Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Multi-agent differential game based cooperative synchronization control using a data-driven method*

Yu SHI[1], Yongzhao HUA[2], Jianglong YU[1], Xiwang DONG‡[1,2], Zhang REN[1]

[1]*School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China*

[2]*Institute of Artificial Intelligence, Beihang University, Beijing 100191, China*

E-mail: shiyu_sasee@buaa.edu.cn; yongzhaohua@buaa.edu.cn; sdjxyjl@buaa.edu.cn; xwdong@buaa.edu.cn; renzhang@buaa.edu.cn

Received Jan. 3, 2022; Revision accepted Apr. 21, 2022; Crosschecked May 10, 2022

Abstract: This paper studies the multi-agent differential game based problem and its application to cooperative synchronization control. A systematized formulation and analysis method for the multi-agent differential game is proposed and a data-driven methodology based on the reinforcement learning (RL) technique is given. First, it is pointed out that typical distributed controllers may not necessarily lead to global Nash equilibrium of the differential game in general cases because of the coupling of networked interactions. Second, to this end, an alternative local Nash solution is derived by defining the best response concept, while the problem is decomposed into local differential games. An off-policy RL algorithm using neighboring interactive data is constructed to update the controller without requiring a system model, while the stability and robustness properties are proved. Third, to further tackle the dilemma, another differential game configuration is investigated based on modified coupling index functions. The distributed solution can achieve global Nash equilibrium in contrast to the previous case while guaranteeing the stability. An equivalent parallel RL method is constructed corresponding to this Nash solution. Finally, the effectiveness of the learning process and the stability of synchronization control are illustrated in simulation results.

Key words: Multi-agent system; Differential game; Synchronization control; Data-driven; Reinforcement learning
https://doi.org/10.1631/FITEE.2200001 **CLC number:** TP273

1 Introduction

Cooperative control of multi-agent systems (MASs) has been a significant part of the networked control field in the past decades due to the wide applications in unmanned vehicles, robotics, and so on. The consensus problem was previously investigated in Olfati-Saber and Murray (2004) and Ren and Beard (2005) based on neighboring information through networks, which provides a fundamental methodology for subsequent studies. The synchronization problem took a step forward, where agents not only reach a common constant value but also track a leader's trajectory using local interactions (Qin et al., 2011; Dong et al., 2014). With the development of related studies, researchers started to pay attention to the optimality in cooperative control. Distributed optimization (Yang T et al., 2019) has attracted sufficient attention, and has been applied in a wide range of practical scenarios such as smart grids (Zheng et al., 2016; Peng and Low, 2018; Wen et al., 2021) and intelligent transportation (Wang

‡ Corresponding author

* Project supported by the Science and Technology Innovation 2030, China (No. 2020AAA0108200), the National Natural Science Foundation of China (Nos. 61873011, 61973013, 61922008, and 61803014), the Defense Industrial Technology Development Program, China (No. JCKY2019601C106), the Innovation Zone Project, China (No. 18-163-00-TS-001-001-34), the Foundation Strengthening Program Technology Field Fund, China (No. 2019-JCJQ-JJ-243), and the Fund from the Key Laboratory of Dependable Service Computing in Cyber Physical Society, China (No. CPSDSC202001)

ORCID: Yu SHI, https://orcid.org/0000-0001-8618-7395; Xiwang DONG, https://orcid.org/0000-0002-4778-248X

MY et al., 2021). This can be decomposed into a cooperative synchronization control problem with a predefined optimal reference. Agents simultaneously execute the consensus and gradient-decent algorithms based on local objectives, while the global objective is achieved. By further considering the local coupled objectives, which reflect the conflicts of interest among agents, the distributed game problem (Sun et al., 2017) was proposed. The distributed optimization method can be extended to the algorithm commonly called the Nash seeking strategy (Ye et al., 2018, 2019).

In contrast to distributed optimization with immediate and static objectives, differential game problems formulate an optimization for the controller of dynamic systems. Originating in the optimal control problem (Lewis et al., 2012), the coupling effects of individual actions as well as index functions in dynamic systems increase the complexity of this problem. There are two general types of differential game problems: (1) zero-sum differential game between two players; (2) nonzero-sum differential game between multiple players. The zero-sum differential game is aimed to find two balanced controllers, with one minimizing an index function and the other maximizing the same. This game model, in which each agent has competitive interests, has been commonly used in H_2 and H_∞ control problems (Modares et al., 2015). The nonzero-sum differential game is a more general case, where each agent holds individual, either competitive or cooperative, coupled index functions (Vamvoudakis and Lewis, 2011). Due to the existence of multiple players, the pursuit of an optimal/balanced solution turns to finding the global Nash equilibrium. Early related work was studied and summarized as game theory (Başar and Olsder, 1982), and some recent results were given in Zhu and Başar (2015), Zhao DB et al. (2015), Wang W et al. (2020), and Zhao JG (2020). However, this differential game is based on a unified system with players being regarded as control inputs, and all players are aware of full system states.

As a combination of distributed MAS control and game theory, the differential game in a topology graph, namely the graphical game, has attracted more attention in recent years. Graphical game is the fundamental in attack-defense, pursuit-evasion problems. Vamvoudakis et al. (2012) built a standard graphical game formulation for linear MAS consensus control, where the index functions were defined according to neighboring errors and neighbors' control actions. Each agent optimized its own index function using the best response method, and the Nash equilibrium was proved to be obtained. In addition, agents can update their optimization results either alternately or simultaneously. Further analysis presented a comparison of the distributed graphical game and the centralized game for discrete-time (DT) MASs (Abouheaf et al., 2014), where it was shown that the distributed graphical game is more challenging due to complex couplings whereas the key to handling this issue is solving coupled Hamilton–Jacobi (HJ) equations. However, HJ equations are rather difficult to solve due to complex interactions. Thus, data-driven methods have been considered to find approximate solutions.

Reinforcement learning (RL) (Sutton and Barto, 1998), inspired by natural biological learning mechanisms, provides an online adaptive methodology for decision and control problems. Past results have shown that RL can equivalently deal with optimal control problems for DT (Tamimi et al., 2008) and continuous-time (CT) (Modares and Lewis, 2014) systems without using actual dynamic models. Apparently, this property is suitable for solving the differential game. A Q-learning-based method was implemented to deal with a linear DT zero-sum game in Yang YJ et al. (2021). H_∞ robust control for DT systems was modeled as a zero-sum game between the controller and disturbance in Modares et al. (2015), and the problem was solved using the integral RL method. Recent work has applied the RL-based data-driven method to graphical game problems. The optimal DT MAS consensus problem was investigated in Zhang et al. (2017) under the graphical game frame, and the solution was directly obtained online using neural networks (NNs) and the policy iteration method. This was extended to the synchronization problem using the Q-learning method in Wang MY et al. (2021), where it was proved that the controller stabilizes leader–follower MASs while achieving respective equilibrium. Off-policy RL was implemented in the CT MAS synchronization problem for the controller to satisfy the Nash equilibrium in the graphical game (Li et al., 2017; Mu et al., 2017).

It is worth pointing out that the Nash property in the graphical game should be further discussed.

Although RL-based data-driven methods provide efficient ways to solve coupled HJ equations, the NNs that represent the controllers were usually designed as related only to the immediate neighboring error. As stated in the latest research (Liu et al., 2021; Qian et al., 2021), the approximate results may not satisfy the global Nash equilibrium condition due to the couplings. A typical min–max method was used to deal with the local optimization problem in Lopez et al. (2020). Li et al. (2017) tried to reduce the couplings between connected agents by choosing local related index functions. However, to the best of our knowledge, there are still few guidelines for the formulation and analysis of graphical game based cooperative synchronization of CT MASs and the solution using an RL method. Finding an institutionalized and systematized MAS graphical game framework remains an open question.

To this end, this paper studies the multi-agent differential graphical game with respect to the cooperative synchronization control problem using a data-driven RL method. First, by analyzing the solvability and solution structure of the HJ equations, a contradiction of distributed control and global Nash equilibrium is proposed. Second, inspired by the original work in the field of graphical game (Vamvoudakis et al., 2012) and dynamic game theory (Zhu and Başar, 2015), two compromised schemes are proposed in this paper: the local Nash game solution using a strict best response method and the global Nash solution with respect to modified index functions. Third, an off-policy RL algorithm is investigated to design model-free controllers for both scenarios with the stability properties proved.

Compared with existing work, the main contributions of this paper are as follows:

1. This paper proposes a general game based framework for the MAS cooperative synchronization control problem, where the optimum or equilibrium is considered in the controller design process instead of stability only.

2. The distributed and Nash properties of the game solution are discussed in detail. It is proved that they may not hold at the same time compared with simplified cases under immediate neighbor-related assumptions (Li et al., 2017; Zhao JG, 2020; Wang MY et al., 2021).

3. A systematized scheme consisting of local and modified global cases is proposed with guaranteed Nash equilibrium using a data-driven method. An off-policy RL for CT MASs is derived to solve this problem. Compared with the conventional model-based and existing on-policy RL methods in Vamvoudakis et al. (2012), this paper provides an adaptive and model-free method that relaxes the dependence on the system model.

2 Preliminaries and problem description

2.1 Notations

Let vec() represent a vector expanded from a matrix with respect to its columns. I_N denotes an identity matrix with dimension N. The notation "$\otimes$" represents the Kronecker product. $\nabla_x f(x)$ stands for the gradient of f along parameter x.

2.2 Graph theory

A graph consisting of $N+1$ nodes can be represented by $\mathcal{G}=\{\mathcal{S},W\}$. $\mathcal{S}=\{s_1,s_2,\dots,s_{N+1}\}$ is the node set and the notation (s_i,s_j) denotes a directed path from a parent node s_i to a child node s_j. The connectivity between nodes is described by non-negative weights w_{ij}, while the associated adjacency matrix is constructed as $W=[w_{ij}]\in\mathbb{R}^{(N+1)\times(N+1)}$. The weight $w_{ij}=1$ if and only if (s_j,s_i) exists, and $w_{ij}=0$ otherwise. $w_{ii}=0$ holds since there exists no self-loop in $\mathcal{G}$. Let $\mathcal{N}_i$ stand for the set of neighbors of node s_i. The Laplacian matrix representing the topology interactions can be defined by $L=D-W$, where $D=\mathrm{diag}\left(\sum_j w_{1j},\sum_j w_{2j},\dots,\sum_j w_{(N+1)j}\right)$ denotes the in-degree matrix of the graph. Moreover, if there is a root node that has at least one directed path to every other node, the graph is said to contain a spanning tree.

Consider MASs that consist of one leader and N followers. The leader is labeled 0, and the followers are labeled by a set $\mathcal{F}=\{1,2,\dots,N\}$. The leader is defined as an agent with no neighbor, while each follower has at least one neighbor. Then the Laplace matrix can be written as $L=\begin{bmatrix} L_1 & L_2 \\ 0_{1\times N} & 0 \end{bmatrix}$, where matrix $L_1\in\mathbb{R}^{N\times N}$, and $L_2\in\mathbb{R}^{N\times 1}$ represents the interactions from the leader to followers. Define the in-degree matrix among the followers as $D_1=\mathrm{diag}\,(d_{\mathrm{F}1},d_{\mathrm{F}2},\cdots,d_{\mathrm{F}N})\in\mathbb{R}^{N\times N}$.

Assumption 1 The interactions among followers are strongly connected and there exists a spanning tree in graph $\mathcal{G}$ with the leader being a root node.

Lemma 1 Under Assumption 1, all the eigenvalues of L_1 have positive real parts and there exists a positive definite matrix $\Omega = \mathrm{diag}(\omega_1, \omega_2, \cdots, \omega_N)$ that satisfies $\Omega L_1 + L_1^{\mathrm{T}}\Omega > 0$.

2.3 Problem description

The dynamics of follower $i \in \mathcal{F}$ is given by

$$\dot{x}_i(t) = Ax_i(t) + Bu_i(t), \tag{1}$$

where $x_i(t) \in \mathbb{R}^n$ and $u_i(t) \in \mathbb{R}^m$ represent the state and control input vectors, respectively. $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are a follower's unknown dynamic and input matrices respectively, where the pair (A, B) is stabilizable.

The dynamics of the leader is defined as

$$\dot{x}_0(t) = Ax_0(t), \tag{2}$$

where $x_0(t) \in \mathbb{R}^n$ is the state of the leader. The leader given in Eq. (2) generates a reference trajectory with which the followers synchronize.

In the graph with local interactions, a neighboring synchronization error that can be directly obtained by follower i is defined as

$$\xi_i(t) = \sum_{j\in\mathcal{N}_i} w_{ij}\left(x_i(t) - x_j(t)\right) + w_{i0}\left(x_i(t) - x_0(t)\right). \tag{3}$$

Differentiating Eq. (3) along with Eqs. (1) and (2) yields the neighboring error dynamics:

$$\dot{\xi}_i(t) = A\xi_i(t) + d_iBu_i - \sum_{j\in\mathcal{N}_i} w_{ij}Bu_j. \tag{4}$$

Define the vectors $\xi(t) = \left[\xi_1^{\mathrm{T}}(t), \xi_2^{\mathrm{T}}(t), \cdots, \xi_N^{\mathrm{T}}(t)\right]^{\mathrm{T}}$, $x(t) = \left[x_1^{\mathrm{T}}(t), x_2^{\mathrm{T}}(t), \cdots, x_N^{\mathrm{T}}(t)\right]^{\mathrm{T}}$, and $\bar{x}_0(t) = 1_N \otimes x_0(t)$. The neighboring error and its dynamics can be written in the following compact form:

$$\begin{cases} \xi(t) = (L_1 \otimes I_n)\left(x(t) - \bar{x}_0(t)\right) = (L_1 \otimes I_n)\,\zeta(t), \\ \dot{\xi}(t) = (I_N \otimes A)\,\xi(t) + (L_1 \otimes B)\,u(t), \end{cases} \tag{5}$$

where $\zeta(t) \triangleq [\zeta_1^{\mathrm{T}}(t), \zeta_2^{\mathrm{T}}(t), \cdots, \zeta_N^{\mathrm{T}}(t)]^{\mathrm{T}}$ with the vector $\zeta_i(t) = x_i(t) - x_0(t)$ standing for the local synchronization error.

Definition 1 (Synchronization control) For MASs with any given bounded initial conditions, if there exists a distributed controller $u_i(t) = u_i(\xi_i(t))$ for follower i such that

$$\lim_{t\to\infty} \zeta_i(t) = \lim_{t\to\infty} \left(x_i(t) - x_0(t)\right) = 0, \tag{6}$$

then the cooperative synchronization control is said to be accomplished.

Definition 2 (Multi-agent differential game) In the MAS synchronization process, follower i maintains a local index function $J_i(t)$ and there exists a corresponding value function $V_i(t)$ given in the following form:

$$\begin{aligned} J_i(t)|_{(u_i,u_{-i})} &= V_i\left(\xi_i(t), \xi_{-i}(t)\right) \\ &= \int_t^\infty r_i\big(\xi_i(\tau), \xi_{-i}(\tau), u_i(\tau), u_{-i}(\tau)\big)\mathrm{d}\tau, \end{aligned} \tag{7}$$

where the notation "$-i$" denotes the neighbors of the i^{th} follower. Then the MASs are said to fulfill the differential game when each agent pursues the local optimal controller:

$$u_i^*(t) = \arg\min_{u_i(t)} V_i\left(\xi_i(t), \xi_{-i}(t)\right). \tag{8}$$

Definition 3 (Nash equilibrium (Başar and Olsder, 1982))

1. Local best response

The local best response controller $u_i^{\#}$ of the i^{th} follower corresponding to J_i and u_j, $j \neq i$, is defined to satisfy

$$J_i|(u_1, \cdots, u_i^{\#}, \cdots, u_N) \leq J_i|(u_1, \cdots, u_i, \cdots, u_N). \tag{9}$$

2. Global Nash equilibrium

Consider the unified solution of the differential game with index functions $\{J_1^*, J_2^*, \cdots, J_N^*\}$ and controller tuples $\{u_1^*, u_2^*, \cdots, u_N^*\}$. The MAS is said to achieve the global Nash equilibrium under the condition

$$J_i^*|(u_1^*, \cdots, u_i^*, \cdots, u_N^*) \leq J_i^*|(u_1^*, \cdots, u_i, \cdots, u_N^*). \tag{10}$$

Remark 1 The purpose of this study is to unify the synchronization control in a differential game frame. This problem is multifaceted due to the coupled index functions which are defined using the neighboring error and control of their immediate neighbors. Similar to the optimal solution in single-agent cases, the Nash equilibrium provides an equivalent concept for the multi-agent differential game; i.e., the best response of each individual follower forms the global

Nash solution. Since L_1 is nonsingular according to Lemma 1, $\xi(t)$ and $\zeta(t)$ have the same convergence. Thus, the graphical game in this study is constructed based on $\xi(t)$.

3 Main results

3.1 Solvability analysis of differential games

Consider the local index function defined in Eq. (7) along the following common quadratic form:

$$\begin{aligned} &r_i\left(\xi_i, u_i, u_{-i}\right) \\ =&\frac{1}{2}\left(\xi_i^{\mathrm{T}} Q_i \xi_i + u_i^{\mathrm{T}} R_{ii} u_i + \sum_{j\in\mathcal{N}_i} u_j^{\mathrm{T}} R_{ij} u_j\right), \end{aligned} \tag{11}$$

where $Q_i \in \mathbb{R}^{n\times n}$, $R_{ii} \in \mathbb{R}^{m\times m}$, and $R_{ij} \in \mathbb{R}^{m\times m}$ are symmetric definite matrices.

For any arbitrary finite value function, differentiating Eq. (7) along the system dynamics yields the Bellman equation and the Hamiltonian function:

$$\begin{aligned} &H_i\left(\xi_i, \nabla_{\xi_i} V_i, u_i, u_{-i}\right) \\ \triangleq&\left(\nabla_{\xi_i} V_i\right)^{\mathrm{T}}\left(A\xi_i + d_i B u_i - \sum_{j\in\mathcal{N}_i} w_{ij} B u_j\right) \\ &+\frac{1}{2}\xi_i^{\mathrm{T}} Q_i \xi_i + \frac{1}{2} u_i^{\mathrm{T}} R_{ii} u_i + \frac{1}{2}\sum_{j\in\mathcal{N}_i} u_j^{\mathrm{T}} R_{ij} u_j = 0, \end{aligned} \tag{12}$$

where V_i stands for the value function corresponding to the current u_i and u_{-i}. Using the stationary condition gives the local best response as

$$u_i^{\#} = u_i^{\#}\left(V_i\right) \triangleq -d_i R_{ii}^{-1} B_i^{\mathrm{T}} \nabla_{\xi_i} V_i. \tag{13}$$

Then one can obtain the coupled HJ equation by substituting Eq. (13) into Eq. (12):

$$\begin{aligned} 0 =&\left(\nabla_{\xi_i} V_i\right)^{\mathrm{T}}\left(A\xi_i - d_i^2 B R_{ii}^{-1} B^{\mathrm{T}} \nabla_{\xi_i} V_i\right) \\ &+\left(\nabla_{\xi_i} V_i\right)^{\mathrm{T}}\left(\sum_{j\in\mathcal{N}_i} w_{ij} d_i B_j R_{jj}^{-1} B_j^{\mathrm{T}} \nabla_{\xi_j} V_j\right) \\ &+\frac{1}{2}\xi_i^{\mathrm{T}} Q_i \xi_i + \frac{1}{2} d_i^2 \left(\nabla_{\xi_i} V_i\right)^{\mathrm{T}} B R_{ii}^{-1} B^{\mathrm{T}} \nabla_{\xi_i} V_i \\ &+\frac{1}{2}\sum_{j\in\mathcal{N}_i} d_j^2 \left(\nabla_{\xi_j} V_j\right)^{\mathrm{T}} B R_{jj}^{-1} R_{ij} R_{jj}^{-1} B^{\mathrm{T}} \nabla_{\xi_j} V_j. \end{aligned} \tag{14}$$

The solution tuple $\left\{\left(\nabla_{\xi_1} V_1\right)^{\mathrm{T}}, \left(\nabla_{\xi_2} V_2\right)^{\mathrm{T}}, \cdots, \left(\nabla_{\xi_N} V_N\right)^{\mathrm{T}}\right\}$ for coupled HJ equations gives the theoretical result to the MAS differential game, which leads to the global Nash equilibrium. Denote the augmented gradient vector as $\nabla V = \left[\left(\nabla_{\xi_1} V_1\right)^{\mathrm{T}}, \left(\nabla_{\xi_2} V_2\right)^{\mathrm{T}}, \cdots, \left(\nabla_{\xi_N} V_N\right)^{\mathrm{T}}\right]^{\mathrm{T}}$. Employing the fact that a symmetric matrix P satisfying $\nabla V = P\xi$ is supposed to exist for a linear quadratic differential game, Eq. (14) can be equivalently written in the matrix form:

$$\begin{aligned} 0 =&PA^i + \left(A^i\right)^{\mathrm{T}} P + PE^i \mathrm{diag}\left(d_k B R_{kk}^{-1} B^{\mathrm{T}}\right) P + Q^i \\ &+P\mathrm{diag}\left(d_k B R_{kk}^{-1}\right)\mathrm{diag}\left(R_{ik}\right)\mathrm{diag}\left(d_k R_{kk}^{-1} B^{\mathrm{T}}\right) P, \end{aligned} \tag{15}$$

where $A^i \triangleq \mathrm{diag}\left([A]^i\right)$ and $Q^i \triangleq \mathrm{diag}\left([Q_i]^i\right)$ are block-wise diagonal matrices with the i^{th} elements being A and Q_i, respectively. Matrix E^i is given by

$$E^i \triangleq \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & 0 & \vdots & \vdots \\ w_{i1} I_n & w_{i2} I_n & -d_i I_n & w_{iN} I_n \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Note that there is a set of N equations of Eq. (15) that require a uniform solution P. To simplify the analysis, adding the equations with the index i from 1 to N yields a reduced necessary condition:

$$\begin{aligned} 0 =&P\left(I_N \otimes A\right) + \left(I_N \otimes A\right)^{\mathrm{T}} P - P\left(L_1 \otimes I_n\right) \\ &\cdot \mathrm{diag}\left(d_k B R_{kk}^{-1} B^{\mathrm{T}}\right) P + \mathrm{diag}\left(Q_k\right) + P d_k^2 \\ &\cdot \mathrm{diag}\left(B R_{kk}^{-1}\right)\mathrm{diag}\left(\sum_{j\in\mathcal{N}_i} R_{ik}\right)\mathrm{diag}\left(R_{kk}^{-1} B^{\mathrm{T}}\right) P. \end{aligned} \tag{16}$$

Because $\left(L_1 \otimes I_n\right)\mathrm{diag}\left(d_k B R_{kk}^{-1} B^{\mathrm{T}}\right)$ is usually not symmetric due to the different values of in-degree d_k, $k \in \mathcal{F}$, Eq. (16) is not a typical algebra Riccati equation and may be difficult or even impossible to solve. This indicates that the Nash solution may not exist. Moreover, as an assumption that has been commonly made in the literature, consider that each controller uses only its respective neighboring information. Then it can be further inferred that $\nabla_{\xi_i} V_i$ is related only to ξ_i so that $P = \mathrm{diag}(P_1, P_2, \cdots, P_N)$ is a block diagonal matrix, where P_i is a symmetric positive-definite matrix that satisfies $V_i = \frac{1}{2}\xi_i^{\mathrm{T}} P_i \xi_i$. Then the necessary condition can be equivalent to

$$\begin{cases} 0 = P_i A + A^{\mathrm{T}} P_i - d_i^2 P_i B R_{ii}^{-1} B^{\mathrm{T}} P_i + Q_{ii} \\ \quad\quad +d_i^2 P_i B R_{ii}^{-1} \sum\limits_{j\in\mathcal{N}_i} R_{ji} R_{ii}^{-1} B^{\mathrm{T}} P_i = 0, \\ 0 = w_{ij} d_i P_i B R_{jj}^{-1} B^{\mathrm{T}} P_j. \end{cases} \tag{17}$$

Since there exists no solution tuple for coupled Eq. (17) according to the positive-definite property

of P_i, the requirements of being distributed and the global Nash cannot be fulfilled at the same time.

Remark 2 Because the index function (11) is defined in quadratic form, it is reasonable to assume the value function to be quadratic. There is no block diagonal or symmetric solution, which indicates that the distributed and global Nash solution does not exist. One can see from Eq. (16) that the coupled term $L_1 \otimes I_n$ prevents the solution from being solved. To dig deeper, the coupled term is introduced by both ξ_i and u_j in Eq. (11), whereas the value function is related to not only the immediate neighbors, but also "neighbors' neighbors." In addition, the reduced index function without u_j (Li et al., 2017) cannot solve this problem due to the use of ξ_i. Thus, the key issue lies on how to cover or decouple the graphical game, which is the motivation of this study.

3.2 Local best response solution using RL

Inspired by the zero-sum game mechanism, in this subsection we propose a local differential game solution using the best response method. Off-policy RL for CT systems based on the state-action function is derived for stable controller design.

Consider the local index function defined in the following quadratic form:

$$\begin{aligned} &r_i\left(\xi_i, u_i, u_{-i}\right) \\ =&\frac{1}{2}\left(\xi_i^{\mathrm{T}} Q_i \xi_i + d_i u_i^{\mathrm{T}} R_{ii} u_i - \sum_{j\in\mathcal{N}_i} u_j^{\mathrm{T}} R_{ij} u_j\right). \end{aligned} \tag{18}$$

The Hamiltonian can be written accordingly as

$$\begin{aligned} &H_i\left(\xi_i, \nabla_{\xi_i} V_i, u_i, u_{-i}\right) \\ \triangleq&\left(\nabla_{\xi_i} V_i\right)^{\mathrm{T}}\left(A\xi_i + d_i B u_i - \sum_{j\in\mathcal{N}_i} w_{ij} B u_j\right) \\ &+\frac{1}{2}\xi_i^{\mathrm{T}} Q_i \xi_i + \frac{1}{2} d_i u_i^{\mathrm{T}} R_{ii} u_i - \frac{1}{2}\sum_{j\in\mathcal{N}_i} u_j^{\mathrm{T}} R_{ij} u_j = 0. \end{aligned} \tag{19}$$

Now the local differential game regards u_j, $j \in \mathcal{F}$, as extra players to help form the local best response results, while the corresponding solutions can be given by

$$u_i^{\#} = -R_{ii}^{-1} B^{\mathrm{T}} \nabla_{\xi_i} V_i,\ u_{ij}^{\#} = -w_{ij} R_{ij}^{-1} B^{\mathrm{T}} \nabla_{\xi_i} V_i, \tag{20}$$

where u_{ij} is the local value of u_j in the i^{th} game. In the local scenario, the value function is assumed to be local-dependent because $\nabla_{\xi_i} V_i = \tilde{P}_i \xi_i$. Note that HJ equations independently hold with any arbitrary trajectory of ξ_i. Substituting Eq. (20) into Eq. (19) yields the decoupled equation:

$$\begin{aligned} &\tilde{P}_i^* A + A^{\mathrm{T}} \tilde{P}_i^* - d_i \tilde{P}_i^* B R_{ii}^{-1} B^{\mathrm{T}} \tilde{P}_i^* \\ &+ w_{ij}^2 \tilde{P}_i^* B \sum_{j\in\mathcal{N}_i} R_{ij}^{-1} B^{\mathrm{T}} \tilde{P}_i^* + Q_i = 0. \end{aligned} \tag{21}$$

Theorem 1 Consider the differential MAS graphical game with dynamics (1) and (2) and index function (18). The local best response solution $u_i^* = -R_{ii}^{-1} B^{\mathrm{T}} \tilde{P}_i^* \xi_i$, where $\tilde{P}_i^*$ is derived in Eq. (21), produces the following results: (1) The local Nash is achieved for the differential game; (2) The distributed controller gives $\mathcal{L}_2$ stability against other players; (3) The asymptotical stability of the MAS synchronization error is guaranteed.

Proof For the best responses $u_i^{\#}$ and $u_{ij}^{\#}$ with respect to any arbitrary V_i, u_i, and u_{ij}, the corresponding Hamiltonian has the following property:

$$\begin{aligned} &H_i\left(\xi_i, u_i, u_{ij}, \tilde{P}_i\right) - H_i\left(\xi_i, u_i^{\#}, u_{ij}^{\#}, \tilde{P}_i\right) \\ =&d_i \xi_i^{\mathrm{T}} \tilde{P}_i B\left(u_i - u_i^{\#}\right) - w_{ij} \xi_i^{\mathrm{T}} \tilde{P}_i B\left(u_{ij} - u_{ij}^{\#}\right) \\ &+\frac{1}{2}\xi_i^{\mathrm{T}} Q_i \xi_i + \frac{1}{2} d_i u_i^{\mathrm{T}} R_{ii} u_i - \frac{1}{2}\sum_{j\in\mathcal{N}_i} u_{ij}^{\mathrm{T}} R_{ij} u_{ij} \\ &-\frac{1}{2}\xi_i^{\mathrm{T}} Q_i \xi_i - \frac{1}{2} d_i \left(u_i^{\#}\right)^{\mathrm{T}} R_{ii} u_i^{\#} \\ &+\frac{1}{2}\sum_{j\in\mathcal{N}_i}\left(u_{ij}^{\#}\right)^{\mathrm{T}} R_{ij} u_{ij}^{\#}. \end{aligned} \tag{22}$$

Note that if $\xi_i^{\mathrm{T}} \tilde{P}_i B = -\left(u_i^{\#}\right)^{\mathrm{T}} R_{ii}$ and $w_{ij} \xi_i^{\mathrm{T}} \tilde{P}_i B = -\left(u_{ij}^{\#}\right)^{\mathrm{T}} R_{ij}$, then Eq. (22) can be further derived as

$$\begin{aligned} &H_i\left(\xi_i, u_i, u_{ij}, \tilde{P}_i\right) - H_i\left(\xi_i, u_i^{\#}, u_{ij}^{\#}, \tilde{P}_i\right) \\ =&-d_i\left(u_i^{\#}\right)^{\mathrm{T}} R_{ii}\left(u_i - u_i^{\#}\right) + \frac{1}{2} d_i u_i^{\mathrm{T}} R_{ii} u_i \\ &-\frac{1}{2} d_i \left(u_i^{\#}\right)^{\mathrm{T}} R_{ii} u_i^{\#} + \sum_{j\in\mathcal{N}_i}\left(u_{ij}^{\#}\right)^{\mathrm{T}} R_{ij}\left(u_{ij} - u_{ij}^{\#}\right) \\ &+\frac{1}{2}\sum_{j\in\mathcal{N}_i}\left(u_{ij}^{\#}\right)^{\mathrm{T}} R_{ij} u_{ij}^{\#} - \frac{1}{2}\sum_{j\in\mathcal{N}_i} u_j^{\mathrm{T}} R_{ij} u_j \\ =&-\frac{1}{2}\sum_{j\in\mathcal{N}_i}\left(u_{ij} - u_{ij}^{\#}\right)^{\mathrm{T}} R_{ij}\left(u_{ij} - u_{ij}^{\#}\right) \\ &+\frac{1}{2} d_i \left(u_i - u_i^{\#}\right)^{\mathrm{T}} R_{ii}\left(u_i - u_i^{\#}\right). \end{aligned} \tag{23}$$

For any arbitrary trajectory with bounded $\xi_i(t)$, $\lim_{t\to\infty} V_i(\xi_i(t)) = \lim_{t\to\infty} V_i(0) = 0$ holds. The index function satisfies

$$\begin{aligned}
&J_i\left(\xi_i, u_i, u_j, \tilde{P}_i\right)\\
=&J_i\left(\xi_i, u_i, u_j, \tilde{P}_i\right) + V_i(\xi_i(0)) + \int_0^\infty \dot{V}_i(\xi_i(t))\mathrm{d}t\\
=&V_i(\xi_i(0)) + \int_0^\infty H_i\left(\xi_i, u_i, u_j, \tilde{P}_i\right)\mathrm{d}t\\
=&V_i(\xi_i(0)) + \frac{1}{2}\int_0^\infty d_i\left(u_i - u_i^{\#}\right)^{\mathrm{T}} R_{ii}\left(u_i - u_i^{\#}\right)\mathrm{d}t\\
&-\frac{1}{2}\int_0^\infty \sum_{j\in\mathcal{N}_i}\left(u_{ij} - u_{ij}^{\#}\right)^{\mathrm{T}} R_{ij}\left(u_{ij} - u_{ij}^{\#}\right)\mathrm{d}t.
\end{aligned} \tag{24}$$

Then the following results can be obtained by substituting $u_i^{\#}$ and $u_{ij}^{\#}$ into Eq. (24):

$$J_i\left(u_i^{\#}, u_{ij}, \tilde{P}_i\right) \le J_i\left(u_i^{\#}, u_{ij}^{\#}, \tilde{P}_i\right) \le J_i\left(u_i, u_{ij}^{\#}, \tilde{P}_i\right). \tag{25}$$

Further consider the local best responses u_i^* and u_{ij}^* corresponding to J_i^* and H_i^*. One can obtain

$$J_i^*\left(u_i^*, u_{ij}, \tilde{P}_i^*\right) \le J_i^*\left(u_i^*, u_{ij}^*, \tilde{P}_i^*\right) \le J_i^*\left(u_i, u_{ij}^*, \tilde{P}_i^*\right). \tag{26}$$

This indicates the local Nash between u_i^* and u_{ij}^*.

Since the Bellman equation is satisfied with $H_i^*(\xi_i, u_i^*, u_{ij}^*, \tilde{P}_i^*) = 0$, it can be further inferred from Eq. (23) that

$$\begin{aligned}
&\frac{1}{2}\xi_i^{\mathrm{T}} Q_i \xi_i + \frac{1}{2} d_i (u_i^*)^{\mathrm{T}} R_{ii} u_i^* - \frac{1}{2}\sum_{j\in\mathcal{N}_i} u_{ij}^{\mathrm{T}} R_{ij} u_{ij}\\
&+\xi_i^{\mathrm{T}}\tilde{P}_i^*\left(A\xi_i + d_i B u_i^* - \sum_{j\in\mathcal{N}_i} w_{ij} B u_j\right)\\
=&H_i^*\left(\xi_i, u_i^*, u_{ij}, \tilde{P}_i^*\right)\\
=&-\frac{1}{2}\left(u_{ij} - u_{ij}^*\right)^{\mathrm{T}} R_{ij}\left(u_{ij} - u_{ij}^*\right) \le 0.
\end{aligned} \tag{27}$$

Integrating both sides of inequiality (27) yields

$$\begin{aligned}
&\int_0^T \left(\xi_i^{\mathrm{T}} Q_i \xi_i + d_i (u_i^*)^{\mathrm{T}} R_{ii} u_i^*\right)\mathrm{d}t + V(\xi_i(T))\\
\le&\int_0^T \sum_{j\in\mathcal{N}_i} u_{ij}^{\mathrm{T}} R_{ij} u_{ij}\mathrm{d}t + V(\xi_i(0)).
\end{aligned} \tag{28}$$

This means that u_i^* gives a bounded response under any u_{ij}; i.e., the $\mathcal{L}_2$ stability is realized.

The local $\mathcal{L}_2$ property cannot lead to global asymptotic stability for MASs. For each agent using the local controller u_i^*, the closed loop dynamics can be obtained by

$$\begin{aligned}
\dot{\xi}_i(t) =& A\xi_i(t) - d_i B R_{ii}^{-1} B^{\mathrm{T}} \tilde{P}_i \xi_i(t)\\
&+ \sum_{j\in\mathcal{N}_i} w_{ij} B R_{jj}^{-1} B^{\mathrm{T}} \tilde{P}_j \xi_j(t).
\end{aligned} \tag{29}$$

Without loss of generality, it is assumed that R_{ii} are uniformly selected as $\tilde{R}$. Then the augmented system can be written as

$$\dot{\xi}(t) = \left(I_N \otimes A - \left(L_1 \otimes B R^{-1} B^{\mathrm{T}}\right)\tilde{P}\right)\xi(t), \tag{30}$$

where matrix $\tilde{P} = \mathrm{diag}(\tilde{P}_1, \tilde{P}_2, \cdots, \tilde{P}_N)$. Consider the global Lyapunov function candidate as $L(\xi) = \xi^{\mathrm{T}}(t)\left(\tilde{P}(\Omega \otimes I_N)\right)\xi(t)$ with its time derivative given by

$$\dot{L}(\xi) = \xi^{\mathrm{T}}(t)\left(\Lambda_1 + \Lambda_2\right)\xi(t), \tag{31}$$

where $\Lambda_1 = \tilde{P}(\Omega \otimes A) + \left(\Omega \otimes A^T\right)\tilde{P}$, $\Lambda_2 = -\tilde{P}\left(\Lambda_3 \otimes B R^{-1} B^{\mathrm{T}}\right)\tilde{P}$, and $\Lambda_3 = \Omega L_1 + L_1^{\mathrm{T}}\Omega$. One can obtain from $\Lambda_3 > 0$ that $\Lambda_2 < 0$. Moreover, under the sufficient condition $Q_i - d_{\mathrm{F}i}\tilde{P}_i B R^{-1} B^{\mathrm{T}} \tilde{P}_i > 0$, $d_{\mathrm{F}i} = \sum_{j\in\mathcal{N}_i} w_{ij}^2$ denotes the in-degree among followers. For the diagonal matrix Ω with positive elements, $\omega_i \tilde{P}_i A + \omega_i A^{\mathrm{T}} \tilde{P}_i < 0$ holds, which indicates that $\Lambda_1 < 0$. Then $\dot{L}(\xi) < 0$ can be derived and the global synchronization error converges to zero asymptotically. This completes the proof.

Remark 3 Theorem 1 gives a sufficient condition of asymptotic stability, whereas uniformly choosing R_{ii} aims only to simplify the proof. This selection is reasonable due to the redundant design variables. To be specific, the controller can be tuned by Q_i and the interaction weights can be compensated for by R_{ij}. It is worth pointing out that inequality (26) holds only in the case of local best response u_{ij}^*. Because the best responses u_{ij}^* and u_j^* are incompatible, the local Nash does not lead to the global Nash.

It can be seen that the coupled HJ equations (19) and (21) involve a complex solution process and require extra communication of other followers' parameters. An off-policy RL algorithm is proposed to provide an online solution using system data.

Denote u_i^k and u_{ij}^k as the updated controllers and V_i^k as the value function in the k^{th} learning

phase, and $u_i^{\#k}$ and $u_{ij}^{\#k}$ as the best response controllers related to V_i^k. According to the property of the Hamiltonian function deduced in Eq. (23), the following equation holds for any arbitrary (u_i, u_{ij}), $j \in \mathcal{N}_i$:

$$\begin{aligned} & H_i^k\left(\xi_i, u_i, u_{ij}, \nabla_{\xi_i} V_i^k\right) - H_i^k\left(\xi_i, u_i^{\#k}, u_{ij}^{\#k}, \nabla_{\xi_i} V_i^k\right) \\ =& -\frac{1}{2} \sum_{j \in \mathcal{N}_i}\left(u_{ij} - u_{ij}^{\#k}\right)^{\mathrm{T}} R_{ij}\left(u_{ij} - u_{ij}^{\#k}\right) \\ & + \frac{1}{2} d_i\left(u_i - u_i^{\#k}\right)^{\mathrm{T}} R_{ii}\left(u_i - u_i^{\#k}\right). \end{aligned} \tag{32}$$

Substituting the control tuples (u_i^0, u_{ij}^0) and (u_i^k, u_{ij}^k) into Eq. (32) yields

$$\begin{aligned} & (\nabla_{\xi_i} V_i^k)^{\mathrm{T}}\left(A\xi_i + d_i B u_i^0 - \sum_{j \in \mathcal{N}_i} w_{ij} B u_{ij}^0\right) + \frac{1}{2}\xi_i^{\mathrm{T}} Q_i \xi_i \\ & + \frac{1}{2} d_i (u_i^k)^{\mathrm{T}} R_{ii} u_i^k - \frac{1}{2} \sum_{j \in \mathcal{N}_i} \left(u_{ij}^k\right)^{\mathrm{T}} R_{ij} u_{ij}^k - d_i \left(u_i^{\#k}\right)^{\mathrm{T}} \\ & \cdot R_{ii}\left(u_i^0 - u_i^k\right) + \sum_{j \in \mathcal{N}_i} \left(u_{ij}^{\#k}\right)^{\mathrm{T}} R_{ij}\left(u_{ij}^0 - u_{ij}^k\right) = 0. \end{aligned} \tag{33}$$

Then the off-policy recursive equation can be derived by integrating Eq. (33) through time interval T as

$$\begin{aligned} & V_i^k\left(\xi_i(t+T)\right) - V_i^k\left(\xi_i(t)\right) \\ & + \int_t^{t+T} \sum_{j \in \mathcal{N}_i} \left(u_{ij}^{\#k}\right)^{\mathrm{T}} R_{ij}\left(u_{ij}^0 - u_{ij}^k\right) \mathrm{d}t \\ & - \int_t^{t+T} d_i \left(u_i^{\#k}\right)^{\mathrm{T}} R_{ii}\left(u_i^0 - u_i^k\right) \\ =& -\int_t^{t+T} r_i(\xi_i, u_i^k, u_{ij}^k) \mathrm{d}t. \end{aligned} \tag{34}$$

One can see that the left-hand side of Eq. (34) contains value function V_i^k and action pairs (u_{ij}^0, u_{ij}^k) and (u_i^0, u_i^k) which act as the action-dependent Q-functions of a CT system. Employ NNs to approximate the value function and the best response controllers as

$$\begin{cases} V_i\left(\xi_i(t)\right) = W_{i1}^{\mathrm{T}} \phi_i(\xi_i), \\ u_i^{\#k} = W_{i2}^{\mathrm{T}} \varphi_i(\xi_i), \\ u_{ij}^{\#k} = W_{ij2}^{\mathrm{T}} \varphi_{ij}(\xi_i), \end{cases} \tag{35}$$

where $W_{i1} \in \mathbb{R}^{p_1}$, $W_{i2} \in \mathbb{R}^{p_2}$, $W_{ij2} \in \mathbb{R}^{p_3}$ are the weights, and $\phi_i(\xi_i) \in \mathbb{R}^{p_1}$, $\varphi_i(\xi_i) \in \mathbb{R}^{p_2}$, $\varphi_{ij}(\xi_i) \in \mathbb{R}^{p_3}$ represent the NN basis functions. Then write the cross terms in the linear form as

$$\begin{cases} (u_i^{\#k})^{\mathrm{T}} R_{ii}\left(u_i^0 - u_i^k\right) \\ = \left[\left(u_i^0 - u_i^k\right)^{\mathrm{T}} R_{ii} \otimes \varphi_i(\xi_i)\right] \mathrm{vec}(W_{i2}), \\ (u_{ij}^{\#k})^{\mathrm{T}} R_{ij}\left(u_{ij}^0 - u_{ij}^k\right) \\ = \left[\left(u_{ij}^0 - u_{ij}^k\right)^{\mathrm{T}} R_{ij} \otimes \varphi_{ij}(\xi_i)\right] \mathrm{vec}(W_{ij2}). \end{cases} \tag{36}$$

Then the approximate error of Eq. (34) is given by

$$e_i = \Phi_i^{\mathrm{T}}(\xi_i) W_i + \int_t^{t+T} r_i\left(\xi_i, u_i^k, u_{ij}^k\right) \mathrm{d}t, \tag{37}$$

where

$$\begin{cases} W_i = \left[W_{i1}^{\mathrm{T}}, \mathrm{vec}(W_{i2}^{\mathrm{T}}), \mathrm{vec}(W_{ij2}^{\mathrm{T}})\right], \\ \Phi_i = [\Phi_{i1}, \Phi_{i2}, \Phi_{i3}], \\ \Phi_{i1} = \phi_i(\xi_i(t+T)) - \phi_i(\xi_i(t)), \\ \Phi_{i2} = -\int_t^{t+T} d_i\left[\left(u_i^0 - u_i^k\right)^{\mathrm{T}} R_{ii} \otimes \varphi_i(\xi_i)\right] \mathrm{d}t, \\ \Phi_{i3} = \int_t^{t+T}\left[\left(u_{ij}^0 - u_{ij}^k\right)^{\mathrm{T}} R_{ij} \otimes \varphi_{ij}(\xi_i)\right] \mathrm{d}t. \end{cases} \tag{38}$$

The NN update law is designed by reducing the approximation error $E_i = \frac{1}{2} e_i^2$ using the gradient-decent method as

$$\begin{cases} \dot{W}_i(t) = -\alpha_i \frac{\Phi_i(\xi_i) S_i(t)}{\left(1 + \Phi_i^{\mathrm{T}}(\xi_i)\Phi_i(\xi_i)\right)^2}, \\ S_i(t) = W_i(\Phi_i(\xi_i)) + \int_t^{t+T} r_i\left(\xi_i, u_i^k, u_{ij}^k\right) \mathrm{d}t. \end{cases} \tag{39}$$

The detailed learning procedure is summarized in Algorithm 1.

Remark 4 The two-loop structure of Algorithm 1 has the same convergence property as that of the policy iteration method; the inner loop equivalently solves the Bellman equation (34) and the outer loop updates the controller to its optimal solution. In practice, Eq. (39) can be launched with an independent rate higher than the data collection rate. The algorithm combines the Q-learning mechanism and the integral RL for a CT system. By deriving the action-dependent value function according to Eq. (34), the model-free and off-policy properties are guaranteed compared with other methods. Thus, the controller pair to be updated, $(u_i^{\#k}, u_{ij}^{\#k})$, and the input pair driving the system, (u_i^0, u_{ij}^0), can be different. The collected data can be stored in a data buffer and reused during the learning process, which helps improve the convergence with higher data efficiency. In the learning phase, each follower can

Algorithm 1 Data-driven solution of the differential game using off-policy RL

1: Initialize a control tuple $(u_1^0, u_2^0, \cdots, u_N^0)$ to drive the MASs. For each agent i, set $u_{ij}^0 = u_j^0$ as the j^{th} local controller.
2: Initialize the online local iterative control tuple $(u_1^k, u_2^k, \cdots, u_N^k)$, and set $k = 1$. Randomly choose the NN weights W_{i1}, W_{i2}, W_{ij2} and the activate functions ϕ_i, φ_i, φ_{ij} with a proper scale for each agent.
3: **loop**
4: Collect system data $\xi_i(t)$ and (u_i^0, u_{ij}^0), calculate Φ_{i1}, Φ_{i2}, Φ_{i3} during time intervals $[t_s, t_s + T]$ for $s = 1$ to N_m, and store the data in a buffer.
5: **loop**
6: Randomly choose a set of data from the buffer, and calculate $\int_t^{t+T} r_i(\xi_i, u_i^k, u_{ij}^k)\mathrm{d}t$ as the index function return accordingly.
7: Update the NN weights online using Eq. (39) until the inner loop converges.
8: **end loop**
9: Update the iterative controller with $u_i^{k+1} = u_i^{\#k}$.
10: Set $k = k + 1$, and iterate until the outer loop converges.
11: **end loop**
12: Set $u_i^* = u_i^k$ as the local best response solution.

be initialized with an arbitrary stable local or distributed controller recorded as u_i^0. This avoids the contradiction between u_j and u_{ij}.

3.3 Global Nash solution of the modified differential game using RL

To tackle the unsolvable problem in Section 3.1 and the local Nash restriction in Section 3.2, a general index function for a differential graphical game is proposed to achieve both distributed and global Nash control in this subsection.

Define the following modified index function in a general case:

$$\begin{aligned} &r_i\left(\xi_i, u_i, u_{-i}\right) \\ =&\frac{1}{2}\left(d_i u_i^{\mathrm{T}} R_{ii} u_i + \sum_{j\in\mathcal{N}_i} u_j^{\mathrm{T}} R_{ij} u_j + \sum_{j\in\mathcal{N}_i} \xi_{ij}^{\mathrm{T}} \tilde{Q}_i \xi_{ij}\right), \end{aligned} \tag{40}$$

where $\xi_{ij} = [\xi_i, \xi_j]^{\mathrm{T}}$ and $\tilde{Q}_i = \begin{bmatrix} d_i^{-1} Q_i & S_{ij} \\ S_{ij}^{\mathrm{T}} & Q_{ij} \end{bmatrix}$.

S_{ij} and Q_{ij} are parameter matrices chosen to decouple the differential game and are to be determined later. Compared with earlier results, one can see from Eq. (40) that the neighboring error ξ_j of the j^{th} agent is included in the quadratic form, which gives a more general case.

Theorem 2 Consider the differential game of MASs with dynamics (1) and (2) and index function (40). Under the sufficient condition that $\tilde{Q}_i$ is given with the following elements:

$$\begin{cases} Q_{ij} = -\bar{P}_j^* B R_{jj}^{-1} R_{ij} R_{jj}^{-1} B^{\mathrm{T}} \bar{P}_j^*, \\ S_{ij} = -\bar{P}_i^* B R_{jj}^{-1} B^{\mathrm{T}} \bar{P}_j^*, \end{cases} \tag{41}$$

the value function of each follower can be decoupled in a distributed form $\bar{V}_i = \frac{1}{2}\xi_i^{\mathrm{T}} \bar{P}_i \xi_i$, while the corresponding distributed controller $u_i^* = -R_{ii}^{-1} B^{\mathrm{T}} \bar{P}_i^* \xi_i$ maintains two properties: (1) global asymptotic stability for synchronization control and (2) achievement of the global Nash equilibrium.

Proof Similarly, the Hamiltonian function can be rewritten as

$$\begin{aligned} 0 =&\bar{H}_i\left(\xi_i, u_i, u_{-i}, \nabla_{\xi_i}\bar{V}_i\right) \\ \triangleq&(\nabla_{\xi_i}\bar{V}_i)^{\mathrm{T}}\left(A\xi_i + d_i B u_i - \sum_{j\in\mathcal{N}_i} w_{ij} B u_j\right) \\ &+ \frac{1}{2} d_i u_i^{\mathrm{T}} R_{ii} u_i + \frac{1}{2}\sum_{j\in\mathcal{N}_i} u_j^{\mathrm{T}} R_{ij} u_j \\ &+ \frac{1}{2}\sum_{j\in\mathcal{N}_i}\left(\xi_i^{\mathrm{T}} Q_i \xi_i + \xi_j^{\mathrm{T}} Q_{ij} \xi_j + 2\xi_i^{\mathrm{T}} S_{ij} \xi_j\right). \end{aligned} \tag{42}$$

The value function corresponding to $\bar{J}_i$ and $\bar{H}_i$ is supposed to be $\nabla_{\xi_i}\bar{V} = \bar{P}_i \xi_i$. Then the distributed controller in the best response form satisfies the stationary condition:

$$u_i^{\#} = -R_{ii}^{-1} B^{\mathrm{T}} \bar{P}_i \xi_i. \tag{43}$$

Substituting Eq. (43) into Eq. (42) and rearranging the terms give the HJ equation:

$$\begin{aligned} &\xi_i^{\mathrm{T}}\left[\bar{P}_i A + A^{\mathrm{T}}\bar{P}_i - d_i \bar{P}_i B R_{ii}^{-1} B^{\mathrm{T}} \bar{P}_i + d_i Q_i\right]\xi_i \\ &+ 2\xi_j^{\mathrm{T}}\left(\sum_{j\in\mathcal{N}_i} \bar{P}_j B R_{jj}^{-1} R_{ij} R_{jj}^{-1} B^{\mathrm{T}} \bar{P}_j + \sum_{j\in\mathcal{N}_i} Q_{ij}\right)\xi_j \\ &+ 2\xi_i^{\mathrm{T}}\left(\sum_{j\in\mathcal{N}_i} S_{ij} + \sum_{j\in\mathcal{N}_i} w_{ij} \bar{P}_i B R_{jj}^{-1} B^{\mathrm{T}} \bar{P}_j\right)\xi_j = 0. \end{aligned} \tag{44}$$

Apparently, when extra parameter matrices are chosen by Eq. (41), the coupled terms can be eliminated and Eq. (44) is equivalent to the local algebra Riccati equation:

$$\bar{P}_i^* A + A^{\mathrm{T}} \bar{P}_i^* - d_i \bar{P}_i^* B R_{ii}^{-1} B^{\mathrm{T}} \bar{P}_i^* + Q_i = 0. \tag{45}$$

This indicates that the HJ equation is solvable compared with Eq. (17) and that the solution

controller is in a distributed form. Note that the controller $u_i^* = -R_{ii}^{-1}B^{\mathrm{T}}\bar{P}_i^*\xi_i$ for the local agent has the same form as in Section 3.2. The stability analysis and condition are similar to those in the local case and are omitted here.

Consider the Hamiltonian function $\bar{H}_i^*$ related to controllers u_i^* and u_{-i}^*, which satisfies

$$\begin{aligned}&\bar{H}_i^*(\xi_i,\xi_{-i},u_i,u_{-i},\bar{P}_i^*)-\bar{H}_i^*(\xi_i,\xi_{-i},u_i^*,u_{-i}^*,\bar{P}_i^*)\\=&\frac{1}{2}d_i(u_i-u_i^*)^{\mathrm{T}}R_{ii}\left(u_i-u_i^*\right)\\&-\frac{1}{2}\sum_{j\in\mathcal{N}_i}w_{ij}\left(u_j-u_j^*\right)^{\mathrm{T}}R_{ij}\left(u_j-u_j^*\right)\\&+\sum_{j\in\mathcal{N}_i}w_{ij}(u_i^*R_{ii}-u_j^*R_{ij})^{\mathrm{T}}\left(u_j-u_j^*\right).\end{aligned}\tag{46}$$

Note that $\lim\limits_{t\to\infty}\bar{V}_i\left(\xi_i(t)\right)=\lim\limits_{t\to\infty}\bar{V}_i(0)=0$. The index function can be derived similar to Eq. (24) as

$$\begin{aligned}&\bar{J}_i^*(\xi_i,\xi_{-i},u_i,u_{-i})\\=&\int_0^\infty\bar{H}_i^*\left(\xi_i,\xi_{-i},,u_i,u_{-i},\bar{P}_i^*\right)\mathrm{d}t+\bar{V}_i^*\left(\xi_i(0)\right)\\=&-\frac{1}{2}\int_0^\infty\sum_{j\in\mathcal{N}_i}w_{ij}\left(u_j-u_j^*\right)^{\mathrm{T}}R_{ij}\left(u_j-u_j^*\right)\mathrm{d}t\\&+\frac{1}{2}\int_0^\infty d_i(u_i-u_i^*)^{\mathrm{T}}R_{ii}\left(u_i-u_i^*\right)\mathrm{d}t+\bar{V}_i^*\left(\xi_i(0)\right)\\&+\int_0^\infty\sum_{j\in\mathcal{N}_i}w_{ij}(u_i^*R_{ii}-u_j^*R_{ij})^{\mathrm{T}}\left(u_j-u_j^*\right)\mathrm{d}t.\end{aligned}\tag{47}$$

Substituting the control tuples $\left(u_i,u_{-i}^*\right)$ and $\left(u_i^*,u_{-i}^*\right)$ in respective index functions yields

$$\begin{aligned}&\bar{J}_i^*(\xi_i,\xi_{-i},u_i,u_{-i}^*)=\bar{J}_i^*(\xi_i,\xi_{-i},u_i^*,u_{-i}^*)\\&+\frac{1}{2}\int_0^\infty d_i(u_i-u_i^*)^{\mathrm{T}}R_{ii}\left(u_i-u_i^{\#}\right)\mathrm{d}t\\\geq&\bar{J}_i^*(\xi_i,\xi_{-i},u_i^*,u_{-i}^*).\end{aligned}\tag{48}$$

Therefore, the optimal controllers given by Eq. (45) build the global Nash equilibrium. This completes the proof.

Since the construction of index functions (40) and (41) involves the equilibrium value $\bar{P}_i^*$, whereas the RL method requires online feedback data from index functions, a contradictory problem arises. Note that the final global Nash solution is equivalent to that in the uncoupled case. To this end, the RL procedure can be constructed in a parallel way as the single-agent case while regarding the neighbor's input as a known vector. The NN approximations are given as $\bar{V}_i\left(\xi_i(t)\right)=\bar{W}_{i1}^{\mathrm{T}}\phi_i(\xi_i)$ and $\bar{u}_i^{\#k}=\bar{W}_{i2}^{\mathrm{T}}\varphi_i(\xi_i)$. The NN approximation error is given by

$$e_i=\Phi_i^{\mathrm{T}}(\xi_i)\bar{W}_i+\int_t^{t+T}r_{mi}\left(\xi_i,\bar{u}_i^k\right)\mathrm{d}t,\tag{49}$$

where

$$\begin{cases}r_{mi}\left(\xi_i,\bar{u}_i^k\right)=\dfrac{1}{2}\left(\xi_i^{\mathrm{T}}Q_i\xi_i+d_i\left(\bar{u}_i^k\right)^{\mathrm{T}}R_{ii}\bar{u}_i^k\right),\\\bar{W}_i=\left[\bar{W}_{i1}^{\mathrm{T}},\left(\mathrm{vec}(\bar{W}_{i2})\right)^{\mathrm{T}}\right],\\\Phi_i=[\Phi_{i1},\Phi_{i2}],\\\Phi_{i1}=\phi_i(\xi_i(t+T))-\phi_i(\xi_i(t)),\\\Phi_{i2}=-\int_t^{t+T}\left[\left(\bar{u}_i^0-\bar{u}_i^k\right)^{\mathrm{T}}R_{ii}\otimes\varphi_i(\xi_i)\right]\mathrm{d}t,\\\bar{u}_i^0=u_i^0-\sum\limits_{j\in\mathcal{N}_i}d_i^{-1}u_j^0.\end{cases}\tag{50}$$

The remaining steps are similar to those in Algorithm 1 and are omitted here.

Remark 5 In summary, the local best response solution considers interactions virtually in the worst case, which indicates $\mathcal{L}_2$ stability in inequality (28). The global Nash solution guarantees the unified equilibrium in Eq. (45) by directly handling real interactions. Although the above two cases are derived under divergent bases, they can be solved uniformly and systematically by the proposed RL-based method because of the same formulation structure.

4 Simulation results

Consider a multi-agent system consisting of one leader labeled 0 and four followers labeled from 1 to 4 with the interaction topology demonstrated in Fig. 1. To include as many attributes as possible, the followers are designed with different configurations. The number of neighboring players and in-degree are given by pairs as $(0,1)$, $(1,2)$, $(1,1)$, $(2,2)$.

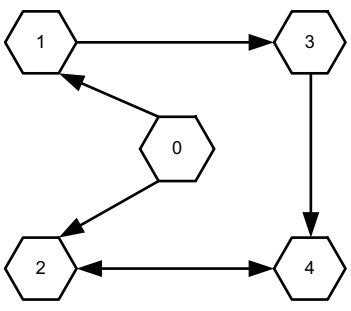

Fig. 1 Communication topology of a multi-agent system (MAS)

The dynamics of MASs is chosen to be naturally unstable as

$$A=\begin{bmatrix}0 & 1\\ 2 & -1\end{bmatrix}, B=\begin{bmatrix}0\\ 1\end{bmatrix}. \tag{51}$$

1. Solution in the local best response case

Define the index functions of the local best response case with parameters set as $Q_i = 5I_2$, $R_{ii} = 1$, $R_{24} = 3$, $R_{42} = 2$, $R_{43} = 3$, $R_{31} = 3$. The activation functions can be chosen by $\phi_i^{\mathrm{T}}(\xi_i) = \left[\xi_{i1}^2, \xi_{i2}^2, \xi_{i1}\xi_{i2}\right]$, $\varphi_i^{\mathrm{T}} = \varphi_{ij}^{\mathrm{T}} = [\xi_{i1}, \xi_{i2}]$ for linear systems. In the RL procedure, let the MASs be persistently excited by an off-policy signal of amplitude 10 and be composed of random frequencies within 30 Hz. The learning rate is set as $\alpha_i = 30$. The convergence curves of the weights for each agent are illustrated in Fig. 2.

The theoretical results of the nominal controllers are listed as follows to show the validity of the RL method and local Nash properties:

$$\begin{aligned}
&K_1^* = [-4.995, -2.998], K_{\mathrm{opt1}} = [-5.000, -3.000],\\
&K_2^* = [-3.303, -2.106], K_{\mathrm{opt2}} = [-3.307, -2.107],\\
&K_3^* = [-7.047, -4.053], K_{\mathrm{opt3}} = [-7.062, -4.062],\\
&K_4^* = [-4.390, -2.683], K_{\mathrm{opt4}} = [-4.402, -2.688].
\end{aligned} \tag{52}$$

One can see that even if the parameters are the same, divergent graph settings bring different solutions. The synchronization error of each agent is given in Fig. 3a to illustrate the stability of local best response based controllers. To further verify the local Nash property stated in Eq. (34), the $\mathcal{L}_2$ gains $\frac{\int_0^t \left(\xi_i^{\mathrm{T}} Q_i \xi_i + d_i (u_i^*)^{\mathrm{T}} R_{ii} u_i^*\right) \mathrm{d}\tau + V(\xi_i(t))}{\int_0^t \sum_{j\in\mathcal{N}_i} u_{ij}^{\mathrm{T}} R_{ij} u_{ij} \mathrm{d}\tau + V(\xi_i(0))}$ for agents 2,

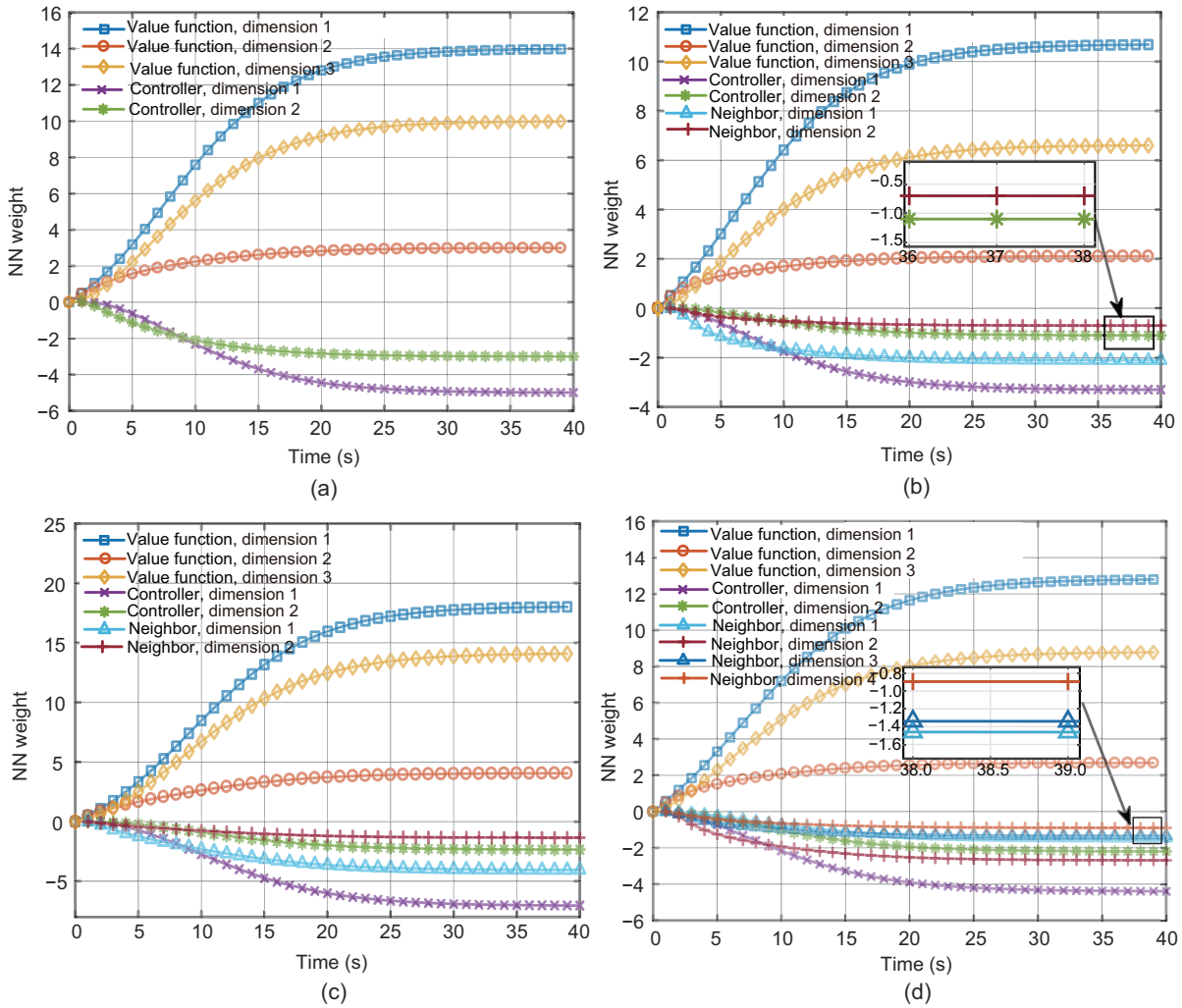

Fig. 2 Convergence curves of the neural network (NN) weights in the local best response case: (a) agent 1; (b) agent 2; (c) agent 3; (d) agent 4

3, and 4 are suppressed under threshold 1.0 as shown in Fig. 4a. Agent 1 is omitted here because it has no neighbor in the follower set.

2. Solution in the global Nash case

The index functions of the global Nash case are set with $Q_i = 5I_2, R_{ii} = 1, R_{24} = R_{31} = R_{42} = R_{43} = 0.1$. Conduct the RL procedure in the same manner as in the single-agent case, where the coupled neighboring control inputs can be regarded as off-policy signals. Taking agents 1 and 3 as examples, the convergence of learning procedures is demonstrated by the error curves of NN weights to their target values in Fig. 5. The distributed global Nash controller and the compensational terms of modified index functions can be given as follows based on the learning results:

$$\bar{K}_1^* = [-4.993, -2.998], \bar{K}_2^* = [-2.867, -1.870],$$
$$\bar{K}_3^* = [-4.994, -2.998], \bar{K}_4^* = [-2.867, -1.870],$$

$$S_{31} = \begin{bmatrix} -24.9 & -14.9 \\ -14.9 & -8.98 \end{bmatrix}, S_{43} = \begin{bmatrix} -14.3 & -8.59 \\ -9.34 & -5.60 \end{bmatrix},$$

$$S_{24} = S_{42} = \begin{bmatrix} -8.22 & -5.36 \\ -5.36 & -3.50 \end{bmatrix},$$

$$Q_{24} = Q_{42} = \begin{bmatrix} -0.822 & -0.536 \\ -0.536 & -0.349 \end{bmatrix},$$

$$Q_{31} = Q_{43} = \begin{bmatrix} -2.494 & -1.496 \\ -1.496 & -0.898 \end{bmatrix}.$$

Similarly, the stabilities in the global Nash case are verified by demonstrating the synchronization error in Fig. 3b.

Moreover, Fig. 4b shows the evolution of index function (40) for global controllers in Section 3.3. The solid lines represent the evolution curves generated by the global Nash equilibrium controllers. The dashed lines denote the results of the above controllers with random biases. As compared in the diagram, one can see that the global Nash controller for each agent gives a smaller index value than the biased controller, which conforms to the global Nash equilibrium.

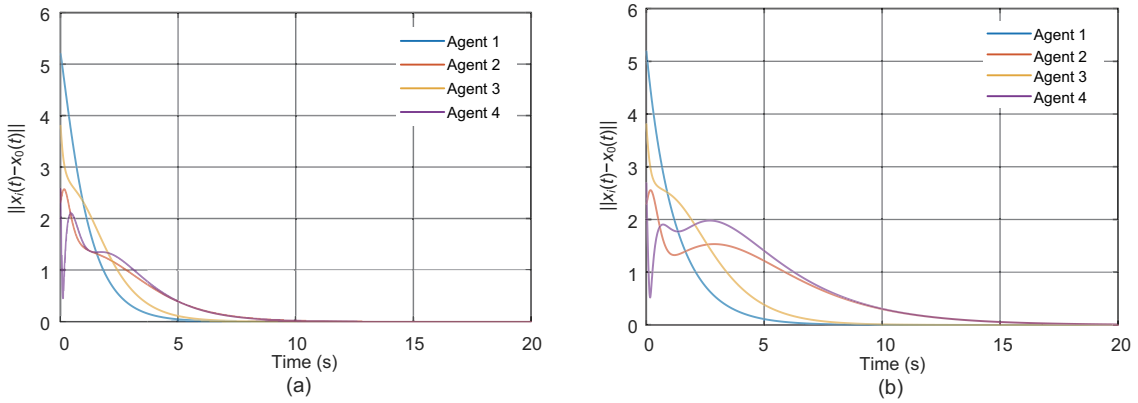

Fig. 3 Multi-agent system (MAS) synchronization errors in the local best response case (a) and the global Nash case (b)

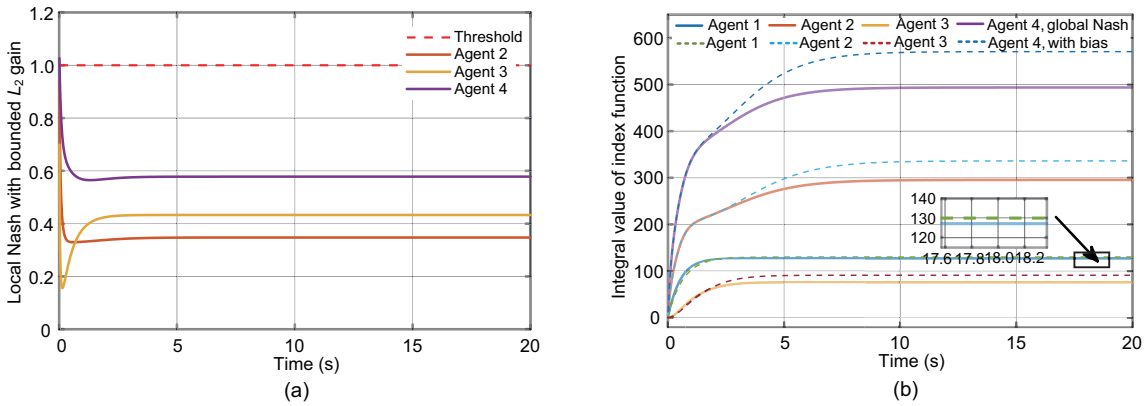

Fig. 4 Evolution of index functions in the local best response case (a) and the global Nash case (b)

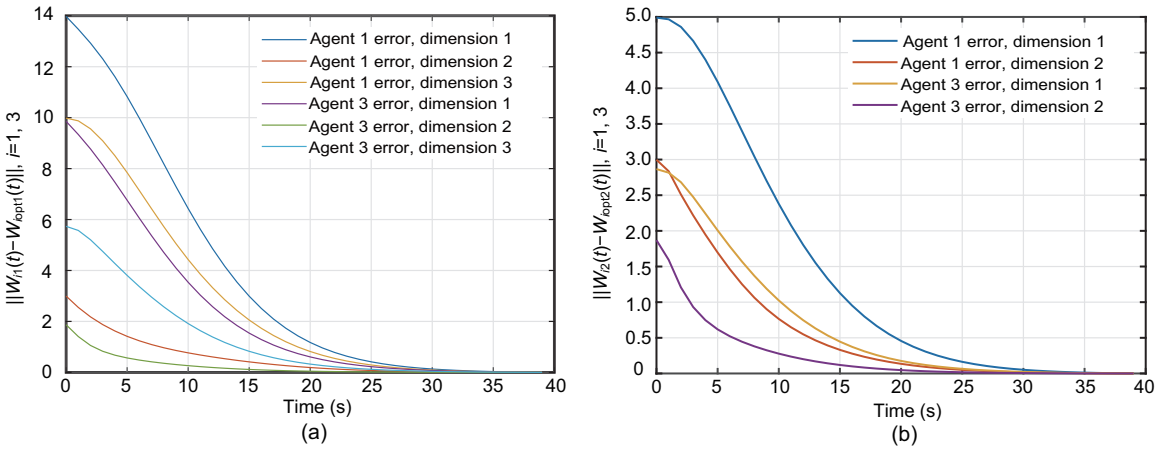

Fig. 5 Convergence error curves of neural network (NN) weights in the global Nash case: (a) value function weight error; (b) controller weight error

5 Conclusions

This paper studied the cooperative synchronization control of MASs from a differential game perspective. The solution was highly coupled and may not exist in general cases. The local best response controller and global Nash controller with modified index functions were successively investigated to deal with the coupling issues. An off-policy RL method was proposed to solve the problem online in a data-driven manner. The configurations of graph topology and system dynamics affected the solutions. Thus, to extend the results to more complex scenarios, such as the existence of switching topologies and disturbances, is quite a significant direction for future research.

Contributors

Yu SHI designed the research, conducted the simulations, and drafted the paper. Yongzhao HUA and Jianglong YU helped organize the paper. Xiwang DONG and Zhang REN revised and finalized the paper.

Compliance with ethics guidelines

Yu SHI, Yongzhao HUA, Jianglong YU, Xiwang DONG, and Zhang REN declare that they have no conflict of interest.

References

Abouheaf MI, Lewis FL, Vamvoudakis KG, et al., 2014. Multi-agent discrete-time graphical games and reinforcement learning solutions. *Automatica*, 50(12):3038-3053. https://doi.org/10.1016/j.automatica.2014.10.047

Başar T, Olsder GJ, 1982. Dynamic Noncooperative Game Theory. Academic Press, New York, USA.

Dong XW, Xi JX, Lu G, et al., 2014. Formation control for high-order linear time-invariant multiagent systems with time delays. *IEEE Trans Contr Netw Syst*, 1(3): 232-240. https://doi.org/10.1109/TCNS.2014.2337972

Lewis FL, Vrabie DL, Syrmos VL, 2012. Optimal Control. John Wiley & Sons, Hoboken, NJ, USA.

Li JN, Modares H, Chai TY, et al., 2017. Off-policy reinforcement learning for synchronization in multiagent graphical games. *IEEE Trans Neur Netw Learn Syst*, 28(10):2434-2445. https://doi.org/10.1109/TNNLS.2016.2609500

Liu MS, Wan Y, Lopez VG, et al., 2021. Differential graphical game with distributed global Nash solution. *IEEE Trans Contr Netw Syst*, 8(3):1371-1382. https://doi.org/10.1109/TCNS.2021.3065654

Lopez VG, Lewis FL, Wan Y, et al., 2020. Stability and robustness analysis of minmax solutions for differential graphical games. *Automatica*, 121:109177. https://doi.org/10.1016/j.automatica.2020.109177

Modares H, Lewis FL, 2014. Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. *IEEE Trans Autom Contr*, 59(11):3051-3056. https://doi.org/10.1109/TAC.2014.2317301

Modares H, Lewis FL, Jiang ZP, 2015. H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning. *IEEE Trans Neur Netw Learn Syst*, 26(10):2550-2562. https://doi.org/10.1109/TNNLS.2015.2441749

Mu CX, Zhen N, Sun CY, et al., 2017. Data-driven tracking control with adaptive dynamic programming for a class of continuous-time nonlinear systems. *IEEE Trans Cybern*, 47(6):1460-1470. https://doi.org/10.1109/TCYB.2016.2548941

Olfati-Saber R, Murray RM, 2004. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Trans Autom Contr*, 49(9):1520-1533. https://doi.org/10.1109/TAC.2004.834113

Peng QY, Low SH, 2018. Distributed optimal power flow algorithm for radial networks, I: balanced single phase

case. *IEEE Trans Smart Grid*, 9(1):111-121. https://doi.org/10.1109/TSG.2016.2546305

Qian YY, Liu MS, Wan Y, et al., 2021. Distributed adaptive Nash equilibrium solution for differential graphical games. *IEEE Trans Cybern*, early access. https://doi.org/10.1109/TCYB.2021.3114749

Qin JH, Gao HJ, Zheng WX, 2011. Second-order consensus for multi-agent systems with switching topology and communication delay. *Syst Contr Lett*, 60(6):390-397. https://doi.org/10.1016/j.sysconle.2011.03.004

Ren W, Beard RW, 2005. Consensus seeking in multiagent systems under dynamically changing interaction topologies. *IEEE Trans Autom Contr*, 50(5):655-661. https://doi.org/10.1109/TAC.2005.846556

Sun C, Ye MJ, Hu GQ, 2017. Distributed time-varying quadratic optimization for multiple agents under undirected graphs. *IEEE Trans Autom Contr*, 62(7):3687-3694. https://doi.org/10.1109/TAC.2017.2673240

Sutton RS, Barto AG, 1998. Reinforcement Learning: an Introduction. MIT Press, Cambridge, MA, USA.

Tamimi A, Lewis FL, Abu-Khalaf M, 2008. Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof. *IEEE Trans Syst Man Cybern B Cybern*, 38(4):943-949. https://doi.org/10.1109/TSMCB.2008.926614

Vamvoudakis KG, Lewis FL, 2011. Multi-player non-zero-sum games: online adaptive learning solution of coupled Hamilton-Jacobi equations. *Automatica*, 47(8):1556-1569. https://doi.org/10.1016/j.automatica.2011.03.005

Vamvoudakis KG, Lewis FL, Hudas GR, 2012. Multi-agent differential graphical games: online adaptive learning solution for synchronization with optimality. *Automatica*, 48(8):1598-1611. https://doi.org/10.1016/j.automatica.2012.05.074

Wang MY, Wang ZJ, Talbot J, et al., 2021. Game-theoretic planning for self-driving cars in multivehicle competitive scenarios. *IEEE Trans Robot*, 37(4):1313-1325. https://doi.org/10.1109/TRO.2020.3047521

Wang W, Chen X, Fu H, et al., 2020. Model-free distributed consensus control based on actor-critic framework for discrete-time nonlinear multiagent systems. *IEEE Trans Syst Man Cybern Syst*, 50(11):4123-4134. https://doi.org/10.1109/tsmc.2018.2883801

Wen GH, Yu XH, Liu ZW, 2021. Recent progress on the study of distributed economic dispatch in smart grid: an overview. *Front Inform Technol Electron Eng*, 22(1):25-39. http://doi.org/10.1631/FITEE.2000205

Yang T, Yi XL, Wu JF, et al., 2019. A survey of distributed optimization. *Ann Rev Contr*, 47:278-305. https://doi.org/10.1016/j.arcontrol.2019.05.006

Yang YJ, Wan Y, Zhu JH, et al., 2021. H_∞ tracking control for linear discrete-time systems: model-free Q-learning designs. *IEEE Contr Syst Lett*, 5(1):175-180. https://doi.org/10.1109/LCSYS.2020.3001241

Ye MJ, Hu GQ, Lewis FL, 2018. Nash equilibrium seeking for N-coalition noncooperative games. *Automatica*, 95:266-272. https://doi.org/10.1016/j.automatica.2018.05.020

Ye MJ, Hu GQ, Lewis FL, et al., 2019. A unified strategy for solution seeking in graphical N-coalition noncooperative games. *IEEE Trans Autom Contr*, 64(11):4645-4652. https://doi.org/10.1109/TAC.2019.2901820

Zhang HG, Jiang H, Luo YH, et al., 2017. Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method. *IEEE Trans Ind Electron*, 64(5):4091-4100. https://doi.org/10.1109/TIE.2016.2542134

Zhao DB, Xia ZP, Wang D, 2015. Model-free optimal control for affine nonlinear systems with convergence analysis. *IEEE Trans Autom Sci Eng*, 12(4):1461-1468. https://doi.org/10.1109/TASE.2014.2348991

Zhao JG, 2020. Neural networks-based optimal tracking control for nonzero-sum games of multi-player continuous-time nonlinear systems via reinforcement learning. *Neurocomputing*, 412:167-176. https://doi.org/10.1016/j.neucom.2020.06.083

Zheng WY, Wu WC, Zhang BM, et al., 2016. A fully distributed reactive power optimization and control method for active distribution networks. *IEEE Trans Smart Grid*, 7(2):1021-1033. https://doi.org/10.1109/TSG.2015.2396493

Zhu QY, Başar T, 2015. Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems. *IEEE Contr Syst*, 35(1):46-65. https://doi.org/10.1109/MCS.2014.2364710