Neurocomputing 412 (2020) 167–176

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Neural networks-based optimal tracking control for nonzero-sum games of multi-player continuous-time nonlinear systems via reinforcement learning

Jingang Zhao

Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Ministry of Education, School of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, PR China
School of Automation, Beijing Institute of Technology, Beijing 100081, PR China

ARTICLE INFO

Article history:
Received 21 August 2019
Revised 23 April 2020
Accepted 20 June 2020
Available online 24 June 2020
Communicated by Y. Yuan

Keywords:
Neural networks
Multi-player nonzero-sum game
Optimal tracking control
Continuous-time nonlinear systems
Coupled Hamilton–Jacobi equations
Reinforcement learning

ABSTRACT

In this paper, optimal tracking control for nonzero-sum games of multi-player continuous-time nonlinear systems is investigated by using a novel reinforcement learning scheme. Based on the multi-player nonlinear systems and reference signal, we firstly formulate the tracking problem by constructing an augmented multi-player nonlinear systems. The optimal tracking control problem for nonzero-sum games of original multi-player nonlinear systems is thus transformed into solving the coupled Hamilton–Jacobi equations of the augmented multi-player nonlinear systems. The novel neural networks (NNs) – based online reinforcement learning (RL) method can learn the solution to coupled Hamilton–Jacobi equations in a forward-in-time manner without requiring any value, policy iterations. In order to relax the dependence of the traditional reinforcement learning method on Persistence of Excitation (PE) conditions, historical data from a period of time has been collected to design NNs tuning laws. The drift dynamic of the augmented system is not required in our scheme. The Uniformly Ultimately Boundedness (UUB) of NNs weight errors and closed-loop augmented system states are rigorous proved. Numerical simulation examples are given to demonstrate the effectiveness of our proposed scheme.

1. Introduction

With the development of artificial intelligence, game theory has played an important role in many fields, such as economics [1,2], multi-agent collision avoidance [3], network security [4,5], cyber-physical security [6–8]. The study of nonzero-sum game theory can be originally traced back to [9]. Game theory provides an ideal environment to investigate multi-player optimal decision and control problems [10–13]. In a multi-player nonzero-sum game, each player chooses an optimal control input to minimize independently its own performance objective, which depends on the actions of itself and all the other players [14]. The set of the optimal control inputs corresponds to the Nash equilibrium. The Nash equilibrium solution for nonlinear systems game can be obtained by solving coupled Hamilton–Jacobi equations, and they get reduced to solve coupled algebraic Riccati equations for linear systems game. However, as we all know, owing to the nonlinear nature and the coupling of players, it is extremely difficult to solve coupled Hamilton–Jacobi equations or coupled algebraic Riccati equations. Therefore, many approximation-based intelligent methods are developed to tackle multi-player non-zero-sum games.

Reinforcement learning is a biologically inspired approximate intelligent method and can handle optimization problems with model uncertainty or unknown dynamics, adaptive dynamic programming [15–19] or approximate dynamic programming (ADP) [20,21] also belongs to the category of reinforcement learning, which overcomes the disadvantage of traditional dynamic programming, such as the curse of modeling and the curse of dimensionality [22–24].

Reinforcement learning-based multi-player nonzero-sum games have been of considerable interest to the control system community during the past few decades. In [25], the near-Nash equilibrium control strategies are investigated for a class of discrete-time nonlinear systems subjected to the round-robin protocol. The authors in [26] develop a novel Q-learning algorithm to solve the problem of N-player non-zero sum Nash games of unknown continuous-time linear systems. In [27], the authors present an online policy iteration-based reinforcement learning algorithm to solve the continuous-time multi-player nonzero-sum game for nonlinear and linear systems. Data-driven reinforcement learning methods

E-mail address: zhaojingang521@126.com

https://doi.org/10.1016/j.neucom.2020.06.083
0925-2312/

are proposed to solve the discrete-time nonzero-sum games for nonlinear systems [28,29] and continuous-time nonzero-sum games for nonlinear systems [30,31]. A novel actor-critic-identifier structure reinforcement learning method is used to approximate N-player nonzero sum game solutions for uncertain continuous-time nonlinear systems in [32]. A novel policy iteration-based ADP method is proposed to tackle the cooperative game issue of discrete-time multi-player systems with control input constraints in [33]. The authors of [34] present an off-policy integral reinforcement learning method to solve nonzero sum games for unknown continuous-time nonlinear systems. In [35,14], optimal control of nonzero-sum game systems with unknown Dynamics is well tackled via ADP method. The authors of [36] develop a single-network ADP scheme to solve the nonzero-sum differential games of continuous-time nonlinear systems. It is easy to see that all of the above studies are about optimal control of multi-player nonzero-sum games. And the RL methods in these studies are mostly based on policy iterations or value iterations. As we all know, policy iterations usually require initial admissible control, while value iterations generally converge very slowly. Moreover, the existing reinforcement learning methods require PE conditions for an initial period of time. These characteristics limit the online practical application of these existing RL methods. In addition, RL methods have been widely used to tackle the multi-player nonzero-sum game and optimal tracking control problem, such as H_∞ optimal tracking control for linear discrete-time systems [37], optimal tracking control of nonlinear partially-unknown systems[38,39], optimal tracking control for nonlinear discrete-time MIMO systems [40], but few results consider solving the optimal tracking control for nonzero-sum games of multi-player systems. However, the optimal tracking control problem based on multiplayer non-zero-sum game has its theoretical and application value [2]. In [41], optimal tracking problem based on multiplayer non-zero-sum games for discrete-time linear systems is addressed by using a model-free off-policy reinforcement learning algorithm. Studying the optimal tracking control for nonzero-sum games of multi-player systems is actually solving the coupled Hamilton–Jacobi equations. As we all know, it is extremely difficult or impossible to obtain the analytical solution of the coupled Hamilton–Jacobi equations.

Therefore, this paper develops a new reinforcement learning method without requiring any value, policy iterations to deal with the optimal tracking control for non-zero-sum games of multi-player nonlinear systems. The main innovations of this note are summarized in the following five aspects.

1) To our best knowledge, optimal tracking control for multi-player non-zero-sum games for nonlinear systems may be the first to be studied.
2) A novel NNs-based online reinforcement learning scheme is proposed to approximately solve the coupled Hamilton–Jacobi equations of augmented continuous-time multi-player systems. In our scheme, historical data from a period of time has been collected to relax the traditional PE condition and the stability of systems is considered during the learning. A new NNs weight tuning laws is thus proposed.
3) The optimal solution of coupled Hamilton–Jacobi equations are learned in a forward-in-time manner instead of traditional value iterations or policy iterations manners. So our designed scheme can be thus better applied online.
4) The system drift dynamics is not required in our developed scheme. That is to say, our proposed scheme allows the multi-player nonlinear systems to be partially unknown.
5) The UUB of NNs weight errors and closed-loop augmented system states are rigorous proved. The value functions and the control inputs for players are also proved to be converged to approximately optimal value functions and optimal control inputs with a small bounded error.

The structure of this note is described as follows. In Section 2, we introduce the problem formulation. In Section 3, a NNs-based online reinforcement learning scheme is presented to solve the coupled Hamilton–Jacobi equations of augmented multi-player systems. In Section 4, the stability and convergence of our scheme are provided by Lyapunov approach. Simulation studies on linear systems and nonlinear systems are given to demonstrate the effectiveness of our design scheme in Section 5. Section 6 concludes this note and gives the direction of future research.

2. Problem formulation

Consider the following N-player continuous-time nonlinear systems game

$$\dot{x}(t) = f(x(t)) + \sum_{j=1}^{N} g_j(x(t))u_j(t) \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the measurable system states, $u_j(t) \in \mathbb{R}^{m_j}$ is each control input or player, $f(x) \in \mathbb{R}^n$ is the system drift dynamics, $g_j(x) \in \mathbb{R}^{n\times m_j}$ are the system input dynamics.

Assumption 1. $f(x), g_j(x)$ is locally Lipschitz and $f(0) = 0$. $\|f(x)\| \leqslant b_f\|x\|, \|g_j(x)\| \leqslant b_{gj}$, where b_f and b_{gj} are constants.

The desired reference signal is generated by the following continuous bounded Lipschitz command

$$\dot{r}(t) = f_d(r(t)) \tag{2}$$

where $r(t) \in \mathbb{R}^n$ is the reference signal and the reference signal needs only to be stable in the sense of Lyapunov, not necessarily asymptotically stable.

The objective of this paper is to find an N-tuple of optimal control inputs $\{u_1^*, u_2^*, \ldots, u_N^*\}$ with $i \in \hat{N} = \{1, 2, \ldots, N\}$ so as to make $x(t)$ follow the reference signal $r(t)$ in an optimal manner which can minimize a predefined cost function for each player i.

In order to achieve tracking of the reference signal, define the tracking error as follow

$$e_r(t) = x(t) - r(t) \tag{3}$$

The infinite-horizon predefined cost function associated with each player i ($i \in \hat{N}$) are given as follow

$$\begin{aligned} &J_i(e_r(0), u_1, u_2, \ldots, u_N) \\ &= \int_0^\infty e^{-\gamma(\tau-t)}\left(e_r^T(\tau)Q_i e_r(\tau) + \sum_{j=1}^{N} u_j^T(\tau)R_{ij}u_j(\tau)\right)d\tau \\ &= \int_0^\infty e^{-\gamma(\tau-t)}U_i(e_r(\tau), u_1(\tau), u_2(\tau), \ldots, u_N(\tau))d\tau \end{aligned} \tag{4}$$

where $Q_i = Q_i^T \geqslant 0,\ R_{ii} = R_{ii}^T > 0,\ R_{ij} = R_{ij}^T \geqslant 0, \gamma$ is discount factor.

According to (3), the tracking error dynamic can be rewritten as

$$\dot{e}_r(t) = f(e_r(t) + r(t)) + \sum_{j=1}^{N} g_j(e_r(t) + r(t))u_j(t) - f_d(r(t)) \tag{5}$$

Then, we can define the augmented state $\xi = \begin{bmatrix} e_r^T & r^T \end{bmatrix}^T \in \mathbb{R}^{2n}$, and the augmented system dynamics comprised of (2) and (5) can be thus written as

$$\dot{\xi}(t) = F(\xi(t)) + \sum_{j=1}^{N} G_j(\xi(t))u_j(t) \tag{6}$$

where $F(\xi(t)) = \begin{bmatrix} f(e_r + r) - f_d(r) \\ f_d(r) \end{bmatrix}$ and $G_j(\xi(t)) = \begin{bmatrix} g_j(e_r + r) \\ 0 \end{bmatrix}$.

Assumption 2. $\|F(\xi)\| \leqslant b_F\|\xi\|, \|G_j(\xi)\| \leqslant b_{Gj}$, where b_F and b_{Gj} are constants.

The infinite-horizon predefined cost function corresponding to (6) associated with each player i ($i \in \hat{N}$) are defined as follow

$$\begin{aligned} &\bar{J}_i(\xi(0), u_1, u_2, \ldots, u_N) \\ &= \int_0^\infty e^{-\gamma(\tau - t)} \left(\xi^T(\tau)\overline{Q}_i\xi(\tau) + \sum_{j=1}^N u_j^T(\tau)R_{ij}u_j(\tau) \right) d\tau \\ &= \int_0^\infty e^{-\gamma(\tau - t)} \overline{U}_i(\xi(\tau), u_1(\tau), u_2(\tau), \ldots, u_N(\tau)) d\tau \end{aligned} \tag{7}$$

where $\bar{Q}_i = \begin{bmatrix} Q_i & 0_{n\times n} \\ 0_{n\times n} & 0_{n\times n} \end{bmatrix}$., $R_{ii} = R_{ii}^T > 0, R_{ij} = R_{ij}^T \geqslant 0$.

Remark 1. Obviously, the cost function (7) is identical to the cost function (4). Therefore, the optimal tracking control for nonzero-sum games of (1) can be obtained by solving the optimal control for nonzero-sum games of (6).

Definition 1. [27] *(Admissible control).* For $i \in \hat{N}$, the feedback control input u_i is admissible with respect to (7) on a compact set $\Omega \in \mathbb{R}^n$, if u_i is continuous on $\Omega, u_i(0) = 0, u_i$ stabilizes (6) on Ω, and (7) is finite $\forall x_0 \in \Omega$.

Given an admissible feedback control input u_i with $i \in \hat{N}$, the value functions are defined as

$$\begin{aligned} &V_i(\xi, u_1, u_2, \ldots, u_N) \\ &= \int_t^\infty e^{-\gamma(\tau - t)} \left(\xi^T\overline{Q}_i\xi + \sum_{j=1}^N u_j^T R_{ij}u_j \right) d\tau \\ &= \int_t^\infty e^{-\gamma(\tau - t)} \overline{U}_i(\xi, u_1, u_2, \ldots, u_N) d\tau \end{aligned} \tag{8}$$

Definition 2. [10] A N-tuple of control inputs $\{u_1^*, u_2^*, \cdots, u_N^*\}$ with $i \in \hat{N}$ is said to constitute a Nash equilibrium solution for an N-player game, if the N inequalities in the following are satisfied

$$\bar{J}_i^* = \bar{J}_i(u_1^*, u_2^*, \ldots, u_N^*) \leqslant \bar{J}_i(u_1^*, u_2^*, \ldots, u_i, \ldots, u_N^*) \tag{9}$$

Assume that value functions (8) are continuously differentiable, for $i \in \hat{N}$. We use Leibniz's rule to differentiate V_i along the augmented system dynamics (6). The infinitesimal version of (8) are thus obtained as

$$0 = \overline{U}_i(\xi, u_1, u_2, \ldots, u_N) - \gamma V_i + \nabla V_i^T \left(F(\xi) + \sum_{j=1}^N G_j(\xi)u_j \right) \tag{10}$$

where $V_i(0) = 0, \nabla V_i = \frac{\partial V_i}{\partial \xi}, \nabla V_i^T$ denotes the transpose of ∇V_i.

Define the Hamiltonian functions as

$$\begin{aligned} H_i(\xi, \nabla V_i, u_1, u_2, \ldots, u_N) &= \overline{U}_i(\xi, u_1, u_2, \ldots, u_N) \\ &-\gamma V_i + \nabla V_i^T \left(F(\xi) + \sum_{j=1}^N G_j(\xi)u_j \right), \quad i \in \hat{N} \end{aligned} \tag{11}$$

According to the stationarity conditions $\frac{\partial H_i}{\partial u_i} = 0$ [27], for $i \in \hat{N}$, the associated state feedback control inputs can be given by

$$u_i(\xi) = -\frac{1}{2}R_{ii}^{-1}G_i^T(\xi)\nabla V_i, i \in \hat{N} \tag{12}$$

By substituting (12) into (10), the coupled Hamilton–Jacobi equations of augmented multi-player nonlinear systems (6) can be thus obtained as

$$\begin{aligned} 0 &= \nabla V_i^T F(\xi) + \xi^T Q_i \xi - \gamma V_i \\ &-\tfrac{1}{2}\nabla V_i^T \sum_{j=1}^N G_j(\xi)R_{jj}^{-1}G_j^T(\xi)\nabla V_j \end{aligned} \tag{13}$$

$$+\frac{1}{4}\sum_{j=1}^N \nabla V_j^T G_j(\xi)R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j^T\nabla V_j, V_i(0) = 0 \tag{14}$$

Now, it is easy to see that optimal tracking control for the continuous-time multi-player non-zero sum game is transformed to solve the coupled Hamilton–Jacobi equations of augmented multi-player nonlinear systems (6). However, due to the nonlinear nature, the coupled Hamilton–Jacobi equations can not generally be solved directly. In the rest of this note, we will be committed to solving the coupled Hamilton–Jacobi equations of augmented multi-player nonlinear systems via reinforcement learning.

3. NNs-based online reinforcement learning scheme

In this section, we will present a novel NNs-based online reinforcement learning scheme to tackle the coupled Hamilton–Jacobi equations of augmented multi-player nonlinear systems.

In line with the Weierstrass high-order approximation theorem [42,43], using a single-layer NNs, for each $i \in \hat{N}$, the value function V_i and ∇V_i can be approximately represented as

$$V_i(\xi) = W_i^T\varphi_i(\xi) + \delta_i(\xi) \tag{15}$$

$$\nabla V_i(\xi) = \nabla\varphi_i(\xi)^T W_i + \nabla\delta_i(\xi) \tag{16}$$

where $\varphi_i(\xi)$ are suitable linearly independent basis function vector including L items. $\delta_i(\xi)$ are the approximate errors. $W_i \in \mathbb{R}^{L\times 1}$ are the ideal weight parameter vector. $\nabla\varphi_i(\xi) = \frac{\partial\varphi_i(\xi)}{\partial\xi}, \nabla\delta_i(\xi) = \frac{\partial\delta_i(\xi)}{\partial\xi}$.

Assumption 3. $\varphi_i(\xi), \nabla\varphi_i(\xi), \delta_i(\xi)$ and $\nabla\delta_i(\xi)$ are bounded such that $\|\varphi_i(\xi)\| \leqslant b_{\varphi_i}, \|\nabla\varphi_i(\xi)\| \leqslant b_{\nabla\varphi_i}, \|\delta_i(\xi)\| \leqslant b_{\delta_i}$ and $\|\nabla\delta_i(\xi)\| \leqslant b_{\nabla\delta_i}$.

By substituting (15) and (16) into (14), and noting (12) and making some mathematical transformation, approximation-based coupled Hamilton–Jacobi equations can be obtained as follow

$$\begin{aligned} 0 &= \xi^T\overline{Q}_i\xi - \tfrac{1}{2}W_i^T\nabla\varphi_i\sum_{j=1}^N G_j(\xi)R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^T W_j \\ &+\tfrac{1}{4}\sum_{j=1}^N W_j^T\nabla\varphi_j G_j(x)R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^T W_j \end{aligned} \tag{17}$$

$$+ W_i^T\nabla\varphi_i F(\xi) - \gamma W_i^T\varphi_i(\xi) + \delta_{Hji}, V_i(0) = 0 \tag{18}$$

where the coupled Hamilton–Jacobi approximation error δ_{Hji} owing to the function approximate error is represented as

$$\begin{aligned}\delta_{HJi} = \ & \nabla\delta_i^T F(\xi) - \gamma\delta_i(\xi) - \tfrac{1}{2}\nabla\delta_i^T\sum_{j=1}^{N} G_j R_{jj}^{-1} G_j^T \nabla\delta_j \\ & -\tfrac{1}{2}\nabla\delta_i^T\sum_{j=1}^{N} G_j R_{jj}^{-1} G_j^T \nabla\varphi_j^T W_j \\ & -\tfrac{1}{2}\nabla\varphi_i^T W_i\sum_{j=1}^{N} G_j R_{jj}^{-1} G_j^T \nabla\delta_j \qquad (19) \\ & +\tfrac{1}{4}\sum_{j=1}^{N}\nabla\delta_j^T G_j R_{jj}^{-T} R_{ij} R_{ij}^{-1} G_j^T \nabla\varphi_j^T W_j \\ & +\tfrac{1}{4}\sum_{j=1}^{N}\nabla\delta_j^T G_j R_{jj}^{-T} R_{ij} R_{ij}^{-1} G_j^T \nabla\delta_j \\ & +\frac{1}{4}\sum_{j=1}^{N} W_j^T\nabla\varphi_j G_j R_{jj}^{-T} R_{ij} R_{jj}^{-1} G_j^T\nabla\delta_j \qquad (20)\end{aligned}$$

The authors of [43,44] have concluded that the δ_{HJi} is bounded on a compact set for fixed L, i.e. $\|\delta_{HJi}\| \leqslant b_{\delta_{HJi}}$.

However, the ideal weight parameter vectors W_i is unknown, therefore, (15) and (16) can be approximated as follows

$$\hat{V}_i(\xi) = \hat{W}_i^T\varphi_i(\xi) \qquad (21)$$

$$\nabla\hat{V}_i = \nabla\varphi_i(\xi)^T\hat{W}_i \qquad (22)$$

where $\hat{W}_i$ is the estimation value of W_i.

The control inputs can be thus represented as

$$u_i(\xi) = -\frac{1}{2}R_{ii}^{-1}G_i^T\nabla\varphi(\xi)^T\widehat{W}_i \qquad (23)$$

By substituting (21)–(23) into (11), we can obtain the approximate Hamiltonian as follow

$$\begin{aligned}\widehat{H}_i\left(\xi,\widehat{W}_i\right) = \ & \xi^T\overline{Q}_i\xi - \tfrac{1}{2}\widehat{W}_i^T\nabla\varphi_i\sum_{j=1}^{N} G_j(x)R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^T\widehat{W}_j \\ & +\tfrac{1}{4}\sum_{j=1}^{N}\widehat{W}_j^T\nabla\varphi_j G_j(x)R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^T\widehat{W}_j \qquad (24) \\ & -\gamma\widehat{W}_i^T\varphi_i(\xi) + \widehat{W}_i^T\nabla\varphi_i F(\xi)\end{aligned}$$

Inspired by [45,46], $\hat{W}_i$ is tuned to minimize $\widehat{H}_i\left(\xi,\widehat{W}_i\right)$ and we choose the error functions as follow

$$E_i = \frac{1}{2}\delta_{\widehat{H}_i}^T\delta_{\widehat{H}_i} \qquad (25)$$

where $\delta_{\widehat{H}_i} = \int_{t-\Delta T}^{t}\widehat{H}_i\left(\xi,\widehat{W}_i\right)d\tau, \Delta T > 0$ is the sampling time.

Remark 2. Choosing $\widehat{H}_i\left(\xi,\widehat{W}_i\right)$ as the goal of minimization does not guarantee the closed-loop stability of the system (6) during the learning process, so we need additional consider its stability in the tuning laws [46].

Inspired by [45], in order to minimize the error functions (25) and guarantee the closed-loop stability of the system (6) simultaneously, we introduce the following tuning laws

$$\begin{aligned}\dot{\widehat{W}}_i = \ & \frac{-a_{1i}\kappa_i}{\left(\kappa_i^T\kappa_i+1\right)^2}\left\{\int_{t-\Delta T}^{t}\left[\xi^T\overline{Q}_i\xi - \gamma\widehat{W}_i^T\varphi_i(\xi)\right.\right. \\ & \left.\left.+\tfrac{1}{4}\sum_{j=1}^{N}\widehat{W}_j^T\nabla\varphi_j G_j R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j\nabla\varphi_j^T\widehat{W}_j\right]d\tau + \Delta\varphi_i\widehat{W}_i\right\} \qquad (26) \\ & -\tfrac{1}{2}\beta(\xi,\hat{u}_i)q_{1i}\sum_{j=1}^{N}\nabla\varphi_j(x)G_jR_{jj}^{-1}G_j^T\xi\end{aligned}$$

where a_{1i} and q_{1i} are constant parameters that need to be designed, $\forall i \in \hat{N}$.

The first item of (26) can be obtained by applying gradient descent in (25), and using the chain rule and normalizing [47]. The second item of (26) is used to guarantee the stability of the closed-loop system, which is defined as

$$\beta(\xi,\hat{u}_i) = \begin{cases} 0, & if \ \ \xi(t)^T\xi(t) - \xi(t-\Delta T)^T\xi(t-\Delta T) \leqslant 0 \\ 1, & else \end{cases}$$

where

$$\begin{aligned}\kappa_i = \ & \int_{t-\Delta T}^{t}\nabla\varphi_i(\xi)\left[F(\xi) + \sum_{j=1}^{N} G_j\hat{u}_j\right]d\tau \\ = \ & \int_{t-\Delta T}^{t}\nabla\varphi_i(\xi)\dot{\xi}d\tau \\ = \ & \int_{t-\Delta T}^{t} d(\varphi_i(\xi)) \\ = \ & \varphi_i(\xi(t)) - \varphi_i(\xi(t-\Delta T)) \\ = \ & \Delta\varphi_i(\xi(t))\end{aligned}$$

In order to ensure the convergence of $\hat{W}_i$, inspired by [48,49], we introduce a concurrent learning technique to modify the tuning laws (26). The concurrent learning technique needs the stored history data and current data to tune $\hat{W}_i$ simultaneously. Then, define the approximate Hamiltonian at the past history time t_k using the current weight's estimation $\hat{W}_i$ as $\widehat{H}_i^k(\xi(t_k),\hat{u}_i), k = 1,\ldots,l$. Use the same derivation as (26), we can obtain the following tuning laws at the past history time t_k

$$\begin{aligned}\dot{\widehat{W}}_i(t_k) = \ & \frac{-a_{1i}\kappa_i^k}{\left(\kappa_i^{kT}\kappa_i^k+1\right)^2}\left\{\int_{t_k-\Delta T}^{t_k}\left[\xi^T\overline{Q}_i\xi - \gamma\widehat{W}_i^T\varphi_i(\xi)\right.\right. \\ & \left.\left.+\tfrac{1}{4}\sum_{j=1}^{N}\widehat{W}_j^T\nabla\varphi_j G_j R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j\nabla\varphi_j^T\widehat{W}_j\right]d\tau + \Delta\varphi_i(t_k)\widehat{W}_i\right\} \qquad (27) \\ & -\tfrac{1}{2}\beta(x(t_k),\hat{u}_i)q_{1i}\sum_{j=1}^{N}\nabla\varphi_j(\xi_k)G_j(\xi(t_k))R_{jj}^{-1}G_j^T(\xi(t_k))\xi(t_k)\end{aligned}$$

where

$$\begin{aligned}&\beta(\xi(t_k),\hat{u}_i) \\ &= \begin{cases} 0, & if \ \ \xi(t_k)^T\xi(t_k) - \xi(t_k-\Delta T)^T\xi(t_k-\Delta T) \ \leqslant 0 \\ 1, & else \end{cases}\end{aligned}$$

and

$$\begin{aligned}\kappa_i^k \ & = \int_{t_k-\Delta T}^{t_k}\nabla\varphi_i(\xi)\left[F(\xi) + \sum_{j=1}^{N} G_j\hat{u}_j\right]d\tau \\ & = \int_{t_k-\Delta T}^{t_k}\nabla\varphi_i(\xi)\dot{\xi}d\tau \\ & = \int_{t_k-\Delta T}^{t_k} d(\varphi_i(\xi)) \\ & = \varphi_i(\xi(t_k)) - \varphi_i(\xi(t_k-\Delta T)) \\ & = \Delta\varphi_i(t_k)\end{aligned}$$

Now, the modified weight tuning laws for $\hat{W}_i$ can be given as follow

$$\begin{aligned}\dot{\widehat{W}}_i = & \frac{-a_{1i}\kappa_i}{\left(\kappa_i^T\kappa_i+1\right)^2}\left\{\int_{t-\Delta T}^{t}\left[\xi^T\overline{Q}_i\xi-\gamma\widehat{W}_i^T\varphi_i(\xi)\right.\right.\\ &\left.+\frac{1}{4}\sum_{j=1}^{N}\widehat{W}_j^T\nabla\varphi_j G_j R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j\nabla\varphi_j^T\widehat{W}_j\right]d\tau+\Delta\varphi_i\widehat{W}_i\Bigg\}\\ &-\frac{1}{2}\beta(\xi,\hat{u}_i)q_{1i}\sum_{j=1}^{N}\nabla\varphi_j(\xi)G_jR_{jj}^{-1}G_j^T\xi\\ &+\sum_{k=1}^{l}\Bigg\langle\frac{-a_{1i}\kappa_i^k}{\left(\kappa_i^{kT}\kappa_i^k+1\right)^2}\left\{\int_{t_k-\Delta T}^{t_k}\left[\xi^T\overline{Q}_i\xi-\gamma\widehat{W}_i^T\varphi_i(\xi)\right.\right.\\ &\left.+\frac{1}{4}\sum_{j=1}^{N}\widehat{W}_j^T\nabla\varphi_j G_j R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j\nabla\varphi_j^T\widehat{W}_j\right]d\tau+\Delta\varphi_i(t_k)\widehat{W}_i\Bigg\}\\ &\left.-\frac{1}{2}\beta(\xi(t_k),\hat{u}_i)q_{1i}\sum_{j=1}^{N}\nabla\varphi_j(t_k)G_j(\xi(t_k))R_{jj}^{-1}G_j^T(\xi(t_k))\xi(t_k)\right\rangle\end{aligned}\tag{28}$$

According to κ_i and κ_i^k, we have

$$\begin{aligned}&\Delta\varphi_i(\xi)^T\widehat{W}_i\\ &=\int_{t-\Delta t}^{t}\widehat{W}_i^T\nabla\varphi_i(\xi)\left[F(\xi)-\frac{1}{2}\sum_{j=1}^{N}G_jR_{jj}^{-1}G_j^T\nabla\varphi_j(\xi)^T\widehat{W}_i\right]d\tau\end{aligned}\tag{29}$$

$$\begin{aligned}&\Delta\varphi_i(t_k)^T\widehat{W}_i\\ &=\int_{t_k-\Delta t}^{t_k}\widehat{W}_i^T\nabla\varphi_i(\xi)\left[F(\xi)-\frac{1}{2}\sum_{j=1}^{N}G_jR_{jj}^{-1}G_j^T\nabla\varphi_j(\xi)^T\widehat{W}_i\right]d\tau\end{aligned}\tag{30}$$

Define the estimation error of weight tuning laws $\hat{W}_i$ as $\widetilde{W}_i=W_i-\widehat{W}_i$, then $\dot{\widetilde{W}}_i=-\dot{\widehat{W}}_i$. Therefore, according to (28), the weight estimation error dynamics can be represented as

$$\begin{aligned}\dot{\widetilde{W}}_i(t)=&-a_{1i}\left(\bar{\kappa}_i\bar{\kappa}_i^T+\sum_{k=1}^{l}\bar{\kappa}_i^k\bar{\kappa}_i^{kT}\right)\widetilde{W}_i(t)\\ &+a_{1i}\frac{\bar{\kappa}_i}{\pi_i}\delta_{Hi}+a_{1i}\sum_{k=1}^{l}\frac{\bar{\kappa}_i^k}{\pi_i^k}\delta_{H_i^k}\end{aligned}\tag{31}$$

where $\bar{\kappa}_i=\frac{\kappa_i}{\kappa_i^T\kappa_i+1}$, $\bar{\kappa}_i^k=\frac{\kappa_i^k}{1+\kappa_i^{kT}\kappa_i^k}$, $\pi_i=\kappa_i^T\kappa_i+1$, $\pi_i^k=\kappa_i^{kT}\kappa_i^k+1$ and

$$\begin{aligned}\delta_{Hi}=&\int_{t-\Delta T}^{t}\left[\xi^T\overline{Q}_i\xi-\gamma\widehat{W}_i^T\varphi_i(\xi)+\widehat{W}_i^T\nabla\varphi_iF(\xi)\right.\\ &-\frac{1}{2}W_i^T\nabla\varphi_i\sum_{j=1}^{N}G_j(\xi)R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^TW_j\\ &\left.+\frac{1}{4}\sum_{j=1}^{N}W_j^T\nabla\varphi_jG_j(\xi)R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^TW_j\right]d\tau\end{aligned}$$

$$\begin{aligned}\delta_{H_i^k}=&\int_{t_k-\Delta T}^{t_k}\left[\xi^T\bar{Q}_i\xi-\gamma\widehat{W}_i^T\varphi_i(\xi)+\widehat{W}_i^T\nabla\varphi_iF(\xi)\right.\\ &-\frac{1}{2}W_i^T\nabla\varphi_i\sum_{j=1}^{N}G_j(\xi)R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^TW_j\\ &\left.+\frac{1}{4}\sum_{j=1}^{N}W_j^T\nabla\varphi_jG_j(\xi)R_{jj}^{-T}R_{ij}R_{jj}^{-1}G_j^T(\xi)\nabla\varphi_j^TW_j\right]d\tau\end{aligned}$$

Assumption 4. Assume that δ_{Hi} and $\delta_{H_i^k}$ are bounded such that $\|\delta_{Hi}\|\leqslant b_{\delta_{Hi}\max}$, $\left\|\delta_{H_i^k}\right\|\leqslant b_{\delta_{H_i^k}\max}$.

Remark 3. Let $\Omega_i=[\bar{\kappa}_i^1,\ldots,\bar{\kappa}_i^l]$ store the past time history data and $rank(\Omega)=L$ and note that the number of sampled data in Ω_i is a fixed value $l>L$.

Remark 4. Define $\Theta_i=\bar{\kappa}_i\bar{\kappa}_i^T+\sum_{k=1}^{l}\bar{\kappa}_i^k\bar{\kappa}_i^{kT}$, if Remark 3 is satisfied, then $\Theta_i>0$.

So far, the NNs-based online reinforcement learning algorithm is given in the following.

Algorithm 1. (NNs-based online reinforcement learning algorithm)

Step 1: Initializaiton: initial ξ_0, learning factors a_{1i} and a_{2i}, basis function vector $\varphi_i(\xi)$, sampling time interval ΔT, NNs initial weight $\widehat{W}_i(0), Q_i, R_{ij}$ for $\forall i,j\in N$, initial time t_0, the maximum running time of the algorithm t_{stop}.

Step 2: Compute $\widehat{V}_i(t-\Delta T),\hat{u}_j(t-\Delta T)$ and apply them to (6). Then, update $\widehat{W}_i(t)$ by (28) and compute $\widehat{V}_i(t),\hat{u}_j(t)$ according to $\widehat{W}_i(t)$.

Step 3: $t=t+\Delta T$, repeat **Step 2** until $t\geqslant t_{stop}$ or $\left\|\widehat{W}_i(t+\Delta T)-\widehat{W}_i(t)\right\|\leqslant\upsilon$, where υ is a sufficiently small positive constant.

Step 4: End.

Remark 5. In our proposed Algorithm 1, the drift dynamics $F(\xi)$ of systems (6) is not required.

4. The analysis of stability and convergence

In this section, the main results of this paper are summarized as the following theorem.

Theorem 1. *Consider the system dynamics (6) with the coupled Hamilton–Jacobi Eqs. (14), let the control inputs be provided by (23). Let the tuning laws for $\hat{W}_i$ be given by (28). The Remark 3 is satisfied. Then, the closed-loop system states ξ and the NNs errors $\widetilde{W}_i$ are UUB for a sufficiently large L. Moreover, the approximate cost function $\hat{V}_i$ and control inputs $\hat{u}_i$ are converged to the optimal value, i.e. $\left\|V_i^*-\widehat{V}_i\right\|<\sigma_{V_i}$, $\|u_i^*-\hat{u}_i\|<\sigma_{u_i}$, $\sigma_{V_i},\sigma_{u_i}$ are small positive constants.*

Proof: Choose the following Lyapunov function

$$L_1(t)=\frac{1}{2}\int_{t-\Delta t}^{t}\xi^T\xi d\tau+\frac{1}{2}\sum_{i=1}^{N}\widetilde{W}_i^T\widetilde{W}_i\tag{32}$$

The derivative of $L(t)$ is given by

$$\dot{L}_1(t)=\int_{t-\Delta t}^{t}\xi^T\dot{\xi}d\tau+\sum_{i=1}^{N}\widetilde{W}_i^T\dot{\widetilde{W}}_i\tag{33}$$

When $\beta(\xi,\hat{u}_j)=0$, that is $\xi(t)^T\xi(t)-\xi(t-\Delta T)^T\xi(t-\Delta T)\leqslant 0$, we have the following relationship

$$\begin{aligned}&\frac{1}{2}\left(\xi(t)^T\xi(t)-\xi(t-\Delta T)^T\xi(t-\Delta T)\right)\\ &=\int_{t-\Delta T}^{t}\xi^T\dot{\xi}d\tau\\ &=\int_{t-\Delta T}^{t}\xi^T\left(F(\xi)+\sum_{j=1}^{N}G_j(\xi)\hat{u}_j\right)d\tau\leqslant 0\end{aligned}\tag{34}$$

According to (34), where exists a positive constant χ for the following relationship satisfied

$$\int_{t-\Delta T}^{t}\xi^T\left(F(\xi)+\sum_{j=1}^{N}G_j(\xi)\hat{u}_j\right)d\tau\leqslant-\chi\int_{t-\Delta T}^{t}\|\xi\|d\tau\tag{35}$$

By substituting (31) and (35) into (33), we have

$$\begin{aligned}\dot{L}_1(t) = & \int_{t-\Delta T}^{t}\xi^T\left(F(\xi)+\sum_{j=1}^{N}G_j(\xi)\hat{u}_j\right)d\tau+\sum_{i=1}^{N}\widetilde{W}_i^T\dot{\widetilde{W}}_i\\ \leqslant & -\chi\int_{t-\Delta T}^{t}\|\xi\|d\tau+\sum_{i=1}^{N}\left[-a_{1i}\widetilde{W}_i^T\left(\bar{\kappa}_i\bar{\kappa}_i^T+\sum_{k=1}^{l}\bar{\kappa}_i^k\bar{\kappa}_i^{kT}\right)\widetilde{W}_i\right.\\ &\left.+a_{1i}\widetilde{W}_i^T\frac{\bar{\kappa}_i}{\pi_i}\delta_{Hi}+a_{1i}\widetilde{W}_i^T\sum_{k=1}^{l}\frac{\bar{\kappa}_i^k}{\pi_i^k}\delta_{H_i^k}\right]\\ = & -\chi\int_{t-\Delta T}^{t}\|\xi\|d\tau-\sum_{i=1}^{N}\left[a_{1i}\widetilde{W}_i^T\Theta_i\widetilde{W}_i\right]\\ &+\sum_{i=1}^{N}\left[a_{1i}\widetilde{W}_i^T\frac{\bar{\kappa}_i}{\pi_i}\delta_{Hi}+a_{1i}\widetilde{W}_i^T\sum_{k=1}^{l}\frac{\bar{\kappa}_i^k}{\pi_i^k}\delta_{H_i^k}\right]\\ = & -\frac{\overbrace{1+\cdots+1}^{N}}{N}\chi\int_{t-\Delta T}^{t}\|\xi\|d\tau\\ &+\sum_{i=1}^{N}\left[\widetilde{W}_i^T\Theta_i\widetilde{W}-(1+a_{1i})\widetilde{W}_i^T\Theta_i\widetilde{W}_i\right]\\ &+\sum_{i=1}^{N}\left[a_{1i}\widetilde{W}_i^T\frac{\bar{\kappa}_i}{\pi_i}\delta_{Hi}+a_{1i}\widetilde{W}_i^T\sum_{k=1}^{l}\frac{\bar{\kappa}_i^k}{\pi_i^k}\delta_{H_i^k}\right]\end{aligned}\tag{36}$$

The derivative of Lyapunov function $\dot{L}(t)$ is negative provided that

$$\int_{t-\Delta T}^{t}\|\xi\|d\tau\geqslant\max_{i\in N}\left[\frac{N\lambda_{\max}(\Theta_i)}{\chi}\left\|\widetilde{W}_i\right\|^2\right]=\hat{U}_{1i}\tag{37}$$

$$\left\|\widetilde{W}_i\right\|\geqslant\frac{Na_{1i}(l+1)b_{\delta_{Hi}}}{(1+a_{1i})\lambda_{\min}(\Theta_i)}=U_{1i},\quad i\in\hat{N}\tag{38}$$

where $\lambda_{\max}(\Theta_i)$ and $\lambda_{\min}(\Theta_i)$ represent the maximum eigenvalue and minimum eigenvalue of Θ_i, respectively.

When $\beta(\xi,\hat{u}_j)=1$, we add $\overline{V}_i(\xi)\geq 0$ to the Lyapunov function (32) as follows

$$L_1(t)=\frac{1}{2}q_{1i}\xi^T\xi+\frac{1}{2}\sum_{i=1}^{N}\widetilde{W}_i^T\widetilde{W}_i+\overline{V}_i(\xi)\tag{39}$$

where $\overline{V}_i(\xi)$ is a local smooth solution of the coupled Hamilton–Jacobi Eqs. (14). Then, taking the derivative of (39) and noting (36)

$$\begin{aligned}\dot{L}_1(t) = & \ q_{1i}\xi^T\dot{\xi}+\sum_{i=1}^{N}\widetilde{W}_i^T\dot{\widetilde{W}}_i-\xi^T\overline{Q}_i\xi-\sum_{j=1}^{N}u_j^TR_{ij}u_j\\ \leqslant & \ q_{1i}b_F\|\xi\|^2+\sum_{i=1}^{N}\left[-a_{1i}\widetilde{W}_i^T\Theta_i\widetilde{W}_i+a_{1i}\widetilde{W}_i^T\frac{\bar{\kappa}_i}{\pi_i}\delta_{Hi}+a_{1i}\widetilde{W}_i^T\sum_{k=1}^{l}\frac{\bar{\kappa}_i^k}{\pi_i^k}\delta_{H_i^k}\right]\\ & -\lambda_{\min}\left(\overline{Q}_i\right)\|\xi\|^2-\frac{1}{2}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\|\xi\|\\ & =\left(q_{1i}b_F-\lambda_{\min}\left(\overline{Q}_i\right)\right)\|\xi\|^2\\ & +\sum_{i=1}^{N}\left[-a_{1i}\widetilde{W}_i^T\Theta_i\widetilde{W}_i+a_{1i}\widetilde{W}_i^T\frac{\bar{\kappa}_i}{\pi_i}\delta_{Hi}+a_{1i}\widetilde{W}_i^T\sum_{k=1}^{l}\frac{\bar{\kappa}_i^k}{\pi_i^k}\delta_{H_i^k}\right]\\ & -\frac{1}{2}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\left[\left(\|\xi\|^2+\frac{1}{2}\right)^2-\|\xi\|^2-\frac{1}{4}\right]\\ = & \left(q_{1i}b_F-\lambda_{\min}\left(\overline{Q}_i\right)+\frac{1}{2}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\right)\|\xi\|^2\\ & -\frac{1}{2}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\left(\|\xi\|^2+\frac{1}{2}\right)^2+\frac{1}{8}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\\ & \sum_{i=1}^{N}\left[-a_{1i}\widetilde{W}_i^T\Theta_i\widetilde{W}_i+a_{1i}\widetilde{W}_i^T\frac{\bar{\kappa}_i}{\pi_i}\delta_{Hi}+a_{1i}\widetilde{W}_i^T\sum_{k=1}^{l}\frac{\bar{\kappa}_i^k}{\pi_i^k}\delta_{H_i^k}\right]\\ \leqslant & \left(q_{1i}b_F-\lambda_{\min}\left(\overline{Q}_i\right)+\frac{1}{2}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\right)\|\xi\|^2\\ & -\frac{1}{2}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\left(\|\xi\|^2+\frac{1}{2}\right)^2+\frac{1}{8}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\\ & \sum_{i=1}^{N}\left[-a_{1i}\widetilde{W}_i^T\Theta_i\widetilde{W}_i+a_{1i}(l+1)b_{\delta_{H_i^k}}\widetilde{W}_i(t)\right]\end{aligned}\tag{40}$$

where $\Lambda=\sum_{j=1}^{N}G_jR_{jj}^{-1}G_j^T$ and assume that $\Lambda_{\min}\leqslant\|\Lambda\|\leqslant\Lambda_{\max}$. The derivative of Lyapunov function $\dot{L}(t)$ is negative provided that

$$\|\xi\|^2\geqslant\max_{i\in N}\left[\frac{q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}}{4\left(2\lambda_{\min}\left(\overline{Q}_i\right)-2q_{1i}b_f+q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}\right)}\right]=\hat{U}_{2i}\tag{41}$$

$$\left\|\widetilde{W}_i\right\|\geqslant\frac{(l+1)b_{\delta_{H_i^k\max}}}{\lambda_{\min}(\Theta_i)}=U_{2i},i\in\hat{N}\tag{42}$$

and $q_{1i}b_F-\lambda_{\min}(\bar{Q}_i)+\frac{1}{2}q_{1i}b_{\nabla\varphi_i}\Lambda_{\min}<0$. In summary, $\|\xi\|^2\geqslant\max\left\{\hat{U}_{1i},\hat{U}_{2i}\right\}$, $\left\|\widetilde{W}_i\right\|\geqslant\max\{U_{1i},U_{2i}\},\forall i\in\hat{N}$. The UUB of augmented system states and the NNs weight estimation error are thus proved. $\left\|V_i^*-\hat{V}_i\right\|\leqslant\left\|\widetilde{W}_i\right\|\|\varphi_i(\xi)\|+\|\delta_i(\xi)\|\leqslant\max\left\{\hat{U}_{1i},\hat{U}_{2i}\right\}$ $b_{\varphi_i}=\sigma_{V_i}$, $\left\|u_j^*-\hat{u}_j\right\|\leqslant\frac{1}{2}\left(\left\|\widetilde{W}_i\right\|\|\nabla\varphi_i(\xi)\|+\|\nabla\delta_i(\xi)\|\right)\lambda_{\min}\left(R_{ii}^{-1}\right)b_{G_j}\leqslant\frac{1}{2}\left(\max\{U_{1i},U_{2i}\}b_{\nabla\varphi_i}+b_{\nabla\delta_i}\right)\lambda_{\min}\left(R_{ii}^{-1}\right)b_{G_j}=\sigma_{u_j}$.

The proof is thus complete.

5. Numerical simulation results

In this section, two simulation examples including linear systems and nonlinear systems are given to verify the effectiveness of the proposed scheme.

Example 1. In this example, consider the following linear systems with two players from [50]

$$\dot{x}=\begin{bmatrix}0&0\\0&-1\end{bmatrix}x+\begin{bmatrix}2\\1\end{bmatrix}u_1+\begin{bmatrix}1\\1\end{bmatrix}u_2$$

The desired reference signal is generated by the following command

$$\dot{r}(t)=\begin{bmatrix}0&2\\-2&0\end{bmatrix}r(t)$$

Let the initial state $x_0=[1,-1]^T,r_0=[0.5,-0.5]^T,Q_1=diag[1,1]$, $Q_2=[1,-1;-1,5]$. $R_{11}=R_{12}=2,R_{21}=R_{22}=1,a_{11}=a_{12}=50,q_{11}=q_{12}=0.001$. $\Delta T=0.05,\gamma=0.5$. For $\forall i=1,2$, the NNs activation functions are selected as

$$\varphi_i(\xi(t))=\left[e_1^2,e_1e_2,e_1r_1,e_1r_2,e_2^2,e_2r_1,e_2r_2,r_1^2,r_1r_2,r_2^2\right]^T$$

and the initial NNs weight vectors $\hat{W}_i$ are randomly taken from interval $[-1\ \ 1]$. Fig. 1 shows the evolution process of critic NNs weights for first player, and its eventually converge to $\widehat{W}=[3.3625,-0.2917,-3.0856,-1.4389,-1.5522,-1.8288,-1.1222,0.2882,0.4685,0.5138]^T$. Fig. 2 shows the evolution process of critic NNs weights for second player, and its eventually converge to $\widehat{W}=[-1.2073,1.5556,4.1568,-0.7976,2.3899,3.8432,0.7976,-0.3229,0.1958,-0.5487]^T$. The evolution of tracking errors are shown in Fig. 3. The evolution of $(x-r)$ between system states and reference signal are depicted in Fig. 4 and Fig. 5. The actual trajectory and the reference trajectory are displayed in Fig. 6. The optimal control inputs are plotted in Fig. 7.

Example 2. In this example, consider the following nonlinear systems with two players from [27]

$$\dot{x}=f(x)+g(x)u_1+g(x)u_2$$

$$f(x)=\begin{bmatrix}x_2\\-\frac{1}{2}x_1-x_2+\frac{1}{4}x_2(\cos(2x_1)+2)^2+\frac{1}{4}x_2(\sin(4x_1^2)+2)^2\end{bmatrix}$$

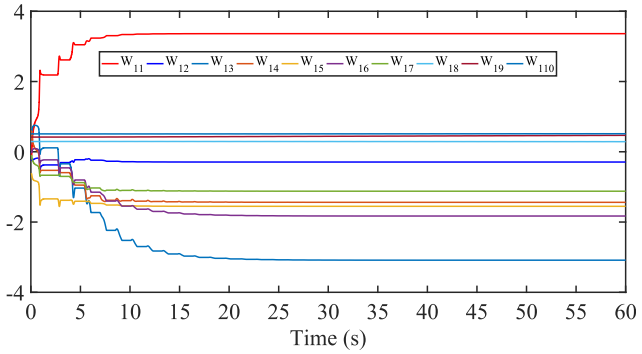

Fig. 1. The evolution process of critic NNs weights for first player.

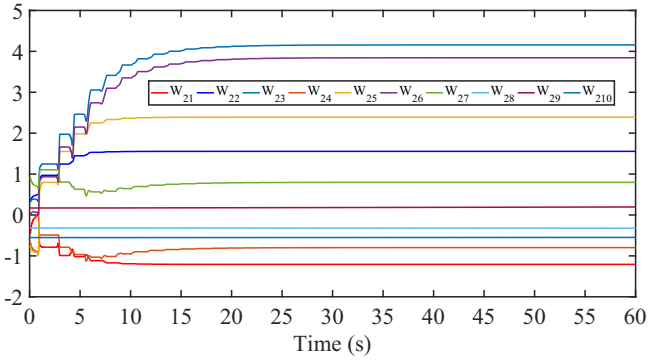

Fig. 2. The evolution process of critic NNs weights for second player.

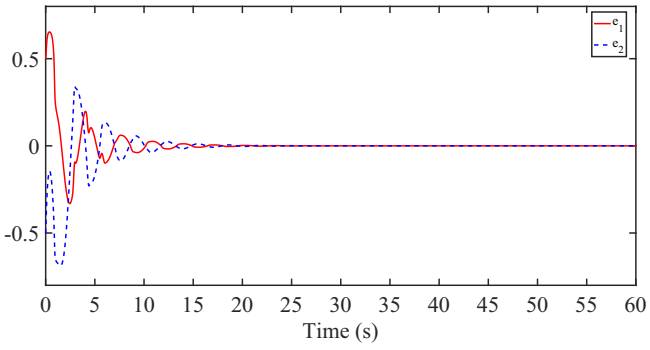

Fig. 3. The evolution of tracking errors.

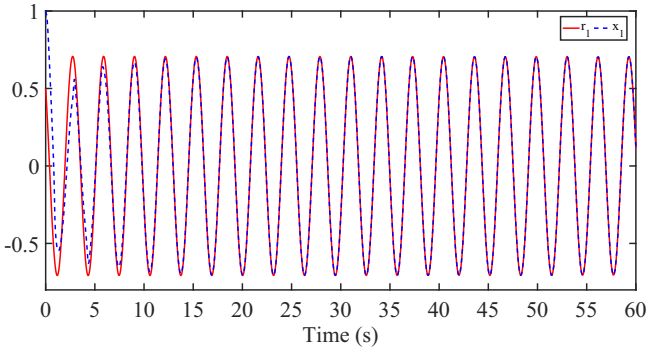

Fig. 4. The evolution of $(x_1 - r_1)$.

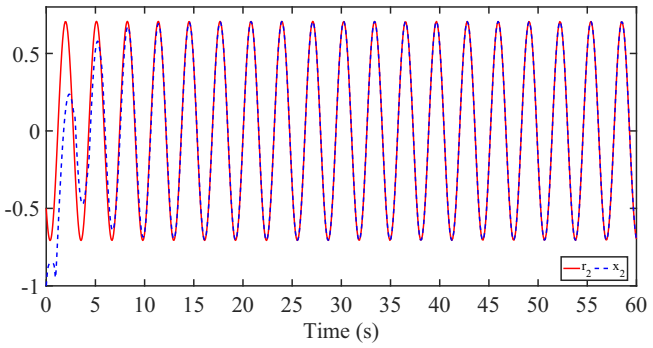

Fig. 5. The evolution of $(x_2 - r_2)$.

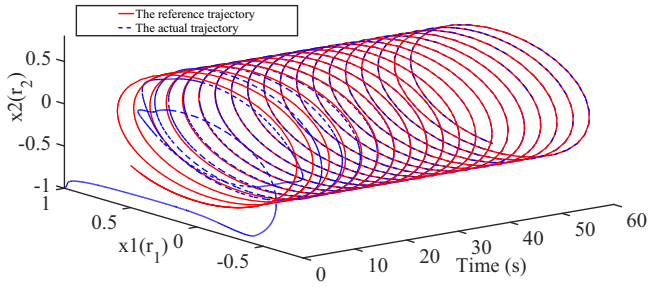

Fig. 6. The actual trajectory and the reference trajectory.

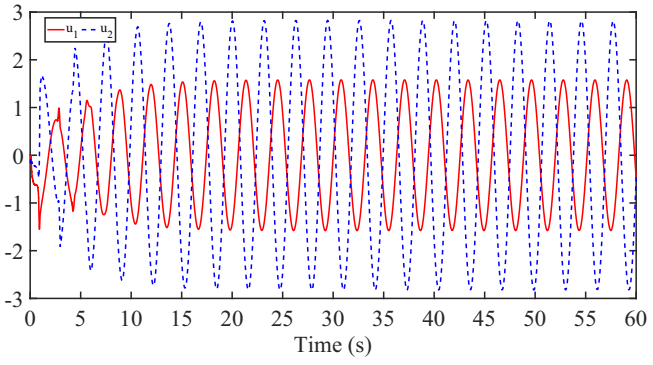

Fig. 7. The optimal control inputs.

$$g_1(x) = \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}$$

$$g_2(x) = \begin{bmatrix} 0 \\ \sin(4x_1^2) + 2 \end{bmatrix}$$

The desired reference signal is generated by the following command

$$\dot{r}(t) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} r(t)$$

Let the initial state $x_0 = [0.5, -0.5]^T, r_0 = [0.1, -0.1]^T, Q_1 = diag[2, 2]$, $Q_2 = diag[1, 1]$. $R_{11} = R_{12} = 2, R_{21} = R_{22} = 1, a_{11} = a_{12} = 20, q_{11} = q_{12} = 0.01$. $\Delta T = 0.05, \gamma = 0.5$. For $\forall i = 1, 2$, the NNs activation functions are selected as

$$\varphi_i(\xi(t)) = \left[e_1^2, e_1e_2, e_1r_1, e_1r_2, e_2^2, e_2r_1, e_2r_2, r_1^2, r_1r_2, r_2^2\right]^T$$

and the initial NNs weight vectors $\hat{W}_i$ are randomly taken from interval $[-1\ \ 1]$. Fig. 8 shows the evolution process of critic NNs weights for first player, and its eventually converge to $\widehat{W} = [0.1057, 0.4586, -0.9353, 0.2288, 0.7231, -0.4167, 0.2924, -0.6151, -0.7539, -0.5889]^T$. Fig. 9 shows the evolution process of critic NNs weights for second player, and its eventually converge to $\widehat{W} = [-0.7052, 0.0481, -0.9140, 0.2701, 1.3626, 0.6699, 0.7884, -0.0018, 0.0715, -0.1096]^T$. The evolution of tracking errors are shown in Fig. 10. The evolution of $(x - r)$ between system states and reference signal are depicted in Fig. 11 and Fig. 12. The actual trajectory and the reference trajectory are depicted in Fig. 13. The optimal control inputs are plotted in Fig. 14.

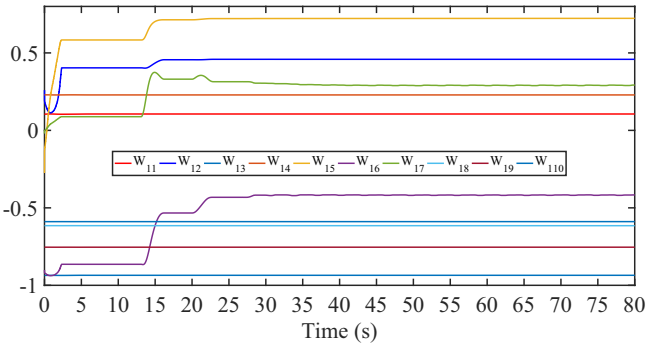

Fig. 8. The evolution process of critic NNs weights for first player.

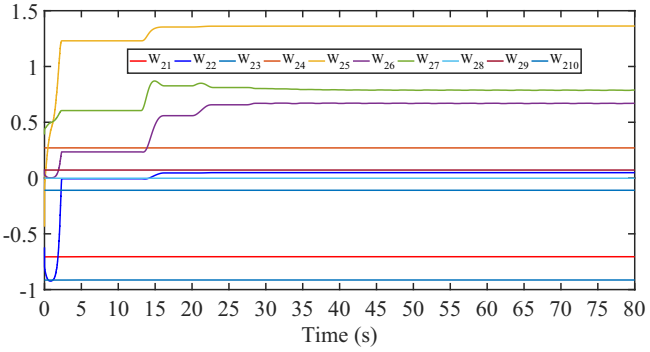

Fig. 9. The evolution process of critic NNs weights for second player.

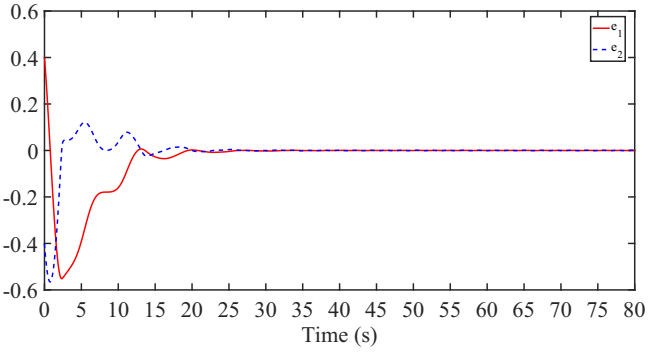

Fig. 10. The evolution of tracking errors.

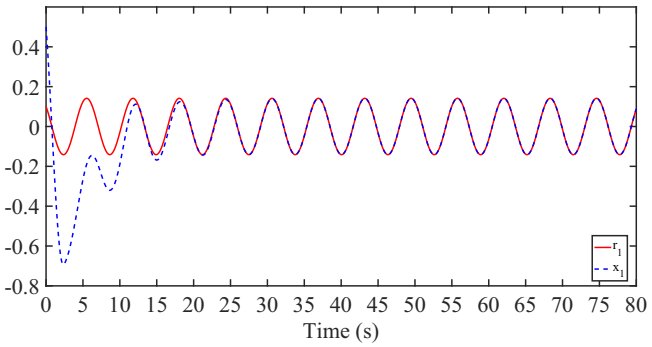

Fig. 11. The evolution of $(x_1 - r_1)$.

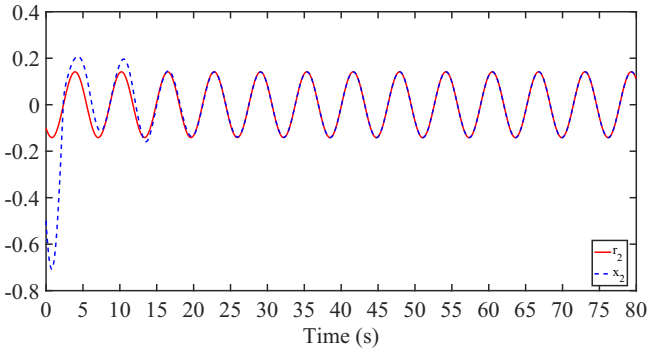

Fig. 12. The evolution of $(x_2 - r_2)$.

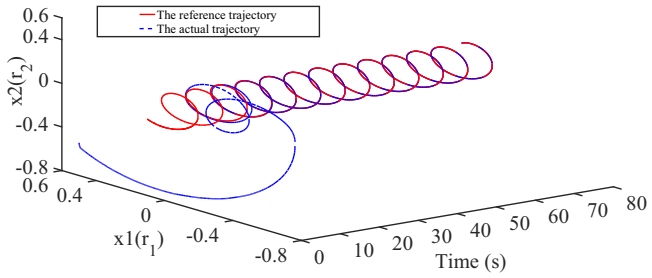

Fig. 13. The actual trajectory and the reference trajectory.

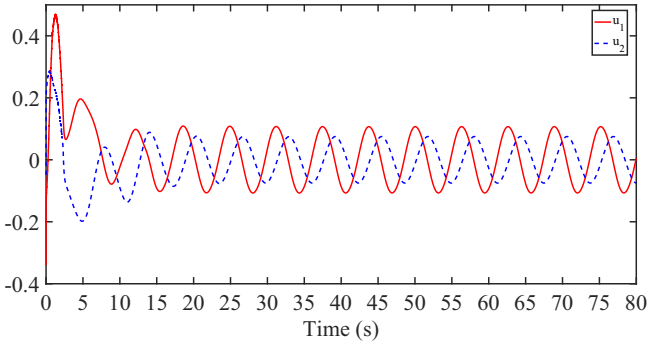

Fig. 14. The optimal control inputs.

Through the simulation results of the above two examples, it is obvious that the systems states can track the desired reference signal well using our designed scheme. Therefore, the effectiveness of our proposed scheme is well demonstrated.

6. Conclusions and future work

This paper proposes a novel NNs-based online reinforcement learning computational intelligent scheme for optimal tracking control of continuous-time multi-player non-zero-sum games. N single-layer NNs are adopted to approximate the value function for each player. To relax traditional PE conditions, historical data from a period of time have been collected to design an adaptive NNs tuning laws. The UUB of NNs weight errors and closed-loop augmented system states are rigorously proved. The value function and the control input for each player are also proved to be converged to approximately optimal value function and optimal control input with a small bounded error. Finally, simulation studies on linear systems and nonlinear systems verify the effectiveness of our design scheme. In this paper, the system inputs dynamics are required, we will investigate the optimal tracking control for continuous-time multi-player non-zero-sum games with completely unknown system dynamics in future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This paper is funded by International Graduate Exchange Program of Beijing Institute of Technology.

References

[1] M.O. Rabin, Effective computability of winning strategies, Ann. Math. Stud. 39 (1958) 147–157.
[2] K.G. Vamvoudakis, H. Modares, B. Kiumarsi, et al., Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online, IEEE Control Syst. Mag. 37 (1) (2017) 33–52.
[3] T. Mylvaganam, M. Sassano, A. Astolfi, A differential game approach to multi-agent collision avoidance, IEEE Trans. Autom. Control 62 (8) (2017) 4229–4235.
[4] K.G. Vamvoudakis, S. Jagannathan, Control of Complex Systems: Theory and Applications, Butterworth-Heinemann, 2016.
[5] Q. Zhu, T. Basar, Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems, IEEE Control Syst. Mag. 35 (1) (2015) 46–65.
[6] S.R. Etesami, T. Bşar, Dynamic games in cyber-physical security: An overview, Dyn. Games Appl. 9 (2019) 884–913.
[7] K.G. Vamvoudakis, A. Kanellopolos, Non-equilibrium dynamics games and cyber-physical security: a cognitive hierarchy approach, Syst. Control Lett. 125 (2019) 59–66.
[8] K.G. Vamvoudakis, Game-theoretic tracking control for actuator attack attenuation in cyber-physical systems, in: International Joint Conference on Neural Networks, 2016, pp. 4233–4240.
[9] A.W. Starr, Y.C. Ho, Nonzero-sum differential games, J. Optim. Theory Appl. 3 (3) (1969) 184–206.
[10] T. Başar, G.J. Olsder, Dynamic noncooperative game theory, Soc. Ind. Appl. Math. (1999).
[11] S.H. Tijs, Introduction to Game Theory, Hindustan Book Agency, 2003.
[12] B. Kiumarsi, K.G. Vamvoudakis, H. Modares, F.L. Lewis, Optimal and autonomous control using reinforcement learning: a survey, IEEE Trans. Neural Networks Learn. Syst. 29 (6) (2018) 2042–2062.
[13] A.A. Mylvaganam T, Sassano M, Constructive ϵ -nash equilibria for nonzero-sum differential games, IEEE Trans. Autom. Control 60 (4) (2015) 950–965. .
[14] D. Liu, H. Li, D. Wang, Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics, IEEE Trans. Syst. Man Cybern. Syst. 44 (8) (2014) 1015–1027.
[15] D. Liu, H. Li, D. Wang, Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm, Neurocomputing 110 (8) (2013) 92–100.
[16] Y. Huang, D. Wang, D. Liu, Bounded robust control design for uncertain nonlinear systems using single-network adaptive dynamic programming, Neurocomputing 266 (2017) 128–140.
[17] Y. Xiong, D. Liu, Q. Wei, D. Wang, Guaranteed cost neural tracking control for a class of uncertain nonlinear systems using adaptive dynamic programming, Neurocomputing 198 (C) (2016) 80–90. .
[18] P. Zhang, Y. Yuan, L. Guo, Fault-tolerant optimal control for discrete-time nonlinear system subjected to input saturation: a dynamic event-triggered approach, IEEE Trans. Cybern. doi:10.1109/TCYB.2019.2923011. .
[19] M. Gan, J. Zhao, C. Zhang, Extended adaptive optimal control of linear systems with unknown dynamics using adaptive dynamic programming, Asian J. Control (2019) 1–10, https://doi.org/10.1002/asjc.2243.
[20] Q. Wei, H. Zhang, D. Jing, Model-free multiobjective approximate dynamic programming for discrete-time nonlinear systems with general performance index functions, IEEE Trans. Cybern. 72 (2009) 1839–1848.
[21] J. Si, A.G. Barto, W.B. Powell, Handbook of Learning and Approximate Dynamic Programming, Wiley-IEEE Press, 2004.
[22] R. Sutton, A. Barto, Reinforcement Learning: An Introduction, The MIT Press, Cambridge, Massachusetts, London, England, 2018.
[23] J. Zhao, M. Gan, C. Zhang, Event-triggered h_∞ optimal control for continuous-time nonlinear systems using neurodynamic programming, Neurocomputing 360 (2019) 14–24.
[24] F.L. Lewis, D. Liu, Reinforcement learning and approximate dynamic programming for feedback control, IEEE Circ. Syst. Mag. 9 (3) (2015) 32–50.
[25] P. Zhang, Y. Yuan, H. Yang, H. Liu, Near-nash equilibrium control strategy for discrete-time nonlinear systems with round-robin protocol, IEEE Trans. Neural Networks Learn. Syst. 30 (8) (2019) 2478–2492.
[26] K.G. Vamvoudakis, Non-zero sum nash q-learning for unknown deterministic continuous-time linear systems, Automatica 61 (C) (2015) 274–281.
[27] K.G. Vamvoudakis, F.L. Lewis, Multi-player non-zero-sum games: online adaptive learning solution of coupled hamilton-jacobi equations, Automatica 47 (8) (2011) 1556–1569.
[28] H. Zhang, H. Jiang, C. Luo, G. Xiao, Discrete-time nonzero-sum games for multiplayer using policy-iteration-based adaptive dynamic programming algorithms, IEEE Trans. Cybern. 47 (10) (2016) 3331–3340.
[29] H. Jiang, H. Zhang, K. Zhang, X. Cui, Data-driven adaptive dynamic programming schemes for non-zero-sum games of unknown discrete-time nonlinear systems, Neurocomputing 275 (2017) 649–658.
[30] Q. Zhang, D. Zhao, Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics, IEEE Trans. Cybern. 49 (8) (2019) 2874–2885.
[31] Y. Lv, X. Ren, J. Na, Online optimal solutions for multi-player nonzero-sum game with completely unknown dynamics, Neurocomputing 283 (2018) 87–97.
[32] M. Johnson, R. Kamalapurkar, S. Bhasin, W.E. Dixon, Approximate n-player nonzero-sum game solution for an uncertain continuous nonlinear system, IEEE Trans. Neural Networks Learn. Syst. 26 (8) (2017) 1645–1658.
[33] H. Jiang, H. Zhang, K. Zhang, X. Cui, Neural-network-based learning algorithms for cooperative games of discrete-time multi-player systems with control constraints via adaptvie dynamic programming, Neurocomputing 344 (2019) 13–19.
[34] R. Song, F.L. Lewis, Q. Wei, Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games, IEEE Trans. Neural Networks Learn. Syst. 28 (3) (2017) 704–713.
[35] D. Zhao, Q. Zhang, D. Wang, Y. Zhu, Experience replay for optimal control of nonzero-sum game systems with unknown dynamics, IEEE Trans. Cybern. 46 (3) (2015) 854–865.
[36] H. Zhang, L. Cui, Y. Luo, Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network adp, IEEE Trans. Cybern. 43 (1) (2013) 206–216.
[37] Y. Liu, Z. Wang, Z. Shi, h_∞ tracking control for linear discrete-time systems via reinforcement learning, Int. J. Robust Nonlinear Control (2019) 1–20, https://doi.org/10.1002/rnc.4762.
[38] H. Modares, F.L. Lewis, Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning, Automatica 50 (2014) 1780–1792.
[39] J. Zhao, Neural network-based optimal tracking control of continuous-time uncertain nonlinear system via reinforcement learning, Neural Process. Lett. (2020) 1–18, https://doi.org/10.1007/s11063-020-10220-z.
[40] Y. Liu, L. Tang, S. Tong, C. Chen, D. Li, Reinforcement learning design-based adaptive tracking control with less learning parametets for nonlinear discrete-time mimo systems, IEEE Trans. Neural Networks Learn. Syst. 26 (1) (2015) 165–176.
[41] Y. Wen, H. Zhang, H. Su, H. Ren, Optimal tracking control for non-zero-sum games of linear discrete-time systems via off-policy reinforcement learning, Opt. Control Appl. Methods (2020) 1–18, https://doi.org/10.1002/oca.2597.
[42] F.A. Bruce, The Method of Weighted Residuals and Variational Principles, Academic Press, New York, 1990.
[43] M. Abu-Khalaf, F.L. Lewis, Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach, Automatica 41 (5) (2005) 779–791.
[44] K.G. Vamvoudakis, F.L. Lewis, Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem, Automatica 46 (5) (2010) 878–888.
[45] N.T. Luy, Reinforcement learning-based tracking control for wheeled mobile robot, Trans. Inst. Measure. Control 36 (7) (2014) 171–176.

[46] H. Zargarzadeh, T. Dierks, S. Jagannathan, Optimal control of nonlinear continuous-time systems in strict-feedback form, IEEE Trans. Neural Networks Learn. Syst. 26 (10) (2015) 2535–2549.
[47] K.G. Vamvoudakis, Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach, Syst. Control Lett. 100 (Complete) (2017) 14–20. .
[48] H. Modares, F.L. Lewis, M.B. Naghibi-Sistani, Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems, Automatica 50 (1) (2014) 193–202.
[49] G.V. Chowdhary, Concurrent learning for convergence in adaptive control without persistency of excitation, in: Proceedings of Decision and Control, 2011, pp. 3674–3679. .
[50] K.G. Vamvoudakis, F.L. Lewis, W.E. Dixon, Open-oop stackelberg learning solution for hierarchical control problems, Int. J. Adapt. Control Signal Process. 33 (2019) 285–299.

Jingang Zhao, received his B.E. degree in Automation from Qingdao University of Technology, China, in 2013, M.Sc. degree in Pattern Recognition and Intelligence System from Beijing Information Science and Technology University, China, in 2016, and Ph.D degree in Control Science and Engineering from Beijing Institute of Technology, China, in 2020. He is currently an assistant professsor with School of Electrical Engineering, Anhui Polytechnic University. During 2018–2019, he was a Visiting Scholar with the Department of Electrical and Computer Engineering at The Ohio State University, United States. His research interests include optimal control, reinforcement learning, event-triggered control and hybrid system.