right)^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)\\ = & -\tilde{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)-\hat{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)\\ \leqslant & -\tilde{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)\end{aligned} \tag{72}$$

where the above inequality results from the fact that $\hat{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)=2\lambda^2\bar{R}\left[\hat{D}_u\tanh\left(\hat{D}_u\right)\right]$ and $x^{\mathrm{T}}\tanh\left(x\right)\geqslant 0$, for arbitrary vector signal x. Considering now the facts (67), (69), (70), (71) and (72), $\dot{V}^*$ further satisfies

$$\begin{aligned}\dot{V}^* \leqslant & -\tilde{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)-s^{\mathrm{T}}qs+b_{d\varepsilon}b_f\left\|s\right\|\\ & +\lambda b_g b_{d\phi}b_*+\frac{1}{4\gamma^2}b_*^2b_{d\phi}^2b_k^2+b_\zeta+\lambda b_{d\varepsilon}b_g+\frac{1}{2\gamma^2}b_{d\varepsilon}b_k^2b_{d\phi}b_*\\ & -\frac{1}{2\gamma^2}\left(\nabla\varepsilon\right)^{\mathrm{T}}K\left(s\right)K(s)^{\mathrm{T}}\left[\nabla\phi\left(s\right)\right]^{\mathrm{T}}\tilde{W}_d-\frac{1}{2\gamma^2}\left(W^*\right)^{\mathrm{T}}D_d\tilde{W}_d\\ = & -\tilde{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right)-s^{\mathrm{T}}qs+M_s\left\|s\right\|+N_s+M_{d1}\tilde{W}_d\end{aligned} \tag{73}$$

where

$$\begin{aligned} M_s &= b_{d\varepsilon}b_f\\ N_s &= \lambda b_g b_{d\phi}b_*+\frac{1}{4\gamma^2}b_*^2b_{d\phi}^2b_k^2+b_\zeta+\lambda b_{d\varepsilon}b_g+\frac{1}{2\gamma^2}b_{d\varepsilon}b_k^2b_{d\phi}b_*\\ M_{d1} &= -\frac{1}{2\gamma^2}\left\{\left(\nabla\varepsilon\right)^{\mathrm{T}}K\left(s\right)K(s)^{\mathrm{T}}\left[\nabla\phi\left(s\right)\right]^{\mathrm{T}}+\left(W^*\right)^{\mathrm{T}}D_d\right\}\end{aligned} \tag{74}$$

Second, for the critic weight error $\tilde{W}_c$, from (41) one has

$$\dot{\tilde{W}}_c(t) = \alpha_c \frac{\sigma_a}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} e_c(t) + \alpha_c \sum_{i=1}^{k} \frac{\sigma_{ai}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2} e_c(t_i, t) \tag{75}$$

where σ_a has been defined in (66) and $\sigma_{ai} = \sigma_a(t_i)$. Differentiating V_c along with (75), one has

$$\begin{aligned} \dot{V}_c &= \tilde{W}_c^{\mathrm{T}} \alpha_c^{-1} \dot{\tilde{W}}_c \\ &= \tilde{W}_c^{\mathrm{T}} \left[\frac{\sigma_a}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} e_c(t) + \sum_{i=1}^{k} \frac{\sigma_{ai}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2} e_c(t_i, t) \right] \end{aligned} \tag{76}$$

From (32), one has

$$-Q(s) - \Theta\left(-\lambda \tanh\left(D_u^*\right)\right) - \left(W^*\right)^{\mathrm{T}} \sigma + \frac{1}{4\gamma^2} \left(W^*\right)^{\mathrm{T}} D_d W^* + \zeta = 0. \tag{77}$$

Therefore, one can obtain

$$\begin{aligned} e_c &= Q(s) + \Theta\left(-\lambda \tanh\left(\hat{D}_u\right)\right) + \hat{W}_c^{\mathrm{T}} \sigma_a - \frac{1}{4\gamma^2} \hat{W}_d^{\mathrm{T}} D_d \hat{W}_d \\ &= Q(s) + \Theta\left(-\lambda \tanh\left(\hat{D}_u\right)\right) + \hat{W}_c^{\mathrm{T}} \sigma_a - \frac{1}{4\gamma^2} \hat{W}_d^{\mathrm{T}} D_d \hat{W}_d \\ &\quad -Q(s) - \Theta\left(-\lambda \tanh\left(D_u^*\right)\right) - \left(W^*\right)^{\mathrm{T}} \sigma + \frac{1}{4\gamma^2} \left(W^*\right)^{\mathrm{T}} D_d W^* + \zeta \end{aligned}$$

Adding and subtracting $\left(W^*\right)^{\mathrm{T}} \sigma_a$ to e_c yields

$$\begin{aligned} e_c &= \Theta\left(-\lambda \tanh\left(\hat{D}_u\right)\right) - \Theta\left(-\lambda \tanh\left(D_u^*\right)\right) - \tilde{W}_c^{\mathrm{T}} \sigma_a + \left(W^*\right)^{\mathrm{T}} \left(\sigma_a - \sigma\right) \\ &\quad - \frac{1}{4\gamma^2} \hat{W}_d^{\mathrm{T}} D_d \hat{W}_d + \frac{1}{4\gamma^2} \left(W^*\right)^{\mathrm{T}} D_d W^* + \zeta \end{aligned} \tag{78}$$

Moreover, note that

$$\begin{aligned} &\Theta\left(-\lambda \tanh\left(\hat{D}_u\right)\right) - \Theta\left(-\lambda \tanh\left(D_u^*\right)\right) \\ =\ &\lambda \hat{W}_a^{\mathrm{T}} \nabla\phi G \tanh\left(\hat{D}_u\right) + \lambda^2 \bar{R} \ln\left(1 - \tanh^2\left(\hat{D}_u\right)\right) \\ &-\lambda W^{\mathrm{T}} \nabla\phi G \tanh\left(D_u^*\right) - \lambda^2 \bar{R} \ln\left(1 - \tanh^2\left(D_u^*\right)\right) \end{aligned} \tag{79}$$

Note that the term $\lambda^2 \bar{R} \ln\left(1 - \tanh^2\left(D_u^*\right)\right)$ in (79) can be rewritten as

$$\lambda^2 \bar{R} \ln\left(1 - \tanh^2\left(D_u^*\right)\right) = \lambda^2 \bar{R} \left[\ln 4 - 2D_u^* - 2\ln\left(1 + e^{-2D_u^*}\right)\right], \tag{80}$$

where $-2\ln\left(1 + e^{-2D_u^*}\right)$ can be approximated using Lemma 1 as

$$-2\ln\left(1 + e^{-2D_u^*}\right) = 2D_u^* - 2D_u^* \operatorname{sgn}\left(D_u^*\right) + \varepsilon_{D_u^*}, \tag{81}$$

where $\left\|\varepsilon_{D_u^*}\right\| \leqslant \ln 4$. Then, inserting (81) into (80) yields

$$\lambda^2 R \ln\left(1-\tanh^2\left(D_u^*\right)\right)=\lambda^2 \bar{R}\left[\ln 4-2 D_u^* \operatorname{sgn}\left(D_u^*\right)+\varepsilon_{D_u^*}\right]. \tag{82}$$

Similarly,

$$\lambda^2 \bar{R} \ln\left(1-\tanh^2\left(\hat{D}_u\right)\right)=\lambda^2 R\left[\ln 4-2 \hat{D}_u \operatorname{sgn}\left(\hat{D}_u\right)+\varepsilon_{\hat{D}_u}\right], \tag{83}$$

where $\left\|\varepsilon_{\hat{D}_u}\right\| \leqslant \ln 4$. Consider (79), (82) and (83), one has

$$\begin{aligned} & \Theta\left(-\lambda \tanh\left(\hat{D}_u\right)\right)-\Theta\left(-\lambda \tanh\left(D_u^*\right)\right) \\ = & \lambda \hat{W}_a^{\mathrm{T}} \nabla \phi G \tanh\left(\hat{D}_u\right)-\lambda W^{\mathrm{T}} \nabla \phi G \tanh\left(D_u^*\right) \\ & +\lambda^2 \bar{R}\left[2 D_u^* \operatorname{sgn}\left(D_u^*\right)-2 \hat{D}_u \operatorname{sgn}\left(\hat{D}_u\right)+\varepsilon_{\hat{D}_u}-\varepsilon_{D_u^*}\right] \end{aligned} \tag{84}$$

The nonsmooth function sgn$(\cdot)$ in (84) can be approximated by the function tanh$(\cdot)$ by using Lemma 2. Then, based on (84), one has

$$\begin{aligned} & \lambda^2 \bar{R}\left(2 D_u^* \operatorname{sgn}\left(D_u^*\right)-2 \hat{D}_u \operatorname{sgn}\left(\hat{D}_u\right)\right) \\ = & \left(W^*\right)^{\mathrm{T}} \nabla \phi G \lambda \tanh\left(\frac{D_u^*}{\rho}\right)-\hat{W}_u^{\mathrm{T}} \nabla \phi G \lambda \tanh\left(\frac{\hat{D}_u}{\rho}\right)+\lambda^2 \bar{R} \varepsilon_\rho \end{aligned} \tag{85}$$

with approximation error satisfying $0 \leqslant \varepsilon_\rho \leqslant 2\kappa\rho$ where $\kappa=0.2785$ is defined in Lemma 2. Based on (84) and (85), adding and substracting $\left(W^*\right)^{\mathrm{T}} \nabla \phi G \lambda \tanh\left(\hat{D}_u\right)$, one has

$$\begin{aligned} e_c = & -\tilde{W}_c^{\mathrm{T}} \sigma_a+\tilde{W}_u^{\mathrm{T}} \nabla \phi G \lambda\left[\tanh\left(\frac{\hat{D}_u}{\rho}\right)-\tanh\left(\hat{D}_u\right)\right] \\ & +\left(W^*\right)^{\mathrm{T}} \nabla \phi G \lambda\left[\tanh\left(\frac{D_u^*}{\rho}\right)-\tanh\left(\frac{\hat{D}_u}{\rho}\right)\right]+\zeta+\lambda^2 \bar{R}\left(\varepsilon_{\hat{D}_u}-\varepsilon_{D_u^*}+\varepsilon_\rho\right)+\epsilon_c \end{aligned} \tag{86}$$

where $\epsilon_c=-\frac{1}{4\gamma^2} \hat{W}_d^{\mathrm{T}} D_d \hat{W}_d+\frac{1}{2\gamma^2}\left(W^*\right)^{\mathrm{T}} D_d \hat{W}_d-\frac{1}{4\gamma^2}\left(W^*\right)^{\mathrm{T}} D_d W^*$, which can be further rewritten as

$$\begin{aligned} \epsilon_c & =-\frac{1}{4\gamma^2} \hat{W}_d^{\mathrm{T}} D_d \hat{W}_d-\frac{1}{4\gamma^2}\left(W^*\right)^{\mathrm{T}} D_d W^*+\frac{1}{4\gamma^2}\left(W^*\right)^{\mathrm{T}} D_d \hat{W}_d+\frac{1}{4\gamma^2}\left(W^*\right)^{\mathrm{T}} D_d \hat{W}_d \\ & =\frac{1}{4\gamma^2} \tilde{W}_d^{\mathrm{T}} D_d \hat{W}_d-\frac{1}{4\gamma^2}\left(W^*\right)^{\mathrm{T}} D_d \tilde{W}_d \\ & =-\frac{1}{4\gamma^2} \tilde{W}_d^{\mathrm{T}} D_d \tilde{W}_d \end{aligned} \tag{87}$$

Denote $\varepsilon_J=\lambda^2 R\left(\varepsilon_{\hat{D}_u}-\varepsilon_{D_u^*}+\varepsilon_\rho\right)+\zeta$, one has

$$e_c(t)=-\tilde{W}_c^{\mathrm{T}}(t) \sigma_a(t)+\tilde{W}_a^{\mathrm{T}}(t) \psi(t)+\pi(t)-\frac{1}{4\gamma^2} \tilde{W}_d^{\mathrm{T}}(t) D_d \tilde{W}_d(t) \tag{88}$$

where ψ and π is defined in (58). Similarly,

$$e_c(t_i,t) = -\tilde{W}_c^{\mathrm{T}}(t)\,\sigma_a(t_i) + \tilde{W}_a^{\mathrm{T}}(t)\,\psi(t) + \pi(t_i) - \frac{1}{4\gamma^2}\tilde{W}_d^{\mathrm{T}}(t)\,D_d\tilde{W}_d(t) \tag{89}$$

where Based on Assumptions 1 and 3, both ψ and π are bounded. Substituting (88) and (89) into (75) yields,

$$\begin{aligned}\dot{\tilde{W}}_c \;=\; & -\alpha_c\left[\frac{\sigma_a\sigma_a^{\mathrm{T}}}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}\sigma_{ai}^{\mathrm{T}}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2}\right]\tilde{W}_c \\ & +\alpha_c\left[\frac{\sigma_a}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2}\right]\psi^{\mathrm{T}}\tilde{W}_a \\ & +\alpha_c\left[\frac{\sigma_a\pi}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}\pi_i}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2}\right] \\ & -\frac{\alpha_c}{4\gamma^2}\left[\frac{\sigma_a}{\left(1+\sigma_a^{\mathrm{T}}\sigma_a\right)^2} + \sum_{i=1}^{k}\frac{\sigma_{ai}}{\left(1+\sigma_{ai}^{\mathrm{T}}\sigma_{ai}\right)^2}\right]\tilde{W}_d^{\mathrm{T}}D_d\tilde{W}_d\end{aligned} \tag{90}$$

Substituting (90) into (76), one has

$$\begin{aligned}\dot{V}_c \;&=\; \tilde{W}_c^{\mathrm{T}}\alpha_c^{-1}\dot{\tilde{W}}_c \\ &= -\tilde{W}_c^{\mathrm{T}}\xi_1\tilde{W}_c + \tilde{W}_c^{\mathrm{T}}\xi_2\psi^{\mathrm{T}}\tilde{W}_u + \tilde{W}_c^{\mathrm{T}}\xi_3 + \tilde{W}_c^{\mathrm{T}}\xi_4\tilde{W}_d^{\mathrm{T}}D_d\tilde{W}_d \\ &\leqslant -\tilde{W}_c^{\mathrm{T}}\xi_1\tilde{W}_c + \frac{r_c}{2}\tilde{W}_c^{\mathrm{T}}\xi_2\xi_2^{\mathrm{T}}\tilde{W}_c + \frac{1}{2r_c}\tilde{W}_u^{\mathrm{T}}\psi\psi^{\mathrm{T}}\tilde{W}_u + \tilde{W}_c^{\mathrm{T}}\xi_3 + \tilde{W}_c^{\mathrm{T}}\xi_4\tilde{W}_d^{\mathrm{T}}D_d\tilde{W}_d \\ &= \tilde{W}_c^{\mathrm{T}}\left[-\xi_1 + \frac{r_c}{2}\xi_2\xi_2^{\mathrm{T}}\right]\tilde{W}_c + \frac{1}{2r_c}\tilde{W}_u^{\mathrm{T}}\psi\psi^{\mathrm{T}}\tilde{W}_u + \tilde{W}_c^{\mathrm{T}}\xi_3 + \tilde{W}_c^{\mathrm{T}}\xi_4\tilde{W}_d^{\mathrm{T}}D_d\tilde{W}_d\end{aligned} \tag{91}$$

where ξ_i for $i = 1,2,3,4$ has been defined in (58).

Next, we give the upper bound of $\dot{V}_u$. Based on (60), differentiating V_a yields

$$\begin{aligned}\dot{V}_u \;&=\; \tilde{W}_u^{\mathrm{T}}\alpha_u^{-1}\dot{\tilde{W}}_u \\ &= -\tilde{W}_u^{\mathrm{T}}\left[\nabla\phi G e_u + \nabla\phi G \tanh^2\left(\hat{D}_u\right)e_u + Y_u\hat{W}_u\right] \\ &= -\tilde{W}_u^{\mathrm{T}}Y_u\tilde{W}_u + \tilde{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right) + \tilde{W}_u^{\mathrm{T}}M_u \\ &\leqslant -\tilde{W}_u^{\mathrm{T}}Y_u\tilde{W}_u + \tilde{W}_u^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(\hat{D}_u\right) + {M_u}^{\mathrm{T}}\tilde{W}_u\end{aligned} \tag{92}$$

where $M_u = \left[-\nabla\phi G\lambda\tanh\left(\hat{D}_c\right) + \nabla\phi G\tanh^2\left(\hat{D}_u\right)e_u + Y_u W^*\right]$

Based on Assumption 1 - 3 and the definition of the actor learning error e_u in (56), M_u is also bounded.

For the derivative of V_d, according to (61) one has

$$\begin{aligned}\dot{V}_d &= -\tilde{W}_d^{\mathrm{T}} Y_{d1} W^* + \tilde{W}_d^{\mathrm{T}} Y_{d1} \tilde{W}_d + \tilde{W}_d^{\mathrm{T}} Y_{d2} W^* - \tilde{W}_d^{\mathrm{T}} Y_{d2} \tilde{W}_c \\ &\quad - \tilde{W}_d^{\mathrm{T}} D_d W^* \xi_4^{\mathrm{T}} W^* + \tilde{W}_d^{\mathrm{T}} D_d \tilde{W}_d \xi_4^{\mathrm{T}} W^* + \tilde{W}_d^{\mathrm{T}} D_d W^* \xi_4^{\mathrm{T}} \tilde{W}_c - \tilde{W}_d^{\mathrm{T}} D_d \tilde{W}_d \xi_4^{\mathrm{T}} \tilde{W}_c \\ &= \tilde{W}_d^{\mathrm{T}} Y_{d1} \tilde{W}_d + \tilde{W}_d^{\mathrm{T}} D_d \tilde{W}_d \xi_4^{\mathrm{T}} W^* - \tilde{W}_d^{\mathrm{T}} Y_{d1} W^* - \tilde{W}_d^{\mathrm{T}} D_d W^* \xi_4^{\mathrm{T}} W^* + \tilde{W}_d^{\mathrm{T}} Y_{d2} W^* \\ &\quad - \tilde{W}_d^{\mathrm{T}} Y_{d2} \tilde{W}_c + \tilde{W}_d^{\mathrm{T}} D_d W^* \xi_4^{\mathrm{T}} \tilde{W}_c - \tilde{W}_d^{\mathrm{T}} D_d \tilde{W}_d \xi_4^{\mathrm{T}} \tilde{W}_c \end{aligned} \tag{93}$$

Using Young's inequality to (93) yields

$$\begin{aligned}\dot{V}_d &\leqslant \tilde{W}_d^{\mathrm{T}} \left(Y_{d1} + D_d \xi_4^{\mathrm{T}} W^*\right) \tilde{W}_d + \tilde{W}_d^{\mathrm{T}} \left[Y_{d2} W^* - Y_{d1} W^* - D_d W^* \xi_4^{\mathrm{T}} W^*\right] \\ &\quad + \frac{r_{d1}}{2} \tilde{W}_d^{\mathrm{T}} Y_{d2} Y_{d2}^{\mathrm{T}} \tilde{W}_d + \frac{1}{2r_{d1}} \left\|\tilde{W}_c\right\|^2 + \frac{1}{2r_{d2}} \tilde{W}_d^{\mathrm{T}} D_d W^* (W^*)^{\mathrm{T}} D_d^{\mathrm{T}} \tilde{W}_d + \frac{r_{d2}}{2} \tilde{W}_c^{\mathrm{T}} \xi_4 \xi_4^{\mathrm{T}} \tilde{W}_c \\ &\quad - \tilde{W}_d^{\mathrm{T}} D_d \tilde{W}_d \xi_4^{\mathrm{T}} \tilde{W}_c \\ &= \frac{1}{2r_{d1}} \left\|\tilde{W}_c\right\|^2 + \frac{r_{d2}}{2} \tilde{W}_c^{\mathrm{T}} \xi_4 \xi_4^{\mathrm{T}} \tilde{W}_c - \tilde{W}_d^{\mathrm{T}} Q_d \tilde{W}_d + M_{d2}^{\mathrm{T}} \tilde{W}_d - \tilde{W}_d^{\mathrm{T}} D_d \tilde{W}_d \xi_4^{\mathrm{T}} \tilde{W}_c \end{aligned} \tag{94}$$

where

$$\begin{aligned} Q_d &= -\left[Y_{d1} + D_d \xi_4^{\mathrm{T}} W^* + \frac{r_{d1}}{2} Y_{d2} Y_{d2}^{\mathrm{T}} + \frac{1}{2r_{d2}} D_d W^* (W^*)^{\mathrm{T}} D_d^{\mathrm{T}}\right] \\ M_{d2} &= Y_{d2} W^* - Y_{d1} W^* - D_d W^* \xi_4^{\mathrm{T}} W^* \end{aligned} \tag{95}$$

Finally, collecting the results in (73), (91), (92) and (94), one has

$$\begin{aligned} \dot{J} &\leqslant -s^{\mathrm{T}} Q_s s + M_s \|s\| + N_s - \tilde{W}_c^{\mathrm{T}} Q_c \tilde{W}_c + \tilde{W}_c^{\mathrm{T}} \xi_3 \\ &\quad - \tilde{W}_u^{\mathrm{T}} Q_u \tilde{W}_u + M_u^{\mathrm{T}} \tilde{W}_u - \tilde{W}_d^{\mathrm{T}} Q_d \tilde{W}_d + \left(M_{d1} + M_{d2}^{\mathrm{T}}\right) \tilde{W}_d \\ &\leqslant -\lambda_{\min}(Q_s) \|s\|^2 + \|M_s\| \|s\| + \|N_s\| - \lambda_{\min}(Q_c) \left\|\tilde{W}_c\right\|^2 + \|\xi_3\| \left\|\tilde{W}_c\right\| \\ &\quad - \lambda_{\min}(Q_u) \left\|\tilde{W}_u\right\|^2 + \|M_u\| \left\|\tilde{W}_u\right\| - \lambda_{\min}(Q_d) \left\|\tilde{W}_d\right\|^2 + \left\|M_{d1} + M_{d2}^{\mathrm{T}}\right\| \left\|\tilde{W}_d\right\| \end{aligned} \tag{96}$$

where

$$\begin{aligned} Q_s &= qI \\ Q_c &= \xi_1 - \frac{r_c}{2} \xi_2 \xi_2^{\mathrm{T}} - \frac{1}{2r_{d1}} I - \frac{r_{d2}}{2} \xi_4 \xi_4^{\mathrm{T}} \\ Q_u &= -\frac{1}{2r_c} \psi \psi^{\mathrm{T}} + Y_u \end{aligned}$$

Based on Assumptions 1 and 3, M_s, M_u, M_{d1}, M_{d2} and N_s are bounded. Note that the parameters design in (62) guarantees that $Q_s > 0$, $Q_c > 0$, $Q_u > 0$ and $Q_d > 0$. Therefore, $\dot{J} < 0$ if

$$\begin{aligned} \|s\| &> \frac{\|s\|}{2\lambda_{\min}(Q_s)} + \sqrt{\frac{\|M_s\|^2}{4\lambda_{\min}^2(Q_s)} + \frac{\|N_s\|}{\lambda_{\min}(Q_s)}} \\ \left\|\tilde{W}_c\right\| &> \frac{\left\|\tilde{W}_c\right\|}{2\lambda_{\min}(Q_c)} + \sqrt{\frac{\|M_c\|^2}{4\lambda_{\min}^2(Q_c)}} \\ \left\|\tilde{W}_u\right\| &> \frac{\left\|\tilde{W}_u\right\|}{2\lambda_{\min}(Q_u)} + \sqrt{\frac{\|M_u\|^2}{4\lambda_{\min}^2(Q_u)}} \\ \left\|\tilde{W}_d\right\| &> \frac{\left\|\tilde{W}_d\right\|}{2\lambda_{\min}(Q_d)} + \sqrt{\frac{\|M_d\|^2}{4\lambda_{\min}^2(Q_d)}} \end{aligned} \tag{97}$$

Then, the augmented state $X = \begin{bmatrix} s^{\mathrm{T}} & \tilde{W}_c^{\mathrm{T}} & \tilde{W}_u^{\mathrm{T}} & \tilde{W}_d^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ converges to the residual set Ω_X defined as

$$\begin{aligned} \Omega_X = \{ X|\, \|s\| &< \frac{\|s\|}{2\lambda_{\min}(Q_s)} + \sqrt{\frac{\|M_s\|^2}{4\lambda_{\min}^2(Q_s)} + \frac{\|N_s\|}{\lambda_{\min}(Q_s)}}, \\ \left\|\tilde{W}_c\right\| &< \frac{\left\|\tilde{W}_c\right\|}{2\lambda_{\min}(Q_c)} + \sqrt{\frac{\|M_c\|^2}{4\lambda_{\min}^2(Q_c)}}, \left\|\tilde{W}_u\right\| < \frac{\left\|\tilde{W}_u\right\|}{2\lambda_{\min}(Q_u)} + \sqrt{\frac{\|M_u\|^2}{4\lambda_{\min}^2(Q_u)}}, \\ \left\|\tilde{W}_d\right\| &< \frac{\left\|\tilde{W}_d\right\|}{2\lambda_{\min}(Q_d)} + \sqrt{\frac{\|M_d\|^2}{4\lambda_{\min}^2(Q_d)}} \Bigg\} \end{aligned} \tag{98}$$

This completes the proof. ■

5. Simulation Study

To verify the effectiveness of the presented online safe RL algorithm with the actor-critic-barrier structure, we consider the following nonlinear systems of a single link robot arm

$$\ddot{\theta}(t) = -\frac{Mgl}{\tilde{G}} \sin(\theta(t)) - \frac{\tilde{D}}{\tilde{G}} \dot{\theta}(t) + \frac{1}{\tilde{G}} u(t) + kd(t) \tag{99}$$

where θ is the angle position, $\dot{\theta}$ is the angle velocity, M is the mass of the payload, g is the acceleration of gravity, l is the length of the arm, $\tilde{D}$ is the viscous friction and $\tilde{G}$ is the moment of inertia. In this experiment, $M = 10kg$, $g = 9.81m/s^2$, $l = 0.5m$, $\tilde{D} = 2N$ and $\tilde{G} = 10kgm^2$. Let $x_1{=}\theta$, $x_2 = \dot{\theta}$ and $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^{\mathrm{T}}$, then the dynamics of x can be written as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ f(x) \end{bmatrix} + \begin{bmatrix} 0 \\ g(x) \end{bmatrix} u + \begin{bmatrix} 0 \\ k(x) \end{bmatrix} d \tag{100}$$

where

$$\begin{aligned} f(x) &= -\frac{Mgl}{\tilde{G}}\sin(x_1) - \frac{\tilde{D}}{\tilde{G}}x_2 \\ g(x) &= \frac{1}{\tilde{G}}, k(x) = k \end{aligned}$$

For Problem 1, the performance output is selected as $L(x,u) = x^{\mathrm{T}}Hx + u^{\mathrm{T}}Ru$ with $H = 50I$, $R = 10I$. In addition, the following safety constraints are considered

$$x_i \in (a_i, A_i), \forall i \in \{1, 2\} \tag{101}$$

where $a_1 = -1.6$, $A_1 = 3$, $a_2 = -4$ and $A_2 = 3$. By using the classical actor-critic reinforcement learning algorithm, the state evolution with respect to time can be shown in Figure 3. The phase portrait of the state evolution in the state space is shown in Figure 5. As can be seen from Figure 3, the full-state constraints cannot be guaranteed by the classical actor-critic reinforcement learning algorithm. The evolution of the actor-critic-disturbance is shown in Figure 4.

To deal with the full-state constraints, the barrier-function-based system transformation (10) is employed. With the barrier function, one can obtain the transformed system as $\dot{s} = F(s) + G(s)u + K(s)d$ with

$$\begin{aligned} F(s) &= \begin{bmatrix} \frac{a_2A_2\left(e^{\frac{s_2}{2}} - e^{-\frac{s_2}{2}}\right)}{a_2e^{\frac{s_2}{2}} - A_2e^{-\frac{s_2}{2}}} \frac{A_1^2e^{-s_1} - 2a_1A_1 + a_1^2e^{s_1}}{A_1a_1^2 - a_1A_1^2} \\ f\left(B^{-1}(s)\right) \frac{A_2^2e^{-s_2} - 2a_2A_2 + a_2^2e^{s_2}}{A_2a_2^2 - a_2A_2^2} \end{bmatrix} \\ G(s) &= \begin{bmatrix} 0 \\ \frac{1}{\tilde{G}} \frac{A_2^2e^{-s_2} - 2a_2A_2 + a_2^2e^{s_2}}{A_2a_2^2 - a_2A_2^2} \end{bmatrix} \\ K(s) &= \begin{bmatrix} 0 \\ k\frac{A_2^2e^{-s_2} - 2a_2A_2 + a_2^2e^{s_2}}{A_2a_2^2 - a_2A_2^2} \end{bmatrix} \end{aligned} \tag{102}$$

with the initial condition

$$\begin{aligned} s_0 &= \begin{bmatrix} s_0(1) & s_0(2) \end{bmatrix}^{\mathrm{T}} \\ s_0(1) &= b(x_0(1); a_1, A_1), s_0(2) = b(x_0(2); a_2, A_2) \end{aligned}$$

Based on the actor-critic-barrier online learning algorithm, the state evolution of state $s(t)$ in system (102) is given in Figure 6. One can observe that the state $s(t)$ of system (102) converges to the origin asymptotically. Based on the state evolution of $s(t)$, by using the barrier function inverse mapping (10), one can obtain the state $x(t)$ as

$$x_1(t) = b^{-1}(s_1(t); a_1, A_1), x_2(t) = b^{-1}(s_2(t); a_2, A_2)$$

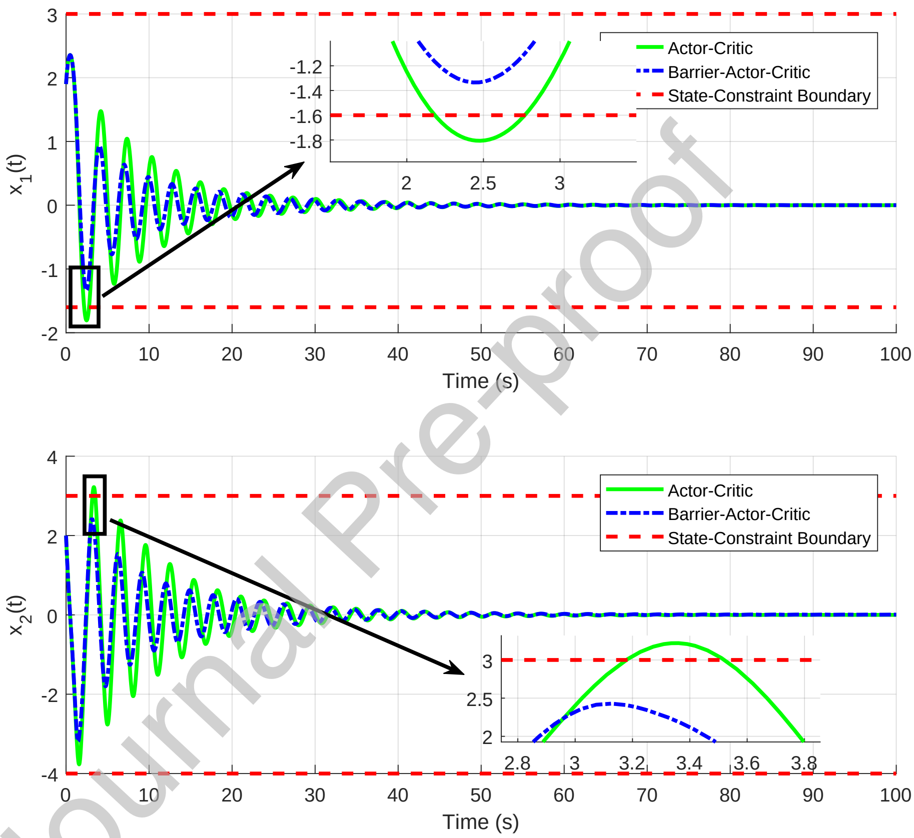


Figure 3: Evolution of the state $x(t)$ by using the presented actor-critic-barrier learning and classical actor-critic learning. The dashed line represents the boundary of the safe region.

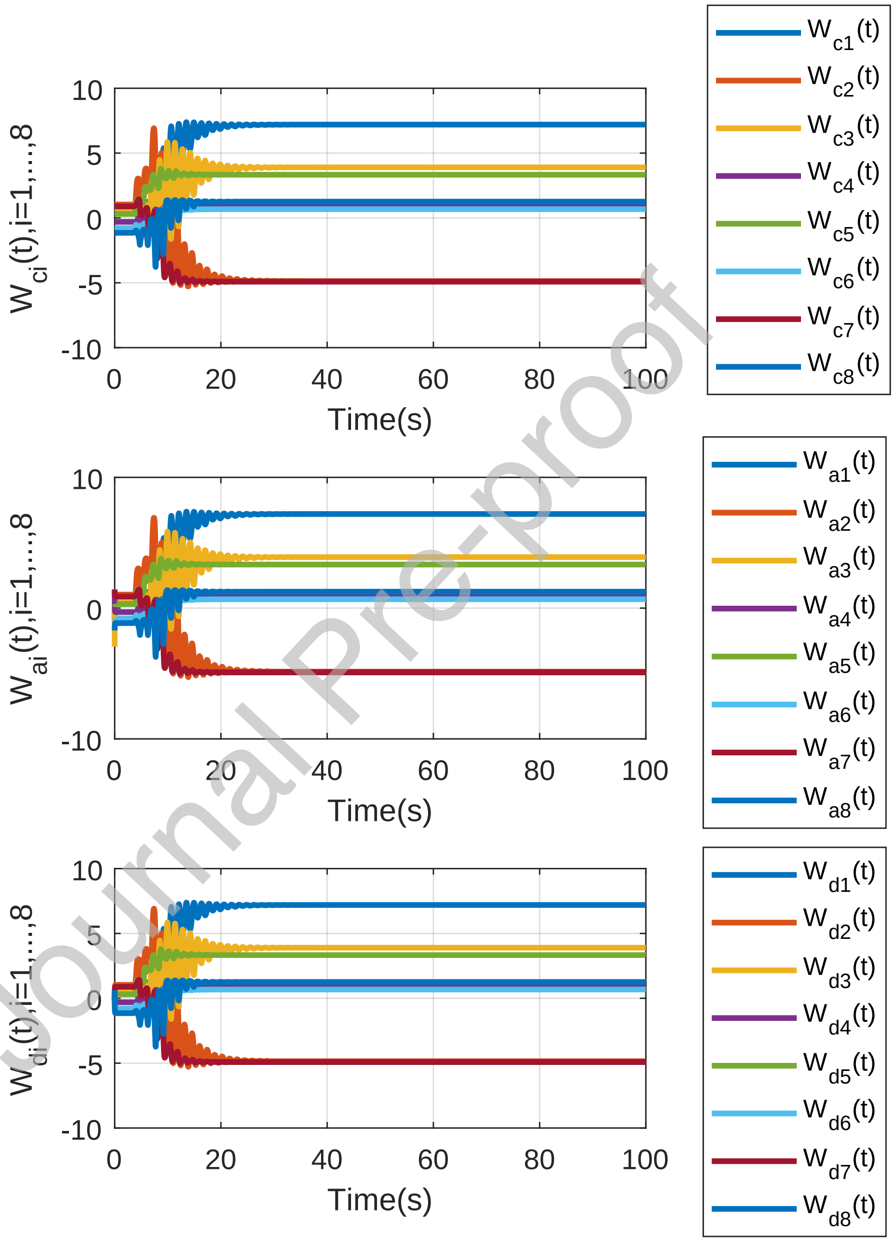


Figure 4: Evolution of the actor and critic weights using classical actor-critic learning.

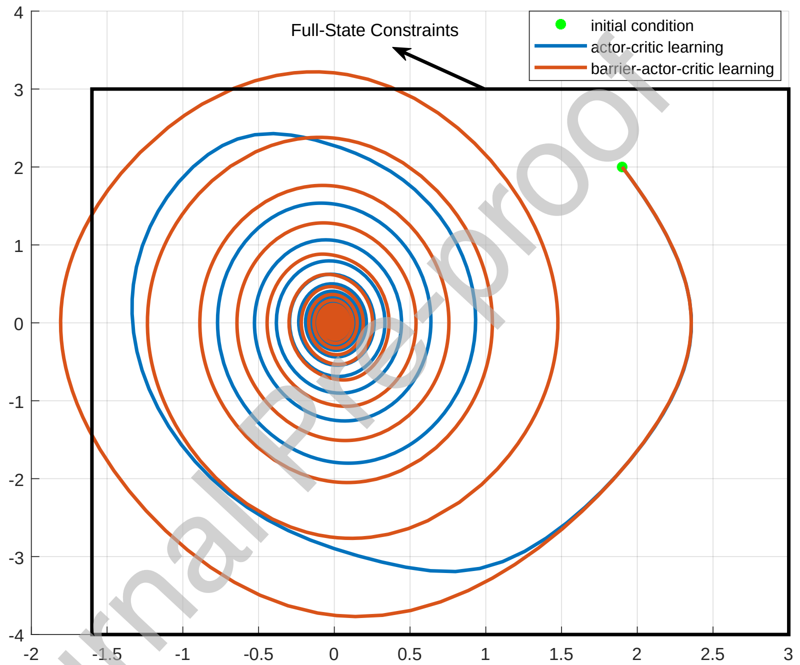


Figure 5: Evolution of the two-dimensional phase plot of the state trajectories $[x_1(t) \quad x_2(t)]$. The black box denotes the safe region.

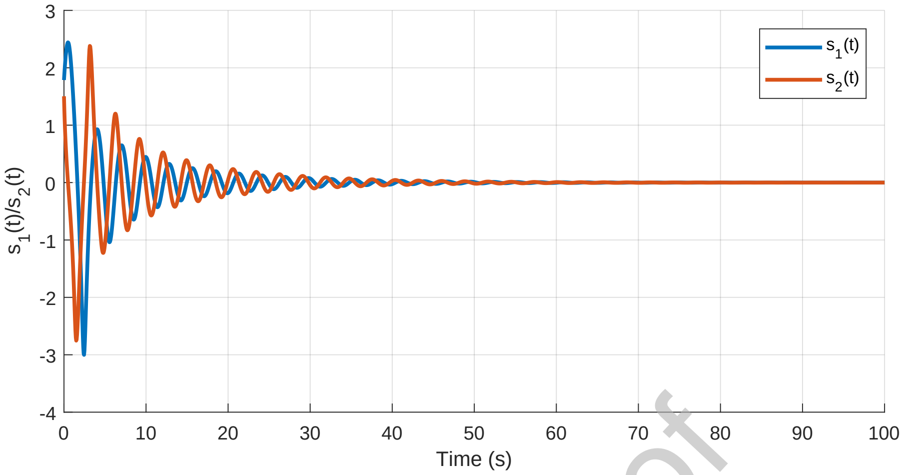


Figure 6: Evolution of the state $s(t)$ by using the presented actor-critic-barrier learning and classical actor-critic learning.

Then, the evolution of the state $x(t)$ is shown in Figure 3. The phase portrait of the state evolution $[x_1(t) \quad x_2(t)]$ is provided in Figure 5. The black box represents the full-state constraints. One can observe that with the barrier-actor-critic learning algorithm, the state evolution does not exceed the boundary of the prescribed region and full-state constraints can be guaranteed. That is, the state $x(t)$ converges to the origin asymptotically while satisfying the safety constraints (101). Finally, the learning process of the barrier-actor-critic networks is shown in Figure 7.

6. Conclusions

In this paper, the disturbance attenuation problem with both full-state constraints and input saturation is considered. An adaptive optimal controller design with the barrier-actor-critic algorithm is developed. First, a novel barrier function is defined to deal with full-state saturation. Based on this barrier function, a novel system transformation is applied to the original system to obtain the transformed system. Second, the barrier-function-based system transformation is then combined with the actor-critic online algorithm to learn the optimal control policy and the worst-case disturbance. To obviate the requirement of PE condition for online critic learning, the experience replay technique is employed to utilize the online and history data concurrently. The stability of the closed-loop system and the convergence of the actor-critic parameters to the optimal condition are discussed in the framework of Lyapunov analysis. The input saturation and

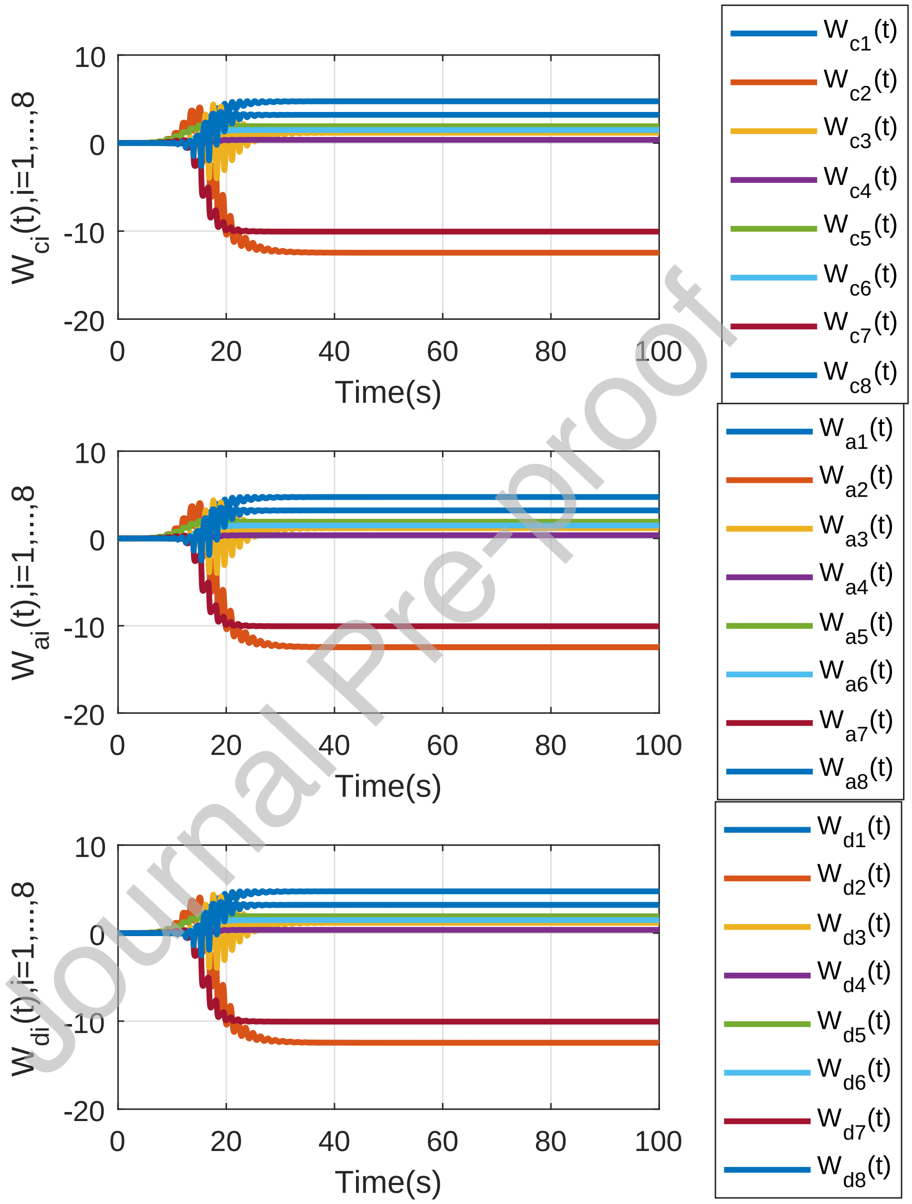


Figure 7: Evolution of the actor and critic weights using barrier-actor-critic learning.

full-state constraints are guaranteed to be satisfied during the learning phase. Finally, simulation studies are conducted to verify the efficacy of the presented barrier-actor-critic online learning.

References

[1] M. Rehan, C. K. Ahn, M. Chadli, Consensus of one-sided lipschitz multi-agents under input saturation, IEEE Transactions on Circuits and Systems II: Express Briefsdoi:10.1109/TCSII.2019.2923721.

[2] Z. Liu, Z. Zhao, C. K. Ahn, Boundary constrained control of flexible string systems subject to disturbances, IEEE Transactions on Circuits and Systems II: Express Briefsdoi:10.1109/TCSII.2019.2901283.

[3] Z. Zhao, Z. Liu, Z. Li, N. Wang, J. Yang, Control design for a vibrating flexible marine riser system, Journal of the Franklin Institute 354 (18) (2017) 8117 – 8133.

[4] R. R. Selmic, F. L. Lewis, Deadzone compensation in motion control systems using neural networks, IEEE Transactions on Automatic Control 45 (4) (2000) 602–613.

[5] W. He, B. Huang, Y. Dong, Z. Li, C. Su, Adaptive neural network control for robotic manipulators with unknown deadzone, IEEE Transactions on Cybernetics 48 (9) (2018) 2670–2682.

[6] Y. Liu, S. Lu, S. Tong, Neural network controller design for an uncertain robot with time-varying output constraint, IEEE Transactions on Systems, Man, and Cybernetics: Systems 47 (8) (2017) 2060–2068.

[7] Q. Zhou, L. Wang, C. Wu, H. Li, H. Du, Adaptive fuzzy control for nonstrict-feedback systems with input saturation and output constraint, IEEE Transactions on Systems, Man, and Cybernetics: Systems 47 (1) (2017) 1–12.

[8] R. R. Selmic, F. L. Lewis, Neural-network approximation of piecewise continuous functions: application to friction compensation, IEEE Transactions on Neural Networks 13 (3) (2002) 745–751.

[9] J. Na, Q. Chen, X. Ren, Y. Guo, Adaptive prescribed performance motion control of servo mechanisms with friction compensation, IEEE Transactions on Industrial Electronics 61 (1) (2014) 486–494.

[10] G. Tao, P. V. Kokotovic, Adaptive control of plants with unknown hystereses, IEEE Transactions on Automatic Control 40 (2) (1995) 200–212.

[11] M. Chen, S. S. Ge, Adaptive neural output feedback control of uncertain nonlinear systems with unknown hysteresis using disturbance observer, IEEE Transactions on Industrial Electronics 62 (12) (2015) 7706–7716.

[12] Z. Zhao, S. Lin, D. Zhu, G. Wen, Vibration control of a riser-vessel system subject to input backlash and extraneous disturbances, IEEE Transactions on Circuits and Systems II: Express Briefsdoi:10.1109/TCSII.2019.2914061.

[13] G. A. Rovithakis, Robust redesign of a neural network controller in the presence of unmodeled dynamics, IEEE Transactions on Neural Networks 15 (6) (2004) 1482–1490.

[14] J. C. Doyle, K. Glover, P. P. Khargonekar, B. A. Francis, State-space solutions to standard H_2 and H_∞ control problems, IEEE Transactions on Automatic Control 34 (8) (1989) 831–847.

[15] A. J. van der Schaft, l_2-gain analysis of nonlinear systems and nonlinear state-feedback h_∞ control, IEEE Transactions on Automatic Control 37 (6) (1992) 770–784.

[16] P. A. Ioannou, J. Sun, Robust Adaptive Control, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995.

[17] M. Krstic, P. V. Kokotovic, I. Kanellakopoulos, Nonlinear and Adaptive Control Design, 1st Edition, John Wiley & Sons, Inc., New York, NY, USA, 1995.

[18] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, NY, USA, 2004.

[19] K. P. Tee, S. S. Ge, E. H. Tay, Barrier lyapunov functions for the control of output-constrained nonlinear systems, Automatica 45 (4) (2009) 918 – 927.

[20] B. Ren, S. S. Ge, K. P. Tee, T. H. Lee, Adaptive neural control for output feedback nonlinear systems using a barrier lyapunov function, IEEE Transactions on Neural Networks 21 (8) (2010) 1339–1345.

[21] Y.-J. Liu, S. Lu, S. Tong, X. Chen, C. P. Chen, D.-J. Li, Adaptive control-based barrier lyapunov functions for a class of stochastic nonlinear systems with full state constraints, Automatica 87 (2018) 83 – 93.

[22] Y.-J. Liu, S. Tong, Barrier lyapunov functions-based adaptive control for a class of nonlinear pure-feedback systems with full state constraints, Automatica 64 (2016) 70 – 75.

[23] W. He, Y. Chen, Z. Yin, Adaptive neural network control of an uncertain robot with full-state constraints, IEEE Transactions on Cybernetics 46 (3) (2016) 620–629.

[24] D. Li, D. Li, Adaptive tracking control for nonlinear time-varying delay systems with full state constraints and unknown control coefficients, Automatica 93 (2018) 444 – 453.

[25] C. P. Bechlioulis, G. A. Rovithakis, Robust adaptive control of feedback linearizable mimo nonlinear systems with prescribed performance, IEEE Transactions on Automatic Control 53 (9) (2008) 2090–2099.

[26] A. K. Kostarigka, G. A. Rovithakis, Adaptive dynamic output feedback neural network control of uncertain mimo nonlinear systems with prescribed performance, IEEE Transactions on Neural Networks and Learning Systems 23 (1) (2012) 138–149.

[27] C. P. Bechlioulis, G. A. Rovithakis, Decentralized robust synchronization of unknown high order nonlinear multi-agent systems with prescribed transient and steady state performance, IEEE Transactions on Automatic Control 62 (1) (2017) 123–134.

[28] A. Theodorakopoulos, G. A. Rovithakis, Guaranteeing preselected tracking quality for uncertain strict-feedback systems with deadzone input nonlinearity and disturbances via low-complexity control, Automatica 54 (2015) 135 – 145.

[29] A. K. Kostarigka, Z. Doulgeri, G. A. Rovithakis, Prescribed performance tracking for flexible joint robots with unknown dynamics and variable elasticity, Automatica 49 (5) (2013) 1137 – 1147.

[30] Y. Yang, C. Ge, H. Wang, X. Li, C. Hua, Adaptive neural network based prescribed performance control for teleoperation system under input saturation, Journal of the Franklin Institute 352 (5) (2015) 1850 – 1866.

[31] E. Arabi, T. Yucelen, B. C. Gruenwald, M. Fravolini, S. Balakrishnan, N. T. Nguyen, A neuroadaptive architecture for model reference control of uncertain dynamical systems with performance guarantees, Systems & Control Letters 125 (2019) 37 – 44.

[32] F. L. Lewis, D. Vrabie, V. L. Syrmos, Optimal control, John Wiley & Sons, Hoboken, NJ, USA, 2012.

[33] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, F. L. Lewis, Optimal and autonomous control using reinforcement learning: A survey, IEEE Transactions on Neural Networks and Learning Systems 29 (6) (2018) 2042–2062.

[34] D. Liu, Q. Wei, Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems, IEEE Transactions on Neural Networks and Learning Systems 25 (3) (2014) 621–634.

[35] Y. Yang, D. Wunsch, Y. Yin, Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems, IEEE Transactions on Neural Networks and Learning Systems 28 (8) (2017) 1929–1940.

[36] Y. Yang, K. G. Vamvoudakis, H. Ferraz, H. Modares, Dynamic intermittent Q-learning-based model-free suboptimal co-design of L_2-stabilization, International Journal of Robust and Nonlinear Control 29 (9) (2019) 2673–2694.

[37] K. G. Vamvoudakis, F. L. Lewis, Online actorcritic algorithm to solve the continuous-time infinite horizon optimal control problem, Automatica 46 (5) (2010) 878 – 888.

[38] H. Modares, F. L. Lewis, Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning, IEEE Transactions on Automatic Control 59 (11) (2014) 3051–3056.

[39] H. Modares, F. L. Lewis, Z. Jiang, H_∞tracking control of completely unknown continuous-time systems via off-policy reinforcement learning, IEEE Transactions on Neural Networks and Learning Systems 26 (10) (2015) 2550–2562.

[40] Y. Yang, Z. Guo, H. Xiong, D. Ding, Y. Yin, 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 30 (12) (2019) 3735 – 3747.

[41] D. Wang, D. Liu, Learning and guaranteed cost control with event-based adaptive critic implementation, IEEE Transactions on Neural Networks and Learning Systems 29 (12) (2018) 6004–6014.

[42] D. Wang, Intelligent critic control with robustness guarantee of disturbed nonlinear plants, IEEE Transactions on Cybernetics.

[43] Y. Yang, H. Modares, D. C. Wunsch, Y. Yin, Optimal containment control of unknown heterogeneous systems with active leaders, IEEE Transactions on Control Systems Technology 27 (3) (2019) 1228–1236.

[44] Y. Yang, H. Modares, D. C. Wunsch, Y. Yin, Leaderfollower output synchronization of linear heterogeneous systems with active leader using reinforcement learning, IEEE Transactions on Neural Networks and Learning Systems 29 (6) (2018) 2139–2153.

[45] D. Wang, D. Liu, Neural robust stabilization via event-triggering mechanism and adaptive learning technique, Neural Networks 102 (2018) 27 – 35.

[46] D. Zhao, Q. Zhang, D. Wang, Y. Zhu, Experience replay for optimal control of nonzero-sum game systems with unknown dynamics, IEEE Transactions on Cybernetics 46 (3) (2016) 854–865.

[47] H. Modares, F. L. Lewis, 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 24 (10) (2013) 1513–1525.

[48] J. Sun, C. Liu, Disturbance observer-based robust missile autopilot design with full-state constraints via adaptive dynamic programming, Journal of the Franklin Institute 355 (5) (2018) 2344 – 2368.

[49] H. Modares, F. L. Lewis, M.-B. N. Sistani, Online solution of nonquadratic two-player zero-sum games arising in the h_∞ control of constrained input systems, International Journal of Adaptive Control and Signal Processing 28 (3-5) (2014) 232–254.

[50] M. Polycarpou, P. Ioannou, A robust adaptive nonlinear control design, Automatica 32 (3) (1996) 423 – 427.

[51] N. us Saqib, M. Rehan, M. Hussain, N. Iqbal, H. ur Rashid, Delay-range-dependent static anti-windup compensator design for nonlinear systems subjected to input-delay and saturation, Journal of the Franklin Institute 354 (14) (2017) 5919 – 5948.

[52] M. Hussain, M. Rehan, C. Ki Ahn, M. Tufail, Robust antiwindup for one-sided lipschitz systems subject to input saturation and applications, IEEE Transactions on Industrial Electronics 65 (12) (2018) 9706–9716.

[53] Z. Zhao, X. He, G. Wen, Boundary robust adaptive anti-saturation control of vibrating flexible riser systems, Ocean Engineering 179 (2019) 298 – 306.

[54] Z. Zhao, X. He, Z. Ren, G. Wen, Boundary adaptive robust control of a flexible riser system with input nonlinearities, IEEE Transactions on Systems, Man, and Cybernetics: Systems 49 (10) (2019) 1971–1980.

[55] H. Modares, F. L. Lewis, Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning, Automatica 50 (7) (2014) 1780 – 1792.

[56] T. Başar, P. Bernhard, H_∞ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach, Springer, Berlin, Germany, 2008.

Conflict of interest statement

The authors declared that they have no conflicts of interest to this work.
We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.