International Journal of Control, Automation, and Systems 22(5) (2024) 1751-1759
http://dx.doi.org/10.1007/s12555-022-1133-1

ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Reinforcement Q-learning and Optimal Tracking Control of Unknown Discrete-time Multi-player Systems Based on Game Theory

Jin-Gang Zhao

Abstract: This paper studies the fully cooperative game tracking control problem (FCGTCP) for a class of discrete-time multi-player linear systems with unknown dynamics. The reference trajectory is generated by a command generator system. An augmented multi-player systems composed of the origin multi-player systems and the command generator system is constructed, and an exponential discounted cost function is introduced to derive an augmented fully cooperative game tracking algebraic Riccati equation (FCGTARE). When the system dynamics are known, a model-based policy iteration (PI) algorithm is proposed to solve the augmented FCGTARE. Furthermore, to relax the system dynamics, an online reinforcement Q-learning algorithm is designed to obtain the solution to the augmented FCGTARE. The convergence of designed online reinforcement Q-learning algorithm is proved. Finally, two simulation examples are given to verify the validity of the model-based PI algorithm and online reinforcement Q-learning algorithm.

Keywords: Discrete-time, fully cooperative game (FCG), multi-player systems, Q-learning, tracking control.

1. INTRODUCTION

Tracking control aims to design a feedback controller such that the output of the control system can track a reference signal while ensuring the closed-loop stability [1]. The past few decades have seen extensive exploration of tracking control [2-5]. It has been widely used in various fields, such as mobile robots [6], quadrotor [7], overhead cranes [8], multiagent systems [9], etc. These wide-ranging applications have greatly promoted the development of optimal tracking control [10-14], which strives to minimize or maximize a predefined performance index function while ensuring that the output tracks a reference signal. Unlike the optimal control problem, the optimal tracking control problem needs to consider both the control system dynamics and the reference signal dynamics, and is more complicated to solve [1]. In addition, the optimal control problem can be essentially regarded as the optimal tracking control problem when the reference signal is zero, which is a special kind of optimal tracking control problem. Therefore, the study of the optimal tracking control problem is of more practical value.

As we all know, for general discrete-time linear systems, it is well known that the solution to optimal tracking control problem can be found by solving an associated algebraic Riccati equation (ARE) [15]. For discrete-time multi-player linear systems, the solution to optimal tracking control problem can be found by solving an associated non-zero-sum game ARE (NZSGARE, from a non-zero sum game perspective) [16] or fully cooperative game ARE (FCGTARE, from a fully cooperative game perspective) [17]. However, it is not easy to directly solve these equations due to the nonlinearity of the unknown parameters. In addition, considering practical applications, we often hope to obtain the solution of the optimal tracking control problem without relying on the accurate model of the system.

In recent years, reinforcement learning (RL), especially adaptive dynamic programming (ADP) based on RL [18-20], has become a powerful tool for solving optimal control problems for complex systems with unknown models, such as generally linear and nonlinear control system [21-23], Helicopter [24], multi-player systems [25], cyber-physical systems [26,27]. For optimal tracking control problems, in general linear and nonlinear systems, reference [15] proposed a reinforcement Q-learning algorithm to solve the ARE. A critic-Only Q-Learning method was proposed to solve the optimal tracking control problem of nonaffine nonlinear discrete-time systems [28]. A model-free policy gradient ADP method is designed for optimal tracking control problem of discrete-time nonlinear systems [29]. Optimal parallel tracking control for general

Manuscript received November 25, 2022; revised April 12, 2023; accepted July 27, 2023. Recommended by Associate Editor Daniela Iacoviello under the direction of Senior Editor Yoshito Ohta. This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2022QF096, and in part by the Weifang University Doctoral Research Startup Fund under Grant 2021BS26. The author would like to thank to his alma mater, Beijing Institute of Technology.

Jin-Gang Zhao is with the School of Machinery and Automation, Institute of Intelligent Perception and Optimization Control of Complex Systems, Weifang University, Weifang, Shandong 261061, China (e-mail: zhaojingang521@126.com).

nonlinear systems was investigated by a new ADP method [30]. A novel value function was proposed to solve the optimal tracking problem of nonlinear discrete-time systems using ADP method in [31]. For linear and nonlinear multi-player systems, references [16,32] respectively developed off-policy reinforcement learning method and Q-learning approach to solve the NZSGARE. References [33,34] have designed RL approach to the optimal tracking control problem of multi-player systems from the non-zero sum game perspective, respectively. In terms of fully cooperative games, scholars mainly focus on the optimal control problem of multi-player linear and nonlinear systems. Reference [17] designed a data-driven ADP method for optimal control problem of multi-player systems with partially constrained inputs. A neural network-based ADP approach was proposed to deal with the cooperative game issues of discrete-time multi-player systems in [35]. Reference [36] investigated the optimal control problem of multi-player systems with completely unknown dynamics using data-driven ADP from the perspective of fully cooperative games. In cooperative games, where all players have the same performance index function and achieve a common goal, it is actually a special case of a non-zero-sum game. In [17,35,36], the optimal control problem for fully cooperative games in different situations was studied using reinforcement learning. However, fewer studies have been conducted for the fully cooperative game optimal tracking control problem, which motivates our research in this paper.

This paper will study the optimal tracking control problem of discrete-time multi-player linear systems from the perspective of fully cooperative game, and considering that the control system mathematical model is usually difficult to obtain in practical applications, we design a reinforcement Q-learning method. The designed method in this paper does not depend on system dynamics and has more practical application value. The main contributions of this paper can be described as follows:

1) The tracking control problem for a class of discrete-time multi-player systems may be the first to be studied from the perspective of fully cooperative game.
2) An exponential discounted cost function is introduced. Accordingly, the corresponding Bellman equation and FCGTARE for FCGTCP are derived.
3) An online reinforcement Q-learning algorithm is proposed to solve the FCGTCP without requiring the system dynamics. The convergence of proposed online reinforcement Q-learning algorithm is proved.

The rest of this paper is organized as follows: Section 2 formulates the FCGTCP of multi-player linear systems. In Section 3, we present the Bellman equation and FCGTARE for FCGTCP. In Section 4, an online reinforcement Q-learning algorithm is designed to solve the augmented FCGTARE. Simulation studies on a discretized F-16 dynamic system model is given to demonstrate the effectiveness of the designed online reinforcement Q-learning algorithm in Section 5. Section 6 concludes this paper and gives the future research directions.

2. PROBLEM FORMULATION

Consider a class of discrete-time multi-player linear systems with two players

$$\begin{aligned} x_{k+1} &= Ax_k + B_1 u_{1k} + B_2 u_{2k}, \\ y_k &= Cx_k, \end{aligned} \tag{1}$$

where $x_k \in \mathbb{R}^n$ denotes the system state, $u_{1k} \in \mathbb{R}^{m_1}$ and $u_{2k} \in \mathbb{R}^{m_2}$ denote the two players or two control inputs, $y_k \in \mathbb{R}^p$ denotes the system output. $A \in \mathbb{R}^{n\times n}$, $B_1 \in \mathbb{R}^{n\times m_1}$, $B_2 \in \mathbb{R}^{n\times m_2}$, and $C \in \mathbb{R}^{p\times n}$ are constant matrices, and it is assumed that A, B_1, B_2 are unknown.

Assumption 1: (A,B_1) and (A,B_2) are controllable and (A,C) is observable.

The goal of FCGTCP is to find a tuple of feedback control inputs (u_{1k},u_{2k}) for the system (1) which ensures that the output y_k tracks a reference trajectory r_k, and the two control inputs take actions together as a team to minimize the following cost function or value function

$$\begin{aligned} &V(x_k,r_k,u_{1k},u_{2k}) \\ &= \sum_{i=k}^{\infty} e^{-\alpha(i-k)} \big[(Cx_i - r_i)^T Q (Cx_i - r_i) \\ &\quad + u_{1i}^T R_1 u_{1i} + u_{2i}^T R_2 u_{2i}\big], \end{aligned} \tag{2}$$

where Q, R_1, R_2 are positive definite matrices with compatible dimensions. $e^{-\alpha} \in (0,\ 1)$ is a discount factor and $\alpha > 0$ is an adjustable parameter.

In other words, the optimal control inputs (u_{1k}^*, u_{2k}^*) can be obtained by solving the minimization problem as

$$V^*(x_k,r_k) = V(u_{1k}^*, u_{2k}^*) = \min_{u_{1k},u_{2k}} V(u_{1k},u_{2k}), \tag{3}$$

and satisfy

$$V(u_{1k}^*, u_{2k}^*) \le \min\{V(u_{1k}, u_{2k}^*), V(u_{1k}^*, u_{2k})\}. \tag{4}$$

The optimal control inputs (u_{1k}^*, u_{2k}^*) obtained by (3) and satisfying (4) constitute a coordination equilibria solution of two-player FCG [17].

The reference trajectory is generated by the following command generator system

$$r_{k+1} = Fr_k, \tag{5}$$

where $F \in \mathbb{R}^{p\times p}$ is a constant matrix. Note that F may not be Hurwitz due to the introduction of the discount factor $e^{-\alpha}$ in the cost function (2).

Defining $X_k = \begin{bmatrix} x_k^T & r_k^T \end{bmatrix}^T$, based on (1) and (5), an augmented system with two players is constructed as follows:

$$\begin{aligned} X_{k+1} &= \begin{bmatrix} x_{k+1} \\ r_{k+1} \end{bmatrix} = \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & F \end{bmatrix} \begin{bmatrix} x_k \\ r_k \end{bmatrix} + \begin{bmatrix} B_1 \\ \mathbf{0} \end{bmatrix} u_{1k} + \begin{bmatrix} B_2 \\ \mathbf{0} \end{bmatrix} u_{2k} \\ &= \bar{A} X_k + \bar{B}_1 u_{1k} + \bar{B}_2 u_{2k}. \end{aligned} \tag{6}$$

According to the augmented system state, the cost function (2) can be rewritten as follows:

$$V(X_k) = \sum_{i=k}^{\infty} e^{-\alpha(i-k)} \left[X_i^T \bar{Q} X_i + u_{1i}^T R_1 u_{1i} + u_{2i}^T R_2 u_{2i} \right], \tag{7}$$

where

$$\bar{Q} = \begin{bmatrix} C^T Q C & -C^T Q \\ -QC & Q \end{bmatrix}.$$

Now, the solution of FCGTCP can be obtained by solving the optimal control problem consisting of augmented system (6) and cost function (7).

3. THE SOLUTION FOR THE FCGTCP

In this section, the Bellman equation and FCGTARE for FCGTCP are firstly presented. Then, when system dynamics A and B are known, a model-based online PI algorithm is given to solve the FCGTARE.

3.1. Derivation of Bellman equation and FCGTARE

The value function (7) can be written in the following recursive form

$$\begin{aligned} &V(X_k) \\ &= X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} \\ &\quad + e^{-\alpha} \sum_{i=k+1}^{\infty} e^{-\alpha(i-k-1)} [X_i^T \bar{Q} X_i + u_{1i}^T R_1 u_{1i} + u_{2i}^T R_2 u_{2i}]. \end{aligned} \tag{8}$$

According to (8), we can obtain the Bellman equation for FCGTCP as follows:

$$\begin{aligned} V(X_k) = {} & X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} \\ & + e^{-\alpha} V(X_{k+1}). \end{aligned} \tag{9}$$

Similar to [15], the value function (7) can be written in a quadratic form as follows:

$$V(X_k) = X_k^T P X_k. \tag{10}$$

Based on (10), the Bellman equation (9) can be rewritten as follows:

$$\begin{aligned} X_k^T P X_k = {} & X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} \\ & + e^{-\alpha} X_{k+1}^T P X_{k+1}. \end{aligned} \tag{11}$$

Define the FCGTCP Hamiltonian equation as

$$\begin{aligned} H(X_k, u_{1k}, u_{2k}) = {} & X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} \\ & + e^{-\alpha} X_{k+1}^T P X_{k+1} - X_k^T P X_k, \end{aligned} \tag{12}$$

or equivalently

$$\begin{aligned} H(X_k, u_{1k}, u_{2k}) = {} & X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} \\ & + e^{-\alpha} V(X_{k+1}) - V(X_k). \end{aligned} \tag{13}$$

The next theorem will show how to solve the FCGTCP by an augmented FCGTARE.

Theorem 1: For the augmented system (6) with the cost function (7), the optimal control inputs u_{1k}^* and u_{2k}^* have the form

$$\begin{aligned} u_{1k}^* &= L_1^* X_k, \\ u_{2k}^* &= L_2^* X_k, \end{aligned}$$

with

$$\begin{aligned} L_1^* = {} & [F_{11}^* - e^{-2\alpha} \bar{B}_1^T P^* \bar{B}_2 (F_{22}^*)^{-1} \bar{B}_2^T P^* \bar{B}_1]^{-1} \\ & \times [e^{-2\alpha} \bar{B}_1^T P^* \bar{B}_2 (F_{22}^*)^{-1} \bar{B}_2^T P^* \bar{A} - e^{-\alpha} \bar{B}_1^T P^* \bar{A}], \end{aligned} \tag{14}$$

$$\begin{aligned} L_2^* = {} & [F_{22}^* - e^{-2\alpha} \bar{B}_2^T P^* \bar{B}_1 (F_{11}^*)^{-1} \bar{B}_1^T P^* \bar{B}_2]^{-1} \\ & \times [e^{-2\alpha} \bar{B}_2^T P^* \bar{B}_1 (F_{11}^*)^{-1} \bar{B}_1^T P^* \bar{A} - e^{-\alpha} \bar{B}_2^T P^* \bar{A}], \end{aligned} \tag{15}$$

and the P^* satisfies the following augmented FCGTARE

$$\begin{aligned} P^* = {} & e^{-\alpha} \bar{A}^T P^* \bar{A} + \bar{Q} - e^{-2\alpha} \begin{bmatrix} \bar{A}^T P^* \bar{B}_1 & \bar{A}^T P^* \bar{B}_2 \end{bmatrix} \\ & \times \begin{bmatrix} F_{11}^* & F_{12}^* \\ F_{21}^* & F_{22}^* \end{bmatrix}^{-1} \begin{bmatrix} \bar{B}_1^T P^* \bar{A} \\ \bar{B}_2^T P^* \bar{A} \end{bmatrix}, \end{aligned} \tag{16}$$

where $F_{11}^* = R_1 + e^{-\alpha} \bar{B}_1^T P^* \bar{B}_1$, $F_{12}^* = e^{-\alpha} \bar{B}_1^T P^* \bar{B}_2$, $F_{21}^* = e^{-\alpha} \bar{B}_2^T P^* \bar{B}_1$, $F_{22}^* = R_2 + e^{-\alpha} \bar{B}_2^T P^* \bar{B}_2$.

Proof: Based on (13), according to the stationary conditions $\frac{\partial H(X_k, u_{1k}, u_{2k})}{\partial u_{1k}} = 0$ and $\frac{\partial H(X_k, u_{1k}, u_{2k})}{\partial u_{2k}} = 0$, we have

$$\begin{aligned} & (R_1 + e^{-\alpha} \bar{B}_1^T P \bar{B}_1) u_{1k} + e^{-\alpha} \bar{B}_1^T P \bar{B}_2 u_{2k} \\ & = -e^{-\alpha} \bar{B}_1^T P \bar{A} X_k, \end{aligned} \tag{17}$$

and

$$\begin{aligned} & (R_2 + e^{-\alpha} \bar{B}_2^T P \bar{B}_2) u_{2k} + e^{-\alpha} \bar{B}_2^T P \bar{B}_1 u_{1k} \\ & = -e^{-\alpha} \bar{B}_2^T P \bar{A} X_k. \end{aligned} \tag{18}$$

By solving (17) and (18) simultaneously, we can obtain the optimal control inputs u_{1k}^* and u_{2k}^* with (14) and (15).

Furthermore, the augmented FCGTARE can be obtained by substituting the obtained optimal control inputs u_{1k}^* and u_{2k}^* into the Bellman equation (11). □

It is worth noting that the system stability is affected by the discount factor $e^{-\alpha}$. In practice, we can always choose a small α or a large Q to guarantee stability [37].

3.2. Model-based online PI algorithm for solving FCGTARE

Since the augmented FCGTARE (16) is a nonlinear equation of P^* and involves matrix inversion, it is difficult to solve the FCGTARE (16) directly. Inspired by [15], a model-based online PI algorithm presented in Algorithm 1 is developed to solve the Bellman equation (11).

Remark 1: In policy evaluation, the LS is employed to implement Algorithm 1 online by using the data tuple X_k, X_{k+1}, u_{1k}, u_{2k} measured along the system trajectories. In fact, (19) is a scalar equation and P is a positive symmetric $(n+p)\times(n+p)$ matrix with $(n+p)\times(n+p+1)/2$ independent element. Therefore, at least $(n+p)\times(n+p+1)/2$ data tuples are required to solve (19) using LS. In addition, to maintain persistence of excitation (PE), probing noises are generally added to the control inputs. The addition of the probe noise may cause Algorithm 1 to produce a biased solution [37].

It should be noted that the complete knowledge of system dynamics A and B are required in Algorithm 1. To eliminate the requirement for system dynamics, a reinforcement Q-learning algorithm is provided to solve the FCGTCP in the next section.

Algorithm 1: Model-based online PI algorithm.

1) **Initialization:** Start with initial admissible control input policies $\{u_1^0, u_2^0\}$ and the iteration number $j=0$.

2) **Policy evaluation:** Solve for P^{j+1} using the least-squares (LS) by

$$X_k^T P^{j+1} X_k = X_k^T \bar{Q} X_k + (u_{1k}^j)^T R_1 u_{1k}^j + (u_{2k}^j)^T R_2 u_{2k}^j + e^{-\alpha} X_{k+1}^T P^{j+1} X_{k+1}. \tag{19}$$

3) **Policy improvement:** Update the control input policies using obtained P^{j+1} by

$$F_{11}^{j+1} = R_1 + e^{-\alpha}\bar{B}_1^T P^{j+1}\bar{B}_1, \tag{20}$$

$$F_{22}^{j+1} = R_2 + e^{-\alpha}\bar{B}_2^T P^{j+1}\bar{B}_2, \tag{21}$$

$$\begin{aligned} L_1^{j+1} = {} & [F_{11}^{j+1} - e^{-2\alpha}\bar{B}_1^T P^{j+1}\bar{B}_2 (F_{22}^{j+1})^{-1}\bar{B}_2^T \\ & \times P^{j+1}\bar{B}_1]^{-1}[e^{-2\alpha}\bar{B}_1^T P^{j+1}\bar{B}_2 (F_{22}^{j+1})^{-1} \\ & \times \bar{B}_2^T P^{j+1}\bar{A} - e^{-\alpha}\bar{B}_1^T P^{j+1}\bar{A}], \end{aligned} \tag{22}$$

$$\begin{aligned} L_2^{j+1} = {} & [F_{22}^{j+1} - e^{-2\alpha}\bar{B}_2^T P^{j+1}\bar{B}_1 (F_{11}^{j+1})^{-1}\bar{B}_1^T \\ & \times P^{j+1}\bar{B}_2]^{-1}[e^{-2\alpha}\bar{B}_2^T P^{j+1}\bar{B}_1 (F_{11}^{j+1})^{-1} \\ & \times \bar{B}_1^T P^{j+1}\bar{A} - e^{-\alpha}\bar{B}_2^T P^{j+1}\bar{A}]. \end{aligned} \tag{23}$$

4) If $\|L_1^{j+1} - L_1^j\| \le \varepsilon$ and $\|L_2^{j+1} - L_2^j\| \le \varepsilon$, stop and use $\{L_1^{j+1}, L_2^{j+1}\}$ as the approximated optimal L_1^*, L_2^*, where ε is a pre-given small positive number; Else, let $j=j+1$, and go to Step 2.

4. REINFORCEMENT Q-LEARNING TO SOLVE THE FCGTARE

In this section, a reinforcement Q-learning algorithm without requiring the system dynamics A, B_1, B_2 and reference trajectory dynamics F is designed to solve the augmented FCGTARE (16).

4.1. Q-function for the FCGTCP

According to the FCGTCP Bellman equation (11), define the FCGTCP Q-function as

$$Q(X_k, u_{1k}, u_{2k}) = X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} + e^{-\alpha} X_{k+1}^T P X_{k+1}. \tag{24}$$

Using augmented system (6), (24) becomes

$$\begin{aligned} & Q(X_k, u_{1k}, u_{2k}) \\ & = X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} + e^{-\alpha} X_{k+1}^T P X_{k+1} \\ & = X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} \\ & \quad + e^{-\alpha}(\bar{A}X_k + \bar{B}u_{1k} + \bar{B}u_{2k})^T P(\bar{A}X_k + \bar{B}u_{1k} + \bar{B}u_{2k}) \\ & = \begin{bmatrix} X_k \\ u_{1k} \\ u_{2k} \end{bmatrix}^T H \begin{bmatrix} X_k \\ u_{1k} \\ u_{2k} \end{bmatrix}, \end{aligned} \tag{25}$$

where the kernel matrix

$$\begin{aligned} & H \\ & = \begin{bmatrix} \bar{Q} + e^{-\alpha}\bar{A}^T P\bar{A} & e^{-\alpha}\bar{A}^T P\bar{B}_1 & e^{-\alpha}\bar{A}^T P\bar{B}_2 \\ e^{-\alpha}\bar{B}_1^T P\bar{A} & R_1 + e^{-\alpha}\bar{B}_1^T P\bar{B}_1 & e^{-\alpha}\bar{B}_1^T P\bar{B}_2 \\ e^{-\alpha}\bar{B}_2^T P\bar{A} & e^{-\alpha}\bar{B}_2^T P\bar{B}_1 & R_2 + e^{-\alpha}\bar{B}_2^T P\bar{B}_2 \end{bmatrix} \\ & = \begin{bmatrix} H_{XX} & H_{Xu_1} & H_{Xu_2} \\ H_{u_1X} & H_{u_1u_1} & H_{u_1u_2} \\ H_{u_2X} & H_{u_2u_1} & H_{u_2u_2} \end{bmatrix} \in \mathbb{R}^{l\times l}, \end{aligned} \tag{26}$$

where $l = n + p + m_1 + m_2$.

Based on the Q-function, the FCGTCP is to derive

$$Q^*(X_k, u_{1k}, u_{2k}) = \min_{u_{1k}, u_{2k}} Q(X_k, u_{1k}, u_{2k}). \tag{27}$$

By applying $\frac{\partial Q(X_k,u_{1k},u_{2k})}{\partial u_{1k}} = 0$ and $\frac{\partial Q(X_k,u_{1k},u_{2k})}{\partial u_{2k}} = 0$ to (25), we can obtain the following optimal control input polices

$$u_{1k}^* = [H_{u_1u_1}^* - H_{u_1u_2}^*(H_{u_2u_2}^*)^{-1}H_{u_2u_1}^*]^{-1} \times [H_{u_1u_2}^*(H_{u_2u_2}^*)^{-1}H_{u_2X}^* - H_{u_1X}^*]X_k, \tag{28}$$

$$u_{2k}^* = [H_{u_2u_2}^* - H_{u_2u_1}^*(H_{u_1u_1}^*)^{-1}H_{u_1u_2}^*]^{-1} \times [H_{u_2u_1}^*(H_{u_1u_1}^*)^{-1}H_{u_1X}^* - H_{u_2X}^*]X_k, \tag{29}$$

and optimal control gains

$$L_1^* = [H^*_{u_1u_1} - H^*_{u_1u_2}(H^*_{u_2u_2})^{-1}H^*_{u_2u_1}]^{-1} \times [H^*_{u_1u_2}(H^*_{u_2u_2})^{-1}H^*_{u_2X} - H^*_{u_1X}], \tag{30}$$

$$L_2^* = [H^*_{u_2u_2} - H^*_{u_2u_1}(H^*_{u_1u_1})^{-1}H^*_{u_1u_2}]^{-1} \times [H^*_{u_2u_1}(H^*_{u_1u_1})^{-1}H^*_{u_1X} - H^*_{u_2X}], \tag{31}$$

which are the same as (14) and (15), respectively.

4.2. Online reinforcement Q-learning algorithm for FCGTCP

According to the Q-function (24), we can develop a reinforcement Q-learning algorithm to solve the FCGTARE (16) online without requiring the augmented system dynamics.

The Q-function (24) satisfies the following Bellman equation

$$Q(X_k, u_{1k}, u_{2k}) = X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} + e^{-\alpha} Q(X_{k+1}, u_{1k+1}, u_{2k+1}). \tag{32}$$

Define

$$Z_k = \begin{bmatrix} X_k & u_{1k} & u_{2k} \end{bmatrix}^T,$$

to rewrite (25) as follows:

$$Q(X_k, u_{1k}, u_{2k}) = Z_k^T H Z_k. \tag{33}$$

By substituting (33) into (32), we can rewrite the Q-function Bellman equation as follows:

$$Z_k^T H Z_k = X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} + e^{-\alpha} Z_{k+1}^T H Z_{k+1}. \tag{34}$$

Furthermore, denote

$$Z_k^T H Z_k = \bar{H}^T \bar{Z}_k, \tag{35}$$

with

$$\bar{H} = vec(H) \in \mathbb{R}^{l(l+1)/2} \triangleq [H_{11},\ 2H_{12},\ \dots,\ 2H_{1l},\ H_{22},\ H_{23},\ \dots,\ H_{2l},\ \dots.,\ H_{ll}]^T, \tag{36}$$

and

$$\bar{Z}_k = Z_k \otimes Z_k \in \mathbb{R}^{l(l+1)/2},$$

where H_{ij}, i, $j = 1, 2, ..., l$ represents the ith row and the jth column element of matrix H. $\otimes$ represents the Kronecker product.

By substituting (35) and (36) into (34), yields the following parameterized Q-function Bellman equation

$$\bar{H}^T \bar{Z}_k = X_k^T \bar{Q} X_k + u_{1k}^T R_1 u_{1k} + u_{2k}^T R_2 u_{2k} + e^{-\alpha} \bar{H}^T \bar{Z}_{k+1}. \tag{37}$$

Based on the parameterized Q-function Bellman equation (37), we can establish an online reinforcement Q-learning algorithm presented in Algorithm 2.

Algorithm 2: Online reinforcement Q-learning algorithm.

1) **Initialization:** Start with initial admissible control input policies $\{u_1^0, u_2^0, \bar{H}^0\}$.

2) **Policy evaluation:** Solve for P^{j+1} using the least-squares (LS) by

$$(\bar{H}^{j+1})^T (\bar{Z}_k - e^{-\alpha} \bar{Z}_{k+1}) = X_k^T \bar{Q} X_k + (u_{1k}^{j+1})^T R_1 u_{1k}^{j+1} + (u_{2k}^{j+1})^T R_2 u_{2k}^{j+1}. \tag{38}$$

3) **Policy improvement:** Update the control input policies

$$u_{1k}^{j+1} = [H^{j+1}_{u_1u_1} - H^{j+1}_{u_1u_2}(H^{j+1}_{u_2u_2})^{-1}H^{j+1}_{u_2u_1}]^{-1} \times [H^{j+1}_{u_1u_2}(H^{j+1}_{u_2u_2})^{-1}H^{j+1}_{u_2X} - H^{j+1}_{u_1X}]X_k, \tag{39}$$

$$u_{2k}^{j+1} = [H^{j+1}_{u_2u_2} - H^{j+1}_{u_2u_1}(H^{j+1}_{u_1u_1})^{-1}H^{j+1}_{u_1u_2}]^{-1} \times [H^{j+1}_{u_2u_1}(H^{j+1}_{u_1u_1})^{-1}H^{j+1}_{u_1X} - H^{j+1}_{u_2X}]X_k. \tag{40}$$

4) If $\left\|\bar{H}^{j+1} - \bar{H}^j\right\| \le \varepsilon$, stop and use $\{u_{1k}^{j+1},\ u_{2k}^{j+1}\}$ as the approximated optimal control inputs u_{1k}^*, u_{2k}^*, where ε is a pre-given small positive number; Else, let $j = j+1$, and go to Step 2.

Remark 2: Similar to Algorithm 1, in policy evaluation of Algorithm 2, the LS is adopted. Since $\bar{H}$ has $l(l+1)/2$ independent elements, we need to collect at least $l(l+1)/2$ data samples. Similarly, to maintain persistence of excitation (PE), probing noises need to be added to the control inputs. Unlike Algorithm 1, in Algorithm 2, the added probing noises do not cause any bias in estimating the Q-function [37].

Theorem 2: The online reinforcement Q-learning algorithm converges to the optimal solution given in Theorem 1, as $j \to \infty$ under the sufficient excitation.

Proof: By substituting (39) and (40) into (38) and doing some math transformations, one has

$$P^{j+1} = e^{-\alpha} \bar{A}^T P^{j+1} \bar{A} + \bar{Q} - e^{-2\alpha} \begin{bmatrix} \bar{A}^T P^{j+1} \bar{B}_1 & \bar{A}^T P^{j+1} \bar{B}_2 \end{bmatrix} \times \begin{bmatrix} F_{11}^{j+1} & F_{12}^{j+1} \\ F_{21}^{j+1} & F_{22}^{j+1} \end{bmatrix}^{-1} \begin{bmatrix} \bar{B}_1^T P^{j+1} \bar{A} \\ \bar{B}_2^T P^{j+1} \bar{A} \end{bmatrix}, \tag{41}$$

where $F_{11}^{j+1} = R_1 + e^{-\alpha} \bar{B}_1^T P^{j+1} \bar{B}_1$, $F_{12}^{j+1} = e^{-\alpha} \bar{B}_1^T P^{j+1} \bar{B}_2$, $F_{21}^{j+1} = e^{-\alpha} \bar{B}_2^T P^{j+1} \bar{B}_1$, $F_{22}^{j+1} = R_2 + e^{-\alpha} \bar{B}_2^T P^{j+1} \bar{B}_2$. □

According to the arguments in [38], we can conclude that iterating on (41) converges to the solution of the augmented FCGTARE (16). This completes the proof.

Remark 3: The developed online reinforcement Q-learning algorithm presented in Algorithm 2 is model-free and can be extended in a straightforward manner to discrete-time multi-players systems with more than two players.

Remark 4: Similar to [15,39], the tracking error e can be made as small as desired by choosing a small adjustable parameter α, R_1, R_2 and/or large Q. Simulation results in Section 5 will confirm this conclusion.

5. SIMULATION

In this section, in order to verify the validity of our proposed scheme, two simulation examples are presented in the following.

5.1. Example 1

Consider a discretized F-16 dynamic system model from [16] as follows:

$$x_{k+1} = Ax_k + B_1 u_{1k} + B_2 u_{2k},$$
$$y_k = Cx_k,$$

where $A = \begin{bmatrix} 0.9065 & 0.0816 & -0.0009 \\ 0.0741 & 0.9012 & -0.0159 \\ 0 & 0 & 0.9048 \end{bmatrix}$, $B_1 = [-0.0002,\ -0.0041,\ 0.4758]^T$, $B_2 = [0.0952,\ 0.0038,\ 0]^T$, $C = [1,\ -1,\ 1]^T$. The reference trajectory dynamic $F = -1$. $\alpha = 0.1$, $Q = 10000$, $R_1 = 0.01$, and $R_2 = 0.05$. The initial admissible control input policies are chosen as $u_1^0 = [-1,\ 0,\ 0,\ 1]$, $u_2^0 = [-1,\ 0,\ 0,\ 1]$. Algorithms 1 and 2 are respectively applied to the discretized F-16 system for simulation experiments. And, some suitable probing noise is added into initial input policies for the first 950 times. The simulation results corresponding to Algorithm 1 are shown in Figs. 1-3. The simulation results corresponding to Algorithm 2 are depicted in Figs. 4-6.

From Figs. 1 and 4, it can be seen that L_1 and L_2 can quickly converge to the optimal value L_1^* and L_2^*. From Figs. 2, 3, 5, and 6 , it can be seen that the designed scheme can achieve good tracking performance.

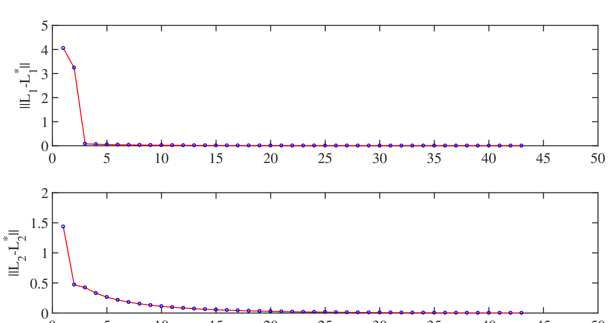

Fig. 1. The evolution of L_1, L_2 under Algorithm 1.

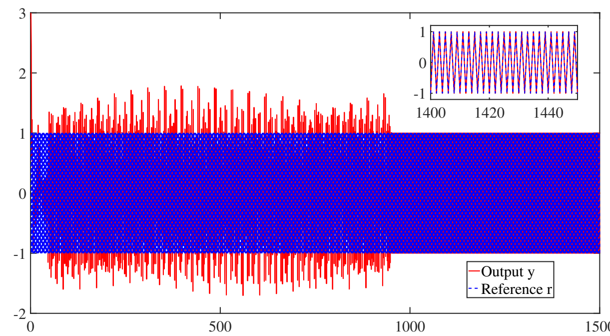

Fig. 2. Output y and reference r under Algorithm 1.

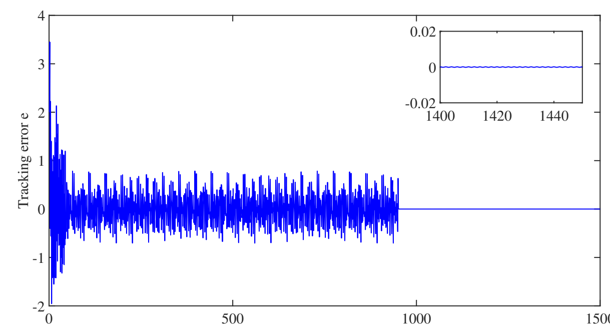

Fig. 3. The tracking error e under Algorithm 1.

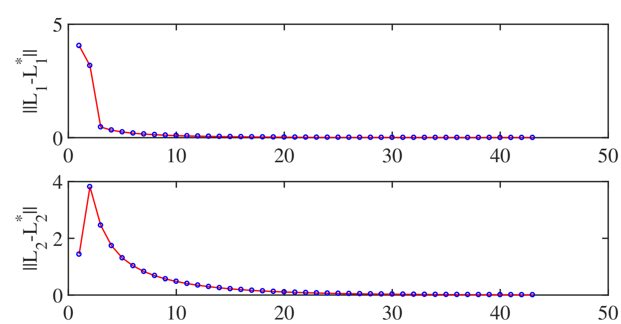

Fig. 4. The evolution of L_1, L_2 under Algorithm 2.

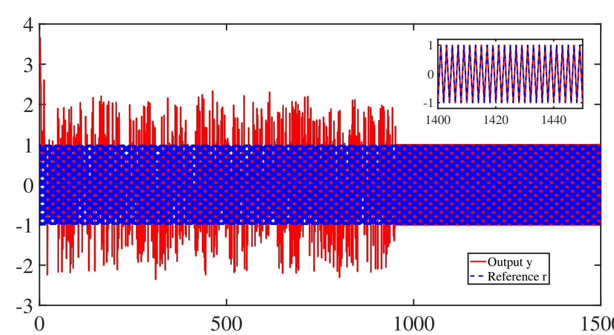

Fig. 5. Output y and reference r under Algorithm 2.

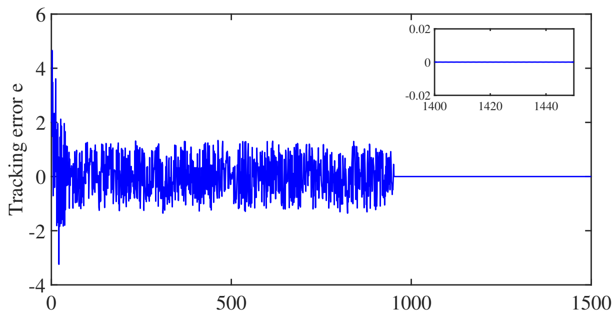

Fig. 6. The tracking error e under Algorithm 2.

5.2. Example 2

Consider the discrete-time linear multi-player systems, where $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \end{bmatrix}$, $B_1 = [0.2, 0, 0.3]^T$, $B_2 = [0.3, 0, 0.2]^T$, $C = [1, 0, 0]^T$. The reference trajectory dynamic $F = -1$. $\alpha = 0.5$, $Q = 10000$, $R_1 = 0.01$, and $R_2 = 0.05$. The simulation results corresponding to Algorithm 2 are presented in Figs. 7-9.

From Fig. 7, it can be seen that L_1 and L_2 can quickly converge to the optimal value L_1^* and L_2^*. From Figs. 8 and 9, it can be seen that the designed scheme can achieve good tracking performance.

To sum up, our designed scheme can achieve good

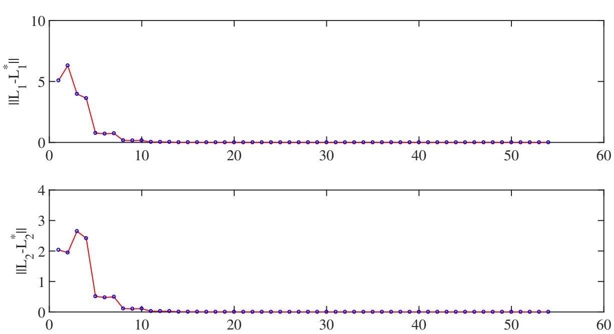

Fig. 7. The evolution of L_1, L_2 under Algorithm 2.

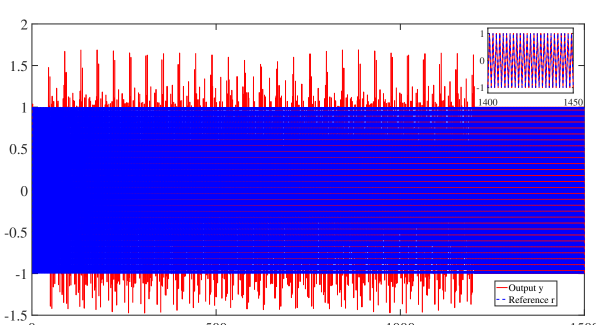

Fig. 8. Output y and reference r under Algorithm 2.

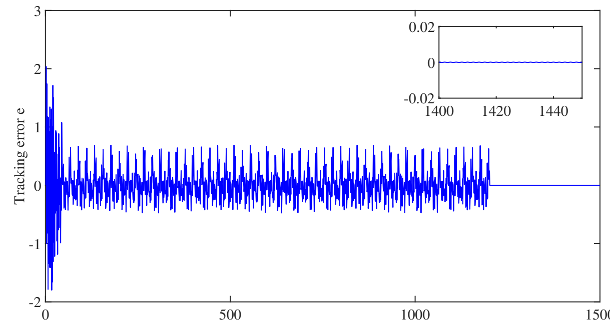

Fig. 9. The tracking error e under Algorithm 2.

tracking performance. The tracking performance is also related to parameters α, Q, R_1 and R_2. When choosing a large Q or/and small R_1, R_2, a better tracking performance can be achieved; as the discount factor α increases, the learning rate will increase, that is, L_1 and L_2 in Figs. 1 and 3 can converge to the optimal value L_1^* and L_2^* faster. In addition, too large a discount factor α may make the tracking performance worse. Therefore, both learning rate and tracking performance should be considered when choosing the discount factor.

6. CONCLUSION

In this paper, the tracking control for a class of discrete-time multi-player linear systems with unknown dynamics is investigated from the perspective of FCG. In order to obtain the solution to the tracking problem, an augmented FCGTARE is derived. An online reinforcement Q-learning algorithm is proposed to solve the augmented FCGTARE without requiring the system dynamics. We infer the impact of the relevant parameters on the tracking performance and analyze the convergence of the proposed online reinforcement Q-learning algorithm. Lastly, a discretized F-16 dynamic system model is simulated to verify the validity of our proposed reinforcement Q-learning algorithm and the influence of relevant parameters on the tracking performance. In future work, we will extend the results of this paper to more complex systems, such as networked control systems, multi-agent systems, etc.

CONFLICTS OF INTERESTS

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

REFERENCES

[1] F. L. Lewis, D. L. Vrabie, and V. L. Syrmos, *Optimal Control*, 3rd ed., John Wiley and Sons, 2015.

[2] C. Deng, C. Wen, W. Wang, X. Li, and D. Yue, "Distributed adaptive tracking control for high-order nonlinear multi-agent systems over event-triggered communication," *IEEE Transactions on Automatic Control*, vol. 68, no. 2, pp. 1176-1183, 2023.

[3] R. Postoyan, N. Van de Wouw, D. Nesic, and W. P. M .H Heemels, "Tracking control for nonlinear networked control systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1539-1554, 2014.

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

[5] H. Chen, Y. C. Fang, and N. Sun, "An adaptive tracking control method with swing suppression for 4-DOF tower crane systems," *Mechanical Systems and Signal Processing*, vol. 123, pp. 426-442, 2019.

[6] Z. P. Jiang and H. Nijmeijer, "Tracking control of mobile robots: A case study in backstepping," *Automatica*, vol. 33, no. 7, pp. 1393-1399, 1997.

[7] Q. Gao, X. T. Wei, D. H. Li, Y. H. Ji, and C. Jia, "Tracking control for a quadrotor via dynamic surface control and adaptive dynamic programming," *International Journal of Control, Automation, and Systems*, vol. 20, pp. 349-363, 2022.

[8] H. Chen, Y. C. Fang, and N. Sun, "Optimal trajectory planning and tracking control method for overhead cranes," *IET Control Theory & Applications*, vol. 10, no. 6, pp. 692-699, 2016.

[9] C. Deng, C. Wen, J. Huang, X. M. Zhang, and Y. Zou, "Distributed observer-based cooperative control approach for uncertain nonlinear MASs under event-triggered communication," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2669-2676, 2022.

[10] D. G. Xu, Q. L. Wang, and Y. Li, "Optimal guaranteed cost tracking of uncertain nonlinear systems using adaptive dynamic programming with concurrent learning," *International Journal of Control, Automation, and Systems*, vol. 18, no. 5, pp. 1116-1127, 2020.

[11] B. Zhao and Y. C. Li, "Model-free adaptive dynamic programming based near-optimal decentralized tracking control of reconfigurable manipulators," *International Journal of Control, Automation, and Systems*, vol. 16, no. 2, pp. 478-490, 2018.

[12] A. Mannava, S. N. Balakrishnan, L. Tang, and R. G. Landers, "Optimal tracking control of motion systems," *IEEE Transactions on Control Systems Technology*, vol. 20, no. 6, pp. 1548-1558, 2012.

[13] J. Zhao, "Neural network-based optimal tracking control of continuous-time uncertai nonlinear system via reinforcement learning," *Neural Processing Letters*, vol. 51, no. 3, pp. 2513-2530, 2020.

[14] Q. Wei and D. Liu, "Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 4, pp. 1020-1036, 2014.

[15] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M. B. Naghibi-Sistani, "Reinforcement q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167-1175, 2014.

[16] Y. Wen, H. Zhang, H. Su, and H. Ren, "Optimal tracking control for non-zero-sum games of linear discrete-time systems via off-policy reinforcement learning," *Optimal Control Applications and Methods*, vol. 41, no. 4, pp. 1233-1250, 2020.

[17] Q. Zhang, D. Zhao, and Y. Zhu, "Data-driven adaptive dynamic programming for continuous-time fully cooperative games with partially constrained inputs," *Neurocomputing*, vol. 238, pp. 377-386, 2017.

[18] K. Zhang, S. L. Ge, and Y. L. Ge, "Adaptive dynamic programming for minimal energy control with guaranteed convergence rate of linear systems," *International Journal of Control, Automation, and Systems*, vol. 17, no. 2, pp. 3140-3148, 2019.

[19] W. N. Gao, Y. Y. Liu, A. Odekunle, Y. J. Yu, and P. L. Lu, "Adaptive dynamic programming and cooperative output regulation of discrete-time multi-agent systems," *International Journal of Control, Automation, and Systems*, vol. 16, no. 5, pp. 2273-2281, 2018.

[20] L. An and G. Yang, "Optimal transmission power scheduling of networked control systems via fuzzy adaptive dynamic programming," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 6, pp. 1629-1639, 2021.

[21] J. Zhao and P. Vishal, "Neural network-based optimal tracking control for partially unknown discrete-time nonlinear systems using reinforcement learning," *IET Control Theory and Applications*, vol. 15, no. 2, pp. 260-271, 2021.

[22] Y. Yang, K. G. Vamvoudakis, H. Modares, Y. Yin, and D. C. Wunsch, "Hamiltonian-driven hybrid adaptive dynamic programming," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 10, pp. 6423-6434, 2021.

[23] A. AI-tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473-481, 2007.

[24] T. Y. Chun, J. B. Park, and Y. H. Choi, "Reinforcement Q-learning based on multirate generalized policy iteration and its application to a 2-DOF helicopter," *International Journal of Control, Automation, and Systems*, vol. 16, pp. 377-386, 2018.

[25] A. Odekunle, W. N. Gao, M. Davari, and Z. P. Jiang, "Reinforcement learning and non-zero-sum game output regulaton for multi-player linear uncertain systems," *Automatica*, vol. 112, 108672, 2020.

[26] L. An and G. Yang, "Opacity enforcement for confidential robust control in linear cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 1234-1241, 2020.

[27] L. An and G. Yang, "Data-driven coordinated attack policy design based on adaptive L2-gain optimal theory," *IEEE Transactions on Automatic Control*, vol. 63, no. 6, pp. 1850-1857, 2018.

[28] B. Luo, D. Liu, T. Huang, and D. Wang, "Model-free optimal tracking control via critic-only q-learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 2134-2144, 2016.

[29] M. Lin, B. Zhao, and D. Liu, "Policy gradient adaptive critic designs for model-free optimal tracking control with experience replay," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 6, pp. 3692-3703, 2022.

[30] J. Lu, Q. Wei, and F. Y. Wang, "Parallel control for optimal tracking via adaptive dynamic programming," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 6, pp. 1662-1674, 2020.

[31] C. Li, J. Ding, F. L. Lewis, and T. Chai, "A novel adaptive dynamic programming based on tracking error for nonlinear discrete-time systems," *Automatica*, vol. 129, 109687, 2021.

[32] J. Li, Z. Xiao, P. Li, and J. Cao, "Robust optimal tracking control for multiplayer systems by off-policy q-learning approach," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 1, pp. 87-106, 2021.

[33] Y. Lv, X. Ren, and J. Na, "Adaptive optimal tracking controls of unknown multi-input systems based on nonzero-sum game theory," *Journal of the Franklin Institute*, vol. 356, no. 15, pp. 8255-8277, 2019.

[34] J. Zhao, "Neural networks-based optimal tracking control for nonzero-sum games of multi-player continuous-time nonlinear systems via reinforcement learning," *Neurocomputing*, vol. 412, pp. 167-176, 2020.

[35] H. Jiang, H. Zhang, X. Xie, and J. Han, "Neural-network-based learning algorithms for cooperative games of discrete-time multi-player systems with control constraints via adaptive dynamic programming," *Neurocomputing*, vol. 344, pp. 13-19, 2019.

[36] J. Zhao, "Data-driven adaptive dynamic programming for optimal control of continuous-time multicontroller systems with unknown dynamics," *IEEE Access*, vol. 10, pp. 41503-41511, 2022.

[37] Y. Yang, Y. Wan, J. Zhu, and F. L. Lewis, "H∞ tracking control for linear discrete-time systems: Model-free q-learning designs," *IEEE Control Systems Letters*, vol. 5, no. 1, pp. 175-180, 2021.

[38] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free q-learning designs for linear discrete-time zero-sum games with application to H∞ control," *Automatica*, vol. 43, no. 3, pp. 473-481, 2007.

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

Jin-Gang Zhao received his B.E. degree in automation from Qingdao University of Technology, Qingdao, China, in 2013, an M.Sc. degree in pattern recognition and intelligence system from Beijing Information Science and Technology University, Beijing, China, in 2016, and a Ph.D degree in control science and engineering from Beijing Institute of Technology, Beijing, China, in 2020. From 2018 to 2019, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA. He is currently a lecturer with School of Machinery and Automation, Weifang University. His research interests include optimal control, reinforcement learning, adaptive dynamic programming, and hybrid system.