Off-Policy Q-Learning for Anti-Interference Control of Multi-Player Systems [⋆]

Jinna Li [*] **Zhenfei Xiao** [*] **Tianyou Chai** [**] **Frank. L. Lewis** [***] **Sarangapani Jagannathan** [****]

[*] *School of Information and Control Engineering, Liaoning Shihua University, Fushun 113001, China, (e-mail: lijinna_721@126.com; Xiaozhenfeiwm@outlook.com).*
[**] *State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China (e-mail: tychai@mail.neu.edu.cn).*
[***] *UTA Research Institute, the University of Texas at Arlington, Arlington, TX 76118 USA (e-mail: lewis@uta.edu).*
[****] *Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail: sarangap@mst.edu).*

Abstract: This paper develops a novel off-policy game Q-learning algorithm to solve the anti-interference control problem for discrete-time linear multi-player systems using only data without requiring system matrices to be known. The primary contribution of this paper lies in that the Q-learning strategy employed in the proposed algorithm is implemented in an off-policy policy iteration approach other than on-policy learning due to the well-known advantages of off-policy Q-learning over on-policy Q-learning. All of the players work hard together for the goal of minimizing their common performance index meanwhile defeating the disturbance that tries to maximize the specific performance index, and finally they reach the Nash equilibrium of the game resulting in satisfying disturbance attenuation condition. In order to find the solution to the Nash equilibrium, the anti-interference control problem is first transformed into an optimal control problem. Then an off-policy Q-learning algorithm is proposed in the framework of typical adaptive dynamic programming (ADP) and game architecture, such that control policies of all players can be learned using only measured data. Comparative simulation results are provided to verify the effectiveness of the proposed method.

Keywords: H_∞ control, off-policy Q-learning, game theory, Nash equilibrium

1. INTRODUCTION

Since the external disturbance is inevitable for systems, then anti-interference has to be considered when controlling systems. However, the truth of more complex and large-scale systems with multiple subsystems and multiple controllers in practical engineering applications makes anti-interference control of multi-player systems valuable and more complicated. Increasing attention of researchers to anti-interference control for multi-player or multi-agent systems has been paid Jiao et al. (2016); Lv and Ren (2018); Wang et al. (2018b,a). As a traditional and practical control method to eliminate interference, H_∞ control can not only keep the system stable but also guarantee the interference attenuation bounded Van de Schaft (1992); Isidori (1994). With the game theory, the H_∞ anti-interference control problem can be transformed into a zero-sum game problem Al-Tamimi et al. (2007); Kiumarsi et al. (2017). Therefore, H_∞ control method is used to solve the anti-interference problem in this paper.

Most of researchers are concerned about model-based H_∞ controller design using zero-sum game Jiang and Jiang (2012); Vamvoudakis and Lewis (2010). The limitation of these methods is that the dynamics of systems should be accurately known a priori, so they cannot directly work for systems with inaccurate or even completely unknown models. Adaptive dynamic programming (ADP) combined with reinforcement learning (RL), which can deal with controller design or decision-making problems in an uncertain or unknown environment, could be an effective method to solve H_∞ control for systems with unknown system dynamics. Al-Tamimi et al. (2007) focused on the design of the model-free Q-learning algorithm for the zero-sum game of linear discrete-time (DT) systems without knowing the system dynamics. Kiumarsi et al. (2017) presented a model-free solution to the H_∞ control of linear DT systems through off-policy RL method. Modares et al. (2015) used the off-policy RL approach to solve the H_∞ tracking control problem for nonlinear continuous-time (CT) systems. Luo et al. (2015) presented a novel off-

[⋆] Research supported by the National Natural Science Foundation of China under Grants 61673280, 61673100, the Open Project of Key Field Alliance of Liaoning Province under Grant 2019-KF-03-06 and the Project of Liaoning Shihua University under Grant 2018XJJ-005.

policy learning method to learn the optimal controller of H_∞ control problem for nonlinear CT distributed parameter systems. In the other hand, the existing reports on multi-player games Li et al. (2017); Liu et al. (2014); Vamvoudakis and Lewis (2011) usually ignore the negative effects caused by disturbances on performance of systems. This is our motivation to design a novel data-driven anti-interference algorithm for multi-player systems.

Most relevant results are H_∞ control for multi-agent and multi-player systems Lv and Ren (2018); Jiao et al. (2016). In Jiao et al. (2016) agents have their individual dynamics, and the anti-interference problem has been investigated for continuous-time multi-player systems where all players share the common system state in Lv and Ren (2018). Like Lv and Ren (2018), the data-driven H_∞ controller design will be taken into account for multi-player systems with unknown system dynamics in this paper, while the difference of nature of discrete-time sampling from continues-time processes makes it more complicated to solve H_∞ control problem from the discrete-time system perspective, and multiple players and completely unknown dynamics of players increase this difficulty. Moreover, given the advantages of off-policy learning over on-policy learning Li et al. (2019a), we aim at developing an off-policy game Q-learning algorithm to solve the anti-interference control problem for discrete-time linear multi-player systems using only measured data.

In this paper, we develop a novel off-policy game Q-learning algorithm based on game theory and ADP method to solve a multi-player zero-sum game using measured data, such that the negative effect caused by the external interference can be eliminated and meanwhile the stability of multi-player systems can be ensured.

This paper is organized as follows. Section 2 devotes to the problem formulation of the multi-player anti-interference control with linear DT systems and transforms it into a zero-sum game. In Section 3, we derive the theoretical solution to the multi-player zero-sum game and develop a data-driven off-policy game Q-learning algorithm. Section 4 gives a simulation example to demonstrate the effectiveness of the proposed algorithm. Conclusions are stated in Section 5.

The following notations will be used in this article: $\mathbb{R}^p$ denotes the p dimensional Euclidean space. $\mathbb{R}^{p\times q}$ is the set of all real p by q matrices. Positive definite matrix is assumed that in the case that Q is a square matrix of order n and x is any non-zero vector, $x^TQx > 0$. If $x^TQx \geq 0$, it is a semi-positive definite matrix. $\|\cdot\|$ donates the vector norm. The superscript T is used for the transpose. $\otimes$ stands for the Kronecker product. $vec(L)$ is used to turn any matrix L into a single column vector.

2. PROBLEM STATEMENT

In this section, we formulate the anti-interference control problem for linear DT systems with multiple players first. Then it is transformed into a zero-sum game problem, where all players have a common goal of minimizing the performance index, while the disturbance is to make the performance worse.

Consider the following linear DT multi-player system subject to exogenous disturbance

$$x_{k+1} = Ax_k + B\sum_{i=1}^{n} u_{ik} + Ed_k \tag{1}$$

where $x_k = x(k) \in \mathbb{R}^p$ is the system state with initial state x_0, $u_{ik} = u_i(k) \in \mathbb{R}^m$ is the control input and $d_k = d(x_k) \in \mathbb{R}^q$ is the external disturbance input. $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{p\times m}$, $E \in \mathbb{R}^{p\times q}$ and k is the sampling time instant.

Definition 1. (Kiumarsi et al. (2017)) System (1) has L_2-gain less than or equal to γ if

$$\sum_{k=0}^{\infty}(x_k^TQx_k + \sum_{i=1}^{n} u_{ik}^TR_iu_{ik}) \leq \gamma^2 \sum_{k=0}^{\infty} \|d_k\|^2 \tag{2}$$

for all $d_k \in L_2[0,\infty)$. $Q \geq 0$, $R_i > 0$ and $\gamma \geq 0$ is a prescribed constant disturbance attenuation level.

For the H_∞ control, it is desired to find the feedback control policies u_i such that the system (1) with $d_k = 0$ is asymptotically stable and the disturbance attenuation condition (2) can be satisfied. By recalling the max-min problem in Van de Schaft (1992), the anti-interference control problem can be written as

$$\begin{aligned} &J(x_0, U, d_k) \\ &= \min_U \max_{d_k} \sum_{k=0}^{\infty}(x_k^TQx_k + \sum_{i=1}^{n} u_{ik}^TR_iu_{ik} - \gamma^2 \|d_k\|^2) \end{aligned} \tag{3}$$

where $U = \{u_{1k}, u_{2k}, \ldots, u_{nk}\}$, which means the set U is composed of n players with each of them is a controller. As one can know from (3), the objective of these players U is to fight with the disturbance for minimizing the performance in (3), while the disturbance d_k could also be viewed as a player that tries to maximize (3). This is a typical zero-sum game problem and it indicates the Nash equilibrium Vamvoudakis et al. (2012) condition holds, that is

$$J(x_0, U^*, d_k) \leq J(x_0, U^*, d_k^*) \leq J(x_0, U, d_k^*)$$

where $U^* = \{u_{1k}^*, u_{2k}^*, \ldots, u_{nk}^*\}$.

The saddle point solution exists for the zero-sum game shown in (3) if and only if there is a value function $V(x_k)$ satisfying the following Hamilton-Jacobi-Bellman (HJB) equation Vamvoudakis et al. (2017).

$$\begin{aligned} V^*(x_k) &= \min_U \max_{d_k} \sum_{l=k}^{\infty}(x_l^TQx_l + \sum_{i=1}^{n} u_{il}^TR_iu_{il} - \gamma^2 \|d_l\|^2) \\ &= \min_U \max_{d_k}(x_k^TQx_k + \sum_{i=1}^{n} u_{ik}^TR_iu_{ik} - \gamma^2 \|d_k\|^2 + V^*(x_{k+1})) \\ &= x_k^TQx_k + \sum_{i=1}^{n} u_{ik}^{*T}R_iu_{ik}^* - \gamma^2 \|d_k^*\|^2 \\ &+ V^*(Ax_k + B\sum_{i=1}^{n} u_{ik}^* + Ed_k^*) \end{aligned} \tag{4}$$

where $V^*(x_k)$ is viewed as the optimal value function.

The arguments provided above have demonstrated that anti-interference control can be fixed out by solving the HJB equation (4). Now, we are in the position to solve HJB (4) for finding the H_∞ controller.

Similar to Al-Tamimi et al. (2007); Li et al. (2017), the optimal Q-function referring to (3) can be defined as

$$\begin{aligned} &Q^*(x_k, U, d_k) \\ &= x_k^T Q x_k + \sum_{i=1}^{n} u_{ik}^T R_i u_{ik} - \gamma^2 \|d_k\|^2 + V^*(x_{k+1}) \end{aligned} \tag{5}$$

Thus, the following relation holds

$$V^*(x_k) = \min_{U} \max_{d_k} Q^*(x_k, U, d_k) = Q^*(x_k, U^*, d_k^*) \tag{6}$$

Lemma 2. Under the assumption that there are admissible control policies $u_i = -K_i x_k$ and $d_k = -K x_k$, the quadratic forms of the value function and Q-function can be expressed as

$$V(x_k) = x_k^T P x_k \tag{7}$$

and

$$Q(x_k, U, d_k) = z_k^T H z_k \tag{8}$$

where P and H are positive definite matrices. And

$$\begin{aligned} z_k &= \begin{bmatrix} x_k^T & u_{1k}^T & u_{2k}^T & \dots & u_{nk}^T & d_k^T \end{bmatrix}^T \\ M &= \begin{bmatrix} I & -K_1^T & -K_2^T & \dots & -K_n^T & -K^T \end{bmatrix}^T \\ P &= M^T H M \end{aligned} \tag{9}$$

$$\begin{aligned} H &= \begin{bmatrix} H_{xx} & H_{xu_1} & H_{xu_2} & \dots & H_{xu_n} & H_{xd} \\ H_{xu_1}^T & H_{u_1u_1} & H_{u_1u_2} & \dots & H_{u_1u_n} & H_{u_1d} \\ H_{xu_2}^T & H_{u_1u_2}^T & H_{u_2u_2} & \dots & H_{u_2u_n} & H_{u_2d} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ H_{xu_n}^T & H_{u_1u_n}^T & H_{u_2u_n}^T & \dots & H_{u_nu_n} & H_{u_nd} \\ H_{xd}^T & H_{u_1d}^T & H_{u_2d}^T & \dots & H_{u_nd}^T & H_{dd} \end{bmatrix} \\ &= \begin{bmatrix} A^TPA+Q & \dots & A^TPB & A^TPE \\ (A^TPB)^T & \dots & B^TPB & B^TPE \\ (A^TPB)^T & \dots & B^TPB & B^TPE \\ \vdots & \dots & \vdots & \vdots \\ (A^TPB)^T & \dots & B^TPB+R_n & B^TPE \\ (A^TPE)^T & \dots & (B^TPE)^T & -\gamma^2 I + E^TPE \end{bmatrix} \end{aligned} \tag{10}$$

3. SOLVING MULTI-PLAYER ZERO-SUM GAME

In this section, the theoretical solution to the zero-sum game for multi-player systems is first obtained. Then, an off-policy game Q-learning algorithm is developed to solve this problem.

3.1 Theoretical Solution

By Lemma 1, we can refer to the HJI equation (4) to get the Q-function based Bellman equation.

$$z_k^T H z_k = x_k^T Q x_k + \sum_{i=1}^{n} u_{ik}^T R_i u_{ik} - \gamma^2 \|d_k\|^2 + z_{k+1}^T H z_{k+1} \tag{11}$$

The optimal control policy u_i^* of each player i and the worst-case disturbance d_k^* should satisfy $\frac{\partial Q^*(x_k,U,d_k)}{\partial u_i} = 0$ and $\frac{\partial Q^*(x_k,U,d_k)}{\partial d_k} = 0$. Therefore, one has

$$u_i^*(k) = -K_i^* x_k,\ d_k^* = -K^* x_k \tag{12}$$

where

$$\begin{aligned} K_i^* = H_{u_iu_i}^{-1} \big[& H_{xu_i}^T - (H_{u_iu_1}K_1 + \dots + H_{u_iu_{i-1}}K_{i-1} \\ & + H_{u_iu_{i+1}}K_{i+1} + \dots + H_{u_iu_n}K_n + H_{u_id}K) \big] \end{aligned} \tag{13}$$

$$K^* = H_{dd}^{-1} \left[H_{xd}^T - (H_{du_1}K_1 + H_{du_2}K_2 + \dots + H_{du_n}K_n \right] \tag{14}$$

Substituting K_i^* in (13) and K^* in (14) into (11) yields the optimal Q-function based game Riccati equation.

$$\begin{aligned} &(z_k)^T H^* z_k \\ &= x_k^T Q x_k + \sum_{i=1}^{n} (u_i^*)^T R_i u_i^* - \gamma^2 \|d_k^*\|^2 + (z_{k+1})^T H^* z_{k+1} \end{aligned} \tag{15}$$

It is worth pointing out that Al-Tamimi et al. (2007) and Li et al. (2019b) have proved that the following $K_i^* (i = 1, 2, ...n)$ and K can keep system (1) stable when $d_k = 0$ and achieve Nash equilibrium.

$$\begin{aligned} K_i^* = (H_{u_iu_i}^*)^{-1} \Big[& (H_{xu_i}^*)^T - (H_{u_iu_1}^* K_1^* + \dots + H_{u_iu_{i-1}}^* \\ & \times K_{i-1}^* + H_{u_iu_{i+1}}^* K_{i+1}^* + \dots + H_{u_iu_n}^* K_n^* + H_{u_id}^* K^* \Big] \end{aligned} \tag{16}$$

$$\begin{aligned} K^* = (H_{dd}^*)^{-1} \big[& (H_{xd}^*)^T - (H_{du_1}^* K_1^* + H_{du_2}^* K_2^* + \dots \\ & + H_{du_n}^* K_n^* \big] \end{aligned} \tag{17}$$

Note that solving (16) and (17), that is, solving the zero-sum game problem defined by (5), is to find the optimal control policies satisfying the disturbance attenuation condition (2).

Remark 3. Since the control policy gains K_i and K are coupled with each other, it is difficult to solve them. In addition, accurately identification of the system models can not come ture in real industry. Therefore, a data-driven off-policy game Q-learning algorithm is going to be presented to overcome these difficulties, such that the control laws $u_i^*(k)$ and disturbance policy d_k^* can be learned.

3.2 Off-Policy Q-Learning Algorithm

Rewrite (11) as the following form of iteration Bellman equation.

$$\begin{aligned} &z_k^T H^{j+1} z_k \\ &= x_k^T Q x_k + \sum_{i=1}^{n} (u_{ik}^j)^T R_i u_{ik}^j - \gamma^2 \left\| d_k^j \right\|^2 + z_{k+1}^T H^{j+1} z_{k+1} \end{aligned} \tag{18}$$

Then, one has

$$\begin{aligned} &(M^j)^T H^{j+1} M^j \\ &= (M^j)^T \Lambda M^j + \left(A - B \sum_{i=1}^{n} K_i^j - E K^j \right)^T \\ &\times (M^j)^T H^{j+1} M^j \left(A - B \sum_{i=1}^{n} K_i^j - E K^j \right) \end{aligned} \tag{19}$$

where

$$\Lambda = diag(Q, R_1, R_2, \dots, R_n, -\gamma^2 I)$$

Introducing auxiliary variables $u_i^j = -K_i^j x_k$ and $d_k^j = -K^j x_k$ to system (1) yields

$$x_{k+1} = A_c x_k + B\sum_{i=1}^{n}(u_{ik} - u_{ik}^j) + E(d_k - d_k^j) \tag{20}$$

where $A_c = A - B\sum_{i=1}^{n} K_i^j - EK^j$, u_i and d_k are called the behavior control policies and the behavior disturbance policy, while u_{ik}^j and d_k^j are called the target control policies and the target disturbance policy. Along the system trajectory (20), one has

$$\begin{aligned}
&Q^{j+1}(x_k, U, d_k) - x_k^T A_c^T (M^j)^T H^{j+1} M^j A_c x_k \\
&= x_k^T (M^j)^T H^{j+1} M^j x_k \\
&- \left(x_{k+1} - B\sum_{i=1}^{n}(u_{ik} - u_{ik}^j) - E(d_k - d_k^j)\right)^T (M^j)^T H^{j+1} \\
&\times M^j \left(x_{k+1} - B\sum_{i=1}^{n}(u_{ik} - u_{ik}^j) - E(d_k - d_k^j)\right) \\
&= x_k^T (M^j)^T \Lambda M^j x_k
\end{aligned} \tag{21}$$

Since P^{j+1} and H^{j+1} have the relationship shown in (9) and (10), then the following holds.

$$\begin{aligned}
&x_k^T (M^j)^T H^{j+1} M^j x_k - x_{k+1}^T (M^j)^T H^{j+1} M^j x_{k+1} \\
&+ 2\left(Ax_k + B\sum_{i=1}^{n} u_{ik} + Ed_k\right)^T P^{j+1} B\sum_{i=1}^{n}(u_{ik} - u_{ik}^j) \\
&+ 2\left(Ax_k + B\sum_{i=1}^{n} u_{ik} + Ed_k\right)^T P^{j+1} E(d_k - d_k^j) \\
&- \sum_{i=1}^{n}(u_{ik} - u_{ik}^j)^T B^T P^{j+1} B\sum_{i=1}^{n}(u_{ik} - u_{ik}^j) \\
&- 2\sum_{i=1}^{n}(u_{ik} - u_{ik}^j)^T B^T P^{j+1} E(d_k - d_k^j) \\
&- (d_k - d_k^j)^T E^T P^{j+1} E(d_k - d_k^j) \\
&= x_k^T (M^j)^T \Lambda M^j x_k
\end{aligned} \tag{22}$$

Further

$$\begin{aligned}
&x_k^T (M^j)^T H^{j+1} M^j x_k - x_{k+1}^T (M^j)^T H^{j+1} M^j x_{k+1} \\
&+ 2x_k^T \left[H_{xu_1}^{j+1} \quad H_{xu_2}^{j+1} \dots H_{xu_n}^{j+1}\right] \sum_{i=1}^{n}(u_{ik} + K_i^j x_k) \\
&+ 2\sum_{i=1}^{n} u_{ik}^T G^{j+1} \sum_{i=1}^{n}(u_{ik} + K_i^j x_k) \\
&+ 2d_k^T (H_{u_i d}^{j+1})^T \sum_{i=1}^{n}(u_{ik} + K_i^j x_k) \\
&+ 2x_k^T (H_{xd}^{j+1})(d_k + K^j x_k) + 2\sum_{i=1}^{n} u_{ik}^T H_{u_i d}^{j+1}(d_k + K^j x_k) \\
&- \sum_{i=1}^{n}\left(u_{ik} + K_i^j x_k\right)^T G^{j+1} \sum_{i=1}^{n}(u_{ik} + K_i^j x_k) \\
&- 2\sum_{i=1}^{n}\left(u_{ik} + K_i^j x_k\right)^T H_{u_i d}^{j+1}(d_k + K^j x_k) \\
&- (d_k + K^j x_k)^T (H_{dd}^{j+1} + \gamma^2 I)(d_k + K^j x_k) \\
&= x_k^T (M^j)^T \Lambda M^j x_k
\end{aligned} \tag{23}$$

where

$$G^{j+1} = \begin{bmatrix}
H_{u_1u_1}^{j+1} - R_1 & H_{u_1u_2}^{j+1} & \dots & H_{u_1u_n}^{j+1} \\
(H_{u_1u_2}^{j+1})^T & H_{u_2u_2}^{j+1} - R_2 & \dots & H_{u_2u_n}^{j+1} \\
(H_{u_1u_3}^{j+1})^T & (H_{u_2u_3}^{j+1})^T & \dots & H_{u_3u_n}^{j+1} \\
\vdots & \vdots & \dots & \vdots \\
(H_{u_1u_n}^{j+1})^T & (H_{u_2u_n}^{j+1})^T & \dots & H_{u_nu_n}^{j+1} - R_n
\end{bmatrix}$$

Rewritten (23) as the following form

$$\hat{\theta}^j(k)\hat{L}^{j+1} = \hat{\rho}_k \tag{24}$$

where

$$\hat{\rho}_k = x_k^T Q x_k + \sum_{i=1}^{n} u_{ik}^T R_i u_{ik} - \gamma^2 d_k^T d_k$$

$$\hat{L}^{j+1} = \left[(vec(\hat{L}_{rz}^{j+1}))^T, \dots, (vec(\hat{L}_{n+1,n+1}^{j+1}))^T\right]^T$$

$$\hat{\theta}^j(k) = \left[\hat{\theta}_{rz}^j, \dots, \hat{\theta}_{n+1,n+1}^j\right]$$

with $r = 0, 1, 2, \dots, n+1$, $z = r, r+1, r+2, \dots, n+1$. Besides,

$$\begin{aligned}
&\hat{\theta}_{00}^j = x_k^T \otimes x_k^T - x_{k+1}^T \otimes x_{k+1}^T \\
&\hat{L}_{00}^{j+1} = H_{xx}^{j+1} \\
&\hat{\theta}_{ss}^j = -(K_s^j x_{k+1})^T \otimes (K_s^j x_{k+1})^T + u_s^T \otimes u_s^T \\
&\hat{L}_{ss}^{j+1} = H_{u_su_s}^{j+1} \\
&\hat{\theta}_{s+1,s+1}^j = -(K^j x_{k+1})^T \otimes (K^j x_{k+1})^T + d_k^T \otimes d_k^T \\
&\hat{L}_{s+1,s+1}^{j+1} = H_{dd}^{j+1} \\
&\hat{\theta}_{0s}^j = 2x_{k+1}^T \otimes (K_s^j x_{k+1})^T + 2x_k^T \otimes u_s^T \\
&\hat{L}_{0s}^{j+1} = H_{xu_s}^{j+1} \\
&\hat{\theta}_{0s+1}^j = 2x_{k+1}^T \otimes (K^j x_{k+1})^T + 2x_k^T \otimes d_k^T \\
&\hat{L}_{0s+1}^{j+1} = H_{xd}^{j+1} \\
&\hat{\theta}_{st}^j = -2(K_s^j x_{k+1})^T \otimes (K_t^j x_{k+1})^T + 2u_s^T \otimes u_t^T \\
&\hat{L}_{st}^{j+1} = H_{u_su_t}^{j+1} \\
&\hat{\theta}_{s,s+1}^j = -2(K_s^j x_{k+1})^T \otimes (K^j x_{k+1})^T + 2u_s^T \otimes d_k^T \\
&\hat{L}_{s,s+1}^{j+1} = H_{u_sd}^{j+1}
\end{aligned}$$

with $s \neq t$ and $s, t = 1, 2, \dots, n$.

Based on the above part, $K_1^{j+1}, K_2^{j+1}, \dots, K_n^{j+1}$ and K^{j+1} can be expressed as the form of $\hat{L}^{j+1}$

$$\begin{aligned}
K_i^{j+1} = (\hat{L}_{ii}^{j+1})^{-1}\Big((\hat{L}_{0i}^{j+1})^T - \Big[(\hat{L}_{i1}^{j+1})^T K_1^j + \dots \\
+(\hat{L}_{(i,i-1)}^{j+1})^T K_{i-1}^j + \hat{L}_{(i,i+1)}^{j+1})^T K_{i+1}^j \\
+ \dots + (\hat{L}_{in}^{j+1})^T K_n^j + (\hat{L}_{i,n+1}^{j+1})^T K^j\Big]\Big)
\end{aligned} \tag{25}$$

$$\begin{aligned}
K^{j+1} =(\hat{L}_{n+1,n+1}^{j+1})^{-1}\Big((\hat{L}_{0,n+1}^{j+1})^T - \Big[(\hat{L}_{n+1,1}^{j+1})^T K_1^j \\
+(\hat{L}_{n+1,2}^{j+1})^T K_2^j + \dots + (\hat{L}_{n+1,n}^{j+1})^T K_n^j\Big]\Big)
\end{aligned} \tag{26}$$

4. SIMULATION RESULTS

In this section, we use a simulation example to demonstrate the effectiveness of the proposed algorithm. Consider the following linear DT system with four players and disturbance input:

$$x_{k+1} = Ax_k + B\sum_{i=1}^{4} u_i + Ed_k \tag{27}$$

Algorithm 1 Off-Policy Game Q-Learning for the Zero-Sum Game

```
1 Data collection: Collect system data $x_k$ and store them in (24) by using (20);
2 Initialize the admissible control policies of multiple players $K_1^0, K_2^0, K_3^0, \dots, K_n^0$ and disturbance policy gain(the $n+1$ player) $K^0$. Set the iteration index $j = 0$ and $i = 1$ represents player $i(i = 1, 2, \dots, n+1)$;
3 Performing the off-policy game Q-learning: use the recursive least-square method to solve the $\hat{L}^{j+1}$ in (24), and then $K_i^{j+1}$ and $K^{j+1}$ can be updated by (25) and
4 If $i < n+1$, then $i = i+1$ and go back to Step 3. Otherwise $j = j+1$, $i = 1$ and go to Step 5;
5 Stop when $\left\|K_i^j - K_i^{j-1}\right\| \le \varepsilon$ $(i = 1, 2, \dots, n+1)$, the optimal control policy is obtained. Otherwise, $i = 1$, and go back to Step 3.
```

where

$$A = \begin{bmatrix} 0.906488 & 0.0816012 & -0.0005 \\ 0.074349 & 0.90121 & -0.000708383 \\ 0 & 0 & 0.132655 \end{bmatrix}$$

$$B = \begin{bmatrix} -0.00563451 \\ -0.08962 \\ 0.356478 \end{bmatrix}, \ E = \begin{bmatrix} 0.0123956 \\ 0.068 \\ -0.05673 \end{bmatrix}$$

Choose $Q = diag(5, 5, 5)$, $R_1 = 1$,$R_2 = 2$,$R_3 = 3$,$R_4 = 4$, and set the disturbance attenuation to be $\gamma = 1$. Rewrite (15) as

$$H^* = \Lambda + \Pi^T H^* \Pi \tag{28}$$

where

$$\Pi = \begin{bmatrix} A & B & B & B & B & E \\ -K_1A & -K_1B & -K_1B & -K_1B & -K_1B & -K_1E \\ -K_2A & -K_2B & -K_2B & -K_2B & -K_2B & -K_2E \\ -K_3A & -K_3B & -K_3B & -K_3B & -K_3B & -K_3E \\ -K_4A & -K_4B & -K_4B & -K_4B & -K_4B & -K_4E \\ -KA & -KB & -KB & -KB & -KB & -KE \end{bmatrix}$$

By Algorithm 1, the controller gains and the worst-case disturbance policy can be obtained below and they converge to the theoretical solution to (28) calculated by using Matlab software based on model (27).

$$\begin{aligned} K_1 &= [0.9828 \ 1.2568 \ -0.0784] \\ K_2 &= [0.4914 \ 0.6284 \ -0.0392] \\ K_3 &= [0.3276 \ 0.4189 \ -0.0261] \\ K_4 &= [0.2457 \ 0.3142 \ -0.0196] \\ K &= [1.8357 \ 2.1312 \ 0.0249] \end{aligned} \tag{29}$$

Then, we assume $E = 0$ which means the external disturbance is not taken into account, and implementing Algorithm 1 yields the optimal controller gains which converge to the optimal control gains calculated by using Matlab software based on model in (27).

$$\begin{aligned} K_1 &= [0.4458 \ 0.6942 \ -0.0877] \\ K_2 &= [0.2229 \ 0.3471 \ -0.0439] \\ K_3 &= [0.1486 \ 0.2314 \ -0.0292] \\ K_4 &= [0.1115 \ 0.1735 \ -0.0219] \end{aligned} \tag{30}$$

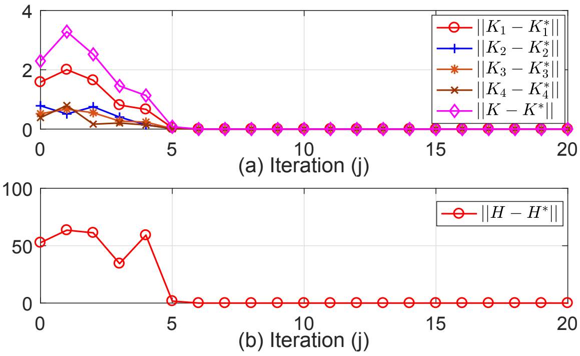

Fig. 1. Convergence of H when implementing the off-policy game Q-learning($E \neq 0$)

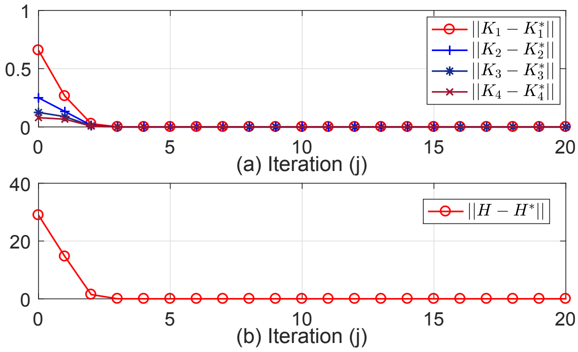

Fig. 2. Convergence of H when implementing the off-policy Q-learning($E = 0$)

Fig. 1 and Fig. 2 respectively show the convergence process of matrix H and controller gains when $E \neq 0$ and $E = 0$ during the execution of the Algorithm 1, thus showing the effectiveness of the proposed algorithm.

Furthermore, simulation comparisons are going to be made under the following external disturbances.

$$d_k = e^{-0.001k} \sin(2.0k) \tag{31}$$

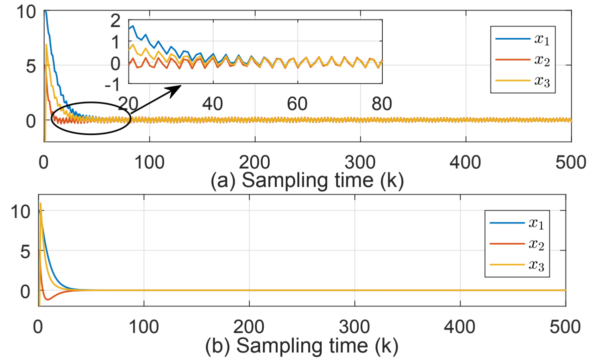

Fig. 3. (a) The states x of system when implementing the off-policy Q-learning algorithm($E = 0$) and (b) The states x of system when implementing the proposed algorithm($E \neq 0$)

Fig. 3(a) and Fig. 3(b) respectively show the system trajectories with and without external interference. It can be seen that the states of the system considering interference always tend to be stable, while the system state without considering the anti-interference will be greatly affected.

5. CONCLUSION

In this paper, we propose a novel off-policy game Q-learning algorithm based on game theory and ADP to learn the Nash equilibrium of the zero-sum game for multi-player linear DT systems. This algorithm was shown to be completely data-driven without requiring system models. The simulation results have demonstrated the effectiveness of the developed algorithm.

REFERENCES

Al-Tamimi, A., Lewis, F.L., and Abu-Khalaf, M. (2007). Model-free Q-learning designs for linear discrete-time zero-sum games with application to H_∞ control. *Automatica*, 43(3), 473–481.

Isidori, A. (1994). H_∞ control via measurement feedback for affine nonlinear systems. *International Journal of Robust and Nonlinear Control*, 4(4), 553–574.

Jiang, Y. and Jiang, Z.P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 48(10), 2699–2704.

Jiao, Q., Modares, H., Xu, S., Lewis, F.L., and Vamvoudakis, K.G. (2016). Multi-agent zero-sum differential graphical games for disturbance rejection in distributed control. *Automatica*, 69, 24–34.

Kiumarsi, B., Lewis, F.L., and Jiang, Z. (2017). H_∞ control of linear discrete-time systems: Off-policy reinforcement learning. *Automatica*, 78, 144–152.

Li, J., Chai, T., Lewis, F.L., Ding, Z., and Jiang, Y. (2019a). Off-policy interleaved Q-learning: Optimal control for affine nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5), 1308–1320.

Li, J., Ding, J., Chai, T., and Lewis, F.L. (2019b). Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes. *IEEE Transactions on Cybernetics.* doi: 10.1109/TCYB.2019.2950262.

Li, J., Modares, H., Chai, T., Lewis, F.L., and Xie, L. (2017). Off-policy reinforcement learning for synchronization in multiagent graphical games. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2434–2445.

Liu, D., Li, H., and Wang, D. (2014). Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(8), 1015–1027.

Luo, B., Huang, T., Wu, H.N., and Yang, X. (2015). Data-driven H_∞ control for nonlinear distributed parameter systems. *IEEE Transactions on Neural Networks and Learning Systems*, 26(11), 2949–2961.

Lv, Y. and Ren, X. (2018). Approximate nash solutions for multi-player mixed-zero-sum game with reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(12), 2739–2750.

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

Vamvoudakis, K.G. and Lewis, F.L. (2010). Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5), 878–888.

Vamvoudakis, K.G. and Lewis, F.L. (2011). Multi-player non-zero-sum games: Online adaptive learning solution of coupled hamilton–jacobi equations. *Automatica*, 47(8), 1556–1569.

Vamvoudakis, K.G., Lewis, F.L., and Hudas, G.R. (2012). Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality. *Automatica*, 48(8), 1598–1611.

Vamvoudakis, K.G., Modares, H., Kiumarsi, B., and Lewis, F.L. (2017). Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online. *IEEE Control Systems Magazine*, 37(1), 33–52.

Van de Schaft, A. (1992). L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control. *IEEE Trans. Autom. Control*, 37(6), 770–784.

Wang, C., Tnunay, H., Zuo, Z., Lennox, B., and Ding, Z. (2018a). Fixed-time formation control of multirobot systems: Design and experiments. *IEEE Transactions on Industrial Electronics*, 66(8), 6292–6301.

Wang, C., Zuo, Z., Qi, Z., and Ding, Z. (2018b). Predictor-based extended-state-observer design for consensus of mass with delays and disturbances. *IEEE Transactions on Cybernetics*, 49(4), 1259–1269.