

RESEARCH ARTICLE

Safe reinforcement learning for dynamical games

Yongliang Yang¹  | Kyriakos G. Vamvoudakis²  | Hamidreza Modares³ 

¹Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

²Daniel Guggenheim School of Aerospace Engineering, Georgia Tech, Atlanta, Georgia, USA

³Mechanical Engineering Department, Michigan State University, East Lansing, Michigan, USA

Correspondence

Yongliang Yang, Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China.
Email: yangyongliang@ieee.org

Funding information

National Natural Science Foundation of China, Grant/Award Number: 61903028; China Post-Doctoral Science Foundation, Grant/Award Number: 2018M641197; Fundamental Research Funds for the Central Universities of China, Grant/Award Numbers: FRF-BD-19-002A, FRF-TP-18-031A1; National Science Foundation, Grant/Award Number: NSF CAREER CPS-1851588; NATO, Grant/Award Number: SPS G5176

Summary

This article presents a novel actor-critic-barrier structure for the multiplayer safety-critical systems. **Non-zero-sum (NVS) games with full-state constraints are first transformed into unconstrained NVS games using a barrier function.**

The barrier function is capable of dealing with both symmetric and asymmetric constraints on the state. It is shown that the Nash equilibrium of the unconstrained NVS guarantees to stabilize the original multiplayer system. The barrier function is combined with an actor-critic structure to learn the Nash equilibrium solution in an online fashion. It is shown that integrating the barrier function with the actor-critic structure guarantees that the constraints will not be violated during learning. Boundedness and stability of the closed-loop signals are analyzed. The efficacy of the presented approach is finally demonstrated by using a simulation example.

KEYWORDS

adaptive optimal learning, barrier-actor-critic structure, Nash equilibrium, safety-aware games

带限制的
零和
博弈问
题转化
成不带
限制的

纳什均
衡

1 | INTRODUCTION

Safety is a critical issue arising in the recent development of learning-based and autonomous system applications.¹⁻⁶ The constraints imposed by the limitation of digital platform and the restriction in the environment that the system is interacting with cannot be neglected. Therefore, decision making problems with safety consideration have wide applications in many fields, such as swarms^{7,8} and robotic manipulators.^{9,10}

集群

Related Work

In a safety critical learning system, it is important to avoid certain unsafe regions during exploration or exploitation. The level surface of the *barrier certificates* serves as a “fence” between safe and unsafe regions.¹¹ Inspired by the control

障碍管理？

这里讲
的数码
平台和
环境的一
些限制？

Lyapunov functions,¹² *control barrier certificates* have been introduced in Reference 13 to design safe controllers to steer the system states from a given set of initial conditions to a desired set of terminal conditions. Computational methods, such as the sum of squares (SOS)^{11,14} and quadratic programming,¹⁵ have also been used to design safe controllers. However, the methods of SOS and quadratic programming are limited to the case of polynomial systems and constraints. Our work will provide a general form of constraints with a **nonpolynomial form**. 提供了一个非多项式的形式

Logarithmic barrier functions originate from the interior point methods for convex optimization.¹⁶ Tee et al developed barrier functions for output constraints, where a *barrier Lyapunov function (BLF)* is designed to guarantee that the output lies within a bounded set.⁴ It is further shown by Ren et al that the boundedness of the BLF yields safety.⁵ The case of state constraints has been considered by Liu and Tong,¹⁷ whereas the case of time-varying constraints has been solved by He et al.⁹

The prescribed performance design guarantees that the controlled state converges to a prescribed residual set with a predetermined convergence rate.¹⁸⁻²¹ To achieve both stability and a user-defined performance when exogenous disturbances and uncertainties exist in the system, a set-theoretic model reference adaptive control framework has been used to guarantee that the norm of the system error stays within a predefined bounded set.²²⁻²⁴ However, in the above existing results, a level of optimality was not provided along with safety. 没有安全的保障

Game theory has been a powerful tool to optimize strategic interactions between rational decision-makers^{25,26} and has been successfully applied to many control problems, such as the H_∞ and the mixed H_2/H_∞ control problem.^{27,28} By using the Hamilton-Jacobi (HJ) theory, several researchers formulated coupled HJ equations to find the feedback Nash equilibrium of differential games.²⁹⁻³⁴ For optimal control problems, learning-based methods have been applied for regulation problem,³⁵⁻³⁸ tracking problem,³⁹⁻⁴¹ output regulation problem,^{42,43} multiagent systems,⁴⁴⁻⁴⁶ networked control,⁴⁷ and robust stabilization problem.^{48,49} In constrained optimal control problem, the input constraints are commonly considered.^{2,50,51} In this article, the full-state constraints is investigated for the nonzero-sum game problem.

Contributions

在博弈论的角度展开去讨论的安全问题 ???

The contribution of this article is threefold. First, we formulate a novel barrier function based system transformation method, which is different from the BLF method, since the full-state constrained system can be transformed to an equivalent system without state constraints. It is guaranteed that given the initial state is within the prescribed bound, the state constraints are not violated. Moreover, we use an actor-critic structure to find the Nash equilibrium in an online fashion. To obviate the requirement of the persistent excitation (PE) condition, an experience replay technique is employed by using recorded and current data concurrently. 另外两个创新是Actor-Critic和replay buffer

Structure

The remainder of the article is structured as follows. Section 2 presents the non-zero-sum (NZS) games with and without full-state constraints. In Section 3, the barrier-function-based transformation is developed to deal with the constraints on the full-state. Then, a novel actor-critic-barrier structure is introduced to present the adaptive optimal learning algorithm. Section 4 conducts the simulation examples to show the effectiveness. Section 5 concludes and talks about future work.

Notations

The following notations used are pretty standard. \mathbb{R} denotes the real numbers. \mathbb{R}^+ is the set of positive real numbers. \mathbb{R}^n denotes the real n -dimensional vectors. $\mathbb{R}^{m \times n}$ denotes the real $m \times n$ matrices. For a scalar v , $|v|$ indicates the absolute value of v . For a vector x , $\|x\|$ indicates the Euclidean norm of x . For a matrix A , $\|A\|$ indicates the induced 2-norm of A . \mathbb{Z}^+ is the set of positive integer numbers. The superscript \star is used to denote the optimal solution of an optimization. $\lambda_{\min}(A)$ is the minimum eigenvalue of a matrix A , and $\mathbf{1}_m$ is the column vector with m ones. The gradient of a scalar-valued function with respect to a vector-valued variable x is defined as a column vector, and is denoted by $\nabla := \frac{\partial}{\partial x}$. A function $\alpha(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}$ said to belong to class \mathcal{K} functions, denoted as $\alpha \in \mathcal{K}$, if it is continuous, strictly increasing, and $\alpha(0) = 0$.

2 | PROBLEM FORMULATION

2.1 | Games

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

$$\begin{aligned}\dot{x}_\ell &= x_{\ell+1}, \ell = 1, \dots, n-1 \\ \dot{x}_n &= f(x) + g_1(x)u_1 + g_2(x)u_2,\end{aligned}\quad (1)$$

where $x = [x_1 \cdots x_n]^T \in \mathbb{R}^n$ with $x_\ell \in \mathbb{R}$, for $\ell = 1, \dots, n$, is the system state, $u_i \in \mathbb{R}^{m_i}$ for $i = 1, 2$ represents each control input or player, $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $g_1: \mathbb{R}^n \rightarrow \mathbb{R}$ and $g_2: \mathbb{R}^n \rightarrow \mathbb{R}$ are Lipschitz continuous functions.

The cost functional associated with each player is defined as,

$$J_i(x(0); u_1, u_2) = \frac{1}{2} \int_0^\infty U_i(x, u_1, u_2) dt, \quad (2)$$

where $U_i(x, u_1, u_2) = H_i(x) + \sum_{j=1}^2 u_j^T R_{ij} u_j$, for $i = 1, 2$, is the reward function.

Definition 1 (Nash equilibrium⁵²). Given the performance without constraints (2), the Nash equilibrium is defined as the pair (u_1^*, u_2^*) that satisfies,

$$\begin{aligned}J_1(x(0); u_1^*, u_2^*) &\leq J_1(x(0); u_1, u_2^*), \quad \forall u_1 \\ J_2(x(0); u_1^*, u_2^*) &\leq J_2(x(0); u_1^*, u_2), \quad \forall u_2.\end{aligned}$$

The problem of interest can be described as follows.

Problem 1 (Safety-aware games). Given a performance of the form (2) and a system given by (1), find the Nash equilibrium (u_1^*, u_2^*) such that the closed-loop system satisfies the constraints expressed as,

$$x_\ell \in (a_\ell, A_\ell), \quad \ell = 1, \dots, n, \quad (3)$$

where $a_\ell < 0$ and $A_\ell > 0$ are lower and upper bounds of the asymmetric constraint on state x_ℓ for $\ell = 1, \dots, n$, respectively.

In the following, we will adopt a barrier-function-based system transformation to convert the safe control problem with constraints given by (3) into a stabilization problem.

Definition 2 (Barrier function⁵⁵). The scalar function $b(\cdot)$ defined on the interval (a, A) is referred to as a barrier function if,

$$b(z; a, A) = \log \left(\frac{A}{a} \frac{a - z}{A - z} \right), \quad \forall z \in (a, A), \quad (4)$$

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

$$b^{-1}(y; a, A) = aA \frac{e^{\frac{y}{2}} - e^{-\frac{y}{2}}}{ae^{\frac{y}{2}} - Ae^{-\frac{y}{2}}}, \quad \forall y \in \mathbb{R}, \quad (5)$$

with a derivative given by,

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

Remark 1. The barrier function (4), developed for zero-sum game problem with safety concern,⁵⁵ is used in this article for nonzero-sum game problem. In addition, existing results on different barrier functions for constrained control problems can be summarized in Table 1. In Table 1, the barrier function is used for Lyapunov candidate design. In contrast, in

TABLE 1 Existing results on barrier-function-based control

Constraints type	Constraints	Barrier function
Symmetric input constraint ²	$u \in \Omega_u = \{u \in R^m \mid u_i \leq \lambda, i = 1, \dots, m\},$	$b(u) = 2 \int_0^u \left[\lambda \beta^{-1} \left(\frac{v}{\lambda} \right) \right]^T R dv$ $R = \text{diag}([r_1 \dots r_m]), r_i > 0$ $i = 1, \dots, m, \beta(\cdot) = \tanh(\cdot)$
Symmetric output constraints ⁴	$y \in \Omega_s = \{y \in R \mid -k \leq y \leq k\}$	$b(y) = \frac{1}{2} \log \frac{k^2}{k^2 - (y - y_d)^2}, y_d: \text{desired output.}$
Asymmetric output constraints ⁴	$y \in \Omega_a = \{y \in R \mid -a \leq y \leq A, 0 < a < A\}$	$b(y) = \frac{1}{p} q(y) \log \frac{A^p}{A^p - (y - y_d)^p}$ $+ \frac{1}{p} [1 - q(y)] \log \frac{a^p}{a^p - (y - y_d)^p}$
Symmetric state constraints ⁵³	$x \in \Omega_x = \{x \in R^n \mid x_\ell \leq k_\ell, k_\ell > 0, \ell = 1, \dots, m\}$	$b(x) = \sum_{\ell=1}^n b_\ell(x_\ell)$ $b_\ell(x_\ell) = \frac{1}{2} \log \frac{k_\ell^2}{k_\ell^2 - (x_\ell - x_{d\ell})^2}$
Symmetric state constraints ⁵⁴	$x \in \Omega_x = \{x \in R^n \mid x_\ell \leq k_\ell, \ell = 1, \dots, n_c\}$ $n_c: \text{number of constrained states.}$	$b(x) = \sum_{\ell=1}^{n_c} \left(\frac{x_\ell}{B_\ell - a_\ell} \right)^{2k}$

this article, the barrier function (4) is adopted to transform the system with full-state constraints into an unconstrained system. The barrier function (4) has the following desired properties:

- 1) It takes a finite value when its arguments are within the prescribed bound, that is,

$$|b(z; a, A)| < +\infty \Leftrightarrow z \in (a, A).$$

- 2) It approaches infinity as the state approaches the boundary of the constraints, that is,

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

- 3) It vanishes at the equilibrium of the system (1), that is,

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

Consider the system (1), and define the barrier-function-based transformation as,

$$\begin{aligned} s_\ell &= b_\ell(x_\ell) = b(x_\ell; a_\ell, A_\ell), \\ x_\ell &= b_\ell^{-1}(s_\ell) = b^{-1}(s_\ell; a_\ell, A_\ell), \quad \ell = 1, \dots, n, \end{aligned} \quad (6)$$

then,

$$\frac{dx_\ell}{dt} = \frac{dx_\ell}{ds_\ell} \frac{ds_\ell}{dt}.$$

Therefore, one can obtain the dynamics of the transformed variable $s = [s_1 \dots s_n]^T$ as,

$$\begin{aligned} \dot{s}_\ell &= \frac{x_{\ell+1}(s_{\ell+1})}{\frac{db^{-1}(y; a_\ell, A_\ell)}{dy} \Big|_{y=s_\ell}} = \frac{a_{\ell+1} A_{\ell+1} \left(e^{\frac{s_{\ell+1}}{2}} - e^{-\frac{s_{\ell+1}}{2}} \right)}{a_{\ell+1} e^{\frac{s_{\ell+1}}{2}} - A_{\ell+1} e^{-\frac{s_{\ell+1}}{2}}} \frac{A_\ell^2 e^{-s_\ell} - 2a_\ell A_\ell + a_\ell^2 e^{s_\ell}}{A_\ell a_\ell^2 - a_\ell A_\ell^2}, \quad F_\ell(s_\ell, s_{\ell+1}), \quad \ell = 1, \dots, n-1 \\ \dot{s}_n &= \frac{f(x) + g_1(x) u_1 + g_2(x) u_2}{\frac{db^{-1}(y; a_n, A_n)}{dy} \Big|_{y=s_n}} = [f(x) + g_1(x) u_1 + g_2(x) u_2] \frac{A_n^2 e^{-s_n} - 2a_n A_n + a_n^2 e^{s_n}}{A_n a_n^2 - a_n A_n^2} \\ &= F_n(s) + g_n^1(s) u_1 + g_n^2(s) u_2, \end{aligned} \quad (7)$$

where

$$\begin{aligned} F_n(s) &= \frac{A_n^2 e^{-s_n} - 2a_n A_n + a_n^2 e^{s_n}}{A_n a_n^2 - a_n A_n^2} \times f\left(\left[b_1^{-1}(s_1) \cdots b_n^{-1}(s_n)\right]^T\right) \\ g_n^1(s) &= \frac{A_n^2 e^{-s_n} - 2a_n A_n + a_n^2 e^{s_n}}{A_n a_n^2 - a_n A_n^2} \times g_1\left(\left[b_1^{-1}(s_1) \cdots b_n^{-1}(s_n)\right]^T\right) \\ g_n^2(s) &= \frac{A_n^2 e^{-s_n} - 2a_n A_n + a_n^2 e^{s_n}}{A_n a_n^2 - a_n A_n^2} \times g_2\left(\left[b_1^{-1}(s_1) \cdots b_n^{-1}(s_n)\right]^T\right). \end{aligned}$$

Then, system (7) can be written in a compact form as,

$$\dot{s} = F(s) + G_1(s)u_1 + G_2(s)u_2, \quad (8)$$

with

$$F(s) = [F_1 \cdots F_n]^T, \quad G_1(s) = [\mathbf{0}_{1 \times n-1} \ g_n^1(s)]^T, \quad G_2(s) = [\mathbf{0}_{1 \times n-1} \ g_n^2(s)]^T. \quad (9)$$

The following assumptions are now needed.

Assumption 1. The system given by (8) satisfies:

1. $F(s)$ is Lipschitz with $F(0) = 0$, and there exists a constant b_f such that, for $s \in \Omega_s$, $\|F(s)\| \leq b_f \|s\|$ where Ω_s is a compact set containing the origin.
2. $G_1(s)$ and $G_2(s)$ are bounded on Ω_s , that is, there exists constants b_{g1} and b_{g2} such that $\|G_1(s)\| \leq b_{g1}$ and $\|G_2(s)\| \leq b_{g2}$.
3. It is stabilizable over the compact set Ω_s .

2.2 | Problem transformation

In the previous subsection, the barrier function (2) is employed in order to transform the original system described by (1) to an equivalent system described by (8). In this subsection, we will show that the Nash equilibrium for (8) provides the optimal feedback policies to Problem 1.

Problem 2 (Safety-aware games). Find policies $u_i^*(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for $i = 1, 2$ such that the performance

$$\mathcal{F}_i(s(0); u_1, u_2) = \frac{1}{2} \int_0^\infty r_i(s, u_1, u_2) dt, \quad (10)$$

is minimized, where $r_i(s, u_1, u_2) = Q_i(s) + \sum_{j=1}^2 u_j^T R_{ij} u_j$ is the reward function.

Definition 3 (Admissible policy²). A feedback control strategy $\mu = \{u_1(s), u_2(s)\}$ for the system (8) is said to be admissible with respect to (10) on a set $\Omega_s \subset \mathbb{R}^n$, denoted as $\mu \in \mathcal{A}(\Omega_s)$ given that the following hold:

- $u_i(0) = 0$;
- $u_i(s)$ is continuous on Ω_s ;
- $u_i(s)$ stabilizes system (8) on Ω_s ;
- the performance (10) is finite $\forall x_0 \in \Omega_s$.

Given an admissible feedback strategy $\mu = \{u_1(t), u_2(t)\}$, then for $i = 1, 2$, the value function is,

$$V_i(s(t), u_1, u_2) = \int_t^\infty \left(Q_i(s) + \sum_{j=1}^2 u_j^T R_{ij} u_j \right) d\tau = \int_t^\infty r_i(s(t), u_1, u_2) d\tau, \quad (11)$$

and the problem to solve becomes,

$$V_1^*(s(t)) = \min_{u_1} \int_t^\infty r_1(s(t), u_1, u_2) d\tau, \quad V_2^*(s(t)) = \min_{u_2} \int_t^\infty r_2(s(t), u_1, u_2) d\tau,$$

given the equality constraint described by (8).

Differentiating now the value function $V_i(s)$ given in (11) yields the following Bellman equations,

$$\begin{aligned} 0 &= r_i(s, u_1, u_2) + (\nabla V_i)^T \left(F(s) + \sum_{j=1}^2 G_j(s) u_j \right), \\ 0 &= V_i(0), \quad i = 1, 2, \end{aligned} \quad (12)$$

where $\nabla V_i(s) := \partial V_i(s) / \partial s \in \mathbb{R}^n$ is the gradient vector.

Assumption 2. For an admissible feedback strategy $\mu \in \mathcal{A}(\Omega_s)$, the nonlinear Lyapunov equations (12) have locally smooth solutions $V_i(s) \geq 0, \forall s \in \Omega_s$.

Define the Hamiltonian function for every player as,

$$\mathcal{H}_i(s, \nabla V_i, u_1, u_2) = r_i(s, u_1, u_2) + (\nabla V_i)^T \left[F(s) + \sum_{j=1}^N G_j(s) u_j \right]. \quad (13)$$

According to the necessary conditions of optimality,⁵² the Nash feedback policy can be determined for $i = 1, 2$ as

$$\frac{\partial \mathcal{H}_i}{\partial u_i} = 0 \Rightarrow u_i^*(s) = -\frac{1}{2} R_{ii}^{-1} G_i^T(s) \nabla V_i^*. \quad (14)$$

Substituting now (14) into (12) yields,

$$\begin{aligned} 0 &= (\nabla V_i^*)^T \left(F(s) - \frac{1}{2} \sum_{j=1}^2 G_j(s) R_{jj}^{-1} G_j^T(s) \nabla V_j \right) + Q_i(s) + \frac{1}{4} \sum_{j=1}^2 \nabla(V_j^*)^T G_j(s) R_{jj}^{-1} R_{ij} R_{jj}^{-1} G_j^T(s) \nabla V_j^*, \\ 0 &= V_i^*(0). \end{aligned} \quad (15)$$

In the following lemma, the condition which guarantees the equivalence between Problems 1 and 2 is discussed.

Lemma 1. Suppose that Assumptions 1 and 2 hold. Given that $\mu^* = \{u_1^*, u_2^*\}$ solves Problem (2) for (8) with (10), then the following hold:

1. The closed-loop system satisfies the constraints (3) provided that the initial state x_0 of system (1) is within the region described by the constants $a_i, A_i, \forall i$.
2. The performance in (10) is equivalent to the one in (2) given that the penalty functions $H_i(\cdot)$ and $Q_i(\cdot)$ satisfy,

$$H_i(x) = Q_i(b(x)) = Q_i(s).$$

Proof. 1. Based on Assumptions 1 and 2, the existence of a positive definite and continuously differentiable optimal value function $V_i^*(s)$ can be guaranteed. From (13), one can obtain that $\dot{V}_i^*(t) \leq 0$, that is,

$$V_i^*(s(t)) \leq V_i^*(s(0)), \quad \forall t \geq 0.$$

Then, $V_i^*(s(t))$ remains bounded if $V_i^*(s(0))$ is bounded, which is guaranteed by the condition that the initial condition $x(0)$ of system (13) satisfies the constraints in (3). Finally, from the discussions in Remark 1, one can infer that

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

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 (6). Then, each element of the state $s = [b_1(x_1) \dots b_n(x_n)]^T$ is finite given that x satisfies the constraints given in (3).

Next, comparing the two performance functions (2) and (10) yields,

$$\mathcal{F}_i(s(0); u_1, u_2) = J_i(x(0); u_1, u_2),$$

provided that $Q_i(s) = H_i(x)$. This completes the proof. ■

Remark 2. In classical NZS games, the penalty function $H_i(x)$ is a quadratic form in the state variable x .^{56,57} However, this reward function design does not guarantee the safe constraints (3). In contrast, in Problem 2, the safe NZS game is transformed into an unconstrained NZS game with the barrier function (4). According to Lemma 1, to consider safety constraints on the state, the state penalty $H_i(x)$ should be properly designed as $H_i(x) = Q_i(s)$.

As shown in (14), the safe Nash equilibrium $u_i^*(s)$ depends on the solution of the HJ equations (15) given $V_i^*(s)$. In order to find the Nash policies for Problem 2, an offline policy iteration (PI) algorithm is formulated in Algorithm 1.

Algorithm 1. Policy Iteration for Problem 2

Require:

- 1: Set the iteration index $\kappa = 0$ and start with admissible initial policies $u_1^0(\cdot), u_2^0(\cdot)$;
- 2: *Policy Evaluation:* Given the strategy $\{u_1(\cdot), u_2(\cdot)\}$, find $V_i(\cdot)$ for each player by solving the Bellman equation

$$0 = r_i(x, u_1^\kappa, u_2^\kappa) + (\nabla V_i^\kappa)^\top \left[F(s) + \sum_{j=1}^2 G_j(s) u_j^\kappa \right],$$

$$0 = V_i^\kappa(0), i = 1, 2$$

- 3: *Policy Improvement:* Update the control policy for each player as

$$u_j^{\kappa+1} = -\frac{1}{2} R_{jj}^{-1} g_j^\top(x) \nabla V_j^\kappa, \quad i = 1, 2.$$

- 4: Stop if convergence is achieved; Otherwise, set $\kappa = \kappa + 1$ and go to Step 1;
-

Note that Algorithm 1 is an off-line algorithm. In the next section, a learning algorithm with an actor-critic-barrier structure will be developed.

3 | ACTOR-CRITIC-BARRIER LEARNING

As discussed in Lemma 1, the barrier-function-based system transformation guarantees that the games with full-state constraints can be solved by finding the Nash equilibrium of Problem 2. Then, we develop an actor-critic online learning algorithm to solve Problem 2. The critic learning employs an experience replay technique to relax the PE condition. To improve the stability of the closed-loop signals during the learning phase, an additional term is introduced into the actor to guarantee the asymptotic stability of the equilibrium point of the closed-loop system. The overall framework of the actor-critic-barrier structure is shown in Figure 1.

3.1 | Value function approximation

Definition 4 (Uniform convergence³⁵). A sequence of functions $\{p_n(x)\}$ converges uniformly to $p(x)$ on a set Ω if $\forall \epsilon > 0$, $\exists N(\epsilon)$ such that $\sup_{x \in \Omega} \|p_n(x) - p(x)\| < \epsilon$ for $n > N(\epsilon)$.

According to the Weierstrass high-order approximation theorem,⁵⁸ for each player, there exists an approximator such that the optimal value function $V_i^*(s)$ and its gradient $\nabla V_i^*(s)$ can be uniformly approximated by a critic network within a set $\Omega \subseteq \mathbb{R}^n$ that contains the origin, as

$$V_i^*(s) = \left(W_{c,i}^* \right)^\top \phi_{c,i}(s) + \varepsilon_{c,i}(s), \quad \nabla V_i^*(s) = \left[\nabla \phi_{c,i}(s) \right]^\top W_{c,i}^* + \nabla \varepsilon_{c,i}(s), \quad (16)$$

where $W_{c,i}^* \in \mathbb{R}^{N_c}$ is the optimal critic weight, $\phi_{c,i}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{N_c}$ is the critic basis, and $\varepsilon_{c,i}(s)$ and $\nabla \varepsilon_{c,i}(s)$ are the residual errors of the critic as well as the gradient for each player.

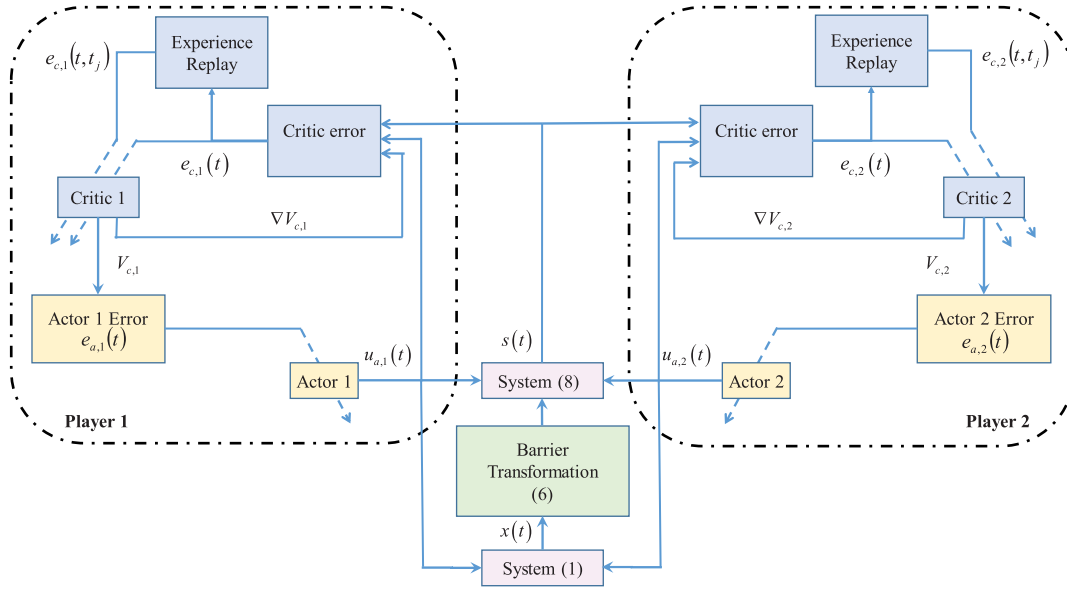


FIGURE 1 Actor-critic-barrier Structure for safe games. The barrier function (4) is used to transform Problem 1 with full-state constraints to Problem 2 without state constraints. Then, the online actor-critic adaptive optimal learning algorithm for system (8) is implemented to consider the input saturation. The critic learning is driven by the critic error $e_{c,i}$ in (20) with its history version $e_{c,i}(t_k, t)$ in (23) and the actor learning is driven by the actor error e_a in (28) [Colour figure can be viewed at wileyonlinelibrary.com]

Given the value function approximation in (16), the Hamiltonian functions can be rewritten in terms of the optimal critic weights $W_{c,i}^*$ as,

$$\begin{aligned} 0 &= H_i(s, \nabla V_i, u_1^*, u_2^*) \\ &= \underbrace{r_i(s, u_1^*, u_2^*) + (W_{c,i}^*)^T \sigma_i^*(s, u_1^*, u_2^*)}_{:= H_i(s, W_{c,i}^*, u_1, u_2)} + \underbrace{\nabla \epsilon_{c,i}^T [F(s) + G_1(s) u_1^* + G_2(s) u_2^*]}_{:= -\epsilon_{\text{ham},i}}, \end{aligned}$$

with σ_i^* being a N -dimensional vector signal defined as

$$\sigma_i^*(s, u_1, u_2) = \nabla \phi_{c,i}(s) [F(s) + G_1(s) u_1^* + G_2(s) u_2^*].$$

As the number of basis $N \rightarrow \infty$, the approximation errors $\epsilon_{c,i} \rightarrow 0$, and $\nabla \epsilon_{c,i} \rightarrow 0$, $i = 1, 2$ uniformly.⁵⁹

Assumption 3. The following hold on the compact set Ω_s .

1. The optimal critic weight $W_{c,i}^*$ is bounded, that is, $\|W_{c,i}^*\| \leq W_{\text{cmax},i}$.
2. The critic residual as well as its gradient are bounded, that is, $\|\epsilon_{c,i}(s)\| \leq \epsilon_{\text{cmax},i}$ and $\|\nabla \epsilon_{c,i}(s)\| \leq \epsilon_{\text{cdmax},i}$.
3. The critic basis as well as its gradient are bounded, that is, $\|\phi_{c,i}(s)\| \leq \phi_{\text{cmax},i}$, $\|\nabla \phi_{c,i}(s)\| \leq \phi_{\text{cdmax},i}$.
4. The Hamiltonian residual is bounded, that is, $\|\epsilon_{\text{ham},i}\| \leq \epsilon_{\text{hmax},i}$.

3.2 | Critic learning using experience replay

For a fixed pair of policies (u_1, u_2) , the ideal critic weight, $W_{c,1}^*$, which provides the best approximation to the value function $V_i(s)$ on the compact set Ω to evaluate (u_1, u_2) , is unknown. Therefore, the estimation of $W_{c,1}^*$ is implemented by a critic network with weights $W_{c,i}$. Then, the critic output as well as its gradient can be represented as

$$\begin{aligned} V_i(s) &= W_{c,i}^T \phi_{c,i}(s) \\ \nabla V_i(s) &= [\nabla \phi_{c,i}(s)]^T W_{c,i}. \end{aligned} \quad (17)$$

The Hamiltonian approximation error using the value function approximation can be expressed as

$$\mathcal{H}_i(s, W_{c,i}, u_1, u_2) = r_i(s, u_1, u_2) + W_{c,i}^T \sigma_i(t) = e_{c,i}. \quad (18)$$

For each player, define the critic approximation error as,

$$\tilde{W}_{c,i} = W_{c,i}^* - W_{c,i}. \quad (19)$$

From (18), and after combining the Bellman residual $e_{c,i}$ and the Bellman equation approximation error $\epsilon_{\text{ham},i}$, one gets,

$$e_{c,i} = \epsilon_{\text{ham},i} - \tilde{W}_{c,i}^T \sigma_i. \quad (20)$$

The policy evaluation for an admissible control policy $u(\cdot)$ can be formulated by adapting the critic weight $W_{c,i}$ to minimize the following objective function,³⁵

$$E_c = E_{c,1} + E_{c,2} = \frac{1}{2} \frac{e_{c,1}^2(t)}{(1 + \sigma_1^T \sigma_1)^2} + \frac{1}{2} \frac{e_{c,2}^2(t)}{(1 + \sigma_2^T \sigma_2)^2}. \quad (21)$$

Then $e_{c,i} \rightarrow \epsilon_{\text{ham},i}$ as $W_{c,i} \rightarrow W_{c,i}^*$. Using the chain rule yields the gradient descent algorithm for minimizing E_c given as,³⁵

$$\dot{W}_{c,i} = -a_{c,i} \frac{\partial E_c}{\partial W_{c,i}} = -a_{c,i} \frac{\sigma_i}{(1 + \sigma_i^T \sigma_i)^2} \left[\sigma_i^{TW_{c,i}} + r_i(x, u_1, u_2) \right],$$

where

$$\sigma_i(s, u_1, u_2) = \nabla \phi_{c,i}(s) [F(s) + G_1(s) u_1 + G_2(s) u_2].$$

Definition 5 (Persistent excitation⁶⁰). The bounded vector signal $y(t) \in \mathbb{R}^p$ is PE over the interval $[t, t + T_{\text{pe}}]$ with $T_{\text{pe}} \in \mathbb{R}^+$ if there exists $\beta_1 \in \mathbb{R}^+$ and $\beta_2 \in \mathbb{R}^+$ such that $\forall t$,

$$\beta_1 I_{p \times p} \leq \int_t^{t+T_{\text{pe}}} y(\tau) y^T(\tau) d\tau \leq \beta_2 I_{p \times p},$$

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

To guarantee the convergence of the critic learning to the ideal critic weight, the signal σ_i is required to be PE.³⁵

To relax this requirement, the following modified objective for the critic learning is presented

$$\bar{E}_c = \bar{E}_{c,1} + \bar{E}_{c,2} = \frac{1}{2} \left[\frac{e_{c,1}^2(t)}{(1 + \sigma_1^T(t) \sigma_1(t))^2} + \sum_{k=1}^p \frac{e_{c,1}^2(t_k, t)}{(1 + \sigma_{1,k}^T \sigma_{1,k})^2} \right] + \frac{1}{2} \left[\frac{e_{c,2}^2(t)}{(1 + \sigma_2^T(t) \sigma_2(t))^2} + \sum_{k=1}^p \frac{e_{c,2}^2(t_k, t)}{(1 + \sigma_{2,k}^T \sigma_{2,k})^2} \right], \quad (22)$$

where,

$$e_{c,i}(t_k, t) := r_{i,k} + W_{c,i}^T(t) \sigma_{i,k}, \quad i = 1, 2, \quad (23)$$

with $\sigma_{i,k} := \sigma_i(t_k)$ and $r_{i,k} := r_i(s(t_k), u_1(t_k), u_2(t_k))$. Then, the critic adaptive law can be obtained as

$$\dot{W}_{c,i} = -a_{c,i} \frac{\partial \bar{E}_c}{\partial W_{c,i}} = -a_{c,i} \frac{\sigma_i e_{c,i}}{(1 + \sigma_i^T \sigma_i)^2} - a_{c,i} \sum_{k=1}^p \frac{\sigma_{i,k} e_{c,i}(t_k, t)}{(1 + \sigma_{i,k}^T \sigma_{i,k})^2}. \quad (24)$$

Condition 1. Let $Z_i = [\sigma_{i,1} \dots \sigma_{i,p}]$ be the history stack. Then, Z_i in the recorded data contains as many linearly independent elements as the number of neurons in (16). That is $\text{rank}(Z_i) = N$.

Fact 1. For an arbitrary vector ω , one has

$$\left\| \frac{\omega}{1 + \omega^T \omega} \right\| \leq \frac{1}{2}, \quad \left\| \frac{1}{1 + \omega^T \omega} \right\| \leq 1, \quad \left\| \frac{\omega \omega^T}{1 + \omega^T \omega} \right\| \leq 1, \quad \left\| \frac{\omega \omega^T}{(1 + \omega^T \omega)^2} \right\| \leq \frac{1}{4}, \quad \forall \omega.$$

Theorem 1. Let $u(\cdot)$ be any admissible control policy. Let the critic network (17) with the experience replay tuning law (24) be used to evaluate the policy $u(\cdot)$. Suppose that the history stack satisfies Condition 1. Then, the critic weight error $\tilde{W}_{c,i}$ is uniformly ultimately bounded (UUB).

Proof. Considering the fact in (20) and $\tilde{W}_{c,i} = W_{c,1}^* - W_{c,i}$, one can obtain the critic error dynamics as,

$$\dot{\tilde{W}}_{c,i}(t) = -a_{c,i} [\Gamma(t) + \Gamma_k] \tilde{W}_{c,i}(t) + a_{c,i} \Lambda,$$

where

$$\Gamma(t) = \frac{\sigma_i(t) \sigma_i^T(t)}{[1 + \sigma_i^T(t) \sigma_i(t)]^2}, \quad \Gamma_k = \sum_{k=1}^p \frac{\sigma_{i,k} \sigma_{i,k}^T}{[1 + \sigma_{i,k}^T \sigma_{i,k}]^2}, \quad \Lambda = \frac{\sigma_i(t) \varepsilon_{\text{ham},i}(t)}{[1 + \sigma_i^T(t) \sigma_i(t)]^2} + \sum_{k=1}^p \frac{\sigma_{i,k} \varepsilon_{\text{ham},i}}{[1 + \sigma_{i,k}^T \sigma_{i,k}]^2}.$$

Consider the Lyapunov candidate

$$V_{c,i}(t) = \frac{1}{2a_{c,i}} \tilde{W}_{c,i}^T(t) \tilde{W}_{c,i}(t).$$

Differentiating $V_{c,i}(t)$ yields

$$\dot{V}_{c,i} = -\tilde{W}_{c,i}^T(t) [\Gamma(t) + \Gamma_k] \tilde{W}_{c,i}(t) + \tilde{W}_{c,i}^T(t) \Lambda.$$

Note that $\Gamma(t) \geq 0$, then

$$\dot{V}_{c,i} \leq -\tilde{W}_{c,i}^T(t) \Gamma_k \tilde{W}_{c,i}(t) + \tilde{W}_{c,i}^T(t) \Lambda.$$

Based on Condition 1, the matrix Γ_k is positive definite. Therefore,

$$\dot{V}_{c,i} \leq -\lambda_{\min}(\Gamma_k) \|\tilde{W}_{c,i}(t)\|^2 + \|\tilde{W}_{c,i}(t)\| \|\Lambda\|.$$

From Fact 1, one has $\|\Lambda\| \leq \frac{p+1}{2} \varepsilon_{h\max,i}$. Therefore,

$$\dot{V}_{c,i} \leq -\lambda_{\min}(\Gamma_k) \|\tilde{W}_{c,i}(t)\|^2 + \|\tilde{W}_{c,i}(t)\| \left(\frac{p+1}{2} \varepsilon_{h\max,i} \right). \quad (25)$$

From (25), it is guaranteed that $\dot{V}_{c,i} < 0$ if $\|\tilde{W}_{c,i}(t)\| > \frac{(p+1)\varepsilon_{h\max,i}}{2\lambda_{\min}(\Gamma_k)}$ is outside the set $\Omega_{\tilde{W}_{c,i}}$ with $\Omega_{\tilde{W}_{c,i}} := \left\{ \tilde{W}_{c,i} \mid \|\tilde{W}_{c,i}\| \leq \frac{(p+1)\varepsilon_{h\max,i}}{2\lambda_{\min}(\Gamma_k)} \right\}$. Therefore, the critic weight error $\tilde{W}_{c,i}(t)$ is UUB.⁶¹ This completes the proof. ■

Remark 3. As discussed by Yang et al,³⁶ the Bellman residual $e_{c,i}$ using the critic network (17) is the counterpart of the temporal difference (TD) for continuous-time system. Compared with (21), the additional term in the modified objective function (22) is the TD errors for the stored samples in the history stack. That is, the modified critic learning (24) tries to minimize both the current TD error and the past errors stored in a history stack. This yields the second term in the critic adaptive law in (24). As shown in Theorem 1, the second term in (24) contributes to the stability guarantee.

Remark 4. The PE condition cannot be verified during learning, especially for complex nonlinear systems. In contrast, the experience replay modification, as shown in (24), is based on the online data collection and verification of rank condition in Condition 1. Therefore, the experience replay based online learning can be implemented and checked in an online fashion.

3.3 | Actor learning guaranteeing asymptotically stability

As shown in (14), the optimal control policy depends on the optimal value gradient $\frac{\partial V^*(s)}{\partial s}$. By using the optimal value function approximation (16), the Nash strategy can be represented in terms of the ideal critic as

$$u_i^* = -\frac{1}{2}R_{ii}^{-1}G_i^T(s) \left\{ [\nabla\phi_{c,i}(s)]^T W_{c,i}^* + \nabla\epsilon_{c,i} \right\},$$

where $W_{c,i}^*$ is the unknown ideal critic weight for player i . With the value gradient approximation using the critic weight $W_{c,i}$ in (17), the policy can be expressed in terms of the adaptive critic weight $W_{c,i}$ as

$$u_{c,i}(s) = -\frac{1}{2}R_{ii}^{-1}G_i^T(s) [\nabla\phi_{c,i}(s)]^T W_{c,i}.$$

However, this control policy does not guarantee stability of the closed-loop system.^{35,50} Therefore, to ensure stability in a Lyapunov sense, the policy applied to the system is implemented by another approximator, named as the actor network. The Nash strategy can be represented in terms of the actor as

$$u_i^*(s) = \left(W_{a,i}^* \right)^T \phi_{a,i}(s) + \epsilon_{a,i}, \quad (26)$$

where $W_{a,i}^* \in \mathbb{R}^{N_a \times m}$ is the unknown ideal actor weight, $\phi_{a,i} \in \mathbb{R}^{N_a}$ is the actor basis, and $\epsilon_{a,i}$ is the residual error for the actor network.

Assumption 4. For the actor network, the followings hold on the compact set Ω_s .

- 1) The actor residual as well as its gradient are bounded, that is, $\|\epsilon_{a,i}(s)\| \leq \epsilon_{\max,i}$ and $\|\nabla\epsilon_{a,i}(s)\| \leq \epsilon_{\max,i}$.
- 2) The actor basis as well as its gradient are bounded, that is, $\|\phi_{a,i}(s)\| \leq \phi_{\max,i}$, $\|\nabla\phi_{a,i}(s)\| \leq \phi_{\max,i}$.

Based on the optimal actor representation (26), the online actor network applied to the system takes the form as

$$u_{a,i}(s(t)) = W_{a,i}^T(t) \phi_{a,i}(s(t)), \quad (27)$$

where $W_{a,i} \in \mathbb{R}^{N_a \times m}$ is the actor approximator. To determine the learning rule for the actor, we define the following error for each actor as

$$\begin{aligned} e_{a,i} &= u_{a,i}(s) - u_{c,i}(s) \\ &= W_{a,i}^T \phi_{a,i}(s) + \frac{1}{2}R_{ii}^{-1}G_i^T(s) [\nabla\phi_{c,i}(s)]^T W_{c,i}. \end{aligned} \quad (28)$$

The actor learning rule for each player can be formulated as adapting the actor weight $W_{a,i}$ to minimize the objective function

$$E_a = \frac{1}{2}e_{a,1}^T e_{a,1} + \frac{1}{2}e_{a,2}^T e_{a,2}. \quad (29)$$

By using the gradient descent algorithm, one can obtain the actor learning as

$$\begin{aligned} \dot{W}_{a,i} &= -a_{a,i} \frac{\partial E_a}{\partial W_{a,i}} \\ &= -a_{a,i} \phi_{a,i} \left[W_{a,i}^T \phi_{a,i} + \frac{1}{2}R_{ii}^{-1}G_i^T(\nabla\phi_{c,i})^T W_{c,i} \right]^T. \end{aligned} \quad (30)$$

Defining the actor weight error as $\tilde{W}_{a,i} := W_{a,i}^* - W_{a,i}$, one can obtain the actor dynamics as

$$\dot{\tilde{W}}_{a,i} = -a_{a,i} \phi_{a,i} \phi_{a,i}^T \tilde{W}_{a,i} - a_{a,i} \phi_{a,i} \epsilon_{a,i}^T - a_{a,i} \phi_{a,i} \left(\frac{1}{2}R_{ii}^{-1}G_i^T \nabla\epsilon_{c,i} \right)^T - a_{a,i} \phi_{a,i} \left[\frac{1}{2}R_{ii}^{-1}G_i^T (\nabla\phi_{c,i})^T \tilde{W}_{c,i} \right]^T. \quad (31)$$

Most existing actor-critic learning only guarantees UUB⁶¹ of the state $s(t)$ and the actor-critic weights $\tilde{W}_{c,i}(t)$ and $\tilde{W}_{a,i}(t)$ during the learning phase.^{35,39,62,63} To guarantee asymptotically stability of the equilibrium point of the closed-loop

system, an additional robust term is added into the actor network to improve the control performance, that is, the control input applied to the system during the online actor-critic learning takes the following form

$$\begin{aligned}\bar{u}_{a,i} &= W_{a,i}^T \phi_{a,i}(s) + \eta, \quad i = 1, 2, \\ \eta &= -B\|s\|^2 \frac{1_m}{(A + s^T s)},\end{aligned}\quad (32)$$

where A and B are positive design parameters. Accordingly, the critic learning error and the reward function can be represented as

$$\begin{aligned}\sigma_i^a &= \nabla \phi_i [F(s) + G_1(s) \bar{u}_{a,1} + G_2(s) \bar{u}_{a,2}], \\ \sigma_{i,k}^a &= \sigma_i^a(t_k), \\ e_{c,i}^a &= r_i^a + W_{c,i}^T \sigma_i^a \\ e_{c,i}^a(t_k, t) &= r_{i,k}^a + W_{c,i}^T \sigma_i^a \\ r_i^a &= r_i(s, \bar{u}_{a,1}, \bar{u}_{a,2}) \\ r_{i,k}^a &= r_i(s(t_k), \bar{u}_{a,1}(t_k), \bar{u}_{a,2}(t_k)), i = 1, 2.\end{aligned}$$

Based on (24), the critic learning when applying the control input (32) for each player can be expressed as

$$\dot{W}_{c,i} = -a_{c,i} \frac{\sigma_i^a e_{c,i}^a}{\left[1 + (\sigma_i^a)^T \sigma_i^a\right]^2} - a_{c,i} \sum_{k=1}^p \frac{\sigma_{i,k}^a e_{c,i}^a(t_k, t)}{\left[1 + (\sigma_{i,k}^a)^T \sigma_{i,k}^a\right]^2}. \quad (33)$$

The following variable definitions are needed.

$$\begin{aligned}\bar{V}_{c,i} &:= W_{cmax,i} \phi_{cdmax,i} + \epsilon_{cdmax,i}, \quad i = 1, 2 \\ \rho &:= \frac{b_{g1} \phi_{amax,1} + b_{g2} \phi_{amax,2}}{2} (\bar{V}_{c,1}^2 + \bar{V}_{c,2}^2) + (b_{g1} + b_{g2}) \left(\frac{\bar{V}_{c,1}^2 + \bar{V}_{c,2}^2}{2} \right) \\ &\quad + \frac{p+1}{2} (\alpha_{c,1} \epsilon_{hmax,1}^2 + \alpha_{c,2} \epsilon_{hmax,2}^2) + \frac{\phi_{amax,1}}{2} \left(\epsilon_{amax,1} + \frac{1}{2} b_{g1} \lambda_{\min}(R_{11}^{-1}) \epsilon_{cdmax,1} \right)^2 \\ &\quad + \frac{\phi_{amax,2}}{2} \left(\epsilon_{amax,2} + \frac{1}{2} b_{g2} \lambda_{\min}(R_{22}^{-1}) \epsilon_{cdmax,2} \right)^2 + b_{g1} \epsilon_{amax,1}^2 + b_{g2} \epsilon_{amax,2}^2.\end{aligned}\quad (34)$$

Theorem 2. Suppose that Assumptions 1-4 and Condition 1 hold. Consider the dynamical system (8) with the control input given by (32) and the actor-critic learning framework described by (30) and (33). Then, Problem 2 can be solved by the presented actor-critic-barrier online learning algorithm in the sense that there exists $\bar{\Omega} := \left\{ \Omega_s \times \Omega_{\bar{W}_{c,1}} \times \Omega_{\bar{W}_{c,2}} \times \Omega_{\bar{W}_{a,1}} \times \Omega_{\bar{W}_{a,2}} \right\} \subset \Omega$ such that the equilibrium point of the closed-loop system with state $\chi := [s^T \bar{W}_{c,1}^T \bar{W}_{c,2}^T \bar{W}_{a,1}^T \bar{W}_{a,2}^T]^T \in \bar{\Omega}$ exists globally and converges asymptotically to zero for all weights $\bar{W}_{c,i}(0)$ inside $\Omega_{\bar{W}_{c,i}}$, $\bar{W}_{a,i}(0)$ inside $\Omega_{\bar{W}_{a,i}}$ and $s(0)$ inside Ω_s given that,

$$\phi_{amax,i} > b_{gi} + \frac{b_{gi} \lambda_{\min}(R_{ii}^{-1}) \phi_{cdmax,i}}{4} + \frac{1}{2} \quad (35)$$

$$\alpha_{c,i} > \frac{b_{gi} \lambda_{\min}(R_{ii}^{-1}) \phi_{amax,i} \phi_{cdmax,i}}{4 \left[2 \lambda_{\min}(\Gamma_k^a) - \frac{p+1}{2} \right]} \quad (36)$$

$$\rho (A + \|s\|^2) < (\bar{V}_{c,2} + \bar{V}_{c,1}) (b_{g1} + b_{g2}) B \|s\|^2. \quad (37)$$

Proof. We consider the Lyapunov candidate

$$\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2, \quad t \geq 0,$$

where

$$\mathcal{V}_i(\chi(t)) = V_i^*(s) + V_{c,i}(s) + V_{a,i}(s),$$

with

$$V_{c,i}(t) = \tilde{W}_{c,i}^T \tilde{W}_{c,i}, V_{a,i}(s) = \frac{1}{2\alpha_{a,i}} \text{trace} \left\{ \tilde{W}_{a,i}^T \tilde{W}_{a,i} \right\}.$$

1. For each player, we have

$$\dot{V}_i^* = \left(\frac{\partial V_i^*(s)}{\partial s} \right)^T \left[F(s) + G_1(s) (W_{a,1}^T \phi_{a,1} + \eta) + G_2(s) (W_{a,2}^T \phi_{a,2} + \eta) \right]. \quad (38)$$

Combining (26) and (32), then (38) can be equivalently rewritten as

$$\begin{aligned} \dot{V}_i^* = & \left(\frac{\partial V_i^*(s)}{\partial s} \right)^T \left[F - G_1(\tilde{W}_{a,1})^T \phi_{a,1} - G_2(\tilde{W}_{a,2})^T \phi_{a,2} + G_1(u_1^* - \varepsilon_{a,1}) + G_2(u_2^* - \varepsilon_{a,2}) \right. \\ & \left. - G_1 B \|s\|^2 \frac{1_m}{(A + s^T s)} - G_2 B \|s\|^2 \frac{1_m}{(A + s^T s)} \right]. \end{aligned} \quad (39)$$

From the Bellman equation (12), one has

$$\left(\frac{\partial V_i^*(s)}{\partial s} \right)^T (F + G_1 u_1^* + G_2 u_2^*) + r_i(s, u_1^*, u_2^*) = 0, \quad (40)$$

and after inserting (40) back into (39), one gets

$$\begin{aligned} \dot{V}_i^* = & -r_i(s, u_1^*, u_2^*) + \left(\frac{\partial V_i^*(s)}{\partial s} \right)^T \left[-G_1(\tilde{W}_{a,1})^T \phi_{a,1} - G_2(\tilde{W}_{a,2})^T \phi_{a,2} \right. \\ & \left. - G_1 B \|s\|^2 \frac{1_m}{(A + s^T s)} - G_2 B \|s\|^2 \frac{1_m}{(A + s^T s)} - G_1 \varepsilon_{a,1} - G_2 \varepsilon_{a,2} \right]. \end{aligned}$$

From Assumption 3, \dot{V}_i^* can be upper bounded as

$$\begin{aligned} \dot{V}_i^* \leq & -r_i(s, u_1^*, u_2^*) - (W_{c\max,i} \phi_{cd\max,i} + \varepsilon_{cd\max,i}) \times [b_{g1} \phi_{a\max,1} \|\tilde{W}_{a,1}\| + b_{g2} \phi_{a\max,2} \|\tilde{W}_{a,2}\| \\ & + b_{g1} B \|s\|^2 \frac{1}{(A + s^T s)} + b_{g2} B \|s\|^2 \frac{1}{(A + s^T s)} + b_{g1} \varepsilon_{a\max,1} + b_{g2} \varepsilon_{a\max,2}]. \end{aligned} \quad (41)$$

Then, (41) further yields

$$\begin{aligned} \dot{V}_i^* \leq & -r_i(s, u_1^*, u_2^*) - \bar{V}_{c,i} b_{g1} \varepsilon_{a\max,1} - \bar{V}_{c,i} b_{g2} \varepsilon_{a\max,2} - \bar{V}_{c,i} b_{g1} \phi_{a\max,1} \|\tilde{W}_{a,1}\| - \bar{V}_{c,i} b_{g2} \phi_{a\max,2} \|\tilde{W}_{a,2}\| \\ & - \bar{V}_{c,i} b_{g1} B \|s\|^2 \frac{1}{(A + s^T s)} - \bar{V}_{c,i} b_{g2} B \|s\|^2 \frac{1}{(A + s^T s)}. \end{aligned} \quad (42)$$

Applying now Young's inequality to (42), one can finally obtain

$$\begin{aligned} \dot{V}_i^* \leq & -r_i(s, u_1^*, u_2^*) + \frac{b_{g1} \phi_{a\max,1}}{2} \|\tilde{W}_{a,1}\|^2 + \frac{b_{g1} \phi_{a\max,1} \bar{V}_{c,i}^2}{2} + \frac{b_{g2} \phi_{a\max,2}}{2} \|\tilde{W}_{a,2}\|^2 + \frac{b_{g2} \phi_{a\max,2} \bar{V}_{c,i}^2}{2} \\ & + b_{g1} \frac{\bar{V}_{c,i}^2 + \varepsilon_{a\max,1}^2}{2} + b_{g2} \frac{\bar{V}_{c,i}^2 + \varepsilon_{a\max,2}^2}{2} - \left(\bar{V}_{c,i} b_{g1} + \bar{V}_{c,i} b_{g2} \right) B \|s\|^2 \frac{1}{(A + s^T s)}. \end{aligned} \quad (43)$$

2. For each player, differentiating $V_{c,i}$ yields

$$\dot{V}_{c,i} = 2\tilde{W}_{c,i}^T \dot{\tilde{W}}_{c,i}. \quad (44)$$

From (20), one has

$$e_{c,i}^a = \varepsilon_{\text{ham},i}^a - \tilde{W}_{c,i}^T \sigma_i^a, \quad \varepsilon_{\text{ham},i}^a = \nabla \varepsilon_{c,i}^T [F + G_1 \bar{u}_{a,1} + G_2 \bar{u}_{a,2}].$$

Then, given the policy described in (32), the critic error dynamics for each player can be written as

$$\dot{\tilde{W}}_{c,i}(t) = -a_{c,i} [\Gamma_a(t) + \Gamma_k^a] \tilde{W}_{c,i}(t) + a_{c,i} \Lambda_a, \quad (45)$$

where

$$\Gamma_a(t) = \frac{\sigma_i^a (\sigma_i^a)^T}{[1 + (\sigma_i^a)^T \sigma_i^a]^2}, \quad \Gamma_k^a = \sum_{k=1}^p \frac{\sigma_{i,k}^a (\sigma_{i,k}^a)^T}{[1 + (\sigma_{i,k}^a)^T \sigma_{i,k}^a]^2}, \quad \Lambda_a = \frac{\sigma_i^a \varepsilon_{\text{ham},i}^a}{[1 + (\sigma_i^a)^T \sigma_i^a]^2} + \sum_{k=1}^p \frac{\sigma_{i,k}^a \varepsilon_{\text{ham},i}^a(t_k)}{[1 + (\sigma_{i,k}^a)^T \sigma_{i,k}^a]^2}.$$

Inserting (45) into (44) yields

$$\begin{aligned} \dot{V}_{c,i} &= -2\alpha_{c,i} \tilde{W}_{c,i}^T [\Gamma_a + \Gamma_k^a] \tilde{W}_{c,i} + 2\alpha_{c,i} \tilde{W}_{c,i}^T \Lambda_a \\ &\leq -2\alpha_{c,i} \lambda_{\min}(\Gamma_k^a) \|\tilde{W}_{c,i}\|^2 + 2\alpha_{c,i} \frac{p+1}{2} \|\tilde{W}_{c,i}\| \varepsilon_{\text{hmax},i}, \end{aligned} \quad (46)$$

and after considering Fact 1, one has

$$\dot{V}_{c,i} \leq \alpha_{c,i} \left[\frac{p+1}{2} - 2\lambda_{\min}(\Gamma_k^a) \right] \|\tilde{W}_{c,i}\|^2 + \alpha_{c,i} \frac{p+1}{2} \varepsilon_{\text{hmax},i}^2. \quad (47)$$

3. Now differentiating $V_{a,i}$ yields

$$\dot{V}_{a,i}(s) = \frac{1}{2\alpha_{a,1}} \text{trace} \left\{ 2\tilde{W}_{a,i}^T \dot{\tilde{W}}_{a,i} \right\}. \quad (48)$$

Define the actor weight error as $\tilde{W}_{a,i} := W_{a,i}^* - W_{a,i}$, and based on (30), one can obtain the actor error dynamics as

$$\dot{W}_{a,i} = -a_{a,i} \phi_{a,i} \phi_{a,i}^T W_{a,i} - a_{a,i} \phi_{a,i} \varepsilon_{a,i}^T - a_{a,i} \phi_{a,i} \left(\frac{1}{2} R_{ii}^{-1} G_i^T \nabla \varepsilon_{c,i} \right)^T. \quad (49)$$

By the actor error dynamics (49) into (48), we have,

$$\dot{V}_{a,i}(s) = \text{trace} \left\{ -\tilde{W}_{a,i}^T \phi_{a,i} \phi_{a,i}^T \tilde{W}_{a,i} - \tilde{W}_{a,i}^T \phi_{a,i} \left(\frac{1}{2} R_{ii}^{-1} G_i^T (\nabla \phi_{c,i})^T \tilde{W}_{c,i} \right)^T - \tilde{W}_{a,i}^T \phi_{a,i} \varepsilon_{a,i}^T - \tilde{W}_{a,i}^T \phi_{a,i} \left(\frac{1}{2} R_{ii}^{-1} G_i^T \nabla \varepsilon_{c,i} \right)^T \right\}. \quad (50)$$

Based on Assumption 3, then (50) further satisfies

$$\begin{aligned} \dot{V}_{a,i}(s) &\leq -\lambda_{\min}(\phi_{a,i} \phi_{a,i}^T) \|\tilde{W}_{a,i}\|^2 + \phi_{\text{amax},i} \varepsilon_{\text{amax},i} \|\tilde{W}_{a,i}\| + \frac{1}{2} b_{\text{gi}} \lambda_{\min}(R_{ii}^{-1}) \phi_{\text{amax},i} \phi_{\text{cdmax},i} \|\tilde{W}_{a,i}\| \|\tilde{W}_{c,i}\| \\ &\quad + \frac{1}{2} b_{\text{gi}} \lambda_{\min}(R_{ii}^{-1}) \phi_{\text{amax},i} \varepsilon_{\text{cdmax},i} \|\tilde{W}_{a,i}\|. \end{aligned}$$

Applying Young's inequality to the above equation, one can further obtain

$$\begin{aligned} \dot{V}_{a,i}(s) &\leq -\phi_{\text{amax},i}^2 \|\tilde{W}_{a,i}\|^2 + \frac{1}{2} b_{\text{gi}} \lambda_{\min}(R_{ii}^{-1}) \phi_{\text{amax},i} \phi_{\text{cdmax},i} \frac{\|\tilde{W}_{a,i}\|^2 + \|\tilde{W}_{c,i}\|^2}{2} \\ &\quad + \phi_{\text{amax},i} \frac{\left(\varepsilon_{\text{amax},i} + \frac{1}{2} b_{\text{gi}} \lambda_{\min}(R_{ii}^{-1}) \varepsilon_{\text{cdmax},i} \right)^2 + \|\tilde{W}_{a,i}\|^2}{2} \\ &= \frac{\phi_{\text{amax},i}}{2} \left(\varepsilon_{\text{amax},i} + \frac{1}{2} b_{\text{gi}} \lambda_{\min}(R_{ii}^{-1}) \varepsilon_{\text{cdmax},i} \right)^2 + \frac{b_{\text{gi}} \lambda_{\min}(R_{ii}^{-1}) \phi_{\text{amax},i} \phi_{\text{cdmax},i}}{4} \|\tilde{W}_{c,i}\|^2 \\ &\quad + \left(\frac{b_{\text{gi}} \lambda_{\min}(R_{ii}^{-1}) \phi_{\text{amax},i} \phi_{\text{cdmax},i}}{4} - \phi_{\text{amax},i}^2 + \frac{\phi_{\text{amax},i}}{2} \right) \|\tilde{W}_{a,i}\|^2. \end{aligned} \quad (51)$$

4. Finally, grouping the above terms in (43), (47), and (51) yields the results in (52) (see top of this page).

$$\begin{aligned}
\dot{\mathcal{V}} \leq & -r_1 (s, u_1^*, u_2^*) - r_2 (s, u_1^*, u_2^*) + \rho - (\bar{V}_{c,2} + \bar{V}_{c,1}) (b_{g1} + b_{g2}) B \|s\|^2 \frac{1}{(A + s^T s)} \\
& + \left(b_{g1} \phi_{\text{amax},1} + \frac{b_{g1} \lambda_{\min}(R_{11}^{-1}) \phi_{\text{amax},1} \phi_{\text{cdmax},1}}{4} - \phi_{\text{amax},1}^2 + \frac{\phi_{\text{amax},1}}{2} \right) \|\tilde{W}_{a,1}\|^2 \\
& + \left(b_{g2} \phi_{\text{amax},2} + \frac{b_{g2} \lambda_{\min}(R_{22}^{-1}) \phi_{\text{amax},2} \phi_{\text{cdmax},2}}{4} - \phi_{\text{amax},2}^2 + \frac{\phi_{\text{amax},2}}{2} \right) \|\tilde{W}_{a,2}\|^2 \\
& + \left\{ \alpha_{c,1} \left[\frac{p+1}{2} - 2\lambda_{\min}(\Gamma_k^a) \right] + \frac{b_{g1} \lambda_{\min}(R_{11}^{-1}) \phi_{\text{amax},1} \phi_{\text{cdmax},1}}{4} \right\} \|\tilde{W}_{c,1}\|^2 \\
& + \left\{ \alpha_{c,2} \left[\frac{p+1}{2} - 2\lambda_{\min}(\Gamma_k^a) \right] + \frac{b_{g2} \lambda_{\min}(R_{22}^{-1}) \phi_{\text{amax},2} \phi_{\text{cdmax},2}}{4} \right\} \|\tilde{W}_{c,2}\|^2. \tag{52}
\end{aligned}$$

Therefore, $\dot{\mathcal{V}} \leq 0$ if

$$\begin{aligned}
b_{gi} \phi_{\text{amax},i} + \frac{b_{gi} \lambda_{\min}(R_{ii}^{-1}) \phi_{\text{amax},i} \phi_{\text{cdmax},i}}{4} - \phi_{\text{amax},i}^2 + \frac{\phi_{\text{amax},i}}{2} & < 0 \\
\alpha_{c,i} \left[\frac{p+1}{2} - 2\lambda_{\min}(\Gamma_k^a) \right] + \frac{b_{gi} \lambda_{\min}(R_{ii}^{-1}) \phi_{\text{amax},i} \phi_{\text{cdmax},i}}{4} & < 0 \\
\rho - (\bar{V}_{c,2} + \bar{V}_{c,1}) (b_{g1} + b_{g2}) B \|s\|^2 \frac{1_m}{(A + s^T s)} & < 0,
\end{aligned}$$

for each player, which is guaranteed by the conditions in (35)-(37). From Barbalat's lemma [62], it follows that $\chi \rightarrow 0$ as $t \rightarrow \infty$.

The result holds as long as we can show that the state $s(t)$ remains in the set $\Omega \in \mathbb{R}^n$ for all times. To this effect, define the following compact set $M := \{s \in \mathbb{R}^n \mid \mathcal{V}(t) \leq m\} \subset \mathbb{R}^n$ where m is chosen as the largest constant, so that $M \subset \Omega$. Since by assuming $s_0 \in \Omega_s$ and $\Omega_s \in \Omega$, we can conclude that $s_0 \in \Omega$. While $s(t)$ remains inside Ω , we have seen that $\dot{\mathcal{V}} \leq 0$, and therefore, $s(t)$ must remain inside $M \subset \Omega$. The fact that $s(t)$ remains inside a compact set also excludes the possibility of a finite escape time, and therefore, one has a global existence of solution. ■

4 | SIMULATIONS

To verify the effectiveness of the presented online safe RL algorithm with the actor-critic-barrier structure, we consider the following nonlinear system used by Vamvoudakis and Lewis for the dynamical system with two players as,⁶²

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= f(x) + g_1(x) u_1 + g_2(x) u_2, \tag{53}
\end{aligned}$$

with

$$\begin{aligned}
f(x) &= -x_2 - \frac{x_1}{2} + \frac{x_2(\cos(2x_1) + 2)^2}{4} + \frac{x_2(\sin(2x_1) + 2)^2}{4} \\
g_1(x) &= \cos(2x_1) + 2, \quad g_2(x) = \sin(4x_1^2) + 2,
\end{aligned}$$

where $x \in \mathbb{R}^2$ is the state and $u_i \in \mathbb{R}$, $i = 1, 2$ is the control input of each player.

For Problem 1, a performance function with the following reward function is considered

$$U_i(x, u_1, u_2) = x^T H_i x + \sum_{j=1}^2 u_j^T R_{ij} u_j, i = 1, 2,$$

with

$$H_1 = 2I_2, H_2 = I_2, R_{11} = 2I_2, R_{12} = 2I_2, R_{21} = I_2, R_{22} = I_2.$$

According to the optimal control theory, for each player, the optimal value function can be determined as,⁶²

$$V_1^*(x) = \frac{1}{2}x_1^2 + x_2^2, V_2^*(x) = \frac{1}{4}x_1^2 + \frac{1}{2}x_2^2, \quad (54)$$

with the optimal policies as,

$$u_1^*(x) = -2[\cos(2x_1) + 2]x_2, u_2^*(x) = -[\sin(4x_1^2) + 2]x_2. \quad (55)$$

For Problem 2, the following safety constraints are considered

$$x_i \in (a_i, A_i), \forall i \in \{1, 2\}, \quad (56)$$

where $a_1 = -1.3, A_1 = 0.5, a_2 = -3.1$, and $A_2 = 0.5$. With the barrier function, one can obtain the transformed system as

$$\begin{aligned} \dot{s}_1 &= \frac{a_2 A_2 \left(e^{\frac{s_2}{2}} - e^{-\frac{s_2}{2}} \right) A_1^2 e^{-s_1} - 2a_1 A_1 + a_1^2 e^{s_1}}{a_2 e^{\frac{s_2}{2}} - A_2 e^{-\frac{s_2}{2}}} \frac{A_1 a_1^2 - a_1 A_1^2}{A_1 a_1^2 - a_1 A_1^2} \\ \dot{s}_2 &= [f(x) + g_1(x)u_1 + g_2(x)u_2] \frac{A_2^2 e^{-s_2} - 2a_2 A_2 + a_2^2 e^{s_2}}{A_2 a_2^2 - a_2 A_2^2}. \end{aligned} \quad (57)$$

The initial condition for the original system (53) is selected as

$$x_0 = [x_0(1) \ x_0(2)]^T = [-1 \ -3]^T.$$

With the barrier transformation (4), one can obtain the initial condition for the transformed system (57) as

$$s_0 = [s_0(1) \ s_0(2)]^T = [-2.5649 \ -5.3799]^T, s_0(1) = b(x_0(1); a_1, A_1), s_0(2) = b(x_0(2); a_2, A_2).$$

For Problem 2, the reward function is selected as

$$r_i(s, u_1, u_2) = s^T Q_i s + \sum_{j=1}^2 u_j^T R_{ij} u_j, i = 1, 2,$$

with

$$\begin{aligned} Q_1 &= 2I_2, \quad Q_2 = I_2 \\ R_{11} &= 2I_2, \quad R_{12} = 2I_2, \quad R_{21} = I_2, \quad R_{22} = I_2. \end{aligned}$$

Based on the actor-critic-barrier online learning algorithm, the state evolution of state $s(t)$ in system (57) is given in Figure 2. One can observe that the state $s(t)$ of system (57) converges to the origin asymptotically. Based on the state

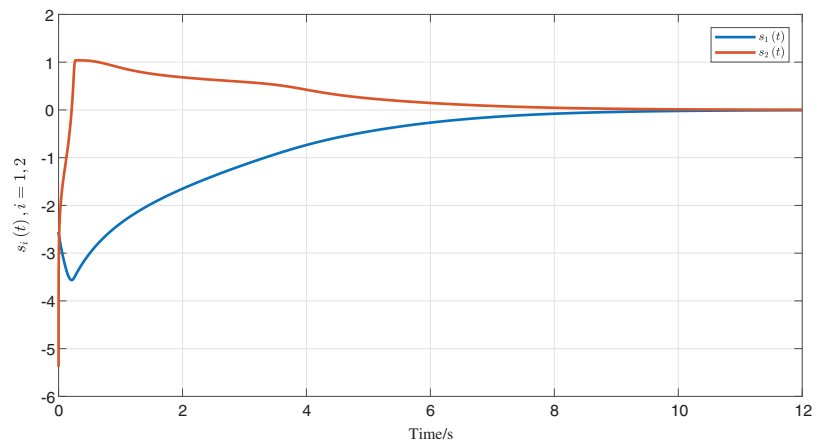


FIGURE 2 Evolution of the state $s(t)$ of the system (57) by using the online actor-critic-barrier safe reinforcement learning algorithm [Colour figure can be viewed at wileyonlinelibrary.com]

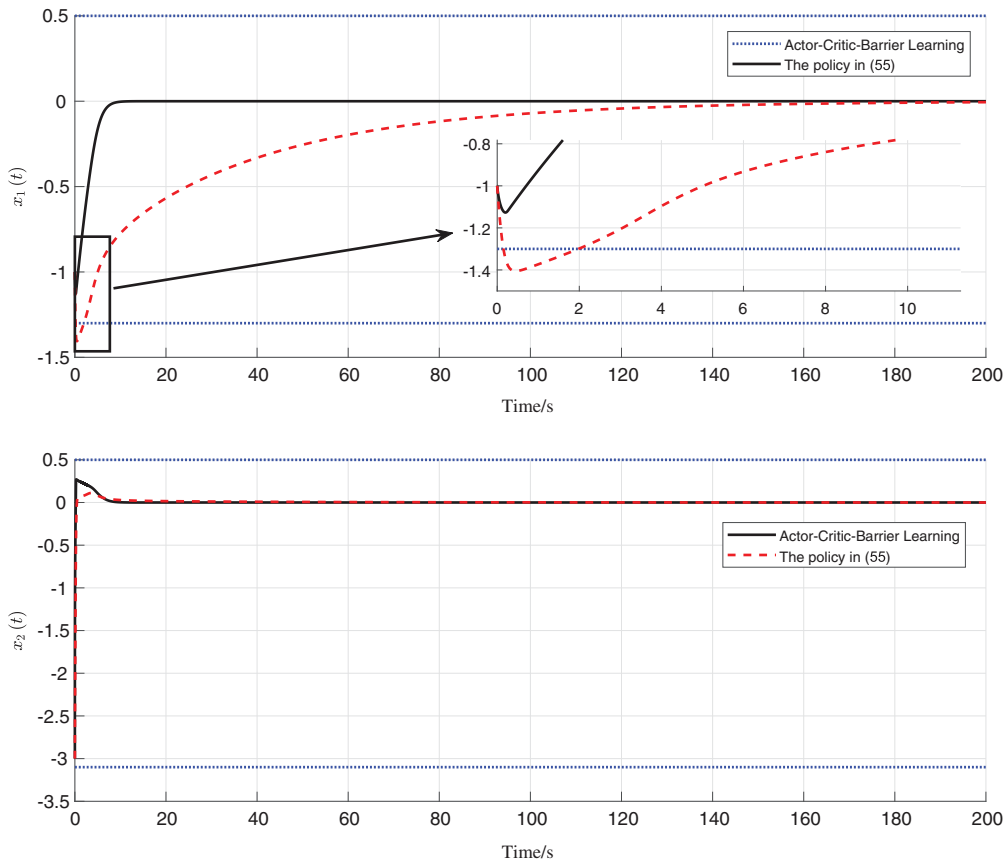


FIGURE 3 Evolution of the state $x(t)$ by using the presented actor-critic-barrier online learning and the policy in (55). The dotted line represents the boundary of the safe region [Colour figure can be viewed at wileyonlinelibrary.com]

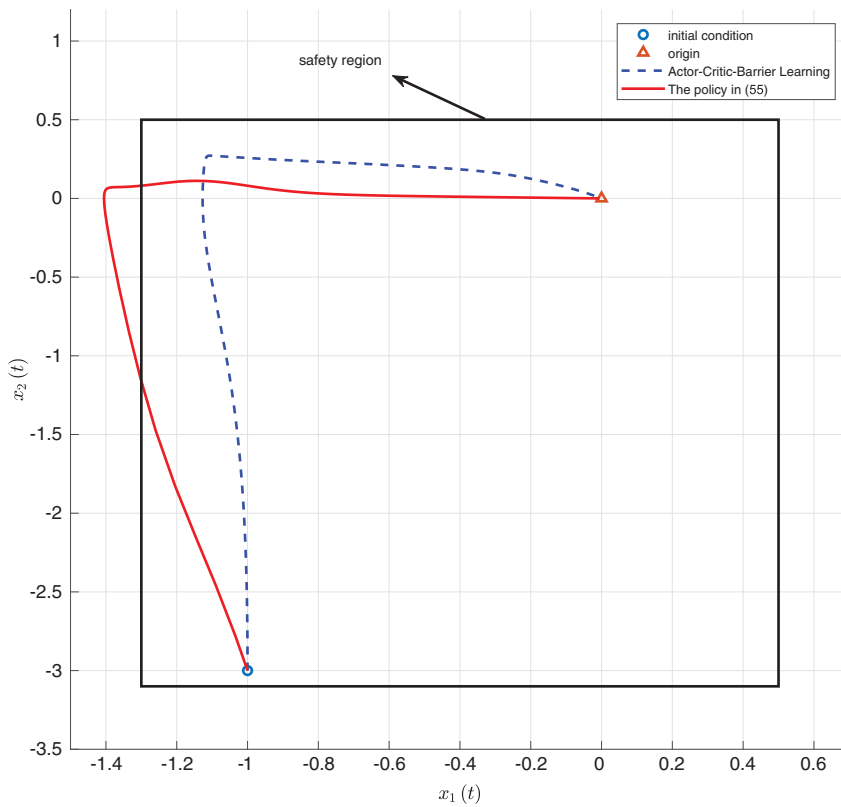
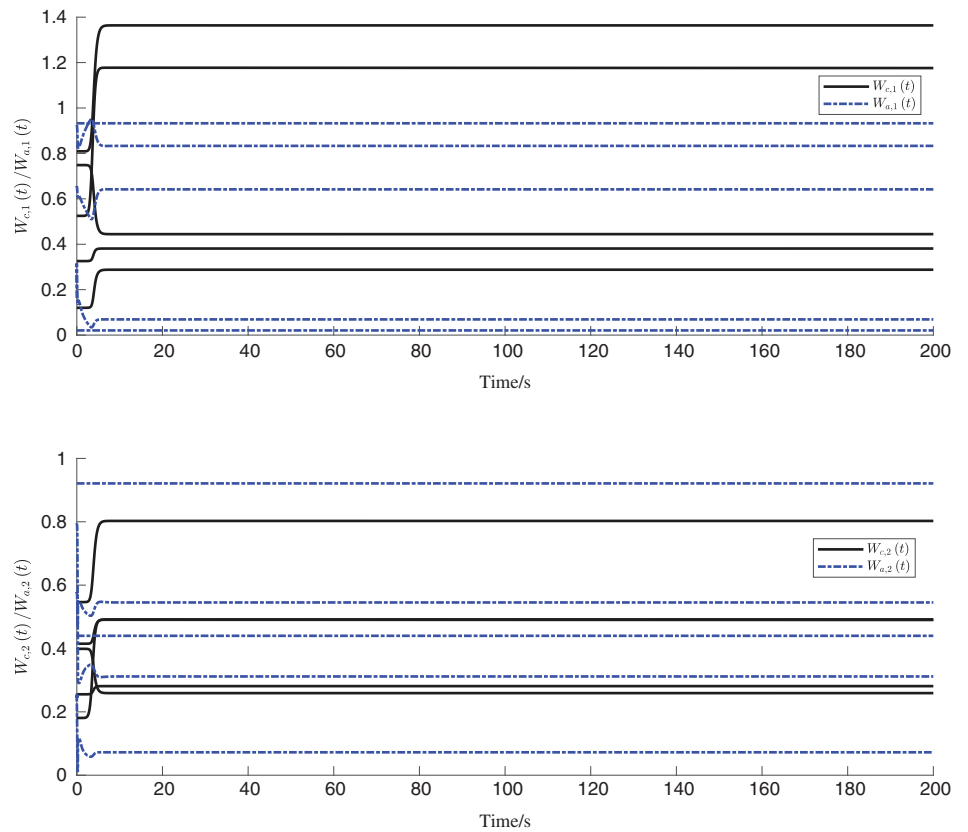


FIGURE 4 Evolution of the two-dimensional phase plot of the state trajectories $[x_1(t), x_2(t)]$. The black box denotes the safe region [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 5 Evolution of the actor and critic weights during the learning process [Colour figure can be viewed at wileyonlinelibrary.com]



evolution of $s(t)$, by using the barrier function inverse mapping (5), one can obtain the state $x(t)$ as

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

Then, the evolution of the state $x(t)$ is shown in Figure 3. One can observe that with the policy (55), the state evolution exceeds the boundary of the safe region and safety can not be guaranteed. When the presented online actor-critic-barrier safe reinforcement learning algorithm is applied, the state $x(t)$ converges to the origin asymptotically while satisfying the safety constraints (56). The phase portrait of the state evolution $[x_1(t) \ x_2(t)]$ is provided in Figure 4. The black box represents the safe region. The same as Figure 3, one can observe that the state $x(t)$ runs moves of the safe region for the optimal policy (55) during the learning process. Finally, the learning process of the actor-critic networks for each player is shown in Figure 5. One can observe that the learning process converges fast within 10 seconds by using the experience replay technique.

5 | CONCLUSION

We presented a novel actor-critic-barrier learning algorithm for the safety-critical control systems. A barrier function was combined with an actor-critic structure to find the Nash equilibrium in an online fashion while guaranteeing safety. It is shown that the addition of the barrier function to the actor-critic structure guarantees that the constraints will not be violated during learning. To obviate the requirement of PE, an experience replay technique is employed by using the recorded and current data concurrently. Boundedness of the closed-loop signals is analyzed.

ACKNOWLEDGEMENTS


This work was supported in part by the National Natural Science Foundation of China under Grant No. 61903028, in part by the China Post-Doctoral Science Foundation under Grant 2018M641197, in part by the Fundamental Research Funds for the Central Universities of China under Grant FRF-TP-18-031A1 and FRF-BD-19-002A, in part by the National

Science Foundation under grant NSF CAREER CPS-1851588, in part by NATO under grant No. SPS G5176 and in part by ONR Minerva under grant No. N00014-18-1-2160.

ORCID

Yongliang Yang  <https://orcid.org/0000-0002-3144-8604>

Kyriakos G. Vamvoudakis  <https://orcid.org/0000-0003-1978-4848>

Hamidreza Modares  <https://orcid.org/0000-0003-0800-5140>

REFERENCES

1. Saberi A, Lin Z, Teel AR. Control of linear systems with saturating actuators. *IEEE Trans Autom Control*. 1996;41(3):368-378.
2. Abu-Khalaf M, Lewis FL. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*. 2005;41(5):779-791.
3. Chen M, Ge SS, Ren B. Adaptive tracking control of uncertain MIMO nonlinear systems with input constraints. *Automatica*. 2011;47(3):452-465.
4. Tee KP, Ge SS, Tay EH. Barrier Lyapunov functions for the control of output-constrained nonlinear systems. *Automatica*. 2009;45(4):918-927.
5. Ren B, Ge SS, Tee KP, Lee TH. adaptive neural control for output feedback nonlinear systems using a barrier Lyapunov function. *IEEE Trans Neural Netw*. 2010;21(8):1339-1345.
6. Fan B, Yang Q, Tang X, Sun Y. Robust ADP design for continuous-time nonlinear systems with output constraints. *IEEE Trans Neural Netw Learn Syst*. 2018;29(6):2127-2138.
7. Wang L, Ames AD, Egerstedt M. Safety barrier certificates for collisions-free multirobot systems. *IEEE Trans Robot*. 2017;33(3):661-674.
8. Panagou D, Stipanović DM, Voulgaris PG. Distributed coordination control for multi-robot networks using Lyapunov-like barrier functions. *IEEE Trans Autom Control*. 2016;61(3):617-632.
9. He W, Huang H, Ge SS. Adaptive neural network control of a robotic manipulator with time-varying output constraints. *IEEE Trans Cybern*. 2017;47(10):3136-3147.
10. He W, Li Z, Chen CLP. A survey of human-centered intelligent robots: issues and challenges. *IEEE/CAA J Autom Sinica*. 2017;4(4):602-609.
11. Prajna S. Barrier certificates for nonlinear model validation. *Automatica*. 2006;42(1):117-126.
12. Sontag ED. A 'universal' construction of Artstein's theorem on nonlinear stabilization. *Syst Control Lett*. 1989;13(2):117-123.
13. Wieland P, Allgöwer F. Constructive safety using control barrier functions. *IFAC Proc Vol*. 2007;40(12):462-467.
14. Wang L, Han D, Egerstedt M. Permissive barrier certificates for safe stabilization using sum-of-squares. Paper presented at: Proceedings of the 2018 Annual American Control Conference (ACC); 2018:585-590; IEEE.
15. Ames AD, Xu X, Grizzle JW, Tabuada P. Control barrier function based quadratic programs for safety critical systems. *IEEE Trans Autom Control*. 2017;62(8):3861-3876.
16. Boyd S, Vandenberghe L. *Convex Optimization*. Cambridge, MA: Cambridge university press; 2004.
17. Liu YJ, Tong S. Barrier Lyapunov functions-based adaptive control for a class of nonlinear pure-feedback systems with full state constraints. *Automatica*. 2016;64:70-75.
18. Bechlioulis CP, Rovithakis GA. Robust adaptive control of feedback linearizable MIMO nonlinear systems with prescribed performance. *IEEE Trans Autom Control*. 2008;53(9):2090-2099.
19. Bechlioulis CP, Rovithakis GA. Prescribed performance adaptive control for multi-input multi-output affine in the control nonlinear systems. *IEEE Trans Autom Control*. 2010;55(5):1220-1226.
20. Na J. Adaptive prescribed performance control of nonlinear systems with unknown dead zone. *Int J Adapt Control Signal Process*. 2013;27(5):426-446.
21. Liu L, Liu Y, Tong S. Fuzzy-based multierror constraint control for switched nonlinear systems and its applications. *IEEE Trans Fuzzy Syst*. 2019;27(8):1519-1531.
22. Arabi E, Yucelen T. Set-theoretic model reference adaptive control with time-varying performance bounds. *Int J Control*. 2019;92(11):2509-2520.
23. Arabi E, Yucelen T, Gruenwald BC, Fravolini M, Balakrishnan S, Nguyen NT. A neuroadaptive architecture for model reference control of uncertain dynamical systems with performance guarantees. *Syst Control Lett*. 2019;125:37-44.
24. Arabi E, Gruenwald BC, Yucelen T, Nguyen NT. A set-theoretic model reference adaptive control architecture for disturbance rejection and uncertainty suppression with strict performance guarantees. *Int J Control*. 2018;91(5):1195-1208.
25. Vamvoudakis KG, Modares H, Kiumarsi B, Lewis FL. Game theory-based control system algorithms with real-time reinforcement learning: how to solve multiplayer games online. *IEEE Control Syst*. 2017;37(1):33-52.
26. He J, Zhang H. Iterative ADP learning algorithms for discrete-time multi-player games. *Artif Intell Rev*. 2018;50(1):1-17.
27. Lin W. Mixed H_2/H_∞ control via state feedback for nonlinear systems. *Int J Control*. 1996;64(5):899-922.
28. Zhang H, Wei Q, Liu D. An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games. *Automatica*. 2011;47(1):207-214.
29. Starr AW, Ho YC. Nonzero-sum differential games. *J Optim Theory Appl*. 1969;3(3):184-206.

30. Kiumarsi B, Vamvoudakis KG, Modares H, Lewis FL. Optimal and autonomous control using reinforcement learning: a survey. *IEEE Trans Neural Netw Learn Syst.* 2018;29(6):2042-2062.
31. Yang Y, Vamvoudakis KG, Ferraz H, Modares H. Dynamic intermittent Q-learning-based model-free suboptimal co-design of \mathcal{L}_2 -stabilization. *Int J Robust Nonlinear Control.* 2019;29(9):2673-2694.
32. Li J, Ding J, Chai T, Lewis FL. Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes. *IEEE Trans Cybern.* 2019. <https://doi.org/10.1109/TCYB.2019.2950262>.
33. Kamalapurkar R, Klotz JR, Dixon WE. Concurrent learning-based approximate feedback-nash equilibrium solution of N-player nonzero-sum differential games. *IEEE/CAA J Autom Sinica.* 2014;1(3):239-247.
34. Johnson M, Kamalapurkar R, Bhasin S, Dixon WE. Approximate N -player nonzero-sum game solution for an uncertain continuous nonlinear system. *IEEE Trans Neural Netw Learn Syst.* 2015;26(8):1645-1658.
35. Vamvoudakis KG, Lewis FL. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica.* 2010;46(5):878-888.
36. Yang Y, Wunsch D, Yin Y. Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems. *IEEE Trans Neural Netw Learn Syst.* 2017;28(8):1929-1940.
37. Zhang H, Luo Y, Liu D. Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Trans Neural Netw.* 2009;20(9):1490-1503.
38. Li J, Chai T, Lewis FL, Ding Z, Jiang Y. Off-policy interleaved Q-learning: optimal control for affine nonlinear discrete-time systems. *IEEE Trans Neural Netw Learn Syst.* 2019;30(5):1308-1320.
39. Modares H, Lewis FL. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica.* 2014;50(7):1780-1792.
40. Kiumarsi B, Lewis FL. Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems. *IEEE Trans Neural Netw Learn Syst.* 2015;26(1):140-151.
41. Jiang Y, Fan J, Chai T, Lewis FL. Dual-rate operational optimal control for flotation industrial process with unknown operational model. *IEEE Trans Ind Electron.* 2019;66(6):4587-4599.
42. Jiang Y, Kiumarsi B, Fan J, Chai T, Li J, Lewis FL. Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning. *IEEE Trans Cybern.* 2019. <https://doi.org/10.1109/TCYB.2018.2890046>.
43. Chen C, Modares H, Xie K, Lewis FL, Wan Y, Xie S. Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics. *IEEE Trans Autom Control.* 2019;64(11):4423-4438.
44. Yang Y, Modares H, Wunsch DC, Yin Y. Leader-follower output synchronization of linear heterogeneous systems with active leader using reinforcement learning. *IEEE Trans Neural Netw Learn Syst.* 2018;29(6):2139-2153.
45. Li J, Modares H, Chai T, Lewis FL, Xie L. Off-policy reinforcement learning for synchronization in multiagent graphical games. *IEEE Trans Neural Netw Learn Syst.* 2017;28(10):2434-2445.
46. Yang Y, Modares H, Wunsch DC, Yin Y. Optimal containment control of unknown heterogeneous systems with active leaders. *IEEE Trans Control Syst Technol.* 2019;27(3):1228-1236.
47. Jiang Y, Fan J, Chai T, Lewis FL, Li J. Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout. *IEEE Trans Neural Netw Learn Syst.* 2018;29(10):4607-4620.
48. Yang Y, Guo Z, Xiong H, Ding D, Yin Y, Wunsch DC. Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning. *IEEE Trans Neural Netw Learn Syst.* 2019;30(12):3735-3747.
49. Liu D, Yang X, Wang D, Wei Q. Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints. *IEEE Trans Cybern.* 2015;45(7):1372-1385.
50. Modares H, Lewis FL, Naghibi-Sistani MB. Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. *IEEE Trans Neural Netw Learn Syst.* 2013;24(10):1513-1525.
51. Vamvoudakis KG, Miranda MF, Hespanha JP. Asymptotically stable adaptive-optimal control algorithm with saturating actuators and relaxed persistence of excitation. *IEEE Trans Neural Netw Learn Syst.* 2016;27(11):2386-2398.
52. Vrabie D, Vamvoudakis KG, Lewis FL. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London: Institution of Engineering and Technology; 2012.
53. He W, Chen Y, Yin Z. Adaptive neural network control of an uncertain robot with full-state constraints. *IEEE Trans Cybern.* 2016;46(3):620-629.
54. Abu-Khalaf M, Lewis FL. Nearly optimal state feedback control of constrained nonlinear systems using a neural networks HJB approach. *Annu Rev Control.* 2004;28(2):239-251.
55. Yang Y, Ding DW, Xiong H, Yin Y, Wunsch DC. Online barrier-actor-critic learning for H_∞ control with full-state constraints and input saturation. *J Franklin Inst.* 2019. <https://doi.org/10.1016/j.jfranklin.2019.12.017>.
56. Zhao D, Zhang Q, Wang D, Zhu Y. Experience replay for optimal control of nonzero-sum game systems with unknown dynamics. *IEEE Trans Cybern.* 2016;46(3):854-865.
57. Zhang Q, Zhao D. Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics. *IEEE Trans Cybern.* 2019;49(8):2874-2885.
58. Finlayson BA. *The Method of Weighted Residuals and Variational Principles*. Vol 73. Thailand, Asia: SIAM; 2013.
59. Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Netw.* 1990;3(5):551-560.
60. Tao G. *Adaptive Control Design and Analysis*. Vol 37. Hoboken, NJ: John Wiley & Sons; 2003.

61. Khalil HK. *Nonlinear Systems*. 3rd ed. Upper Saddle River, NJ: Prentice Hall; 2002.
62. Vamvoudakis KG, Lewis FL. Multi-player non-zero-sum games: online adaptive learning solution of coupled Hamilton-Jacobi equations. *Automatica*. 2011;47(8):1556-1569.
63. Modares H, Lewis FL, Naghibi-Sistani MB. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica*. 2014;50(1):193-202.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Yang Y, Vamvoudakis KG, Modares H. Safe reinforcement learning for dynamical games. *Int J Robust Nonlinear Control*. 2020;30:3706–3726. <https://doi.org/10.1002/rnc.4962>