2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Off-Policy Reinforcement-Learning Algorithm to Solve Minimax Games on Graphs

Victor G. Lopez[1], Kyriakos G. Vamvoudakis[2], Yan Wan[1], and Frank L. Lewis[3]

Abstract— **In this paper, we formulate and find distributed minimax strategies as an alternative to Nash equilibrium strategies for multi-agent systems communicating via graph topologies, i.e., communication restrictions are taken into account for the distributed design. We provide the conditions that guarantee the existence of the minimax solutions in the game. Finally, we present an off-policy Integral Reinforcement Learning (IRL) method to solve the minimax Riccati equations and determine the optimal and worst-case policies of the agents by measuring data along the system trajectories.**

Index Terms— **Games, integral reinforcement learning, graphs.**

I. Introduction

Analyzing the performance and the decision-making processes of groups of dynamical systems has become indispensable as the number of autonomous systems increases in industrial and urban areas. Applications of multi-agent systems with individual goals include intelligent transportation systems, wireless sensor networks and machine interactions in industrial processes. In most practical applications, a system must use incomplete information available from the environment to determine her best possible strategy to achieve the global goals. Differential graphical games is a branch of game theory that studies the interplay of a set of dynamical systems with limited sensing capabilities [25], [26]. Each player of the game, regarded as an *agent*, observes the state information of only a subset of other players, i.e., her neighbors. The agents are said to form a Nash equilibrium when all of them use their best policies simultaneously. [8], [4].

Related Work

There is an extensive research on cooperative control of networked systems [15], [20], [23], [21], [11], [24], [7], [16]. Game-theoretic approaches have been formulated to provide optimality, resilience, and robustness to the cooperative ([26], [28]) and noncooperative ([22], [5]) behaviors of the agents. Every admissible solution for a graphical game requires the use of distributed control policies by the agents. This means that the agents are allowed to use only local information received through the communication graph to design their strategies. The distributed-policy requirement, however, makes Nash equilibrium generally unattainable among the agents. This fact can be intuitively explained by noticing that an agent needs to know her neighbors' best strategies to determine her own best response towards them, but the neighbors' best policies are unknown. Thus, such information restriction imposed by the graph topology prevents the multi-agent system from reaching a Nash equilibrium.

The unattainability of Nash equilibrium in graphical games can be addressed by leveraging alternative solution concepts. In this paper, we analyze the behavior of the agents in a communication graph when they use their minimax strategies [25], [3] to achieve their goals, whereas, each agent develops best policies towards the worst possible behavior from her neighbors. From the perspective of an individual agent, the resulting formulation of this graphical game is the same as an H_∞ control problem [29]. It is known that, the H_∞ problem can be solved as as zero-sum game where the control input acts as a minimizing player and an adversarial/disturbance input acts as a maximizing one [14], [3], [13], [17], [6].

An additional requirement for a practical solution of graphical games is the consideration of uncertainties in the system dynamics. The usual minimax and H_∞ designs require complete knowledge of the physics of the system. Different RL algorithms have been proposed to solve multi-agent problems with partial or without any information about the system dynamics [1], [12], [2], [10]. Off-policy reinforcement learning algorithms have been proposed [9], [19], [18] to provide a solution to the minimax problem without any information of the agents.

Contributions: The main contributions of this paper are as follows. Minimax strategies are developed to solve non-adversarial differential graphical games. Different from the Nash equilibrium solution, the minimax strategies are proven to provide distributed policies under minor conditions in the system dynamics and the performance functions. Finally, an off-policy reinforcement learning algorithm is designed to solve the minimax problem without any knowledge of the system dynamics.

Structure: The paper is structured as follows. Section II presents an overview of graphical games and the Nash equilibrium solution. In Section III, the minimax strategies problem is presented and its solution is obtained. An off-policy RL algorithm is presented in Section IV to solve the minimax control problem, and simulation results of this

[1] V. G. Lopez, Y. Wan are with the Department of Electrical Engineering, University of Texas at Arlington, TX 76010, USA, victor.lopezmejia@mavs.uta.edu, yan.wan@uta.edu

[2] K. G. Vamvoudakis is with The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, GA, USA, kyriakos@gatech.edu

[3] F. L. Lewis is with the Department of Electrical Engineering University of Texas at Arlington, USA, the State Key Laboratory of Synthetical Automation for Process Industries, China, and Northeastern University, China, lewis@uta.edu

This work was supported by ONR under grant number N00014-18-1-2221, and by NSF under grant numbers CPS-1851588 and ECCS-1839804.

978-1-7281-1398-2/19/$31.00 ©2019 IEEE

algorithm are presented in Section V.

Notation: The space $\mathcal{L}_2^n$ is defined as the set of all piecewise continuous functions $x : [0,\infty) \to \mathbb{R}^n$ such that

$$\|x\|_{\mathcal{L}_2} = \left(\int_0^\infty x^{\mathrm{T}}(t)x(t)\right)^{1/2} \mathrm{d}t < \infty,$$

i.e., the space $\mathcal{L}_2^n$ defines the set of all square-integrable functions $x(t)$. Define the inner-product in the space $\mathcal{L}_2^n[0,\infty)$ as

$$\langle x, y\rangle = \int_0^\infty x^{\mathrm{T}}(t)y(t)\mathrm{d}t \tag{1}$$

where $x,\, y \in \mathcal{L}_2^n[0,\infty)$.

II. Background and Problem Formulation

Consider a set of N agents (players) connected by a directed communication graph $\mathcal{G}_r = (V, E)$ where V is the set of nodes and E the set of edges. The edge weights of the graph are represented as a_{ij}, with $a_{ij} > 0$ if $(v_j, v_i) \in E$ and $a_{ij} = 0$ otherwise. The set of neighbors of node v_i is $\mathcal{N}_i = \{v_j : a_{ij} > 0\}$. The graph is assumed to have no self-loops, i.e., $a_{ii} = 0$ for all agents i. Define the graph adjacency matrix as $\mathcal{A} = [a_{ij}]$. The weighted in-degree of node i is defined as $d_i = \sum_{j=1}^N a_{ij}$, and the in-degree matrix of the graph is $D = \mathrm{diag}\{d_i\}$. The Laplacian matrix is finally defined as $L = D - \mathcal{A}$.

A. Agent Dynamics

Consider that the local dynamics of each agent i, $i = 1, ..., N$, are given by

$$\dot{x}_i = Ax_i + Bu_i,\ t \geq 0 \tag{2}$$

where $x_i(t) \in \mathbb{R}^n$ and $u_i \in \mathbb{R}^m$ is the state and the control input of agent i, respectively.

Define an additional agent, regarded as the leader or target node, with uncontrolled dynamics as,

$$\dot{x}_0 = Ax_0,\ t \geq 0, \tag{3}$$

where the eigenvalues of A have non-positive real parts. The communication links between the leader and the other agents is represented by the pinning gains $g_i \geq 0$, which must be non-zero for at least one agent.

The local synchronization error of agent i is thus defined to be

$$\delta_i = \sum_{j=1}^N a_{ij}\left(x_i - x_j\right) + g_i\left(x_i - x_0\right),$$

and the local error dynamics are

$$\begin{aligned}\dot{\delta}_i &= \sum_{j=1}^N a_{ij}\left(\dot{x}_i - \dot{x}_j\right) + g_i\left(\dot{x}_i - \dot{x}_0\right) \\ &= A\delta_i + \left(d_i + g_i\right)Bu_i - \sum_{j=1}^N a_{ij}Bu_j.\end{aligned} \tag{4}$$

Each agent i expresses her individual objectives in a local game by means of a cost function,

$$J_i := J_i\left(\delta_i, \delta_{-i}, u_i, u_{-i}\right),$$

where $J_i\left(\delta_i, \delta_{-i}, u_i, u_{-i}\right)$ is a positive definite scalar function of the variables expected to be minimized by agent i, and δ_{-i} and u_{-i} represent the local errors and control inputs of the neighbors of agent i, respectively. For synchronization games, the cost function

$$J_i = \int_0^\infty \left(\delta_i^{\mathrm{T}}Q_i\delta_i + u_i^{\mathrm{T}}R_iu_i + \sum_{j=1}^N a_{ij}u_j^{\mathrm{T}}R_{ij}u_j\right)\mathrm{d}t, \tag{5}$$

with $Q_i \succeq 0$, $R_i \succ 0$ and $R_{ij} \succeq 0$ is usually employed.

B. Nash Equilibrium

The best response of agent i given fixed neighboring policies u_{-i} is defined as the control policy $u_i^\star$ such that the inequality $J_i(\delta, u_i^\star, u_{-i}) \leq J_i(\delta, u_i, u_{-i})$ holds for all policies u_i.

A Nash equilibrium is achieved given that every agent plays her best response towards her neighbors, i.e.,

$$J_i\left(\delta, u_i^\star, u_{-i}^\star\right) \leq J_i\left(\delta, u_i, u_{-i}^\star\right),\ \forall i.$$

It is proven in [26] that the best response of agent i with cost functional (5) is given by

$$u_i^\star = -\frac{1}{2}\left(d_i + g_i\right)R_i^{-1}B^{\mathrm{T}}\nabla V_i\left(\delta_i\right), \tag{6}$$

with $\nabla V_i\left(\delta_i\right) := \frac{\partial V_i}{\partial \delta_i}$, where the functions $V_i(\delta_i) \geq 0$ solve the following Hamilton-Jacobi (HJ) equations

$$\begin{aligned}\delta_i^{\mathrm{T}}Q_i\delta_i + \nabla V_i^{\mathrm{T}}A\delta_i - \frac{(d_i+g_i)^2}{4}\nabla V_i^{\mathrm{T}}BR_i^{-1}B^{\mathrm{T}}\nabla V_i \\ + \frac{1}{4}\sum_{j=1}^N a_{ij}(d_j+g_j)^2\nabla V_j^{\mathrm{T}}BR_j^{-1}B^{\mathrm{T}}\nabla V_j \\ + \frac{1}{2}\sum_{j=1}^N a_{ij}(d_j+g_j)\nabla V_i^{\mathrm{T}}BR_j^{-1}B^{\mathrm{T}}\nabla V_j = 0.\end{aligned}$$

Suppose now that the value function has a quadratic form as follows,

$$V_i\left(\delta_i\right) = \delta_i^{\mathrm{T}}P_i\delta_i, \tag{7}$$

then the optimal policy of agent i is given by

$$u_i^\star = -\left(d_i + g_i\right)R_i^{-1}B^{\mathrm{T}}P_i\delta_i \tag{8}$$

which is distributed in the sense that it only uses local information δ_i and $P_i = P_i^{\mathrm{T}} \succ 0$ in (8) is the solution to the following coupled HJ equations

$$\begin{aligned}\delta_i^{\mathrm{T}}\left(Q_i + P_iA + A^{\mathrm{T}}P_i - (d_i+g_i)^2P_iBR_i^{-1}B^{\mathrm{T}}P_i\right)\delta_i \\ + \sum_{j=1}^N a_{ij}(d_j+g_j)^2\delta_jP_jBR_j^{-1}R_{ij}R_j^{-1}B^{\mathrm{T}}P_j\delta_j \\ + 2\sum_{j=1}^N a_{ij}(d_j+g_j)\delta_iP_iBR_j^{-1}B^{\mathrm{T}}P_j\delta_j = 0.\end{aligned} \tag{9}$$

Note now that the Nash equilibrium solution for the differential graphical games presents, however, a significant drawback. Because (9) must hold for all values of δ_i and

δ_j, then the matrices P_i and P_j, $j \in \mathcal{N}_i$, must solve simultaneously the matrix equations

$$\begin{aligned} & Q_i + P_i A + A^{\mathrm{T}} P_i - (d_i + g_i)^2 P_i B R_i^{-1} B^{\mathrm{T}} P_i = 0, \\ & P_j B R_j^{-1} R_{ij} R_j^{-1} B^{\mathrm{T}} P_j = 0, \\ & P_i B R_j^{-1} B^{\mathrm{T}} P_j = 0. \end{aligned} \tag{10}$$

Note that there are N sets of equations of the form (10) that need to be solved simultaneously. It is clear that these equations do not necessarily have positive definite solutions. This is an expected result due to the limited knowledge of the agents connected in the communication graph. Given that the agent i does not know the local information of her neighbors, then it cannot determine the best response to the game.

In the following section, *minimax strategies* are proposed as a practical alternative to the Nash equilibrium solution of the graphical games.

III. Minimax Strategies

Despite the lack of global information about the state of the agents, we can still expect the agent i to determine a best policy for the information it has available from her neighbors. Intuitively, minimax strategies are obtained when each agent prepares herself for the worst behavior of her neighbors. As it is shown below, the corresponding equations for the distributed minimax strategies are generally solvable for linear systems.

A. Formulation

Assume that agent i derives her minimax strategy by making the conservative assumption that the goal of her neighbors is to maximize her own performance index, i.e., J_i. The following definition formalizes such a concept.

Definition 1 (minimax strategies.): In a differential graphical game, the minimax strategy of agent i is given by

$$u_i^\star = \arg\min_{u_i} \max_{u_{-i}} J_i\left(\delta_i, u_i, u_{-i}\right).$$

□

Now the performance index that was defined in (5) needs to be modified. To this end, define the function

$$J_i = \int_0^\infty \left(\delta_i^{\mathrm{T}} Q_i \delta_i + \left(d_i + g_i\right) u_i^{\mathrm{T}} R_i u_i - \gamma^2 \sum_{j=1}^N a_{ij} u_j^{\mathrm{T}} R_j u_j \right) \mathrm{d}t \tag{11}$$

where $Q_i \succeq 0$, $R_i, R_j \succ 0$ and $\gamma \in \mathbb{R}^+$. To determine her minimax strategy, agent i assumes that the goal of her neighbors is to maximize her own performance index as given in (11).

Define now the Hamiltonian function associated with the cost index (11) as

$$H_i := \delta_i^{\mathrm{T}} Q_i \delta_i + \left(d_i + g_i\right) u_i^{\mathrm{T}} R_i u_i - \gamma^2 \sum_{j=1}^N a_{ij} u_j^{\mathrm{T}} R_j u_j + \nabla V_i^{\mathrm{T}}(\delta_i) \dot{\delta}, \tag{12}$$

with $\dot{\delta}_i$ given from (4). Assume now that the value function V_i has a quadratic form, i.e., (12) can be expressed as

$$H_i = \delta_i^{\mathrm{T}} Q_i \delta_i + \left(d_i + g_i\right) u_i^{\mathrm{T}} R_i u_i - \gamma^2 \sum_{j=1}^N a_{ij} u_j^{\mathrm{T}} R_j u_j + 2\delta_i^{\mathrm{T}} P_i \left(A\delta_i + \left(d_i + g_i\right) B u_i - \sum_{j=1}^N a_{ij} B u_j \right). \tag{13}$$

The optimal control policy for agent i is now obtained by using the stationary condition $\frac{\partial H_i}{\partial u_i} = 0$, that yields

$$u_i^\star = -R_i^{-1} B^{\mathrm{T}} P_i \delta_i. \tag{14}$$

Similarly, the worst-case policy of the neighbors of agent i can be obtained as

$$\upsilon_j^\star = \frac{1}{\gamma^2} R_j^{-1} B^{\mathrm{T}} P_i \delta_i. \tag{15}$$

Note that $\upsilon_j^\star$ is not necessarily the actual control policy employed by agent j, i.e., u_j.

The coupled HJ equations to be solved for the matrix P_i are finally obtained by substituting the policies (14) and (15) in (13). This procedure yields the following algebraic Riccati equations (ARE)

$$Q_i + P_i A + A^{\mathrm{T}} P_i - \left(d_i + g_i\right) P_i B R_i^{-1} B^{\mathrm{T}} P_i + \frac{1}{\gamma^2} \sum_{j=1}^N a_{ij} P_i B R_j^{-1} B^{\mathrm{T}} P_i = 0. \tag{16}$$

The following theorem shows that the control policy (14) with P_i the solution of (16) provides the minimax strategy for agent i.

Theorem 1: Let the agents of a differential graphical game with dynamics (2) and a leader with dynamics (3) use the control policies (14) where matrices P_i are the solutions of the coupled AREs (16). Moreover, assume that these control policies stabilize the local synchronization error dynamics (4) for all agents i . Then, all agents form their minimax strategies and the minimax value of the game is $V_i(\delta_i(0))$.

Proof: Consider the value function (7) and express the performance index (11) as

$$\begin{aligned} J_i = & \int_0^\infty \left(\delta_i^{\mathrm{T}} Q_i \delta_i + \left(d_i + g_i\right) u_i^{\mathrm{T}} R_i u_i - \gamma^2 \sum_{j=1}^N a_{ij} u_j^{\mathrm{T}} R_j u_j \right) \mathrm{d}t - \int_0^\infty \dot{V}_i(\delta_i) \mathrm{d}t \\ & + \int_0^\infty 2\delta_i^{\mathrm{T}} P_i \left(A\delta_i + \left(d_i + g_i\right) B u_i - \sum_{j=1}^N a_{ij} B u_j \right) \mathrm{d}t. \end{aligned}$$

Using the inner-product notation (1) we can express J_i as

$$\begin{aligned}J_i &= \langle \delta_i, Q_i\delta_i\rangle + (d_i+g_i)\langle u_i, R_iu_i\rangle \\ &\quad - \gamma^2\sum_{j=1}^{N} a_{ij}\langle u_j, R_ju_j\rangle + V_i(\delta(0)) + 2\langle \delta_i, P_iA\delta_i\rangle \\ &\quad + 2(d_i+g_i)\langle \delta_i, P_iBu_i\rangle - 2\sum_{j=1}^{N} a_{ij}\langle \delta_i, P_iBu_j\rangle,\end{aligned}$$

where we have used the fact that $\int_0^\tau \dot{V}_i(\delta_i)\mathrm{d}t = V_i(\delta_i(\tau)) - V_i(\delta_i(0))$, and that, since the system has stable equilibrium point, then $V_i(\delta_i(\tau)) = 0$ in the limit as $\tau \to \infty$. Since P_i is the solution to the ARE (16), we can write,

$$\begin{aligned}J_i =&(d_i+g_i)\langle u_i^\star, R_iu_i^\star\rangle - \gamma^2\sum_{j=1}^{N} a_{ij}\langle \upsilon_i^\star, R_j\upsilon_i^\star\rangle \\ &+ (d_i+g_i)\langle u_i, R_iu_i\rangle - \gamma^2\sum_{j=1}^{N} a_{ij}\langle u_j, R_ju_j\rangle \\ &- 2(d_i+g_i)\langle u_i^\star, R_iu_i\rangle + 2\gamma^2\sum_{j=1}^{N} a_{ij}\langle \upsilon_i^\star, R_ju_j\rangle \\ &+ V_i(\delta(0)) \\ =&(d_i+g_i)\langle u_i - u_i^\star, R_i(u_i - u_i^\star)\rangle \\ &- \gamma^2\sum_{j=1}^{N} a_{ij}\langle u_j - \upsilon_j^\star, R_j(u_j - \upsilon_j^\star)\rangle + V_i(\delta(0)).\end{aligned}$$

Therefore, $u_i^\star$ in (14) with P_i as in (16) is the minimax strategy of agent i, and (15) represents the worst-case policies of the neighbors, with the value of the game given by $V_i(\delta_i(0))$. ■

Remark 1: Control policies (14) are always distributed, in contrast to the policies based in the Nash solution given by (6). □

Remark 2: The equations of the form as in (16) are known to have solutions for P_i if $\left(A, \sqrt{Q_i}\right)$ is observable, (A,B) is stabilizable, and $(d_i+g_i)R_i^{-1} - \frac{1}{\gamma^2}\sum_{j=1}^{N} R_j^{-1} \succ 0$. □

In the following section we shall analyze the stability properties of the minimax policies (14).

IV. Off-Policy Learning

In this section, an off-policy reinforcement learning algorithm is proposed to determine the solutions of the Riccati equations (16) and obtain the control policies (14) that solve the minimax strategies problem. This method is designed such that the agents learn their optimal policies using only data measured from their environment, without any knowledge of the system dynamics (2).

The subsequent design procedure is similar to the one used in [19], where an off-policy algorithm to solve the H_∞ control problem was proposed. Start defining the variables u_i^k and υ_j^k as auxiliary control policies and express the system dynamics (4) as

$$\begin{aligned}\dot{\delta}_i = A\delta_i + (d_i+g_i)Bu_i^k - \sum_{j=1}^{N} a_{ij}B\upsilon_j^k \\ + (d_i+g_i)B\left(u_i - u_i^k\right) - \sum_{j=1}^{N} a_{ij}B\left(u_j - \upsilon_j^k\right).\end{aligned} \tag{17}$$

Here, the variables u_i^k and υ_j^k are the policies to be updated. Notice here that the input u_j corresponds to the actual policy employed by agent j, while υ_j^k is agent i estimation of the worst-case neighbor policy.

Let $V_i^k = \delta_i^{\mathrm{T}}P_i^k\delta_i$ represent the value function V_i at the kth iteration of our algorithm, and note that the expression,

$$\begin{aligned}V_i^k(\delta_i(t+T)) - V_i^k(\delta_i(t)) &= \int_t^{t+T}\dot{V}_i^k(\delta_i)\mathrm{d}\tau \\ &= 2\int_t^{t+T}\delta_i^{\mathrm{T}}P_i^k\dot{\delta}_i\mathrm{d}\tau,\end{aligned} \tag{18}$$

holds. Using the dynamics (17) in (18), we obtain

$$\begin{aligned}&\frac{1}{2}V_i^k(\delta_i(t+T)) - \frac{1}{2}V_i^k(\delta_i(t)) \\ &= \int_t^{t+T}\delta_i^{\mathrm{T}}P_i^k\left[A\delta_i + (d_i+g_i)Bu_i^k - \sum_{j=1}^{N} a_{ij}B\upsilon_j^k\right]\mathrm{d}\tau \\ &\quad + (d_i+g_i)\int_t^{t+T}\delta_i^{\mathrm{T}}P_i^kB\left(u_i - u_i^k\right)\mathrm{d}\tau \\ &\quad - \int_t^{t+T}\delta_i^{\mathrm{T}}P_i^k\sum_{j=1}^{N} a_{ij}B\left(u_j - \upsilon_j^k\right)\mathrm{d}\tau.\end{aligned} \tag{19}$$

From the Hamiltonian (13), we can obtain the kth-iteration for the Bellman equation as

$$\begin{aligned}0 = \delta_i^{\mathrm{T}}Q_i\delta_i + (d_i+g_i)\,u_i^{kT}R_iu_i^k - \gamma^2\sum_{j=1}^{N} a_{ij}\upsilon_j^{kT}R_j\upsilon_j^k \\ + 2\delta_i^{\mathrm{T}}P_i^k\left(A\delta_i + (d_i+g_i)\,Bu_i^k - \sum_{j=1}^{N} a_{ij}B\upsilon_j^k\right).\end{aligned} \tag{20}$$

Using (20) in (19) yields

$$\begin{aligned}V_i^k(\delta_i(t+T)) - V_i^k(\delta_i(t)) = -\int_t^{t+T}\Bigg(\delta_i^{\mathrm{T}}Q_i\delta_i + \\ + (d_i+g_i)u_i^{kT}R_iu_i^k - \gamma^2\sum_{j=1}^{N} a_{ij}\upsilon_j^{kT}R_j\upsilon_j^k\Bigg)\mathrm{d}\tau \\ - 2(d_i+g_i)\int_t^{t+T}u_i^{k+1T}R_i\left(u_i - u_i^k\right)\mathrm{d}\tau \\ + 2\int_t^{t+T}\sum_{j=1}^{N} a_{ij}\upsilon_j^{k+1T}R_j\left(u_j - \upsilon_j^k\right)\mathrm{d}\tau,\end{aligned} \tag{21}$$

where we have also used the fact that the control policies u_i^{k+1} and υ_j^{k+1} at iteration $k+1$ are defined as

$$u_i^{k+1} = -R_i^{-1}B^{\mathrm{T}}P_i^k\delta_i \tag{22}$$

and

$$\upsilon_j^{k+1} = -R_j^{-1} B^{\mathrm{T}} P_i^k \delta_i, \tag{23}$$

respectively. Lemma 1 shows the equivalence between (21) and (20).

Lemma 1: The solution $V_i = \delta_i^{\mathrm{T}} P_i \delta_i$ of (21) is the same as the solution of the Bellman equation (20).

Proof: Using (18), we can express (21) as

$$\int_t^{t+T} \left(\dot{V}_i^k(\delta_i) + 2(d_i + g_i) u_i^{k+1T} R_i \left(u_i - u_i^k\right) - 2\sum_{j=1}^{N} a_{ij} \upsilon_j^{k+1T} R_j \left(u_j - \upsilon_j^k\right) \right) \mathrm{d}\tau = -\int_t^{t+T} \left(\delta_i^{\mathrm{T}} Q_i \delta_i + (d_i + g_i) u_i^{kT} R_i u_i^k - \gamma^2 \sum_{j=1}^{N} a_{ij} \upsilon_j^{kT} R_j \upsilon_j^k \right) \mathrm{d}\tau.$$

Letting $\dot{V}_i^k = \nabla V_i^k \dot{\delta}_i$ and using (17), (22) and (23) we get

$$\int_t^{t+T} \nabla V_i^k \left(A\delta_i + (d_i + g_i) B u_i^k - \sum_{j=1}^{N} a_{ij} B \upsilon_j^k \right) \mathrm{d}\tau = -\int_t^{t+T} \left(\delta_i^{\mathrm{T}} Q_i \delta_i + (d_i + g_i) u_i^{kT} R_i u_i^k - \gamma^2 \sum_{j=1}^{N} a_{ij} \upsilon_j^{kT} R_j \upsilon_j^k \right) \mathrm{d}\tau,$$

which clearly holds if and only if (20) holds. ■

Now, the off-policy RL algorithm consists of solving (21) for V_i^k, u_i^{k+1} and υ_j^{k+1}. A useful method to solve (21) using measured data along the system trajectories is described in [19].

Algorithm 1 presents am iterative procedure for each agent i to determine her minimax policy (14).

Algorithm 1: Off-policy RL for Minimax Strategies

```
1: procedure
2:     Select initial stabilizing control policies $u_i^0$ and $\upsilon_j^0$ for all $j \in \mathcal{N}_i$.
3:     Apply a fixed policy $u_i \neq u_i^k$ and collect the required system information at $M$ different sampling intervals.
4:     Use the information collected in Step 1 to solve the Bellman equation (21) for $V_i^k$, $u_i^{k+1}$ and $\upsilon_j^{k+1}$ for all $j \in \mathcal{N}_i$.
5:     Go to Step 3. On convergence, stop.
6: end procedure
```

The following theorem shows the convergence properties of Algorithm 1.

Theorem 2: Algorithm 1 converges to the policies (14) and policies (15), where the matrix P_i solves the AREs (16).

Proof: Follows directly from Lemma 1 and the proof of convergence of the iterative procedure of solving the Bellman equation (20) and updating the policies (22)-(23) presented in [27]. ■

V. Simulations

A numerical example is presented to show the validity of our theoretical results. Consider a set of 5 agents and one leader node connected in a communication graph as shown in Figure 1. If $j \in \mathcal{N}_i$, let $a_{ij} = 1$. Each agent has linear dynamics given by (2), with

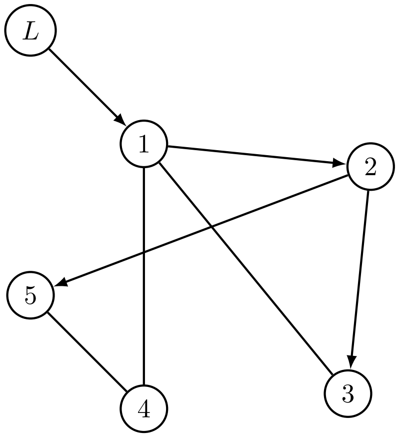

Fig. 1. Graph topology.

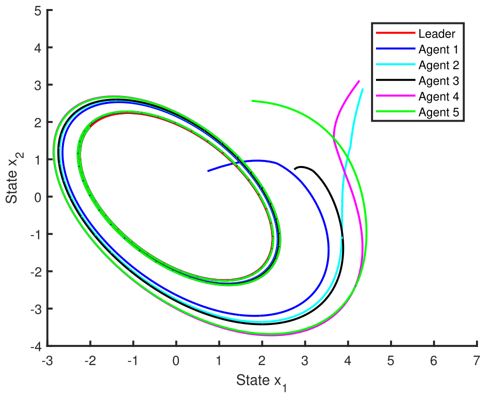

Fig. 2. Evolution of the state trajectories.

$$A = \begin{bmatrix} 1 & 2 \\ -2 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}.$$

The minimax performance indices of the agents are defined by (11) with $Q_1 = Q_3 = 2I$, $Q_2 = Q_5 = 3I$ and $Q_4 = I$, where I is the identity matrix. Let all agents use the same values for $R = 2I$ and $\gamma = 2$.

Algorithm 1 is used to learn the solution of the AREs (16). The resulting matrices P_i are shown below.

$$P_1 = \begin{bmatrix} 0.5103 & 0.0305 \\ 0.0305 & 0.3537 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1.2004 & 0.1316 \\ 0.1316 & 0.7588 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 0.6756 & 0.0589 \\ 0.0589 & 0.4440 \end{bmatrix}, \quad P_4 = \begin{bmatrix} 0.4993 & 0.0705 \\ 0.0705 & 0.3073 \end{bmatrix},$$

$$P_5 = \begin{bmatrix} 0.8051 & 0.0547 \\ 0.0547 & 0.5558 \end{bmatrix}.$$

The minimax control policies are now given by (14). Using these policies, the agents successfully achieve synchronization with the trajectories shown in Figure 2, 3. Figure 4 shows the convergence of the estimated matrices P_i for all agents to the optimal values.

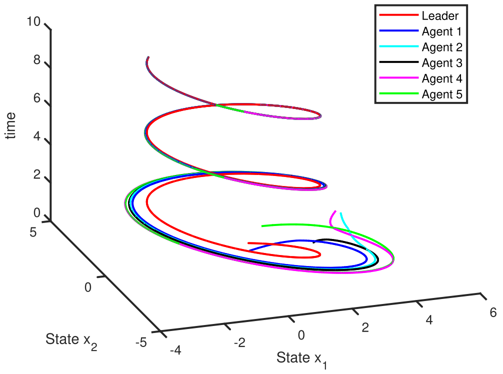

Fig. 3. Synchronization in time.

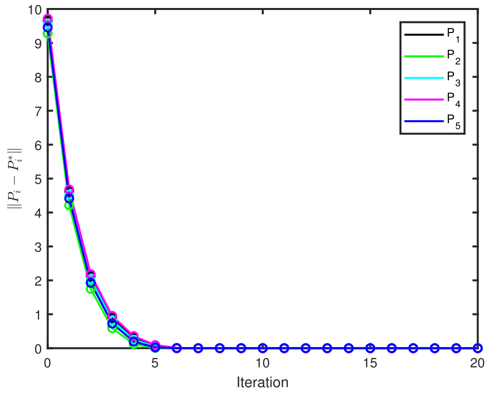

Fig. 4. Convergence of the estimated matrices P_i to their optimal values.

VI. Conclusion

Minimax strategies were designed and analyzed as an alternative solution concept for differential graphical games. The resulting control policies are always distributed in the sense that the agents use only local information obtained from the graph topology. The proposed off-policy RL algorithm is a practical method to determine the minimax strategies of the agents without any knowledge about the system dynamics; moreover, this algorithm allows to compute the worst-case neighbor policy v_j even when this is not the control policy used by the neighbors. Future work will focus on extending the results to nonlinear systems.

VII. Acknowledgements

V. G. Lopez thanks the Mexican Council of Science and Technology (Conacyt) for the support.

References

[1] M.I. Abouheaf and M.S. Mahmoud. Online policy iteration solution for dynamic graphical games. *13th International Multi-Conference on Systems, Signals & Devices*, pages 787–797, 2016.

[2] 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-infinity control. *Automatica*, 43:473–481, 2007.

[3] T. Basar and P. Bernhard. *H_∞-Optimal Control and Related Minmax Design Problems*. Birhäuser, Boston, MA, 1995.

[4] T. Basar and G.J. Olsder. *Dynamic Noncooperative Game Theory*. SIAM, Philadelphia, PA, 2 edition, 1999.

[5] H. Cao, E. Ertin, and A. Arora. Minimax equilibrium of networked differential games. *ACM Transactions on Autonomous and Adaptive Systems*, 3(4):1–21, 2008.

[6] J.C. Doyle, K. Glover, and P.P. Khargonekar. State-space solutions to standard H_2 and H_∞ control problems. *IEEE Trans. on Automatic Control*, 34(8):831–847, 1998.

[7] Y. Hong, J. Hu, and L. Gao. Tracking control for multi-agent consensus with an active leader and variable topology. *Automatica*, 47(7):1177–1182, 2006.

[8] R. Isaacs. *Differential games. A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. John Wiley & Sons, inc., New York, USA, 1965.

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

[10] M. Johnson, S. Bhasin, and W.E. Dixon. Nonlinear two-player zero-sum game approximate solution using a policy iteration algorithm. *Proc. IEEE Conf. Decis. Control*, pages 142–147, 2011.

[11] R. Kamalapurkar, T. Dinh, P. Walters, and W.E. Dixon. Approximate optimal cooperative decentralized control for consensus in a topological network of agents with uncertain nonlinear dynamics. *Proc. American Control Conf.*, 1:1322–1327, 2013.

[12] R. Kamalapurkar, J.R. Klotz, P. Walters, and W.E. Dixon. Model-based reinforcement learning in differential graphical games. *IEEE Trans. on Control of Network Systems*, 5(1):423–433, 2018.

[13] H. Kwakernaak. Robust control and H_∞-optimization. tutorial paper. *Automatica*, 29(2):255–273, 1993.

[14] F.L. Lewis, D. Vrabie, and V.L. Syrmos. *Optimal Control*. John Wiley & Sons, inc., New Jersey, USA, 3 edition, 2012.

[15] F.L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das. *Cooperative Control of Multi-agent systems: Optimal and Adaptive Design Approaches*. Springer-Verlag, New York, USA, 2013.

[16] X. Li, X. Wang, and G. Chen. Pinning a complex dynamical network to its equilibrium. *IEEE Transactions on Circuits and Systems I. Regular papers*, 51(10):2074–2087, 2004.

[17] Z. Li, Z. Duan, and G. Chen. On H_∞ and H_2 performance regions of multi-agent systems. *Automatica*, 47:797–803, 2011.

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

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

[20] R. Olfati-Saber, J. Fax, and R. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[21] Z. Qu. *Cooperative control of dynamical systems: applications to autonomous vehicles*. Springer-Verlag, New York, USA, 2009.

[22] Z. Qu and M. A. Simaan. A design of distributed game strategies for networked agents. *IFAC Proceeding Volumes*, 42(20):270–275, 2009.

[23] W. Ren, R. Beard, and E. Atkins. A survey of consensus problems in multi-agent coordination. *Proc. Amer. control conf.*, pages 1859–1864, 2005.

[24] W. Ren, K. Moore, and Y. Chen. High-order and model reference consensus algorithms in cooperative control of multivehicle systems. *Journal of Dynamic Systems, Measurement and Control*, 129(5):678–688, 2007.

[25] Y. Shoham and K. Leyton-Brown. *Multiagent Systems. Algorithmic, Game-theoretic and Logical Foundations*. Cambridge University Press, New York, NY, 2008.

[26] K.G. Vamvoudakis, F.L. Lewis, and G.R. Hudas. Multiagent differential graphical games: Online adaptive learning solution for synchronization with optimality. *Automatica*, 48:1598–1611, 2012.

[27] H. N. Wu and B. Luo. Neural network based online simultaneous policy update algorithm for solving the hji equation in nonlinear H_∞ control. *IEEE Trans. Neural Networks and Learning Systems*, 23(12):1884–1895, 2012.

[28] F. A. Yaghmaie, F. L. Lewis, and R. Su. Output regulation of heterogeneous linear multi-agent systems with differential graphical game. *International Journal of Robust and Nonlinear Control*, 26:2256–2278, 2016.

[29] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *Proc. 17th Allerton Conf*, pages 744–752, 1979.