非零和微分博弈系统的事件触发最优跟踪控制

石义博[1], 王朝立[2†]

(1. 上海理工大学 理学院, 上海 200093; 2. 上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘要: 近年来, 对于具有未知动态的非零和微分博弈系统的跟踪问题, 已经得到了讨论, 然而这些方法是时间触发的, 在传输带宽和计算资源有限的环境下并不适用. 针对具有未知动态的连续时间非线性非零和微分博弈系统, 本文提出了一种基于积分强化学习的事件触发自适应动态规划方法. 该策略受梯度下降法和经验重放技术的启发, 利用历史和当前数据更新神经网络权值. 该方法提高了神经网络权值的收敛速度, 消除了一般文献设计中常用的初始容许控制假设. 同时, 该算法提出了一种易于在线检查的持续激励条件(通常称为PE), 避免了传统的不容易检查的持续激励条件. 基于李亚普诺夫理论, 证明了跟踪误差和评价神经网络估计误差的一致最终有界性. 最后, 通过一个数值仿真实例验证了该方法的可行性.

关键词: 非零和博弈; 积分强化学习; 最优跟踪控制; 神经网络; 事件触发

引用格式: 石义博, 王朝立. 非零和微分博弈系统的事件触发最优跟踪控制. 控制理论与应用, 2023, 40(2): 220 – 230

DOI: 10.7641/CTA.2022.11292

Event-triggered optimal tracking control for nonzero-sum differential game systems

SHI Yi-bo[1], WANG Chao-li[2†]

(1. College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Recently, for the tracking problem of nonzero-sum differential game systems with unknown dynamics, it has been discussed that these methods are time-triggered, which is not ideal in an environment with limited transmission bandwidth and computing resources. In this paper, an integral reinforcement learning based event-triggered adaptive dynamic programming scheme is developed for continuous-time nonlinear nonzero-sum differential game systems with unknown dynamics. The strategy is inspired by the gradient descent method and the experience replay technique and uses the historical and current data to update the neural network weight. This method can improve the convergence speed of neural network weight and remove the assumption of initial admissible control often used in general literature design. In the meantime, the algorithm proposes a persistent excitation condition (commonly called PE) that is easy to check online, which avoids the traditional PE condition that is not easy to check. Based on the Lyapunov theory, the uniform ultimate boundedness (UUB) properties of the tracking error and the critic neural network estimation error have been proved. Finally, a numerical simulation example is given to verify the feasibility of the proposed method.

Key words: nonzero-sum games; integral reinforcement learning; optimal tracking control; neural network; event-triggered

Citation: SHI Yibo, WANG Chaoli. Event-triggered optimal tracking control for nonzero-sum differential game systems. *Control Theory & Applications*, 2023, 40(2): 220 – 230

1 Introduction

Optimal control is to design a control law to guarantee the stability of the system while minimizing the predetermined performance index function. In some practical applications, a large number of systems are controlled by multiple controllers, each of which can be regarded as a player, and each player minimizes its own cost function by influencing the state of the system, such as power system in [1], military in [2], and automatic driving in [3]. In this case, the optimal control problem of each player is coupled with the optimal control problem of other players. Therefore, such an optimal solution may not exist in the general case, which prompts researchers to find a new alternative form of the optimal standard. Game theory provides a solution to the optimal control problem with multiple players, named

Received 29 December 2021; accepted 24 June 2022.
†Corresponding author. E-mail: clwang@usst.edu.cn; Tel.: +86 21-55271422.
Recommended by Associate Editor: LONG Li-jun.
Supported by the National Defense Basic Research Program (JCKY2019413D001), the Natural Science Foundation (6217023627, 62003214, 62173054) and the Shanghai Natural Science Foundation (19ZR1436000).

Nash equilibrium. A Nash equilibrium is a combination of strategies that contains the optimal strategy for all players. That is, given the strategies of other players, no individual has an incentive to choose another control strategy, so no one will try to break the balance. Such a set of strategies is called a Nash equilibrium.

Differential game theory is an important part of the game theory which has received a lot of research attention in various fields. Differential games can be divided into fully cooperative games in [4], zero-sum games in [2, 5–6] and nonzero-sum (NZS) games in [7–11] according to the relationships between the players. In a fully cooperative game, all players complete an overall task and pursue team interests through complete cooperation. In a two-player zero-sum game, players compete with each other to pursue their own interests, one player's gain is the other's loss, and their control strategies are independent of each other. For the problem of H_∞ control, many scholars treat it as a zero-sum game. For a NZS game, players can minimize their cost function by cooperating or competing. Solving the Nash equilibrium of a NZS game ultimately comes down to solving a coupled Hamilton-Jacobi (HJ) equation, but because the coupled HJ equation is a nonlinear partial differential equation, the problem of "dimension disaster" will occur as the dimension increases.

Therefore, lots of scholars used the adaptive dynamic programming (ADP) based on the neural networks (NNs) to approximate the Nash equilibrium of NZS game. Vamvoudakis and Lewis in [8] used the critic-actor NNs based on policy iteration to solve the Nash equilibrium for NZS game systems, where the critic NNs and the actor NNs were used to approximate value functions and control strategies, respectively. Zhang et al. in [9] used the critic NNs to solve the Nash equilibrium of the NZS game system, which reduced one layer of NN compared with [8], reduced the calculation cost and without need the initial admissible control. However, both [8] and [9] require complete knowledge of system dynamics, which is not applicable to partially unknown systems. Kamalapurkar, Klotz and Dixon in [10] used identifier NNs to identify unknown system knowledge with partial unknown NZS game systems. However, the training of identifiers is often time-consuming and inevitably introduces detrimental identification errors. Zhang and Zhao in [11] used the data-driven integral reinforcement learning (IRL) method to solve partial unknown optimal control problems, avoiding the identification process, that is, avoiding the identification error.

The optimal control problems discussed above are all time-triggered, that is, the sampled data needs to be transmitted to the controller at every moment, and the control input needs to be updated at every moment. Generally speaking, the higher the sampling frequency of the controlled object, the more information can be collected to design the control input, and the corresponding control can work on the controlled object in time, so as to obtain better control performance. However, in some specific applications, such as networked control systems with geographically distributed sensors, controllers, and actuators, transmission bandwidth and computing resources are always limited. In this case, a higher sampling frequency may lead to network congestion and even more task delay. Therefore, it is of great significance to realize less control actions and less communication while ensuring the system performance. Therefore, a non-periodic event-triggered strategy is proposed to replace the traditional time-triggered strategy. Unlike time-triggered systems, the control input of an event-triggered system is updated only at trigger times determined by appropriately designed trigger conditions. In this way, event-triggered can significantly reduce the network bandwidth and computing burden.

For multi-player NZS games, event-based ADP has become the most common method used to approximate the control input of each player in [12–16]. An event-triggered ADP algorithm for solving discrete time multi-player game is proposed based on a model NN and critic NN in [12]. Su et al. in [13] proposed an actor-critic structure to solve the discrete time NZS games, which avoids unnecessary information transmission and computation. [12] and [13] have their respective advantages. Wang et al. [12] proved that the state of the closed-loop system is asymptotically stable, while the state of the closed-loop system in [13] is UUB. However, compared with [12–13] consider the saturation of control input more and only need to use local state measurement information. These are all about discrete time. Su et al. in [14] used the identifier-critic NN to solve continuous time for partially unknown NZS games. Su et al. in [15] used IRL to solve the optimal control for the partially unknown NZS game. Compared with [14], the introduction of identifier NN was avoided, i.e., the identification error was avoided. These are all about optimal regulation problems, but there are few studies on optimal tracking control problem (OTCP) in NZS games. However, in a real system, it is necessary to design a control input to track the state or output of the system to an ideal reference signal. There's a part of this research that is about optimal tracking control for the NZS games in [17–18], but it's all about time triggered. In the limited bandwidth constraints, these methods will not be applicable. For the above motivation, we study the OTCP for partially unknown NZS game event-triggered without the requirement of initial stabilizing control policies.

The main contributions of this paper are listed as follows.

1) In this paper, the OTCP based on event triggered is successfully extended from one control input[19–21] to N control inputs. Therefore, the problems and models considered in this paper are more general and more widely applicable than those in these literatures [19–21].

2) Compared with the critic NNs weight updating rules in [19–21], which only use the current data to update, this paper adopts the experience replay (ER) technology, that is, using the current data and historical data to update the weight. The traditional PE condition can be removed by using the ER technique, and the PE condition that is easier to check online can be obtained, and the convergence rate of weight is faster. Compared with [20], this paper does not need the assumption of initial admissible control.

3) Under the same model and problem, references [17] and [18] are time-triggered, which are not applicable in the environment with limited bandwidth and computing resources. However, this paper is event-triggered, which is not only applicable in the environment with limited bandwidth and computing resources, but also can significantly reduce the bandwidth occupation and computing burden.

The rest of this paper is organized as follows. In section 2, some basic knowledge of optimal control and event triggered mechanism are introduced. A single-critic network structure is proposed to approximate the optimal value function in Section 3. In Section 4, an online iterative algorithm is proposed and the stability of the closed-loop system is analyzed. In Section 5, a simulation example is given. Sections 6 concludes the paper.

Notions: $\mathbb{R}$ is the set of real numbers. $\mathbb{R}^+$ is all nonnegative real numbers. $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ denote the set of the real n-vectors and the $n\times m$ matrices, respectively. Let $\mathbb{N}=\{1,2,\cdots,N\}$ and $u_{-i}=\{u_1,u_2,\cdots,u_{i-1},u_{i+1},\cdots,u_N\}$. T is the transposition symbol. ∇ is the gradient operator. $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of a matrix. ξ is a vector or a matrix, $\|\xi\|$ represents the Euclidean norm or the 2-norm of ξ. Ω is a compact set, and $f(\cdot)\in C^1(\Omega)$ means $f(\cdot)$ is continuous first derivatives on Ω. A continuous function α will be of class-$\mathcal{K}$ if it strictly increasing with initial value being $\alpha(0)=0$; in addition, a class-$\mathcal{K}$ function α can be viewed as the class-$\mathcal{K}_\infty$ if it satisfies $\alpha(r)\to\infty$ as $r\to\infty$. Define $\beta(\xi^-)$ as the left limit of a function $\beta(r)$ when $r\to\xi$ from the left, i.e., $\beta(\xi^-)=\lim\limits_{\varepsilon\to 0}\beta(\xi-\varepsilon)$. The function $f(x)$ is Lipschitz continuous on Ω if the relation $\|f(x_1)-f(x_2)\|\leqslant\mathcal{D}\|x_1-x_2\|$ exists for all $x_1,x_2\in\Omega$ with the constant $\mathcal{D}>0$.

2 Preliminary

2.1 Problem statement

Consider the general N-player NZS differential games[8, 17–18]

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

where $x\in\mathbb{R}^n$ is system state, $u_j\in\mathbb{R}^{m_j}$ is control for player j. $f(x)\in\mathbb{R}^n$ and $g_j(x)\in\mathbb{R}^{n\times m_j}$ represent the drift dynamics and input dynamics of the system respectively. In this paper, we assume that $f(x)$ is unknown and $g_j(x)$ is known.

Assumption 1[22] $f(x)$ and $g_j(x)$ are Lipschitz continuous on a compact set $\bar{\Omega}\subset\mathbb{R}^n$ with $f(0)=0$, $f(\cdot)\leqslant b_f\|x\|$, and $\|g_j(x)\|\leqslant b_{g,j}$, where b_f and $b_{g,j}$ are positive constants.

Remark 1 The Lipschitz continuity of $f(x)$ and $g_j(x)$ is to ensure that system (1) has a unique solution for any initial state x_0. Although the boundedness of $g_j(x)$ is a little harsh, in practice there are still many systems that satisfy this condition, for example: aircraft systems.

The reference signal $r(t)$ is generated by a command generator

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

where $f_d(r(t))$ is the Lipschitz continuous with $f_d(0)=0$ and $r(t)\in\mathbb{R}^n$ is bounded. Note that the reference dynamics only need to be stable in the Lyapunov sense and are not necessarily asymptotically stable. Sine and cosine waves are some examples of such signals.

The tracking error is defined as

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

Using Eq. (3), the tracking error dynamic can be deduced as

$$\dot{e}_r(t)=f(e_r(t)+r(t))+\sum_{j=1}^{N}g_j(e_r(t)+r(t))u_j(t)-f_d(r(t)).\tag{4}$$

Then, construct an augmented system expressed as $\ell(t)=[e_r^{\mathrm{T}}(t)\ \ r^{\mathrm{T}}(t)]^{\mathrm{T}}\in\mathbb{R}^{2n}$. According to Eq. (2) and Eq. (4), the following augmented system can be obtained

$$\dot{\ell}(t)=F(\ell(t))+\sum_{j=1}^{N}G_j(\ell(t))u_j(t),\tag{5}$$

where

$$F(\ell(t))=\begin{bmatrix}f(r(t)+e_r(t))-f_d(r(t))\\ f_d(r(t))\end{bmatrix},$$

$$G_j(\ell(t))=\begin{bmatrix}g_j(r(t)+e_r(t))\\ \mathbf{0}_{n\times m_j}\end{bmatrix}.$$

According to Assumption 1 and the definition of G_j, one has $\|G_j\|\leqslant\lambda_{j,G}$, where $\lambda_{j,G}$ is a positive constant.

The cost function of system (5) is defined as follows:

$$\bar{J}_i(\ell(t), u_1, u_2, \cdots, u_N) = \int_t^{\infty} \mathrm{e}^{-\lambda(\tau-t)}(\ell^{\mathrm{T}}(\tau)\bar{Q}_i\ell(\tau) + \sum_{j=1}^{N} u_j^{\mathrm{T}}(\tau)R_{ij}u_j(\tau))\mathrm{d}\tau, \tag{6}$$

where

$$\bar{Q}_i = \begin{bmatrix} Q_i & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{bmatrix},$$

$Q_i = Q_i^{\mathrm{T}} \geqslant 0, R_{ii} = R_{ii}^{\mathrm{T}} > 0, R_{ij} = R_{ij}^{\mathrm{T}} \geqslant 0, \lambda > 0$ is a discount factor.

Definition 1 (Admissible Control)[22] The feedback control policy $u_i = u_i(\ell(t)) \in \Phi(\Omega)$ is admissible on with respect to Eq. (6) on a set $\Omega \subset \mathbb{R}^{2n}$, if $u_i(\ell(t))$ is continuous on Ω, $u_i(0) = 0$, $u_i(\ell(t))$ stabilizes the tracking error dynamics (4) on Ω, and Eq. (6) is finite $\forall \ell(t) \in \Omega$.

For the sake of description, let $u_i = u_i(\ell(t))$. For a given set of control input $\{u_1, \cdots, u_i, \cdots, u_N\}$, the value function for player i can be expressed as

$$V_i(\ell(t)) = \int_t^{\infty} \mathrm{e}^{-\lambda(\tau-t)}(\ell^{\mathrm{T}}(\tau)\bar{Q}_i\ell(\tau) + \sum_{j=1}^{N} u_j^{\mathrm{T}}R_{ij}u_j)\,\mathrm{d}\tau,\ i \in \mathbb{N}. \tag{7}$$

The purpose of OTCP is to design a set of control $\{u_1^*, u_2^*, \cdots, u_N^*\}$ so that the tracking error converges to zero while minimizing the value function (7). This control combination $\{u_1^*, u_2^*, \cdots, u_N^*\}$ corresponds to the Nash equilibrium of NZS games.

Definition 2 (Nash Equilibrium Strategies)[23] An N-tuple of control policies $\{u_1^*, u_2^*, \cdots, u_N^*\}$, $i \in \mathbb{N}$ is said to constitute a Nash equilibrium solution for an N-player game, if the following N inequalities are satisfied

$$\bar{J}_i(u_1^*, u_2^*, \cdots, u_N^*) \leqslant \bar{J}_i(u_1^*, u_2^*, \cdots, u_i, \cdots, u_N^*). \tag{8}$$

Remark 2 [22,24] The reason for the discount factor in the value function (7) is that $r(t)$ in this paper is not asymptotically stable, that is, when $t \to \infty$, $r(t) \not\equiv 0$, so $u_i \not\equiv 0$ at this time, which leads to the unbounded value function. Therefore, a discount factor should be added to the value function.

Assume that the value function $V_i(\ell(t)) \in C^1(\Omega)$. By differentiating V_i along the system trajectories (5), we can write Eq. (7) as

$$0 = U_i(\ell(t), u_1, u_2, \cdots, u_N) - \lambda V_i(\ell(t)) + \nabla V_i^{\mathrm{T}}(\ell(t))(F(\ell(t)) + \sum_{j=1}^{N} G_j(\ell(t))u_j),\ i \in \mathbb{N}, \tag{9}$$

where

$$U_i(\ell(t), u_1, u_2, \cdots, u_N) = \ell^{\mathrm{T}}\bar{Q}_i\ell(t) + \sum_{j=1}^{N} u_j^{\mathrm{T}}R_{ij}u_j,\ \nabla V_i = \frac{\partial V_i}{\partial \ell}.$$

The optimal value function V_i^* can be expressed as

$$V_i^*(\ell(t)) = \min_{u_i} \int_t^{\infty} \mathrm{e}^{-\lambda(\tau-t)}(\ell(\tau)^{\mathrm{T}}\bar{Q}_i\ell(\tau) + \sum_{j=1}^{N} u_j^{\mathrm{T}}R_{ij}u_j)\mathrm{d}\tau,\ i \in \mathbb{N}. \tag{10}$$

$V_i^*(\ell(t))$ is the solution of the Hamilton-Jacobi-Bellman (HJB) equation

$$\min_{u_i} H_i(\ell, \nabla V_i^*(\ell(t)), u_1, \cdots, u_i, \cdots, u_N) = 0. \tag{11}$$

where

$$H_i(\ell, \nabla V_i^*(\ell(t)), u_1, \cdots, u_i, \cdots, u_N) = U_i(\ell(t), u_1, u_2, \cdots, u_N) - \lambda V_i^*(\ell(t)) + (\nabla V_i^*(\ell(t)))^{\mathrm{T}}(F(\ell(t)) + \sum_{j=1}^{N} G_j(\ell(t))u_j).$$

Using the stationarity conditions $\dfrac{\partial H_i}{\partial u_i} = 0$, the optimal control input for player i is

$$u_i^*(\ell(t)) = -\frac{1}{2}R_{ii}^{-1}G_i^{\mathrm{T}}(\ell(t))\nabla V_i^*(\ell(t)),\ i \in \mathbb{N}. \tag{12}$$

The equivalent transformation of Eq. (7) is

$$V_i(\ell(t-\Delta t)) = \mathrm{e}^{-\lambda\Delta t}V_i(\ell(t)) + \int_{t-\Delta t}^{t} \mathrm{e}^{-\lambda(\tau-t+\Delta t)}U_i(\ell(\tau), u_i, u_{-i})\mathrm{d}\tau, \tag{13}$$

where $\Delta t > 0$ is a time interval.

According to Eq. (13), we have

$$V_i^*(\ell(t-\Delta t)) - \mathrm{e}^{-\lambda\Delta t}V_i^*(\ell(t)) = \int_{t-\Delta t}^{t} \mathrm{e}^{-\lambda(\tau-t+\Delta t)}U_i(\ell(\tau), u_i^*, u_{-i}^*)\mathrm{d}\tau. \tag{14}$$

It is easy to see from Eq. (14) that there are no more unknown dynamics. Therefore, the identification process of unknown dynamics $f(x)$ is no longer needed, that is, identification error is avoided.

In the above description, the control of the system needs to be updated in real time, and in some limited bandwidth environments, this approach is not suitable, and the calculation cost is too high. Therefore, to save communication and computing resources, this paper adopts a control method based on event triggered, which is sampled and updated by defined events.

2.2 Event-triggered control method

In time-triggered control, the N-tuple control input $\{u_1, \cdots, u_N\}$ is a feedback form of system state updated at each sample time. In event-triggered control, the N-tuple control input $\{u_1, \cdots, u_N\}$ is updated on-

ly if the state of the system breaks a preset threshold. In this case, a zero-order holder (ZOH) can be used to ensure that the control input is continuous at the trigger time. Define the triggering instants of events as τ_k, where $\{\tau_k\}_{k=0}^{\infty}$ is a monotonically increasing sequence of time instants with $\tau_0 = 0$. Define the event-triggered error as

$$e_k(t) = \breve{\ell}_k - \ell(t),\ t \in [\tau_k, \tau_{k+1}), \tag{15}$$

where $\breve{\ell}_k = \ell(\tau_k)$ is the event-trigged state.

According to $e_k(t)$ in Eq. (15), we elaborate the event-triggered mechanism as follows. When the event does not trigger, i.e., $t \neq \tau_k$, and $e_k(t) \neq 0$. In this case, the control input will remain constant for two adjacent trigger instants. When the event triggers at trigger time τ_k, i.e., $t = \tau_k$ and $e_k(\tau_k) = 0$. In this case, the control input will be updated.

In the framework of event triggered, the optimal control input Eq. (12) can be written as

$$u_i^*(\breve{\ell}_k) = -\frac{1}{2}R_{ii}^{-1}G_i^{\mathrm{T}}(\breve{\ell}_k)\nabla V_i^*(\breve{\ell}_k), \tag{16}$$

where $\nabla V_i^*(\breve{\ell}_k) = \dfrac{\partial V_i^*}{\partial \ell}|_{\ell=\breve{\ell}_k}$, $i \in \mathbb{N}$.

The piecewise continuous control signal can be expressed by a ZOH

$$u_i^*(t) = \begin{cases} u_i^*(\breve{\ell}_k),\ t \in [\tau_k, \tau_{k+1}), \\ -\dfrac{1}{2}R_{ii}^{-1}G_i^{\mathrm{T}}(\breve{\ell}_{k+1})\nabla V_i^*(\breve{\ell}_{k+1}), \\ \qquad t = \tau_{k+1}. \end{cases} \tag{17}$$

In this part, we mainly introduce the research questions and the basic knowledge of event triggered. In the next part of this article, we will use the critic NNs to learn value functions.

3 Single-critic structure

In the above analysis, we have concluded that the solution of optimal control (16) ultimately comes down to the HJ equation, so this section will use a critic NN to approximate the solution of Eq. (14). According to the Weierstrass high-order approximation theorem, we can get

$$V_i^*(\ell) = \omega_i^{*\mathrm{T}}\phi_i(\ell) + \varepsilon_i(\ell), \tag{18}$$

$$\nabla V_i^*(\ell) = \nabla\phi_i^{\mathrm{T}}(\ell)\omega_i^* + \nabla\varepsilon_i(\ell), \tag{19}$$

where $\omega_i^* \in \mathbb{R}^{K_i}$ is the unknown ideal weight, $\phi_i : \mathbb{R}^{2n} \to \mathbb{R}^{K_i}$ are linearly independent activation functions, K_i denotes the number of neurons, and ε_i is the approximation error.

Assumption 2[25] 1) The approximation error $\varepsilon_i(\ell)$ and its gradient $\nabla\varepsilon_i(\ell)$ are bounded on Ω, i.e., $\|\varepsilon_i(\ell)\| \leqslant b_{i,\varepsilon}$ and $\|\nabla\varepsilon_i(\ell)\| \leqslant b_{i,\nabla\varepsilon}$, with $b_{i,\varepsilon}, b_{i,\nabla\varepsilon}$, being positive constants. 2) The activation function $\phi_i(\ell)$ and its gradient $\nabla\phi_i(\ell)$ are bounded on Ω, i.e., $\|\phi_i(\ell)\| \leqslant b_{i,\phi}$ and $\|\nabla\phi_i(\ell)\| \leqslant b_{i,\nabla\phi}$, with $b_{i,\phi}, b_{i,\nabla\phi}$, being positive constants.

Remark 3 For Assumption 2 2), this condition is mild in practice since many activation functions, such as the sigmoid function and tanh function, satisfy Assumption 2 2).

Substituting Eq. (18) into Eq. (14), the Bellman equation (14) can be written

$$\begin{aligned} &\mathrm{e}_i(t) = \\ &\omega_i^{*\mathrm{T}}\left[\mathrm{e}^{-\lambda\Delta t}\phi_i(\ell(t)) - \phi_i(\ell(t-\Delta t))\right] + \\ &\int_{t-\Delta t}^{t}\mathrm{e}^{-\lambda(\tau-t+\Delta t)}U_i(\ell(\tau), u_i^*(\breve{\ell}_k), u_{-i}^*(\breve{\ell}_k))\mathrm{d}\tau, \end{aligned} \tag{20}$$

where $e_i(t) = \varepsilon_i(\ell(t-\Delta t)) - \mathrm{e}^{-\lambda\Delta t}\varepsilon_i(\ell(t))$ is error from the NN approximation error. According to Assumption 2, $e_i(t)$ is bound on Ω, i.e., $\|e_i(t)\| \leqslant b_{e,\mathrm{imax}}$ where $b_{e,\mathrm{imax}}$ is a positive constant.

Denote $\hat{\omega}_i$ as the estimations of ω_i^*. Then the value function can be approximated as

$$\hat{V}_i(\ell) = \hat{\omega}_i^{\mathrm{T}}\phi_i(\ell). \tag{21}$$

Based on Eq. (16), the approximate control inputs are

$$\hat{u}_i(\breve{\ell}_k) = -\frac{1}{2}R_{ii}^{-1}G_i^{\mathrm{T}}(\breve{\ell}_k)\nabla\phi_i^{\mathrm{T}}(\breve{\ell}_k)\hat{\omega}_i,\ i \in \mathbb{N}. \tag{22}$$

Using $\hat{V}_i(\ell)$ to replace $V_i(\ell)$ in Eq. (13). Therefore, the Bellman equation (13) can be written

$$\begin{aligned} &\hat{e}_i(t) = \\ &\hat{\omega}_i^{\mathrm{T}}\left[\mathrm{e}^{-\lambda\Delta t}\phi_i(\ell(t) - \phi_i(\ell(t-\Delta t))\right] + \\ &\int_{t-\Delta t}^{t}\mathrm{e}^{-\lambda(\tau-t+\Delta t)}U_i(\ell(\tau), \hat{u}_i(\breve{\ell}_k), \hat{u}_{-i}(\breve{\ell}_k))\mathrm{d}\tau. \end{aligned} \tag{23}$$

Eq. (23) can be written as

$$\hat{e}_i(t) = \hat{\omega}_i^{\mathrm{T}}\rho_i(t) + s_i(t), \tag{24}$$

where

$$\rho_i(t) = \mathrm{e}^{-\lambda\Delta t}\phi_i(\ell(t)) - \phi_i(\ell(t-\Delta t)), \tag{25}$$

$$s_i(t) = \int_{t-\Delta t}^{t}\mathrm{e}^{-\lambda(\tau-t+\Delta t)}U_i(\ell(\tau), \hat{u}_i(\breve{\ell}_k), \hat{u}_{-i}(\breve{\ell}_k))\mathrm{d}\tau. \tag{26}$$

It is worth noting that Eq. (24) is highly important for the proposed IRL method. From Eq. (24), it is clear that adjusting $\hat{\omega}_i$ will directly affect $\hat{e}_i(t)$. Therefore, the problem of solving the value function is transformed to adjusting the weight $\hat{\omega}_i$ to minimize the error $\hat{e}_i(t)$. Consider the objective function

$$E_i(t) = \frac{1}{2}\hat{e}_i^{\mathrm{T}}(t)\hat{e}_i(t). \tag{27}$$

In the following section, an online iterative learning scheme is proposed to update $\hat{\omega}_i$ by minimizing $E_i(t)$.

4 Online iterative learning

4.1 Online iterative learning algorithm

In this section, the gradient descent method is used for updating the estimated critic weight. This article

uses the ER technique to update weight. This method uses both historical and current data to update weight. Compared with traditional gradient descent using only the current data, the method adopted in this paper converges faster and obtains a PE condition that is easier to check.

Note $d \in \{1, \cdots, l\}$ is the index of the marked historical state $\ell(t_d)$, $t_d \in [\tau_k, \tau_{k+1})$, l is the number of marked historical states.

$\hat{\omega}_i$ is updated by minimizing the following error:

$$\mathcal{E}_i(t) = \frac{1}{2}\hat{e}_i^{\mathrm{T}}(t)\hat{e}_i(t) + \frac{1}{2}\sum_{d=1}^{l}\hat{e}_i^{\mathrm{T}}(t_d)\hat{e}_i(t_d). \quad (28)$$

Condition 1 Let $Z_i = [\rho_i(t_1), \cdots, \rho_i(t_l)]$ for player i. Then, Z_i in the recorded data contains as many linearly independent elements as the number of neurons in Eq. (18), i.e., rank(Z_i)=K_i.

Remark 4 Condition 1 is actually like a PE condition, but unlike the PE condition, it is easier to check in engineering practice[25]. It should be noted that in condition 1, the number of historical data to be collected l is greater than K_i. The amount of historical data l is constant, which means that as new data are added, old data are removed.

According to the gradient descent method and ER, the update rule of $\hat{\omega}_i$ can be obtained as

$$\begin{aligned}&\dot{\hat{\omega}}_i(t) = \\ &-\alpha_i\frac{\rho_i(t)}{\left(1+\rho_i^{\mathrm{T}}\rho_i(t)\right)^2}\left(s_i(t)+\rho_i^{\mathrm{T}}(t)\hat{\omega}_i(t)\right) - \\ &\alpha_i\sum_{d=1}^{l}\frac{\rho_i(t_d)}{\left(1+\rho_i^{\mathrm{T}}(t_d)\rho_i(t_d)\right)^2}\left(s_i(t_d)+\rho_i^{\mathrm{T}}(t_d)\hat{\omega}_i(t)\right),\end{aligned} \quad (29)$$

where α_i is the learning rate.

However, this learning rule also needs a prerequisite, that is, the initial admissible control. This condition prevents us from using update rule (29) directly. Therefore, the rest of this article discusses how to avoid this condition. The following assumption is generally employed in the study of the stability of closed-loop systems.

Assumption 3 [9,26] It is assumed that there exists a continuously differentiable radially unbounded Lyapunov candidate $J_i(\ell)$ such that $\dot{J}_i = \nabla J_i^{\mathrm{T}}\dot{\ell} = \nabla J_i^{\mathrm{T}}(F(\ell) + \sum_{j=1}^{N} G_j(\ell)u_j(\ell)) < 0$ with ∇J_i being the partial derivative of $J_i(\ell)$ with respect to ℓ. In addition, it holds that

$$\nabla J_i^{\mathrm{T}}(F(\ell) + \sum_{j=1}^{N} G_j(\ell)u_j^*(\breve{\ell}_k)) = -\nabla J_i^{\mathrm{T}}\bar{M}_i(\ell)\nabla J_i, \quad (30)$$

where the matrix $\bar{M}_i(\ell) \in \mathbb{R}^{2n\times 2n}$ is symmetric and positive definite.

Define an index of stability as

$$P(\ell, \hat{u}_1, \cdots, \hat{u}_N) = \begin{cases} 0, & \text{when } \mathcal{J}_i < 0, \\ 1, & \text{else.} \end{cases} \quad (31)$$

where $\mathcal{J}_i = (\nabla J_i(\ell))^{\mathrm{T}}(F(\ell) + \sum_{j=1}^{N} G_j(\ell)\hat{u}_j(\breve{\ell}_k))$.

Let $P(\ell,\hat{\mathcal{U}}) = P(\ell, \hat{u}_1, \cdots, \hat{u}_N)$. According to Eq. (29), the proposed updating rule for $\hat{\omega}_i$ is given by

$$\begin{aligned}&\dot{\hat{\omega}}_i(t) = \\ &-\alpha_i\frac{\rho_i(t)}{\left(1+\rho_i^{\mathrm{T}}\rho_i(t)\right)^2}\left(s_i(t)+\rho_i^{\mathrm{T}}(t)\hat{\omega}_i(t)\right) - \\ &\alpha_i\sum_{d=1}^{l}\frac{\rho_i(t_d)}{\left(1+\rho_i^{\mathrm{T}}(t_d)\rho_i(t_d)\right)^2}\left(s_i(t_d)+\rho_i^{\mathrm{T}}(t_d)\hat{\omega}_i(t)\right) + \\ &\frac{q_i}{2}P(\ell,\hat{\mathcal{U}})\nabla\phi_i(\breve{\ell}_k)G_i(\breve{\ell}_k)R_{ii}^{-1}G_i^{\mathrm{T}}(\ell)(\sum_{j=1}^{N}\nabla J_j),\end{aligned} \quad (32)$$

where q_i is the learning rate.

Remark 5 The first term in Eq. (32) was obtained by the standard gradient descent method. The second term in Eq. (32) is the recorded data based on the ER technique. According to [11, 25], the gradient descent method based on the ER technique has a faster weight convergence speed than the traditional gradient descent method. The last term in Eq. (32) is derived from the Lyapunov stability analysis and is used to ensure the stability of the system during value function learning. The choice of $P(\ell,\hat{\mathcal{U}})$ depends on the stability of the system in the learning process. When system Eq. (5) is stable, the operator $P(\ell,\hat{\mathcal{U}}) = 0$, and it will not work. When the system (5) is unstable, the operator $P(\ell,\hat{\mathcal{U}}) = 1$, and it will be activated. It is worth noting that $P(\ell,\hat{\mathcal{U}}) = 0$ holds only if the condition $\mathcal{J}_i < 0$ is met for all the players. Therefore, by introducing operator $P(\ell,\hat{\mathcal{U}})$, the requirement of initial admissible control in the learning process is eliminated.

4.2 Main results

Denote $\tilde{\omega}_i = \omega_i^* - \hat{\omega}_i$ is the critic weight estimation error and find that $\dot{\tilde{\omega}}_i = -\dot{\hat{\omega}}_i$.

$$\begin{aligned}&\dot{\tilde{\omega}}_i(t) = \\ &-\frac{\alpha_i\rho_i(t)\rho_i^{\mathrm{T}}(t)\tilde{\omega}_i(t)}{\left(1+\rho_i^{\mathrm{T}}\rho_i(t)\right)^2} - \alpha_i\sum_{d=1}^{l}\frac{\rho_i(t_d)\rho_i^{\mathrm{T}}(t_d)\tilde{\omega}_i(t)}{\left(1+\rho_i^{\mathrm{T}}(t_d)\rho_i(t_d)\right)^2} + \\ &\alpha_i\frac{\rho_i(t)}{\left(1+\rho_i^{\mathrm{T}}\rho_i(t)\right)^2}(s_i(t)+\rho_i^{\mathrm{T}}(t)\omega_i^*(t)) + \alpha_i\times \\ &\sum_{d=1}^{l}\frac{\rho_i(t_d)}{\left(1+\rho_i^{\mathrm{T}}(t_d)\rho_i(t_d)\right)^2}(s_i(t_d)+\rho_i^{\mathrm{T}}(t_d)\omega_i^*(t)) - \\ &\frac{q_i}{2}P(\ell,\hat{\mathcal{U}})\nabla\phi_i(\breve{\ell}_k)G_i(\breve{\ell}_k)R_{ii}^{-1}G_i^{\mathrm{T}}(\ell)(\sum_{j=1}^{N}\nabla J_j).\end{aligned} \quad (33)$$

Before we discuss the stability of closed-loop systems, we introduce the following assumptions in [15, 27].

Assumption 4 For $\forall i \in \mathbb{N}$, the control input u_i^* is locally Lipschitz with respect to $e_k(t)$. That is, there exists a constant $L_{u,i} > 0$ satisfying that

$$\|u_i^*(\ell) - u_i^*(\breve{\ell}_k)\|^2 \leqslant L_{u,i}\|e_k(t)\|^2.$$

Theorem 1 For the augmented system (5), suppose that Assumptions 1–4 and Condition 1 holds. Let the critic NN is updating by Eq. (32) and the following event-triggered condition

$$\|e_k(t)\| \leqslant \sqrt{\frac{(1-\eta^2)\lambda_{\min}(Q)\|e_r(t)\|^2 + \mathcal{U}(\breve{\ell}_k)}{\mathcal{L}}} \tag{34}$$

is adopted. Then, the tracking error $e_r(t)$ and the critic NN weight estimation error $\tilde{\omega}_i$ are all UUB, where $0 < \eta < 1$, $Q = \sum\limits_{i=1}^{N} Q_i$, $\mathcal{L} = \sum\limits_{i=1}^{N} \lambda_{\max}(R_{ii})L_{u,i} + \frac{1}{2}(N-1)\sum\limits_{i=1}^{N} \lambda_{i,G}^2 L_{u,i}$, $\mathcal{U}(\breve{\ell}_k) = \sum\limits_{i=1}^{N} \lambda_{\min}(R_{ii})\|\hat{u}_i(\breve{\ell}_k)\|^2$.

Proof. The Lyapunov function is defined as follows:

$$L(X) = L_1 + L_2 + L_3 + L_4, \tag{35}$$

where

$$\begin{cases} L_1 = \sum\limits_{i=1}^{N} V_i^*(\ell),\ L_2 = \sum\limits_{i=1}^{N} V_i^*(\breve{\ell}_k), \\ L_3 = \sum\limits_{i=1}^{N} \frac{1}{2}\tilde{\omega}_i^{\mathrm{T}}\tilde{\omega}_i,\ L_4 = \sum\limits_{i=1}^{N} q_i J_i(\ell). \end{cases} \tag{36}$$

For the convenience of description and analysis, in the following, L_m^i is denoted as the i-th term of L_m, where $m = 1, 2, 3, 4$. The whole proof is divided into two cases according to whether the events are triggered or not.

Case 1 $\big(t \in [\tau_k, \tau_{k+1})\big)$ When the event is not triggered, the derivative of the Lyapunov function with respect to t can be derived and obtained first

$$\dot{L}_1 = \sum_{i=1}^{N} \dot{V}_i^*(\ell) = \sum_{i=1}^{N} (\nabla V_i^*(\ell))^{\mathrm{T}}[F(\ell) + \sum_{j=1}^{N} G_j(\ell)\hat{u}_j(\breve{\ell}_k)]. \tag{37}$$

The derivative of the second term is $\dot{L}_2 = 0$ while the derivative of the third term for player i is

$$\begin{aligned} &\dot{L}_3^i = \\ &-\frac{\alpha_i\tilde{\omega}_i^{\mathrm{T}}\rho_i(t)\rho_i^{\mathrm{T}}\tilde{\omega}_i}{\left(1+\rho_i^{\mathrm{T}}\rho_i(t)\right)^2} - \alpha_i\sum_{d=1}^{l}\frac{\tilde{\omega}_i^{\mathrm{T}}\rho_i(t_d)\rho_i^{\mathrm{T}}(t_d)\tilde{\omega}_i}{\left(1+\rho_i^{\mathrm{T}}(t_d)\rho_i(t_d)\right)^2} + \\ &\alpha_i\frac{\tilde{\omega}_i^{\mathrm{T}}\rho_i(t)}{\left(1+\rho_i^{\mathrm{T}}\rho_i(t)\right)^2}(s_i(t) + \rho_i^{\mathrm{T}}\omega_i^*(t)) + \alpha_i \times \\ &\sum_{d=1}^{l}\frac{\tilde{\omega}_i^{\mathrm{T}}\rho_i(t_d)}{\left(1+\rho_i^{\mathrm{T}}(t_d)\rho_i(t_d)\right)^2}(s_i(t_d) + \rho_i^{\mathrm{T}}(t_d)\omega_i^*(t)) - \\ &\frac{q_i\tilde{\omega}_i^{\mathrm{T}}}{2}P(\ell,\hat{\mathcal{U}})\nabla\phi_i(\breve{\ell}_k)G_i(\breve{\ell}_k)R_{ii}^{-1}G_i^{\mathrm{T}}(\ell)(\sum_{j=1}^{N}\nabla J_j). \end{aligned} \tag{38}$$

Besides, the derivative of the last term is

$$\dot{L}_4 = \sum_{i=1}^{N} q_i(\nabla J_i(\ell))^{\mathrm{T}}[F(\ell) + \sum_{j=1}^{N} G_j(\ell)\hat{u}_j(\breve{\ell}_k)]. \tag{39}$$

For the sake of clarity, we analyze i-th term in Eq. (37) individually, and the transformation of the rest of the terms are analogous. According to Eq. (9) and Eq. (12), we can get

$$\begin{aligned} &\dot{L}_1^i = (\nabla V_i^*(\ell))^{\mathrm{T}}[F(\ell) + \sum_{j=1}^{N} G_j(\ell)\hat{u}_j(\breve{\ell}_k)] = \\ &(\nabla V_i^*(\ell))^{\mathrm{T}}F(\ell) + (\nabla V_i^*(\ell))^{\mathrm{T}}G_i(\ell)\hat{u}_i(\breve{\ell}_k) + \\ &(\nabla V_i^*(\ell))^{\mathrm{T}}\sum_{j\neq i}^{N} G_j(\ell)\hat{u}_j(\breve{\ell}_k) = \\ &\lambda V_i^*(\ell) - \ell^{\mathrm{T}}\bar{Q}_i\ell - \sum_{j=1}^{N}(u_j^*(\ell))^{\mathrm{T}}R_{ij}u_j^*(\ell) - \\ &(\nabla V_i^*(\ell))^{\mathrm{T}}\sum_{j=1}^{N} G_j u_j^*(\ell) + (\nabla V_i^*(\ell))^{\mathrm{T}}G_i(\ell)\hat{u}_i(\breve{\ell}_k) + \\ &(\nabla V_i^*(\ell))^{\mathrm{T}}\sum_{j\neq i}^{N} G_j(\ell)\hat{u}_j(\breve{\ell}_k) \leqslant \\ &\lambda V_i^*(\ell) + (\nabla V_i^*(\ell))^{\mathrm{T}}\sum_{j\neq i}^{N} G_j(\ell)(\hat{u}_j(\breve{\ell}_k) - u_j^*(\ell)) - \\ &2(u_i^*(\ell))^{\mathrm{T}}R_{ii}\hat{u}_i(\breve{\ell}_k) - e_r^{\mathrm{T}}Q_i e_r + \\ &(u_i^*(\ell))^{\mathrm{T}}R_{ii}u_i^*(\ell) \leqslant \\ &\lambda V_i^*(\ell) + (\nabla V_i^*(\ell))^{\mathrm{T}}\sum_{j\neq i}^{N} G_j(\ell)(\hat{u}_j(\breve{\ell}_k) - u_j^*(\ell)) + \\ &(u_i^*(\ell) - \hat{u}_i(\breve{\ell}_k))^{\mathrm{T}}R_{ii}(u_i^*(\ell) - \hat{u}_i(\breve{\ell}_k)) - e_r^{\mathrm{T}}Q_i e_r - \\ &\hat{u}_i^{\mathrm{T}}(\breve{\ell}_k)R_{ii}\hat{u}_i(\breve{\ell}_k). \end{aligned} \tag{40}$$

Thus, for the term $\dot{L}_1$, one can obtain

$$\begin{aligned} &\dot{L}_1 \leqslant \\ &\sum_{i=1}^{N}[\lambda V_i^*(\ell) - e_r^{\mathrm{T}}Q_i e_r + \lambda_{\max}(R_{ii})\|u_i^*(\ell) - \\ &\hat{u}_i(\breve{\ell}_k)\|^2 - \lambda_{\min}(R_{ii})\|\hat{u}_i(\breve{\ell}_k)\|^2 + \frac{1}{2}\|\nabla V_i^*(\ell)\|^2 + \\ &\frac{1}{2}\|\sum_{j\neq i}^{N} G_j(\ell)(u_j^*(\ell) - u_j(\breve{\ell}_k))\|^2] \leqslant \\ &\sum_{i=1}^{N}(\lambda V_i^*(\ell) + \frac{1}{2}\|\nabla V_i^*(\ell)\|^2) - \eta^2 e_r^{\mathrm{T}}Q e_r - \\ &(1-\eta^2)e_r^{\mathrm{T}}Q e_r + \mathcal{L}\|e_k(t)\|^2 - \mathcal{U}(\breve{\ell}_k). \end{aligned} \tag{41}$$

According to Assumption 2 and Eq. (18), the optimal value function $V_i^*(\ell)$ is bound by a positive constant b_{i,V_i^*} and its gradient $\nabla V_i^*(\ell)$ also is bound by a positive constant $b_{i,\nabla V_i^*}$.

For $\dot{L}_3^i$, we apply Young's inequality to the third and fourth terms on the right of Eq. (39)

$$\begin{aligned} &\dot{L}_3^i \leqslant \\ &-\alpha_i\lambda_{\min}(\Phi_i)\|\tilde{\omega}_i\|^2 + \alpha_i\frac{\tilde{\omega}_i^{\mathrm{T}}\rho_i(t)\rho_i^{\mathrm{T}}\tilde{\omega}_i}{2} + \end{aligned}$$

$$\alpha_i\frac{e_i^{\mathrm{T}}e_i(t)}{2}+\alpha_i\frac{\sum\limits_{d=1}^{l}\tilde{\omega}_i^{\mathrm{T}}\rho_i(t_d)\rho_i^{\mathrm{T}}(t_d)\tilde{\omega}_i}{2}+$$
$$\alpha_i\frac{\sum\limits_{d=1}^{l}e_i^{\mathrm{T}}(t_d)e_i(t_d)}{2}-\frac{q_i\tilde{\omega}_i^{\mathrm{T}}}{2}P(\ell,\hat{\mathcal{U}})\nabla\phi_i(\breve{\ell}_k)\times$$
$$G_i(\breve{\ell}_k)R_{ii}^{-1}G_i^{\mathrm{T}}(\ell)(\sum_{j=1}^{N}\nabla J_j)\leqslant$$
$$-\frac{\alpha_i}{2}\lambda_{\min}(\Phi_i)\|\tilde{\omega}_i\|^2+\frac{\alpha_i}{2}(1+l)b_{e,\mathrm{imax}}^2-$$
$$\frac{q_i\tilde{\omega}_i^{\mathrm{T}}}{2}P(\ell,\hat{\mathcal{U}})\nabla\phi_i(\breve{\ell}_k)G_i(\breve{\ell}_k)R_{ii}^{-1}G_i^{\mathrm{T}}(\ell)\times$$
$$(\sum_{j=1}^{N}\nabla J_j),\tag{42}$$

where

$$\Phi_i(\rho_i,\rho_i(t_d))=\frac{\rho_i(t)\rho_i^{\mathrm{T}}(t)}{(1+\rho_i^{\mathrm{T}}\rho_i(t))^2}+\sum_{d=1}^{l}\frac{\rho_i(t_d)\rho_i^{\mathrm{T}}(t_d)}{(1+\rho_i^{\mathrm{T}}(t_d)\rho_i(t_d))^2}.\tag{43}$$

When $P(\ell,\hat{\mathcal{U}})=0$, then for $\dot{L}_3$

$$\dot{L}_3\leqslant\sum_{i=1}^{N}(-\frac{\alpha_i}{2}\lambda_{\min}(\Phi_i)\|\tilde{\omega}_i\|^2+\frac{\alpha_i}{2}(1+l)b_{e,\mathrm{imax}}^2).\tag{44}$$

In this case, we can deduce that $\dot{L}_4^i$ is negative. From the dense property of $\mathbb{R}$[28], we can conclude that there is a constant $\tau_i>0$ such that

$$q_i(\nabla J_i(\ell))^{\mathrm{T}}\dot{\ell}<q_i\tau_i\|\nabla J_i(\ell)\|\leqslant 0.\tag{45}$$

Then

$$\dot{L}(X)\leqslant$$
$$\sum_{i=1}^{N}(\lambda V_i^*+\frac{1}{2}\|\nabla V_i^*\|^2-\frac{\alpha_i}{2}\lambda_{\min}(\Phi_i)\|\tilde{\omega}_i\|^2+$$
$$\frac{\alpha_i}{2}(1+l)b_{e,\mathrm{imax}}^2)-\eta^2e_r^{\mathrm{T}}Qe_r+$$
$$\sum_{i=1}^{N}(q_i\tau_i\|\nabla J_i(\ell)\|)-(1-\eta^2)e_r^{\mathrm{T}}Qe_r+$$
$$\mathcal{L}\|e_k(t)\|^2-\mathcal{U}(\breve{\ell}_k).\tag{46}$$

According to the event-triggered condition (34), one can obtain

$$\dot{L}(X)\leqslant-\sum_{i=1}^{N}(\frac{\alpha_i}{2}\lambda_{\min}(\Phi_i)\|\tilde{\omega}_i\|^2)-\eta^2e_r^{\mathrm{T}}Qe_r-\sum_{i=1}^{N}(q_i\tau_i\|\nabla J_i(\ell)\|)+Z,\tag{47}$$

where

$$Z=\sum_{i=1}^{N}[\lambda b_{i,V_i^*}+\frac{1}{2}b_{i,\nabla V_i^*}^2+\frac{\alpha_i}{2}(1+l)b_{e,\mathrm{imax}}^2].$$

Therefore, Eq. (47) produces $\dot{L}(X)<0$ as long as one of the following conditions holds:

$$\|e_r\|\geqslant\sqrt{\frac{Z}{\eta^2\lambda_{\min}(Q)}}=B_{e_r1},\tag{48}$$

or

$$\|\tilde{\omega}_i\|\geqslant\sqrt{\frac{2Z}{\alpha_i\lambda_{\min}(\Phi_i)}}=B_{\tilde{\omega}_i,1},\tag{49}$$

or

$$\|\nabla J_i(\ell)\|\geqslant\frac{Z}{q_i\tau_i}=B_{\nabla J_i,1}.\tag{50}$$

Thus, according to the Lyapunov extension theorem in [29], this proves the UUB stability of e_r and $\tilde{\omega}_i$.

When $P(\ell,\hat{\mathcal{U}})=1$, combine the last term of $\dot{L}_3$ with $\dot{L}_4$, one can obtain

$$\sum_{i=1}^{N}q_i(-\frac{1}{2}(\sum_{j=1}^{N}\nabla J_j(\ell))^{\mathrm{T}}G_i(\ell)R_{ii}^{-1}G_j^{\mathrm{T}}(\breve{\ell}_k)\tilde{\omega}_i(t)+$$
$$(\nabla J_i(\ell))^{\mathrm{T}}[F(\ell)+\sum_{j=1}^{N}G_j(\ell)\hat{u}_j(\breve{\ell}_k)])=$$
$$\sum_{i=1}^{N}q_i(-\frac{1}{2}(\sum_{j=1}^{N}\nabla J_j(\ell))^{\mathrm{T}}G_i(\ell)R_{ii}^{-1}G_j^{\mathrm{T}}(\breve{\ell}_k)\tilde{\omega}_i(t)+$$
$$(\nabla J_i(\ell))^{\mathrm{T}}[F(\ell)-\frac{1}{2}\sum_{j=1}^{N}G_j(\ell)R_{jj}^{-1}G_j^{\mathrm{T}}(\breve{\ell}_k)\nabla\phi_j^{\mathrm{T}}\times$$
$$(\omega_j-\tilde{\omega}_j)])=$$
$$\sum_{i=1}^{N}q_i(\nabla J_i(\ell))^{\mathrm{T}}[F(\ell)-\frac{1}{2}\sum_{j=1}^{N}G_j(\ell)R_{jj}^{-1}G_j^{\mathrm{T}}(\breve{\ell}_k)\times$$
$$\nabla\phi_j^{\mathrm{T}}\omega_j]=$$
$$\sum_{i=1}^{N}(q_i(\nabla J_i(\ell))^{\mathrm{T}}[F(\ell)+\sum_{j=1}^{N}G_j(\ell)u_j^*(\breve{\ell}_k)]+$$
$$\frac{1}{2}q_i(\nabla J_i(\ell))^{\mathrm{T}}\sum_{j=1}^{N}G_j(\ell)R_{jj}^{-1}G_j^{\mathrm{T}}(\breve{\ell}_k)\nabla\varepsilon_i)\leqslant$$
$$\sum_{i=1}^{N}q_i(-\lambda_{\min}(\bar{M}_i)\|\nabla J_i\|^2+\frac{1}{2}D_i\|\nabla J_i\|)=$$
$$\sum_{i=1}^{N}q_i(-\lambda_{\min}(\bar{M}_i)(\|\nabla J_i\|-\frac{D_i}{4\lambda_{\min}(\bar{M}_i)})^2+$$
$$\frac{D_i^2}{16\lambda_{\min}(\bar{M}_i)}),\tag{51}$$

where $D_i=q_ib_{i,\nabla\varepsilon}\sum\limits_{j=1}^{N}\lambda_{j,G}^2\|R_{jj}^{-1}\|$.

Now,

$$\dot{L}(X)\leqslant-\sum_{i=1}^{N}\frac{\alpha_i}{2}\lambda_{\min}(\Phi_i)\|\tilde{\omega}_i\|^2-\sum_{i=1}^{N}q_i(\lambda_{\min}(\bar{M}_i)\times(\|\nabla J_i\|-\frac{D_i}{4\lambda_{\min}(\bar{M}_i)})^2)+\mathcal{Z}-\eta^2e_r^{\mathrm{T}}Qe_r,\tag{52}$$

where $\mathcal{Z}=Z+\sum\limits_{i=1}^{N}\frac{q_iD_i^2}{16\lambda_{\min}(\bar{M}_i)}$.

If at least one of the following inequalities holds:

$$\|e_r\|\geqslant\sqrt{\frac{\mathcal{Z}}{\eta^2\lambda_{\min}(Q)}}=B_{e_r2},\tag{53}$$

or

$$\|\tilde{\omega}_i\|\geqslant\sqrt{\frac{2\mathcal{Z}}{\alpha_i\lambda_{\min}(\Phi_i)}}=B_{\tilde{\omega}_i,2},\tag{54}$$

or

$$\|\nabla J_i\| \geqslant \sqrt{\frac{\mathcal{Z}}{q_i\lambda_i(\bar{M}_i)}} + \frac{D_i}{4\lambda_{\min}(\bar{M}_i)} = B_{\nabla J_i,2}, \quad (55)$$

then, $\dot{L}(X) < 0$. Thus, according to the Lyapunov extension theorem in [29], this proves the UUB stability of e_r and $\tilde{\omega}_i$.

In summary, for the case $P(\ell,\hat{\mathcal{U}}) = 0$ or 1 , if the condition $\|e_r\| \geqslant \max\{B_{e_r,1}, B_{e_r,2}\} = \bar{B}_{e_r}$ or $\|\tilde{\omega}_i\| \geqslant \max\{B_{\tilde{\omega}_i,1}, B_{\tilde{\omega}_i,2}\} = \bar{B}_{\tilde{\omega}_i}$ or $\|\nabla J_i(\ell)\| \geqslant \max\{B_{\nabla J_i,1}, B_{\nabla J_i,2}\} = \bar{B}_{\nabla J_i}$ holds, then $\dot{L}(X) < 0$. According to the standard Lyapunov extension theorem, one can conclude that the tacking error e_r and NN weight estimation error $\tilde{\omega}_i$ are bounded by $\bar{B}_{e_r}$, $\bar{B}_{\tilde{\omega}_i}$, respectively.

Case 2 ($t = \tau_{k+1}$) The event is triggered. Thus, the differential form of Eq. (35)

$$\begin{aligned}\Delta L(X(\tau_{k+1})) = \Delta L(\breve{\ell}_{k+1}) - \Delta L(\ell(\tau_{k+1}^-)) = \\ \Delta L_1 + \Delta L_2 + \Delta L_3 + \Delta L_4.\end{aligned} \quad (56)$$

Since the state and value functions of the system are continuous, it follows that $\Delta L_1 \leqslant 0$, $\Delta L_3 \leqslant 0$, and $\Delta L_4 \leqslant 0$, where

$$\begin{aligned}\Delta L_1 =& \sum_{i=1}^{N}(V_i^*(\breve{\ell}_{k+1}) - V_i^*(\ell(\tau_{k+1}^-))),\\ \Delta L_3 =& \frac{1}{2}\sum_{i=1}^{N}\left[\tilde{\omega}_i^{\mathrm{T}}(\tau_{k+1})\tilde{\omega}_i(\tau_{k+1}) - \right.\\ & \left.\tilde{\omega}_i^{\mathrm{T}}(\ell(\tau_{k+1}^-))\tilde{\omega}_i(\ell(\tau_{k+1}^-))\right],\\ \Delta L_4 =& \sum_{i=1}^{N} q_i[J_i(\breve{\ell}_{k+1}) - J_i(\ell(\tau_{k+1}^-))].\end{aligned} \quad (57)$$

Combining these time difference terms, one can obtain

$$\begin{aligned}&\Delta L(X(\tau_{k+1})) \leqslant \Delta L_2 = \\ &\sum_{i=1}^{N}(V_i^*(\breve{\ell}_{k+1}) - V_i^*(\breve{\ell}_k)) \leqslant \\ &-\sum_{i=1}^{N}\mathcal{K}_i\|\breve{\ell}_{k+1} - \breve{\ell}_k\|,\end{aligned} \quad (58)$$

where $\mathcal{K}_i$ are class-$\mathcal{K}$. That means that the Lyapunov function (35) is decreasing when $\forall t = \tau_{k+1}$.

According to the above two case, the triggering condition (34) and the inequalities (48)–(50) or (53)–(55) guarantee that the tracking error e_r and the weight error $\tilde{\omega}_i$ of the critic NN are all UUB, which ends of the proof. □

Remark 6 From the expression of $\bar{B}_{e_r}$, together with Eq. (48) and Eq. (53), it can be seen that $\bar{B}_{e_r}$ can be reduced by increasing $\lambda_{\min}(Q)$. It can be seen from Eq. (49) and Eq. (54) that the convergence rate of the weight of NN depends on the minimum eigenvalue of matrix Φ_i, which means that the convergence speed of the weight can be increased by maximizing the minimum eigenvalue of Φ_i. Because the existence of the third term in Eq. (32) eliminates the need for initial admissible control. So the initial weight of the input in learning precess, for convenience, can be selected as zeros. The above design still needs to solve the important problem of how to avoid the Zeno behavior. According to the proof similar to that in [16, 30], it can be concluded that trigger rule (34) is Zeno-free.

5 Simulation

In this section, we simulate the OTCP of nonlinear two-person differential game system using time-triggered and event-triggered respectively and then verify the effectiveness of our proposed method through comparison. Note that time-triggered is run with a fixed sampling period of 0.005 s.

Consider the following nonlinear differential games with two-player[8, 18]:

$$\dot{x} = f(x) + g_1(x)u_1 + g_2(x)u_2, \quad (59)$$

where

$$f(x) = \begin{bmatrix} x_2 \\ -x_2 - 0.5x_1 + 0.25x_2(\cos(2x_1)+2)^2 \\ -0.25(\sin(4x_1^2)+2)^2 \end{bmatrix},$$

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

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

$x = [x_1\ x_2]^{\mathrm{T}} \in \mathbb{R}^2$ is the system state, and $u_1, u_2 \in \mathbb{R}$ are the control inputs.

The reference signal is generated by the following command:

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

Select $Q_1 = 2I, Q_2 = I, R_{11} = R_{12} = 2I$, and $R_{21} = R_{22} = I$, I is an identity matrix. The parameters in the learning process are set as $\alpha_1 = \alpha_2 = 3$, $q_1 = q_2 = 3$, $L_{u,1} = L_{u,2} = 13$, $\lambda_{1,G} = \lambda_{2,G} = 7$, $T = 0.005$, and $J_i = 0.5\ell^{\mathrm{T}}\ell$. The augmented system states are $\ell(t) = [\ell_1\ \ \ell_2\ \ \ell_3\ \ \ell_4]^{\mathrm{T}} = [e_1\ \ e_2\ \ r_1\ \ r_2]^{\mathrm{T}}$, and the NN activation functions are selected as

$$\begin{aligned}&\phi_1(\ell(t)) = \phi_2(\ell(t)) = \\ &[e_1^2\ \ e_1e_2\ \ e_1r_1\ \ e_1r_2\ \ e_2^2\ \ e_2r_1\ \ e_2r_2\ \ r_1^2\ \ r_1r_2\ \ r_2^2]^{\mathrm{T}}.\end{aligned}$$

Since the initial admissible control strategy is eliminated in this paper, the initial weight can be selected as zero for convenience.

Fig.1 shows the critic NN weight convergence curve of player 1, which finally converges to

$$\begin{aligned}\hat{\omega}_1 = [&0.0016\ \ 0.6721\ \ 0.0003\ \ 0\ \ 3.5978\ \ 0.2255,\\ &0.4646\ \ 0.0001\ \ -0.0001\ \ 6.9762]^{\mathrm{T}}.\end{aligned}$$

Fig.2 shows the critic NN weight convergence curve

of player 2, which finally converges to

$$\hat{\omega}_2 = [0.0012\ \ 0.5298\ \ 0.0008\ \ 0.0001\ \ 8.9317 \\ -0.0472\ \ 0.6102\ \ 0.0001\ \ -0.0004\ \ 0]^{\mathrm{T}}.$$

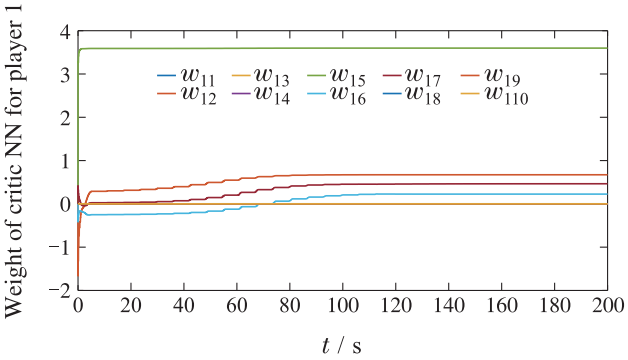

Fig. 1 The evolution process of the critic NNs weight of the first player.

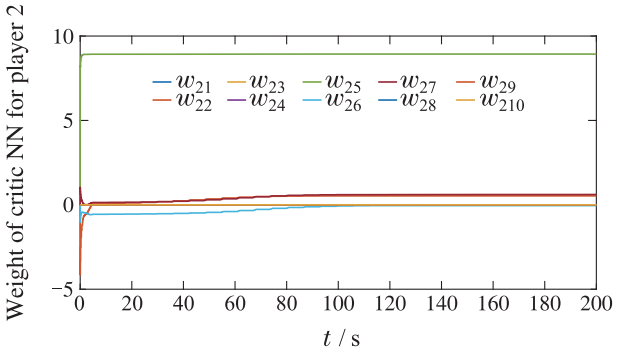

Fig. 2 The evolution process of the critic NNs weight of the second player.

Fig. 3 shows a three-dimensional actual state trajectory and reference trajectory. Fig. 4 shows the evolution of tracking error in the whole learning process. It can be observed that the tracking error converges gradually to zero. According to Fig. 5, we can observe that the minimum event triggered interval is 0.01 s (avoiding the Zeno behavior), which is larger than the time triggering interval, which can effectively reduce communication. During the whole learning process, the time-based controller needs to be updated 40,000 times, while the event-based controller only needs to be updated 8624 times. In other words, the recalculation and transmission of control inputs during the adaptive process are reduced. Therefore, more system resources can be saved by using our method.

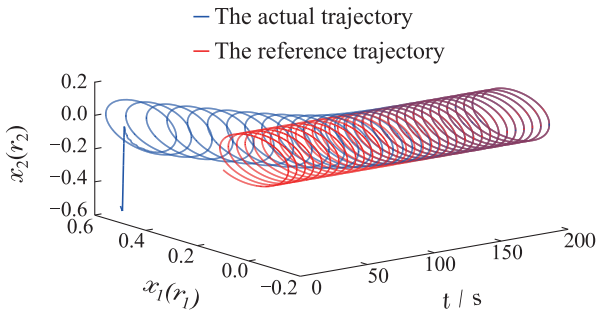

Fig. 3 The actual trajectory and the reference trajectory.

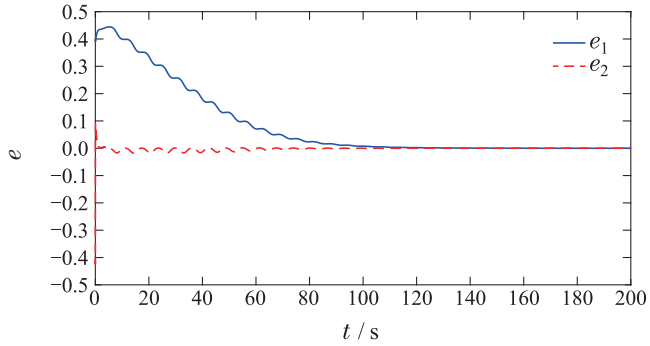

Fig. 4 The evolution of tracking errors.

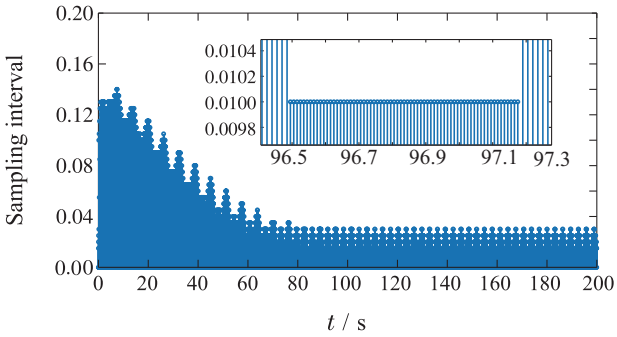

Fig. 5 The triggering interval of event-triggered.

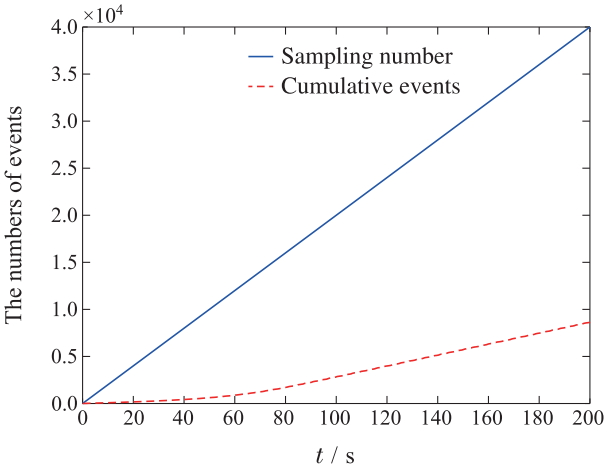

Fig. 6 The cumulative number of the events.

6 Conclusion

In this paper, we studied the OTCP for N-player NZS game systems with unknown drift dynamics. An IRL method is used to avoid unknown drift dynamics systems. The solution of Nash equilibrium is obtained by constructing a single layer critic NN. By improving the updating rules of standard gradient descent weight, the PE conditions are easier to check online, convergence speed is faster and initial admissible control is no longer required. By designing a reasonable trigger condition, the calculation and communication burden in the whole control process are reduced. The UUB properties of the tracking error and the critic NN estimation error are proved. Finally, the effectiveness of the proposed method was demonstrated by a numerical example.

References:

[1] WANG D, HE H B, MU C X, et al. Intelligent critic control with disturbance attenuation for affine dynamics including an application to a microgrid system. *IEEE Transactions on Industrial Electronics*, 2017, 64(6): 4935 – 4944.

[2] SUN J L, LIU C S, ZHAO X. Backstepping-based zero-sum differential games for missile-target interception systems with input and output constraints. *IET Control Theory and Applications*, 2018, 12(2): 243 – 253.

[3] KODAGODA K, WIJESOMA W, TEOH E. Fuzzy speed and steering control of an AGV. *IEEE Transactions on Control Systems Technology*, 2002, 10(1): 112 – 120.

[4] ZHANG Q C, ZHAO D B, ZHU Y H. Data-driven adaptive dynamic programming for continuous-time fully cooperative games with partially constrained inputs cooperative games with partially constrained inputs. *Neurocomputing*, 2017, 238: 377 – 386.

[5] MU C X, WANG K. Single-network ADP for near optimal control of continuous-time zero-sum games without using initial stabilising control laws. *IET Control Theory and Applications*, 2018, 12(18): 2449 – 2458.

[6] WEI Q L, LIU D R, LIN Q, et al. Adaptive dynamic programming for discrete-time zero-sum games. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(4): 957 – 969.

[7] LIU D R, LI H L, WANG D. Online synchronous approximate optimal learning algorithm for multi-player nonzero-sum games with unknown dynamics. *IEEE Transactions on Systems Man and Cybernetics Systems*, 2014, 44(8): 1015 – 1027.

[8] VAMVOUDAKIS K, LEWIS F. Multi-player nonzero-sum games: Online adaptive learning solution of coupled Hamilton-Jacobi equations. *Automatica*, 2011, 47(8): 1556 – 1569.

[9] ZHANG H G, CUI L L, LUO Y H. Near optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP. *IEEE Transactions on Cybernetics*, 2012, 43(1): 206 – 216.

[10] KAMALPURKAR R, KLOTZ J, DIXON W. Concurrent learning-based approximate feedback-nash equilibrium solution of N-player nonzero-sum differential games. *IEEE/CAA Journal of Automatica Sinica*, 2014, 1(3): 239 – 247.

[11] ZHANG Q C, ZHAO D B. Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics. *IEEE Transactions on Cybernetics*, 2019, 49(8): 2874 – 2885.

[12] WANG Z Y, WEI Q L, LIU D R. Event-triggered adaptive dynamic programming for discrete-time multi-player games. *Information Sciences*, 2020, 506: 457 – 470.

[13] SU H G, ZHANG H G, JIANG H, et al. Decentralized event-triggered adaptive control of discrete-time nonzero-sum games over wireless sensor-actuator networks with input constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(10): 4254 – 4266.

[14] SU H G, ZHANG H G, LIANG Y L, et al. Online event-triggered adaptive critic design for non-zero-sum games of partially unknown networked systems. *Neurocomputing*, 2019, 368: 84 – 98.

[15] SU H G, ZHANG H G, SUN S X, et al. Integral reinforcement learning-based online adaptive event-triggered control for nonzero-sum games of partially unknown nonlinear systems. *Neurocomputing*, 2020, 377: 243 – 255.

[16] ZHAO Q T, SUN J, WANG G, et al. Event-triggered ADP for nonzero-sum games of unknown nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 33(5): 1905 – 1913.

[17] LÜ Y F, REN X M, NA J. Adaptive optimal tracking controls of unknown multi-input systems based on nonzero-sum game theory. *Journal of the Franklin Institute*, 2019, 356(15): 8255 – 8277.

[18] ZHAO J G. Neural networks-based optimal tracking control for nonzero-sum games of multi-player continuous-time nonlinear systems via reinforcement learning. *Neurocomputing*, 2020, 412: 167 – 176.

[19] CUI L L, XIE X P, WANG X W, et al. Event-triggered single-network ADP method for constrained optimal tracking control of continuous-time nonlinear systems. *Applied Mathematics and Computation*, 2019, 352: 220 – 234.

[20] VAMVOUDAKIS K, MOJOODI A, FERRAZ H. Event-triggered optimal tracking control of nonlinear systems: Event-triggered optimal tracking control of nonlinear systems. *International Journal of Robust and Nonlinear Control*, 2017, 27: 598 – 619.

[21] XUE S, LUO B, LIU D R, et al. Adaptive dynamic programming-based event-triggered optimal tracking control. *International Journal of Robust and Nonlinear Control*, 2021, 31(15): 7480 – 7497.

[22] MODARES H, LEWIS F. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica*, 2014, 50(7): 1780 – 1792.

[23] BASAR, T. Dynamic noncooperative game theory (2nd ed.). *Philadelphia*, PA: SIAM, 1999.

[24] LIU C, ZHANG H G, REN H, et al. An analysis of IRL-based optimal tracking control of unknown nonlinear systems with constrained input. *Neural Processing Letters*, 2019, 50(3): 2681 – 2700.

[25] MODARES H, LEWIS F, SISTANI N. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica*, 2014, 50(1): 193 – 202.

[26] LIU P D, ZHANG H G, LIU C, et al. Online dual-network-based adaptive dynamic programming for solving partially unknown multi-player nonzero-sum games with control constraints. *IEEE Access*, 2020, 8: 182295 – 182306.

[27] YANG X, WEI Q L. Adaptive critic learning for constrained optimal event-triggered control with discounted cost. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 32(1): 91 – 104.

[28] RUDIN W. *Principles of mathematical analysis*. New York: McGraw-Hill, 1976.

[29] LEWIS F, JAGANNATHAN S, YESILDIREK A. *Neural network control of robot manipulators and nonlinear systems*. USA: Taylor & Francis Inc., 1999.

[30] XUE S, LUO B, LIU D R. Integral reinforcement learning based event-triggered control with input saturation. *Neural Networks*, 2020, 131: 144 – 153.

作者简介:

石义博 硕士研究生, 目前研究方向为非线性控制、最优控制及博弈论, E-mail: syb19971008@163.com;

王朝立 博士, 教授, 博士生导师, 目前研究方向为非线性控制、机器人控制、模糊控制, E-mail: clwang@usst.edu.cn.