Cooperative Q-learning for Rejection of Persistent Adversarial Inputs in Networked Linear Quadratic Systems

Kyriakos G. Vamvoudakis[1], *Senior Member, IEEE,* João P. Hespanha[2], *Fellow, IEEE*

***Abstract*—In this paper, a cooperative Q-learning approach is proposed to enable the agents in large networks to synchronize to the behavior of an unknown leader by each optimizing a distributed performance criterion that depends only on a subset of the agents in the network. The novel distributed Q-functions are parametrized as functions of the tracking error, control and adversarial inputs in the neighborhood. In the proposed approach, the agents coordinate with the neighbors in order to pick their minimizing model-free policies in such a way to guarantee convergence to a graphical Nash equilibrium and also attenuation of maximizing worst-case adversarial inputs. A structure of 2-actors and a single critic approximators are used for each agent in the network. This eventually solves the complexity issues that arise in Q-learning. The 2-actors are used to approximate the optimal control input and the worst-case adversarial input while the critic approximator is used to approximate the optimal cost of each of the coupled optimizations. Effective tuning laws are proposed to solve the model-free cooperative game problem while also guaranteeing closed-loop stability with the use of rigorous Lyapunov-based stability proofs. Finally, a numerical example is used to illustrate the effectiveness of the proposed approach.**

***Index Terms*—Cooperative Q-learning, large complex networks, games on graphs.**

I. Introduction

Modeling large networks and complex systems [18], [27] is very difficult and can become very tedious and time consuming for engineers nowadays in order to create coordinating policies with cooperative or even conflicting goals. There are numerous examples of complex networked systems, ranging from traffic/transportation systems [33], network security [1], [24], human/robot interaction systems [13], internet [7], social networks [21] and health-care [26]. Reinforcement learning [28] algorithms have been shown to overcome several limitations of supervised and model-based methods.

Game theory [2] is widely used to model multi-agent learning problems and create cooperative or conflicting policies [25]. The concept of *graphical games* was introduced by [14], as a compact representation scheme for games on graphs where the action of each self-interested agent is associated only to the actions of the agents in her neighborhood. Now, the complexity of finding a globally optimal joint action can be reduced assuming that agents do not need to coordinate with all the agents but with a smaller subset of them. Graphical games, have gained a lot of attention recently e.g. [8], mainly because of their interesting theoretic characteristics in networks with selfish and not familiar with other nodes strategies and because they have found application to several real world environments that involve large networks. One of the main disadvantages of graphical games and of most game-based [2] control system design techniques is that they rely on complete knowledge of the systems to be controlled. This is clearly not the case in large networks with interacting components since the subsystems are not modeled exactly or some parameters are uncertain or completely unknown and for that reason one has to find more intelligent techniques that can still guarantee optimal performance, robustness and closed-loop stability.

One of the most popular reinforcement learning algorithms is known as Q-learning and has been proposed by Watkins in [35]. The main advantages of Q-learning are its simplicity, its model-free nature and that convergence to the optimal solution can be guaranteed in environments that are described by Markov decision processes. The main problem of Q-learning is that it is well established only for systems with discrete-time dynamics however all the events around us are evolving in a continuous manner. Other major problems of Q-learning are that the state and action space, scales exponentially with the number of agents and it suffers from the combinatorial growth in the time for convergence due to large look-up tables that encode global action-pair values. This complexity can be solved by using universal approximators with adaptive characteristics in the context of approximate dynamic programming [37] or neuro-dynamic programming [4] that allow us to eventually solve difficult distributed optimization problems online and forward in time.

Related work

There is an extensive research on multi-agent coordination from several scientific societies including control systems society [20] and artificial intelligence society [36]. The main disadvantages of most of the existing work though is that it requires complete knowledge of the system dynamics and it cannot provide any formal optimality guarantees. Two recent surveys are given in [6] from the control system and [5] from

[1]K. G. Vamvoudakis, is with the Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA 24061-0203 USA, e-mail: kyriakos@vt.edu.

[2]J. P. Hespanha is with the Center for Control, Dynamical-systems and Computation (CCDC), University of California, Santa Barbara, CA 93106-9560 USA, e-mail: hespanha@ece.ucsb.edu.

This material is based upon work supported by NATO under grant No. SPS G5176 and U.S. Office of Naval Research under MURI grant No. N00014-16-1-2710.

the computational intelligence perspective respectively, where the authors state that distributed multi-agent optimization is a challenging task due to computational complexity issues.

The work of [16] extended the work of [10] and proposed a Q-learning algorithm to learn the coordinated actions of a group of cooperative agents, using a sparse representation of the joint state-action space of the agents. The authors of [9] introduced cooperative learners formed by combining planners with reinforcement learning algorithms and intelligent cooperative control architectures shown through empirical results and without any theoretical guarantees. In [34] it is shown that cooperation remains stable by applying a centralized Q-learning on a variety of random, regular, small-word, scale-free and modular network models in repeated, multi-agent Prisoner's Dilemma and Hawk-Dove games. Also the combination of learning and innovation may help to preserve cooperation during the re-organization of complex networks. The authors in [23] have used a form of decentralized Q-learning, to manage the aggregated interference generated by multiple wireless regional area networks systems by considering complete and partial information. In [19], the authors propose a sparse Q-learning approach by augmenting the action space of each agent with a coordination action that attempts to use information from the other agents. However they do not provide any formal guarantees on the performance of the algorithm. In our previous work in [30] and [31] we have proposed an adaptive learning technique to learn the graphical Nash equilibrium but its dependence on the exact model limits its application to only small networks with interacting agents with known dynamics and without any adversarial inputs.

Contributions

The contributions are fourfold. First the synchronization problem is formulated as a $\min\max$ optimization for each agent, where each control input is minimizing a distributed performance index while the adversarial input is maximizing it. Hence, we are eventually interested in finding plug-n-play model free saddle point solutions for every agent. In order to avoid the knowledge of the system dynamics (input, control and adversarial input matrices), and pinning gains (leader connection) we shall adopt the framework of machine learning and especially Q-learning to parameterize the problem as functions of the neighborhood control, adversarial inputs and the neighborhood tracking errors. Third, we propose a *model-free* 2-actor/single-critic approximator structure for every agent based on integral reinforcement learning along with appropriate tuning laws. A critic is used to approximate the distributed Q-function and two actor approximators to approximate the optimal control and the worst case adversarial input of each agent. Finally, rigorous proofs provide guarantees for asymptotic stability of the equilibrium point of the closed-loop systems, robustness and convergence of the controls and adversarial inputs to a graphical Nash-equilibrium.

Structure

The remainder of the paper is structured as follows. In Section 2 we formulate the problem of synchronization agents in large networks under persistent adversarial inputs by simultaneously satisfying distributed performances. Since the dynamics of the agents are unknown, Section 3 provides a distributed Q-learning approach along with a graphical Nash equilibrium existence theorem. In Section 4, we provide a sparse cooperative learning solution based on a 2-actor/single-critic integral reinforcement learning technique with rigorous Lyapunov-based stability proofs. Section 5, provides a numerical example of the proposed model-free distributed Q-learning algorithm in a network of 10 agents of unknown dynamics. Finally, Section 5 concludes and discusses future work challenges.

Notation: The notation used here is standard. $\mathbb{R}^+$ is the set of positive real numbers. The superscript $*$ is used to denote the optimal solution of an optimization, $\underline{\lambda}(A)$ is the minimum eigenvalue of a matrix A and $\bar{\lambda}(A)$ the maximum eigenvalue of A, $\|.\|$ denotes the Euclidean norm for a vector and the Frobenius norm for a matrix, $(.)^T$ denotes the transpose of a matrix and $\mathrm{diag}[\cdot]$ a diagonal matrix. Finally $A \geq 0$ denotes that the matrix A is positive semidefinite and the notation $A > 0$ to denote that the matrix A is positive definite. The half-vectorization, $\mathrm{vech}(A)$, of a symmetric $n \times n$ matrix A is the $n(n+1)/2$ column vector obtained by vectorizing only the upper (or lower) triangular part of A, $\mathrm{vech}(A) = [A_{1,1} \ \cdots \ A_{n,1} \ A_{2,2} \ \cdots \ A_{n,2} \ \cdots \ A_{n-1,n-1} \ A_{n-1,n} \ A_{n,n}]^T$ and $\mathrm{mat}(\ \cdot\ )$ is the inverse vech operation known as matrization. The colon symbol : can be used to form implicit vectors from a matrix or vector, i.e. $A(j : k), k > j$ is $[A(j) \ A(j+1) \ \cdots \ A(k)]$. Finally $\otimes$ denotes the Kronecker product, and for brevity when the implicit time dependence is obvious we write x instead of $x(t)$.

Background on Networks and Graphs: Before we proceed to the problem formulation, we shall give a brief background on graphs. The multi-agent networked system interactions are modeled by a fixed topology graph $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ consisting of nonempty finite set of N nodes $\mathcal{V}_\mathcal{G} = \{n_1, ..., n_N\}$ that represent agents and a set of edges or arcs $\mathcal{E}_\mathcal{G} \subseteq \mathcal{V}_\mathcal{G} \times \mathcal{V}_\mathcal{G}$ that represent the inter-agent information exchange links. Denote the connectivity matrix $\mathcal{A}_\mathcal{G} = [a_{ij}]$ of $\mathcal{G}$ as

$$a_{ij} \begin{cases} > 0, \text{ if } (n_j, n_i) \in \mathcal{E}_\mathcal{G} \\ = 0, \text{ otherwise.} \end{cases}$$

We assume the graph is simple, e.g. no repeated edges and $(n_i, n_i) \notin \mathcal{E}_\mathcal{G}, \forall i \in \mathcal{N}$ no self-loops ($a_{ii} = 0$). The set of neighbors of a node n_i is $\mathcal{N}_i = \{n_j : (n_j, n_i) \in \mathcal{E}_\mathcal{G}\}$, i.e. the set of nodes with arcs incoming to n_i. Define the *degree* matrix as $\mathcal{D} = diag(d_i)$ with $d_i = \sum_{j \in \mathcal{N}_i} a_{ij}$ the weighted degree of node i (i.e. ith row sum of $\mathcal{A}_\mathcal{G}$). The number of neighbors of agent i is $|\mathcal{N}_i|$ which is equal to d_i.

II. Problem Formulation

We consider a networked-system $\mathcal{G}$, consisting of N agents each modeled $\forall i \in \mathcal{N} := \{1, \ldots, N\}$ by the following dynamics,

$$\frac{d}{dt} x_i(t) =: \dot{x}_i(t) = A x_i(t) + B_i u_i(t) + D_i v_i(t), \ x_i(0) = x_{i0}, \ t \geqslant 0, \tag{1}$$

where $x_i(t) \in \mathbb{R}^n$ is a measurable state vector available for feedback by each agent and known initial conditions $x_i(0)$, $u_i(t) \in \mathbb{R}^{m_i}$, $i \in \mathcal{N} := \{1, \ldots, N\}$ is each control input (or minimizing player as we shall see later), $v_i(t) \in \mathbb{R}^{l_i}$, $i \in \mathcal{N} := \{1, \ldots, N\}$ is each adversarial input (or maximizing player as we shall see later), and $A \in \mathbb{R}^{n\times n}$, $B_i \in \mathbb{R}^{n\times m_i}$, $D_i \in \mathbb{R}^{n\times l_i}$, $i \in \mathcal{N}$ are the plant, control input and adversarial input matrices respectively that will be considered uncertain/unknown. It is assumed that the pairs (A, B_i), $\forall i \in \mathcal{N}$ are stabilizable. We have a total of $2N$ players/controllers that select values for $u_i(t)$, $t \geqslant 0$, $i \in \mathcal{N}$ and $v_i(t)$, $t \geqslant 0$, $i \in \mathcal{N}$. The agents in the network seek to cooperatively asymptotically track the state of a leader node/exosystem with dynamics $\dot{x}_L = Ax_L$, i.e. $x_i(t) - x_L(t) \to 0, \forall i \in \mathcal{N}$ while *simultaneously satisfying user-defined distributed performances*.

Figure 1 shows one such networked system $\mathcal{G}$ consisting of 10 agents with a leader node L pinned to node number 6.

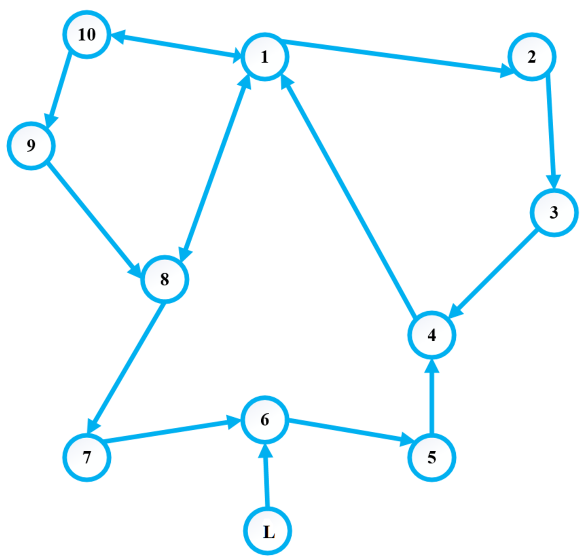

Fig. 1. A networked system $\mathcal{G}$ with 10 agents and a leader node L pinned to agent 6.

Now we shall proceed to the design of the user-defined *distributed performances*. For that reason, we shall define the following *neighborhood tracking error* for every agent,

$$e_i := \sum_{j\in\mathcal{N}_i} a_{ij}(x_i - x_j) + g_i(x_i - x_L), \forall i \in \mathcal{N}, \quad (2)$$

where $g_i \in \mathbb{R}^+$ is the pinning gain that shows if an agent is pinned to the leader node (i.e. $g_i \neq 0$) and it is nonzero for at least one node.

Definition 1. All agents in a connected graph, asymptotically synchronize with a leader which is connected to at least one agent, when $e_i \to 0 \Rightarrow x_i - x_L \to 0, \forall i \in \mathcal{N}$. □

The dynamics of (2) are given by,

$$\dot{e}_i = Ae_i + (d_i + g_i)(B_i u_i + D_i v_i) - \sum_{j\in\mathcal{N}_i} a_{ij}(B_j u_j + D_j v_j), \forall i \in \mathcal{N}, \quad (3)$$

with $e_i \in \mathbb{R}^n$.

The cost functionals associated to each agent $i \in \mathcal{N}$, that depend on the tracking error e_i, the control u_i, the controls in the neighborhood of agent i given as, $u_{\mathcal{N}_i} := \{u_j : j \in \mathcal{N}_i\}$, the adversarial input v_i and the adversarial inputs in the neighborhood of agent i given as $v_{\mathcal{N}_i} := \{v_j : j \in \mathcal{N}_i\}$, have the following form,

$$\mathcal{J}_i(e_i(0); u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) = \frac{1}{2}\int_0^\infty \big(e_i^T H_i e_i + (u_i^T R_{ii} u_i - \gamma_{ii}^2 v_i^T v_i) + \sum_{j\in\mathcal{N}_i}(u_j^T R_{ij} u_j - \gamma_{ij}^2 v_j^T v_j)\big)dt, \ \forall i \in \mathcal{N}, \quad (4)$$

with matrices $H_i \succeq 0$, $R_{ii} \succ 0$, $R_{ij} \succeq 0$, $\forall i, j \in \mathcal{N}$ of appropriate dimensions, $\gamma_{ii}, \gamma_{ij} \in \mathbb{R}^+$, $\forall i \in \mathcal{N}$ and $(\sqrt{H_i}, A)$, $\forall i \in \mathcal{N}$ are detectable.

Hence, given a strongly connected graph $\mathcal{G}$, we are interested in finding a graphical Nash equilibrium [2], [30], that is translated to a saddle point $u_i^\star, v_i^\star$, for every agent $i \in \mathcal{N}$ in the sense that,

$$\mathcal{J}_i(e_i(0); u_i^\star, u_{\mathcal{N}_i}^\star, v_i, v_{\mathcal{N}_i}^\star) \leqslant \mathcal{J}_i(e_i(0); u_i^\star, u_{\mathcal{N}_i}^\star, v_i^\star, v_{\mathcal{N}_i}^\star) \leqslant \mathcal{J}_i(e_i(0); u_i, u_{\mathcal{N}_i}^\star, v_i^\star, v_{\mathcal{N}_i}^\star), \ \forall u_i, v_i, \ \forall i \in \mathcal{N}.$$

This can be expressed by the following coupled distributed optimization problems,

$$\mathcal{J}_i(e_i(0); u_i^\star, u_{\mathcal{N}_i}^\star, v_i^\star, v_{\mathcal{N}_i}^\star) = \min_{u_i} \mathcal{J}_i(e_i(0); u_i, u_{\mathcal{N}_i}^\star, v_i^\star, v_{\mathcal{N}_i}^\star) = \max_{v_i} \mathcal{J}_i(e_i(0); u_i^\star, u_{\mathcal{N}_i}^\star, v_i, v_{\mathcal{N}_i}^\star), \quad (5)$$

given the dynamics in (3).

Thus, the ultimate goal is to find the distributed optimal value functions $V_i^\star$, $\forall i \in \mathcal{N}$ defined by,

$$V_i^\star(e_i(t)) := \min_{u_i}\max_{v_i} \int_t^\infty \frac{1}{2}\big(e_i^T H_i e_i + (u_i^T R_{ii} u_i - \gamma_{ii}^2 v_i^T v_i) + \sum_{j\in\mathcal{N}_i}(u_j^T R_{ij} u_j - \gamma_{ij}^2 v_j^T v_j)\big)dt, \ \forall t, \forall i \in \mathcal{N}, \quad (6)$$

but without any information of the system matrices A, B_i, $D_i, \forall i \in \mathcal{N}$ and pinning gains g_i, $\forall i \in \mathcal{N}$.

Remark 1. Note that the assumptions that the pairs (A, B_i), $\forall i \in \mathcal{N}$ are stabilizable and the pairs $(\sqrt{H_i}, A)$, $\forall i \in \mathcal{N}$ are detectable, will guarantee that the coupled algebraic Riccati equations will have a unique non-negative solution [3]. □

III. Graphical Nash Equilibrium

In this section we shall give a brief discussion about the existence of the graphical Nash equilibrium before we proceed to the model free Q-learning approach with the use of $3N$ distributed adaptation structures, i.e. N critic approximators to approximate the costs and $2N$ actor approximators to approximate the optimal control inputs and worst case adversarial inputs.

First we will define the Hamiltonian associated with each agent's neighborhood tracking error (3) and each $V_i^\star$ given in (6) as follows,

$$\mathcal{H}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}, \frac{\partial V_i^\star}{\partial e_i}) = \frac{\partial V_i^\star}{\partial e_i}^T \Big(Ae_i$$

$$+ (d_i+g_i)(B_iu_i+D_iv_i) - \sum_{j\in\mathcal{N}_i} a_{ij}(B_ju_j+D_jv_j)\Bigg)$$
$$+ \frac{1}{2}\Bigg(e_i^T H_i e_i + (u_i^T R_{ii} u_i - \gamma_{ii}^2 v_i^T v_i)$$
$$+ \sum_{j\in\mathcal{N}_i}(u_j^T R_{ij}u_j - \gamma_{ij}^2 v_j^T v_j)\Bigg), \forall e_i, u_i, v_i \forall i\in\mathcal{N}. \quad (7)$$

After employing the stationarity conditions, in the Hamiltonian (7) we can find the saddle-point solution, i.e. $\frac{\partial \mathcal{H}_i(\cdot)}{\partial u_i}=0$, and $\frac{\partial \mathcal{H}_i(\cdot)}{\partial v_i}=0$. Hence, the optimal control for each $i\in\mathcal{N}$ can be found to be,

$$u_i^*(e_i) = \arg\min_{u_i} \mathcal{H}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}, \frac{\partial V_i^*}{\partial e_i})$$
$$= -(d_i+g_i)R_{ii}^{-1}B_i^T\frac{\partial V_i^*}{\partial e_i},\ \forall e_i, \quad (8)$$

and the worst case adversarial input can be found to be,

$$v_i^*(e_i) = \arg\max_{v_i} \mathcal{H}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}, \frac{\partial V_i^*}{\partial e_i})$$
$$= \frac{(d_i+g_i)}{\gamma_{ii}^2}D_i^T\frac{\partial V_i^*}{\partial e_i},\ \forall e_i. \quad (9)$$

The saddle-point solution (8)-(9) should satisfy the appropriate coupled Hamilton-Jacobi equations,

$$\mathcal{H}_i(e_i, u_i^*, u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*, \frac{\partial V_i^*}{\partial e_i}) = 0, \forall i\in\mathcal{N}. \quad (10)$$

Remark 2. For every fixed $u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*$, we have $\frac{\partial^2 H_i}{\partial u_i^2} = R_{ii} > 0$ and therefore the Hamiltonian is minimized at the stationarity point (8). Conversely, for every fixed $u_i^*, u_{\mathcal{N}_i}^*, v_{\mathcal{N}_i}^*$, we have $\frac{\partial^2 H_i}{\partial v_i^2} = -\gamma_{ii}^2 < 0$ and therefore the Hamiltonian is maximized at the stationarity point (9). □

The value functions can be represented as quadratic in the neighborhood tracking error, i.e. $V_i^*(e_i): \mathbb{R}^n \to \mathbb{R}$,

$$V_i^*(e_i) = \frac{1}{2}e_i^T P_i e_i,\ \forall e_i, \forall i\in\mathcal{N}, \quad (11)$$

where $P_i \in \mathbb{R}^{n\times n}$, $\forall i \in \mathcal{N}$ are the unique symmetric positive definite matrices that solve the following complicated distributed coupled equations,

$$\mathcal{H}_i(e_i, u_i^*, u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*, e_i^T P_i) = 0, \forall i\in\mathcal{N}. \quad (12)$$

By using (11), the optimal control (8) for every agent $i\in\mathcal{N}$ can be written as,

$$u_i^*(e_i) = -(d_i+g_i)R_{ii}^{-1}B_i^T P_i e_i,\ \forall e_i, \quad (13)$$

and the worst case adversarial input (9) for every agent $i\in\mathcal{N}$ can be written as,

$$v_i^*(e_i) = \frac{(d_i+g_i)}{\gamma_{ii}^2}D_i^T P_i e_i,\ \forall e_i. \quad (14)$$

Remark 3. It is important to note that the equations (12), (13), (14), are highly coupled, difficult to solve due to the cross terms $e_i^T e_j$ and require complete knowledge of the system matrix A, the input matrices B_i, D_i, $i \in \mathcal{N}$ and leader connection information g_i. □

Remark 4. Note that the adversarial input $v_i^*(e_i)$ in (14) is not the actual input by the adversary but the worst case input that she can introduce to our system. The actual adversary has the freedom to do anything she likes, as long as she remains stealthy. □

In the following section, we shall propose a new *cooperative Q-learning* based approach to solve the graphical Nash game problem without any information of the system dynamics while attenuating persistent adversarial inputs, by adjusting parameters in an adaptive way.

Before we proceed to the model free formulation we shall provide the following existence theorem for a graphical Nash equilibrium.

Theorem 1. *Let the pairs (A, B_i), $\forall i \in \mathcal{N}$ be stabilizable and $(\sqrt{H}_i, A)$, $\forall i\in\mathcal{N}$ be detectable. Suppose that the graph $\mathcal{G}$ is strongly connected and that there exist unique symmetric positive definite matrices P_i, $\forall i\in\mathcal{N}$ that satisfy the following coupled Riccati inequalities,*

$$\mathcal{H}_i(e_i, \bar{u}_i, \bar{u}_{\mathcal{N}_i}, \bar{v}_i, \bar{v}_{\mathcal{N}_i}, e_i^T P_i) = 0 \quad (15)$$

with $V_i^*(0) = 0$, $\forall i\in\mathcal{N}$,

$$B_i R_{ii}^{-1} B_i^T > \frac{1}{\gamma_{ii}^2} D_i D_i^T,\ \forall i\in\mathcal{N} \quad (16)$$

$$B_j R_{jj}^{-T} R_{ij} R_{jj}^{-1} B_j^T > \frac{\gamma_{ij}^2}{\gamma_{jj}^4} D_j D_j^T,\ \forall i,j\in\mathcal{N}, \quad (17)$$

with optimal control input,

$$\bar{u}_i(e_i) = -(d_i+g_i)R_{ii}^{-1}B_i^T P_i e_i, \forall e_i, \quad (18)$$

and worst case adversarial input,

$$\bar{v}_i(e_i) = \frac{(d_i+g_i)}{\gamma_{ii}^2}D_i^T P_i e_i, \forall e_i. \quad (19)$$

Then all the agents synchronize asymptotically to the state of the leader (exosystem), i.e. the equilibrium point (origin) of the closed-loop systems (3)*, with $u_i = \bar{u}_i(e_i)$ and $v_i = \bar{v}_i(e_i)$ where $\bar{u}_i(e_i)$ and $\bar{v}_i(e_i)$ are given by* (18) *and* (19) *$\forall i\in\mathcal{N}$ is globally asymptotically stable and the control and adversarial policies form a graphical Nash equilibrium by solving the coupled distributed optimization problems as defined in* (6).

Proof: Because of (7), (11) and (12) the time derivative of V_i, $\forall i\in\mathcal{N}$ for $t \geqslant 0$ along the closed-loop solutions satisfies,

$$\dot{V}_i = \frac{\partial V_i}{\partial e_i}\Bigg(Ae_i + (d_i+g_i)(B_i\bar{u}_i + D_i\bar{v}_i) - \sum_{j\in\mathcal{N}_i} a_{ij}(B_j\bar{u}_j + D_j\bar{v}_j)\Bigg)$$
$$= -\frac{1}{2}\Bigg(e_i^T H_i e_i + (\bar{u}_i^T R_{ii}\bar{u}_i - \gamma_{ii}^2 \bar{v}_i^T \bar{v}_i)$$
$$+ \sum_{j\in\mathcal{N}_i}(\bar{u}_j^T R_{ij}\bar{u}_j - \gamma_{ij}^2\bar{v}_j^T\bar{v}_j)\Bigg),\ \forall i\in\mathcal{N}, \quad (20)$$

with $\bar{u}_i$ given by (18) and $\bar{v}_i$ given by (19) and since P_i, $\forall i \in \mathcal{N}$ satisfy (12) and after taking into consideration the conditions (16) and (17), we can upper bound (20) as,

$$\dot{V}_i \leqslant -\frac{1}{2}\underline{\lambda}(H_i)\left\|e_i\right\|^2,\ \forall i\in\mathcal{N}.$$

Using V_i as a Lyapunov function for every agent, we conclude that the origin is a globally asymptotically stable equilibrium point of (3). Since the functions $V_i^* := V_i,\ \forall i \in \mathcal{N}$ are smooth, zero at zero and converge to zero as $t \to \infty$ we can write (4) with arbitrary $\bar{u}_i,\ \bar{v}_i$ as follows,

$$\begin{aligned}
&\mathcal{J}_i(e_i(0); \bar{u}_i, \bar{u}_{\mathcal{N}_i}, \bar{v}_i, \bar{v}_{\mathcal{N}_i}) = \int_0^\infty \frac{1}{2}\Big(e_i^T H_i e_i + (\bar{u}_i^T R_{ii} \bar{u}_i - \gamma_{ii}^2 \bar{v}_i^T \bar{v}_i) \\
&\quad + \sum_{j \in \mathcal{N}_i} (\bar{u}_j^T R_{ij} \bar{u}_j - \gamma_{ij}^2 \bar{v}_j^T \bar{v}_j)\Big) dt \\
&\quad\quad + \frac{1}{2} e_i(0)^T P_i e_i(0) + \frac{1}{2}\int_0^\infty \dot{V}_i^* dt \\
&= \int_0^\infty \frac{1}{2}\Big(e_i^T H_i e_i + (\bar{u}_i^T R_{ii} \bar{u}_i - \gamma_{ii}^2 \bar{v}_i^T \bar{v}_i) \\
&\quad + \sum_{j \in \mathcal{N}_i} (\bar{u}_j^T R_{ij} \bar{u}_j - \gamma_{ij}^2 \bar{v}_j^T \bar{v}_j)\Big) dt \\
&+ \frac{1}{2} e_i(0)^T P_i e_i(0) + \frac{1}{2}\int_0^\infty \frac{\partial V_i^*}{\partial e_i}\Big(A e_i + (d_i + g_i)(B_i \bar{u}_i + D_i \bar{v}_i) \\
&\quad - \sum_{j \in \mathcal{N}_i} a_{ij}(B_j \bar{u}_j + D_j \bar{v}_j)\Big) dt,\ \forall i \in \mathcal{N},
\end{aligned}$$

where V_i^* is defined in (6). Since $V_i^*,\ i \in \mathcal{N}$ derived in (11) satisfy the coupled Riccati equations (12) given the conditions in (16)-(17), and by denoting as u_i^* the optimal controls given by (13) and by denoting as v_i^* the worst case adversarial input given by (14) we can subtract zero (by using (12)) and complete the squares to write $\forall i \in \mathcal{N}$,

$$\begin{aligned}
&\mathcal{J}_i(e_i(0); \bar{u}_i, \bar{u}_{\mathcal{N}_i}, \bar{v}_i, \bar{v}_{\mathcal{N}_i}) = \int_0^\infty \frac{1}{2}\Big((\bar{u}_i - u_i^*)^T R_{ii} (\bar{u}_i - u_i^*) \\
&\quad - \gamma_{ii}^2 (\bar{v}_i - v_i^*)^T (\bar{v}_i - v_i^*) \\
&+ \sum_{j \in \mathcal{N}_i} (\bar{u}_j - u_j^*)^T R_{ij} (\bar{u}_j - u_j^*) - \sum_{j \in \mathcal{N}_i} \gamma_{ij}^2 (\bar{v}_j - v_j^*)^T (\bar{v}_j - v_j^*)\Big) dt \\
&\quad\quad + \frac{1}{2} e_i(0)^T P_i e_i(0) \\
&+ \frac{1}{2}\int_0^\infty \Big(\sum_{j \in \mathcal{N}} a_{ij} {u_j^*}^T R_{ij} (\bar{u}_j - u_j^*) - \frac{\partial V_i^*}{\partial e_i} \sum_{j \in \mathcal{N}_i} B_j (\bar{u}_j - u_j^*)\Big) dt \\
&- \frac{1}{2}\int_0^\infty \Big(\sum_{j \in \mathcal{N}} a_{ij} {v_j^*}^T \gamma_{ij}^2 (\bar{v}_j - v_j^*) - \frac{\partial V_i^*}{\partial e_i} \sum_{j \in \mathcal{N}_i} D_j (\bar{v}_j - v_j^*)\Big) dt.
\end{aligned}$$

After setting $\bar{u}_{\mathcal{N}_i} = u_{\mathcal{N}_i}^*$ and $\bar{v}_{\mathcal{N}_i} = v_{\mathcal{N}_i}^*$ leads to

$$\begin{aligned}
&\mathcal{J}_i(e_i(0); \bar{u}_i, u_{\mathcal{N}_i}^*, \bar{v}_i, v_{\mathcal{N}_i}^*) = \int_0^\infty \frac{1}{2}\Big((\bar{u}_i - u_i^*)^T R_{ii} (\bar{u}_i - u_i^*) \\
&\quad - \gamma_{ii}^2 (\bar{v}_i - v_i^*)^T (\bar{v}_i - v_i^*)\Big) dt \\
&\quad\quad + \frac{1}{2} e_i(0)^T P_i e_i(0),\ \forall i \in \mathcal{N}.
\end{aligned} \tag{21}$$

Now, setting $\bar{u}_i = u_i^*$, $\bar{u}_{\mathcal{N}_i} = u_{\mathcal{N}_i}^*$ and $\bar{v}_{\mathcal{N}_i} = v_{\mathcal{N}_i}^*$ in (21) leads to

$$\begin{aligned}
&\mathcal{J}_i(e_i(0); u_i^*, u_{\mathcal{N}_i}^*, \bar{v}_i, v_{\mathcal{N}_i}^*) = -\frac{1}{2}\gamma_{ii}^2 \int_0^\infty (\bar{v}_i - v_i^*)^T (\bar{v}_i - v_i^*) dt \\
&\quad + \frac{1}{2} e_i(0)^T P_i e_i(0), \forall i \in \mathcal{N}.
\end{aligned} \tag{22}$$

Setting $\bar{u}_{\mathcal{N}_i} = u_{\mathcal{N}_i}^*$, $\bar{v}_i = v_i^*$, and $\bar{v}_{\mathcal{N}_i} = v_{\mathcal{N}_i}^*$ in (21) leads to,

$$\begin{aligned}
&\mathcal{J}_i(e_i(0); \bar{u}_i, u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*) = \frac{1}{2}\int_0^\infty (\bar{u}_i - u_i^*)^T R_{ii} (\bar{u}_i - u_i^*) dt \\
&\quad + \frac{1}{2} e_i(0)^T P_i e_i(0),\ \forall i \in \mathcal{N}.
\end{aligned} \tag{23}$$

Finally setting $\bar{u}_i = u_i^*$, $\bar{u}_{\mathcal{N}_i} = u_{\mathcal{N}_i}^*$, $\bar{v}_i = v_i^*$, and $\bar{v}_{\mathcal{N}_i} = v_{\mathcal{N}_i}^*$ in (21) leads to,

$$\mathcal{J}_i(e_i(0); u_i^*, u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*) = \frac{1}{2} e_i(0)^T P_i e_i(0),\ \forall i \in \mathcal{N}. \tag{24}$$

From equations (21), (22), (23), (24) the inequalities (5) follow $\forall i \in \mathcal{N}$.

Let us take two different non negative matrix solutions P_i and $\mathcal{S}_i$. Since both satisfy (12) both differential games have values equal to $\frac{1}{2} e_i(0)^T P_i e_i(0),\ \forall i \in \mathcal{N}$ and $\frac{1}{2} e_i(0)^T \mathcal{S}_i e_i(0),\ \forall i \in \mathcal{N}$ respectively, independent of the length of the time horizon. Under the corresponding policies, $\|e_i(t)\| \to 0,\ \forall i \in \mathcal{N}$ as $t \to \infty$ and hence in the limit the value is obtained. Since the value has to be unique for each $e_i(0)$, this contradicts the initial hypothesis that $P_i \neq \mathcal{S}_i,\ \forall i \in \mathcal{N}$. ■

Remark 5. The cooperative Q-learning problem can be tackled by an $\mathcal{L}_2$ gain synchronization problem where we consider the systems (3), and performance outputs $\|z_i\|^2 = \big(e_i^T H_i e_i + u_i^T R_{ii} u_i + \sum_{j \in \mathcal{N}_i} (u_j^T R_{ij} u_j - \gamma_{ij}^2 v_j^T v_j)\big)$. The adversarial input attenuation is generally defined by following [2] as,

$$\frac{\int_t^\infty \|z_i\|^2}{\int_t^\infty \|v_i\|^2} dt \leqslant \gamma_{ii}^2,\ \forall i \in \mathcal{N},$$

which is the same with (15) in other words we are interested in finding u_i such that the system is stable the the $\mathcal{L}_2$ gain is bounded by a prescribed value γ_{ii}, see Theorem 7 of [22] and Section 9.2 in [3]. This condition is a special case of the general definition of dissipativity and such inequalities are widely encountered in quadratic optimal control, estimation theory, and H_∞ control. □

IV. Q-FUNCTION

The value functions (11) need to be parametrized as functions of the neighborhood tracking error e_i, the controls u_i and $u_{\mathcal{N}_i}$ and the adversarial inputs v_i and $v_{\mathcal{N}_i}$ to represent the distributed Q-function, i.e. sparse cooperative learning, for each agent in the game. The optimal value given by (11) after adding the Hamiltonian from (7) can be written as the following distributed Q-function or *action-dependent value* $Q_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) : \mathbb{R}^{n + m_i + l_i + \sum_{j \in \mathcal{N}_i}(m_j + l_j)} \to \mathbb{R}$,

$$\begin{aligned}
&Q_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) := V_i^*(e_i) \\
&\quad + \mathcal{H}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}, \frac{\partial V_i^*}{\partial e_i}) \\
&= V_i^*(e_i) + \frac{\partial V_i^*}{\partial e_i}^T \Big(A e_i + (d_i + g_i)(B_i u_i + D_i v_i) \\
&- \sum_{j \in \mathcal{N}_i} a_{ij}(B_j u_j + D_j v_j)\Big) \\
&+ \frac{1}{2}\Big(e_i^T H_i e_i + (u_i^T R_{ii} u_i - \gamma_{ii}^2 v_i^T v_i) \\
&+ \sum_{j \in \mathcal{N}_i} (u_j^T R_{ij} u_j - \gamma_{ij}^2 v_j^T v_j)\Big),\ \forall e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i},\ \forall i \in \mathcal{N},
\end{aligned} \tag{25}$$

where $\mathcal{H}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}, \frac{\partial V_i^*}{\partial e_i})$ is given by (7) and the optimal cost is $V_i^*(e_i) = \frac{1}{2}e_i^T P_i e_i, \forall i \in \mathcal{N}$. Each agent's distributed Q-function (25) can be written in a compact quadratic in the neighborhood tracking error e_i, controls $u_i, u_{\mathcal{N}_i}$ and adversarial inputs $v_i, v_{\mathcal{N}_i}$ distributed form as in (26) (next page), where, $U_i := \begin{bmatrix} e_i^T & u_i^T & u_{\mathcal{N}_i}^T & v_i^T & v_{\mathcal{N}_i}^T \end{bmatrix}^T$, $\mathbf{0}$ are zero matrices of appropriate dimensions, $\operatorname{diag}\left(R_{ij}\right)_{j\in\mathcal{N}_i}$, $\operatorname{diag}\left(\gamma_{ij}\right)_{j\in\mathcal{N}_i}$ are stacked diagonal neighborhood matrices, $\operatorname{col}\left(a_{ij}B_j^T P_i\right)_{j\in\mathcal{N}_i}$, $\operatorname{col}\left(a_{ij}D_j^T P_i\right)_{j\in\mathcal{N}_i}$ are stacked column neighborhood matrices, and $\operatorname{row}\left(a_{ij}P_i B_j\right)_{j\in\mathcal{N}_i}$, $\operatorname{row}\left(a_{ij}P_i D_j\right)_{j\in\mathcal{N}_i}$ are stacked row neighborhood matrices.

In (26), the equivalences of $Q^i_{(\cdot)}$ are straightforward e.g. $Q^i_{\mathrm{e_ie_i}} = P_i + H_i + P_i A + A^T P_i$, $Q^i_{\mathrm{e_iu_i}} = (d_i + g_i)P_i B_i$, $Q^i_{\mathrm{e_iv_i}} = (d_i + g_i)P_i D_i$, $Q^i_{\mathrm{u_iu_i}} = R_{ii}$, $Q^i_{\mathrm{v_iv_i}} = \gamma_{ii}^2$ etc. and positive definite matrices $\bar{Q}^i \in \mathbb{R}^{(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j))\times(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j))}$, $\forall i \in \mathcal{N}$.

Remark 6. Note that in (26) and especially in the matrix $\bar{Q}^i$, $\forall i \in \mathcal{N}$ which is the unknown, to be estimated quantity, we have the information about the system dynamics, the pinning gains g_i (where the leader is pinned to) and the in-degree information d_i of each agent $i \in \mathcal{N}$. □

The next lemma proves that if $u_i = u_i^*,\ v_i = v_i^*,\ \forall i \in \mathcal{N}$ with u_i^* given by (13) and v_i^* given by (14), we can show that (26) and (6) have the same value.

Lemma 1. *Assume that the agents in the graph $\mathcal{G}$ are strongly connected, the neighborhood agents are already optimal, i.e. $u_{\mathcal{N}_i} = u_{\mathcal{N}_i}^*$, $v_{\mathcal{N}_i} = v_{\mathcal{N}_i}^*$ and that the distributed Q-function for every agent is given as in* (26)*. Then, the optimal value $V_i^*(e_i) = \frac{1}{2}e_i^T P_i e_i$ of each one of the distributed coupled optimization problems given by* (6)*, where $P_i > 0,\ \forall i \in \mathcal{N}$ is the solution to the coupled Hamilton-Jacobi equations* (10)*, is the same with the values of the following coupled optimization problems,*

$$\begin{aligned} Q_i^*(e_i, u_i^*, u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*) &= \min_{u_i} Q_i(e_i, u_i, u_{\mathcal{N}_i}^*, v_i^*, u_{\mathcal{N}_i}^*) \\ &= \max_{v_i} Q_i(e_i, u_i^*, u_{\mathcal{N}_i}^*, v_i, u_{\mathcal{N}_i}^*),\ \forall i \in \mathcal{N} \end{aligned}$$

and as a result all the policies $u_i,\ v_i,\ \forall i \in \mathcal{N}$ form a graphical Nash equilibrium.

Proof: The proof is a substitution of (13) into (25). Furthermore, since $P_i,\ \forall i \in \mathcal{N}$ are given by the solution of the coupled Hamilton-Jacobi equations (10) we know that $\mathcal{H}_i(e_i, u_i^*, u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*, \frac{\partial V_i^*}{\partial e_i}) = 0$ and hence this yields,

$$Q_i^*(e_i, u_i^*, u_{\mathcal{N}_i}^*, v_i^*, v_{\mathcal{N}_i}^*) = V_i^*(e_i),\ \forall e_i,\ \forall i \in \mathcal{N},$$

from which the result and (5) follow. ■

A model-free formulation of (13) can be found by solving $\frac{\partial Q_i(e_i,u_i,u_{\mathcal{N}_i},v_i,v_{\mathcal{N}_i})}{\partial u_i} = 0$ to write,

$$\begin{aligned} u_i^*(e_i) &= \arg\min_{u_i} Q_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) \\ &= -(Q^i_{\mathrm{u_iu_i}})^{-1} Q^i_{\mathrm{u_ie_i}} e_i,\ \forall i \in \mathcal{N}, \end{aligned} \tag{27}$$

and a model-free formulation of (14) can be found by solving $\frac{\partial Q_i(e_i,u_i,u_{\mathcal{N}_i},v_i,v_{\mathcal{N}_i})}{\partial v_i} = 0$ to write,

$$\begin{aligned} v_i^*(e_i) &= \arg\max_{v_i} Q_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) \\ &= -(Q^i_{\mathrm{v_iv_i}})^{-1} Q^i_{\mathrm{v_ie_i}} e_i,\ \forall i \in \mathcal{N}, \end{aligned} \tag{28}$$

Remark 7. The control input (27) and the adversarial input (28) are completely model free. They do not depend on any system matrices and furthermore does not need any information on the in-degree d_i and any information on where the leader is connected in the network, i.e. g_i. All of this information will be learned online as will be shown in the next subsection. □

Since we would like to develop a model-free online tuning algorithm to learn the parameters of the action-dependent value function, the next subsection will introduce 2-actor/single-critic structures to achieve that.

A. Control/Adversarial Actors and Critic Approximators

The critic approximator will approximate the distributed Q-function (26), the control actor approximator will approximate the optimal controller (27) and the adversarial input actor approximator will approximate the worst-case adversarial input (28) of each agent $i \in \mathcal{N}$. We shall see that each agent uses only local in formation for the three approximators. Specifically $Q_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i})$ can be written as,

$$Q_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) = \frac{1}{2}\operatorname{vech}\left(\bar{Q}^i\right)^T\left(U_i \otimes U_i\right),\ \forall i \in \mathcal{N}. \tag{29}$$

Now denote the $\operatorname{vech}\left(\bar{Q}^i\right) \in \mathbb{R}^{\frac{1}{2}(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j))(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j)+1)}$ as a half-vectorization of the matrix $\bar{Q}^i$.

By denoting as $W_{ic} := \frac{1}{2}\operatorname{vech}\left(\bar{Q}^i\right)$ we can write (29) in a compact form as,

$$Q_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) = W_{ic}^T\left(U_i \otimes U_i\right),\ \forall i \in \mathcal{N}, \tag{30}$$

with $W_{ic} \in \mathbb{R}^{\frac{1}{2}(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j))(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j)+1)}$ the ideal weights. The critic approximator is given by,

$$\hat{Q}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}) = \hat{W}_{ic}^T\left(U_i \otimes U_i\right),\ \forall i \in \mathcal{N}, \tag{31}$$

where $\hat{W}_{ic}$ are estimated critic weights of the optimal weights in (30); the control actor approximator is given as

$$\hat{u}_i(e_i) = \hat{W}_{ia}^T e_i,\ \forall i \in \mathcal{N}, \tag{32}$$

where $\hat{W}_{ia} \in \mathbb{R}^{n\times m_i}$ are estimated actor weights of the optimal weights in (27); and the adversarial actor approximator is given by,

$$\hat{v}_i(e_i) = \hat{W}_{id}^T e_i,\ \forall i \in \mathcal{N}, \tag{33}$$

where $\hat{W}_{id} \in \mathbb{R}^{n\times l_i}$ are estimated actor weights of the optimal weights in (28).

We have showed in [32] that the value function (6) for each agent satisfies the following Bellman equation,

$$V_i^*(e_i(t)) = V_i^*(e_i(t-T))$$

$$Q_i(e_i,u_i,u_{\mathcal{N}_i},v_i,v_{\mathcal{N}_i})=\frac{1}{2}U_i^T$$
$$\begin{bmatrix} P_i+H_i+P_iA+A^TP_i & (d_i+g_i)P_iB_i & -\mathrm{row}\big(a_{ij}P_iB_j\big) & (d_i+g_i)P_iD_i & -\mathrm{row}\big(a_{ij}P_iD_j\big)_{j\in\mathcal{N}_i} \\ (d_i+g_i)B_i^TP_i & R_{ii} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathrm{col}\big(a_{ij}B_j^TP_i\big)_{j\in\mathcal{N}_i} & \mathbf{0} & \mathrm{diag}\left(R_{ij}\right)_{j\in\mathcal{N}_i} & \mathbf{0} & \mathbf{0} \\ (d_i+g_i)D_i^TP_i & \mathbf{0} & \mathbf{0} & -\gamma_{ii}^2 & \mathbf{0} \\ -\mathrm{col}\big(a_{ij}D_j^TP_i\big)_{j\in\mathcal{N}_i} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathrm{diag}\left(\gamma_{ij}^2\right)_{j\in\mathcal{N}_i} \end{bmatrix} U_i$$
$$:=\frac{1}{2}U_i^T\begin{bmatrix} Q^i_{\mathrm{e_ie_i}} & Q^i_{\mathrm{e_iu_i}} & Q^i_{\mathrm{e_iu_{\mathcal{N}_i}}} & Q^i_{\mathrm{e_iv_i}} & Q^i_{\mathrm{e_iv_{\mathcal{N}_i}}} \\ Q^i_{\mathrm{u_ie_i}} & Q^i_{\mathrm{u_iu_i}} & Q^i_{\mathrm{u_iu_{\mathcal{N}_i}}} & Q^i_{\mathrm{u_iv_i}} & Q^i_{\mathrm{u_iv_{\mathcal{N}_i}}} \\ Q^i_{\mathrm{u_{\mathcal{N}_i}e_i}} & Q^i_{\mathrm{u_{\mathcal{N}_i}u_i}} & Q^i_{\mathrm{u_{\mathcal{N}_i}u_{\mathcal{N}_i}}} & Q^i_{\mathrm{u_{\mathcal{N}_i}v_i}} & Q^i_{\mathrm{u_{\mathcal{N}_i}v_{\mathcal{N}_i}}} \\ Q^i_{\mathrm{v_ie_i}} & Q^i_{\mathrm{v_iu_i}} & Q^i_{\mathrm{v_iu_{\mathcal{N}_i}}} & Q^i_{\mathrm{v_iv_i}} & Q^i_{\mathrm{v_iv_{\mathcal{N}_i}}} \\ Q^i_{\mathrm{v_{\mathcal{N}_i}e_i}} & Q^i_{\mathrm{v_{\mathcal{N}_i}u_i}} & Q^i_{\mathrm{v_{\mathcal{N}_i}u_{\mathcal{N}_i}}} & Q^i_{\mathrm{v_{\mathcal{N}_i}v_i}} & Q^i_{\mathrm{v_{\mathcal{N}_i}v_{\mathcal{N}_i}}} \end{bmatrix} U_i := \frac{1}{2}U_i^T\bar{Q}^iU_i,\ \forall i\in\mathcal{N}, \tag{26}$$

$$-\frac{1}{2}\int_{t-T}^{t}\big(e_i^TH_ie_i+({u_i^*}^TR_{ii}u_i^*-\gamma_{ii}^2{v_i^*}^Tv_i^*) + \sum_{j\in\mathcal{N}_i}({u_j^*}^TR_{ij}u_j^*-\gamma_{ij}^2{v_j^*}^Tv_j^*)\big)d\tau,\ \forall i\in\mathcal{N}, \tag{34}$$

where $T\in\mathbb{R}^+$ is a small fixed time interval that defines how fast the agent measures her own state and the neighboring states.

By using the result from Lemma 1, we can write (34) as,

$$\begin{aligned} &Q_i^*(e_i(t),u_i^*(t),u_{\mathcal{N}_i}^*(t),v_i^*(t),v_{\mathcal{N}_i}^*(t)) = \\ &Q^*(e_i(t-T),u_i^*(t-T),u_{\mathcal{N}_i}^*(t-T),v_i^*(t-T),v_{\mathcal{N}_i}^*(t-T)) \\ &-\frac{1}{2}\int_{t-T}^{t}\big(e_i^TH_ie_i+({u_i^*}^TR_{ii}u_i^*-\gamma_{ii}^2{v_i^*}^Tv_i^*) \\ &+\sum_{j\in\mathcal{N}_i}({u_j^*}^TR_{ij}u_j^*-\gamma_{ij}^2{v_j^*}^Tv_j^*)\big)d\tau,\ \forall i\in\mathcal{N}. \end{aligned} \tag{35}$$

Remark 8. Note that although the value of the small fixed time interval T does not affect in any way the convergence property of the algorithm, it is related to the excitation condition necessary in the setup of a numerically posed least squares problem and the least squares solution. For more details, please see [32]. □

B. Cooperative Learning Solution

To find the update law for the critic approximator of each agent $i\in\mathcal{N}$ we define the following error $E_i\in\mathbb{R}$ that we would like to eventually drive to zero by tuning $\hat{W}_{ic},\forall i\in\mathcal{N}$ appropriately.This error is defined by the difference between the left and the right-hand sides of (35) with the actual weights replaced by $\hat{W}_{ic}$.

$$\begin{aligned} E_i := {}& \hat{W}_{ic}^T\big(U_i(t)\otimes U_i(t)\big)+\frac{1}{2}\int_{t-T}^{t}\big(e_i^TH_ie_i+(\hat{u}_i^TR_{ii}\hat{u}_i \\ &-\gamma_{ii}^2\hat{v}_i^T\hat{v}_i)+\sum_{j\in\mathcal{N}_i}(\hat{u}_j^TR_{ij}\hat{u}_j-\gamma_{ij}^2\hat{v}_j^T\hat{v}_j)\big)d\tau \\ &-\hat{W}_{ic}^T\big(U_i(t-T)\otimes U_i(t-T)\big),\ \forall i\in\mathcal{N}. \end{aligned}$$

For each agent's control actor approximator we can define the error $E_{ia}\in\mathbb{R}^{m_i}$ as the difference between the control (32) that uses the estimated weights and the optimal control (27) with the matrices $Q^i_{\mathrm{u_iu_i}}$, and $Q^i_{\mathrm{u_ie_i}}$ replaced by the critic estimates extracted from $\hat{W}_{ic}$:

$$E_{ia}:=\hat{W}_{ia}^Te_i+(\hat{Q}^i_{\mathrm{u_iu_i}})^{-1}\hat{Q}^i_{\mathrm{u_ie_i}}e_i,\ \forall i\in\mathcal{N}.$$

Similarly for each agent's adversarial actor approximator we can define the error $E_{id}\in\mathbb{R}^{l_i}$ as the difference between the control (33) that uses the estimated weights and the optimal control (28) with the matrices $Q^i_{\mathrm{v_iv_i}}$, and $Q^i_{\mathrm{v_ie_i}}$ replaced by the critic estimates extracted from $\hat{W}_{ic}$:

$$E_{id}:=\hat{W}_{id}^Te_i+(\hat{Q}^i_{\mathrm{v_iv_i}})^{-1}\hat{Q}^i_{\mathrm{v_ie_i}}e_i,\ \forall i\in\mathcal{N}.$$

To find tuning updates for $\hat{W}_{ic}$, $\hat{W}_{ia}\in\mathbb{R}^{n\times m_i}$, and $\hat{W}_{id}\in\mathbb{R}^{n\times l_i}$ such that the errors $E_i\in\mathbb{R}$, $E_{ia}\in\mathbb{R}^{m_i}$ and $E_{id}\in\mathbb{R}^{l_i}$ go to zero $\forall i\in\mathcal{N}$ we follow adaptive control techniques as in [11] and define the squared-norm of errors of each agent $i\in\mathcal{N}$, E_i, E_{ia}, and E_{id} as,

$$K_{i1}=\frac{1}{2}\left\|E_i\right\|^2,\ \forall i\in\mathcal{N}, \tag{36}$$

$$K_{i2}=\frac{1}{2}\left\|E_{ia}\right\|^2,\ \forall i\in\mathcal{N}, \tag{37}$$

and

$$K_{ie}=\frac{1}{2}\left\|E_{id}\right\|^2,\ \forall i\in\mathcal{N}. \tag{38}$$

The gradient descent estimate of $\hat{W}_{ic}$ for the critic weights of each agent can be constructed by differentiating (36) to yield (39) (next page), where $\alpha_{ic}\in\mathbb{R}^+$ is a constant gain that determines the speed of the critic approximator convergence. Similarly by using (37), the gradient descent estimate of $\hat{W}_{ia}$ for the control actor weights can be constructed as,

$$\dot{\hat{W}}_{ia}=-\alpha_{ia}\frac{\partial K_{i2}}{\partial\hat{W}_{ia}}=-\alpha_{ia}e_iE_{ia}^T,\ \forall i\in\mathcal{N}, \tag{40}$$

where $\alpha_{ia}\in\mathbb{R}^+$ is a constant gain that determines the speed of the actor approximator convergence, and finally by using (38) the gradient descent estimate of $\hat{W}_{id}$ for the adversarial actor weights can be constructed as,

$$\dot{\hat{W}}_{id}=-\alpha_{id}\frac{\partial K_{i2}}{\partial\hat{W}_{id}}=-\alpha_{id}e_iE_{id}^T,\ \forall i\in\mathcal{N}, \tag{41}$$

$$
\begin{aligned}
\dot{\hat{W}}_{ic} &= -\alpha_{ic}\frac{\partial K_{i1}}{\partial \hat{W}_{ic}} \\
&= -\alpha_{ic}\frac{\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)}{\Big(1+\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)^T\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)\Big)^2}E_i,\ \forall i\in\mathcal{N}, \quad (39)
\end{aligned}
$$

where $\alpha_{id}\in\mathbb{R}^+$ is a constant gain that determines the speed of the actor approximator convergence.

The dynamics of the weight estimation errors $\tilde{W}_{ic} := W_{ic}-\hat{W}_{ic}$, $\tilde{W}_{ia} := -Q^i_{\mathrm{e_i u_i}}(Q^i_{\mathrm{u_i u_i}})^{-1}-\hat{W}_{ia}$ and $\tilde{W}_{id} := -Q^i_{\mathrm{e_i v_i}}(Q^i_{\mathrm{v_i v_i}})^{-1}-\hat{W}_{id}$ are given by (42) (next page),

$$
\dot{\tilde{W}}_{ia} = -\alpha_{ia}e_ie_i^T\tilde{W}_{ia}-\alpha_{ia}e_ie_i^T\tilde{Q}^i_{\mathrm{e_i u_i}}(\hat{Q}^i_{\mathrm{u_i u_i}})^{-1},\ \forall i\in\mathcal{N}, \quad (43)
$$

$$
\dot{\tilde{W}}_{id} = -\alpha_{id}e_ie_i^T\tilde{W}_{id}-\alpha_{id}e_ie_i^T\tilde{Q}^i_{\mathrm{e_i v_i}}(\hat{Q}^i_{\mathrm{v_i v_i}})^{-1},\ \forall i\in\mathcal{N}, \quad (44)
$$

with $\tilde{Q}^i_{(\cdot)}$ coming from the matrization (in a sense of an "inverse" half-vectorization that defined before) of the critic error vector $\tilde{W}_{ic}$ of each agent. Note that, we actually know the values of $\hat{Q}^i_{\mathrm{u_i u_i}},\forall i\in\mathcal{N}$ and $\hat{Q}^i_{\mathrm{v_i v_i}},\forall i\in\mathcal{N}$ and therefore we can write (43), (44) as,

$$
\dot{\tilde{W}}_{ia} = -\alpha_{ia}e_ie_i^T\tilde{W}_{ia}-\alpha_{ia}e_ie_i^T\tilde{Q}^i_{\mathrm{e_i u_i}}R_{ii}^{-1},\ \forall i\in\mathcal{N}, \quad (45)
$$

and

$$
\dot{\tilde{W}}_{id} = -\alpha_d e_ie_i^T\tilde{W}_{id}-\frac{\alpha_{id}}{\gamma_{ii}^2}e_ie_i^T\tilde{Q}^i_{\mathrm{e_i v_i}},\ \forall i\in\mathcal{N}. \quad (46)
$$

The next result shows that the tuning algorithm for the critic approximator given by (39) is effective under a persistence of excitation condition, in that the estimated weights $\hat{W}_{ic}$ converge to the optimal unknown values W_{ic} exponentially fast.

Definition 2. We define that the vector signal $\Delta_i(t) := \frac{\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)}{1+\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)^T\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)}$ with $T\in\mathbb{R}^+$ is *persistently exciting* over the interval $[t,t+T_{\mathrm{PE}}]$, with $T_{\mathrm{PE}}\in\mathbb{R}^+$ if there exists $\beta_i,\sigma_i\in\mathbb{R}^+$ such that $\beta_iI\leqslant\int_t^{t+T}\Delta_i(\tau)\Delta_i^T(\tau)d\tau\leqslant\sigma_iI,\forall t$ with I an identity matrix of appropriate dimensions. This is equivalent to requiring that the matrix $\int_t^{t+T_{\mathrm{PE}}}\Delta_i(\tau)\Delta_i^T(\tau)d\tau$ be positive definite over any finite interval. □

Lemma 2. *Let the tuning of the critic approximator of each agent $i\in\mathcal{N}$ be given by (39) with fixed control inputs $u_i,u_{\mathcal{N}_i}$, then the critic error dynamics given by (42) have an exponentially stable equilibrium point satisfying $\left\|\tilde{W}_{ic}(t)\right\|\leqslant\left\|\tilde{W}_{ic}(t_0)\right\|\rho_{i1}e^{-\rho_{i2}(t-t_0)}$, for $t>t_0\geqslant 0$, for some $\rho_{i1},\rho_{i2}\in\mathbb{R}^+$ provided that the signal $\Delta_i(t)$ with $T\in\mathbb{R}^+$, is persistently exciting over the interval $[t,t+T_{\mathrm{PE}}]$.*

Proof: Consider the following Lyapunov function $\mathcal{L}:\mathbb{R}^{\frac{1}{2}(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j))(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j)+1)}\to\mathbb{R}$,

$$
\mathcal{L}_i = \frac{1}{2\alpha_{ic}}\left\|\tilde{W}_{ic}\right\|^2,\ t\geqslant 0,\ \forall i\in\mathcal{N}. \quad (47)
$$

The time derivative of (47) is given by

$$
\dot{\mathcal{L}}_i = \frac{1}{\alpha_{ic}}\tilde{W}_{ic}^T\dot{\tilde{W}}_{ic},\ \forall i\in\mathcal{N}
$$

and after substituting (42) becomes,

$$
\dot{\mathcal{L}}_i = -\tilde{W}_{ic}^T\Delta_i\Delta_i^T\tilde{W}_{ic},\forall i\in\mathcal{N}.
$$

Viewing the (42) as a linear-time varying system, the solution $\tilde{W}_{ic}$ is given as,

$$
\tilde{W}_{ic}(t) = \Phi_i(t,t_0)\tilde{W}_{ic}(t_0),\ \forall t,t_0>0,\ \forall i\in\mathcal{N}, \quad (48)
$$

where the state transition matrix is defined as $\frac{\partial\Phi_i(t,t_0)}{\partial t} := -\alpha_{ic}\Delta_i(t)\Delta_i^T(t)\Phi_i(t,t_0)$. Then we can prove by following [11] and [15], that the equilibrium point is exponentially stable provided that $\Delta_i(t)$ is persistently exciting and therefore for some $\rho_{i1},\rho_{i2}\in\mathbb{R}^+$ we can write,

$$
\|\Phi_i(t,t_0)\| = \rho_{i1}e^{-\rho_{i2}(t-t_0)},\ \forall t,t_0>0,\forall i\in\mathcal{N}. \quad (49)
$$

Finally, by combining (48) and (49) we have,

$$
\left\|\tilde{W}_{ic}(t)\right\|\leqslant\left\|\tilde{W}_{ic}(t_0)\right\|\rho_{i1}e^{-\rho_{i2}(t-t_0)},\forall t,t_0>0,\forall i\in\mathcal{N},
$$

from which the result follows. ■

Remark 9. One can ensure that the signal Δ_i is persistently exciting over the interval $[t,t+T_{\mathrm{PE}}]$ by adding exploration noise in the control inputs. The exploration noise should be sinusoids of $\frac{1}{2}(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j))(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j)+1)$ different frequencies according to [11]. □

A pseudocode that describes the algorithm has the following form,

Algorithm 1: Proposed Q-learning

procedure
  Start with initial conditions for every agent $x_i(0)$, and random initial weights $\hat{W}_{ic}(0),\hat{W}_{ia}(0),\hat{W}_{id}(0)$ for the critic, actor and worst case adversarial approximator for each agent.
  Propagate $t,x_i(t)$ and $x_j(t)$ in the neighborhood to compute e_i from (2).
  Propagate $\hat{W}_{ic}(t),\hat{W}_{ia}(t),\hat{W}_{id}(t)$ ▷ $\dot{\hat{W}}_{ic}$ as in (39), $\dot{\hat{W}}_{ia}$ as in (40) and $\dot{\hat{W}}_{id}$ as in (41).
  Compute the Q function $\hat{Q}_i$ from (31), the optimal control $\hat{u}_i$ from (32) and the worst case adversarial input $\hat{v}_i$ from (33) for each agent.
end procedure

$$\dot{\tilde{W}}_{ic} = -\alpha_{ic} \frac{\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)^T}{\left(1+\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)^T\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)\right)^2}\tilde{W}_{ic},\ \forall i \in \mathcal{N}, \tag{42}$$

The main Theorem proved in the next subsection, is presented next and guarantees the closed-loop stability of the origin of $\psi_i := \begin{bmatrix} e_i^T & \tilde{W}_{ic}^T & \tilde{W}_{ia}^T & \tilde{W}_{\mathcal{N}_i a}^T & \tilde{W}_{id}^T & \tilde{W}_{\mathcal{N}_i d}^T \end{bmatrix}^T$ where $\tilde{W}_{\mathcal{N}_i a}^T$ and $\tilde{W}_{\mathcal{N}_i d}^T$ are the control actor and adversarial error estimation weights of the neighbors of agent i.

Property 1. According to [12], [17], [29], [32], using the universal function approximation property for on-line parameter estimation, there exist known positive constants that upper bound the components $Q^i_\cdot$ of $\bar{Q}^i$. □

Theorem 2. *Suppose that the hypotheses and the statements of Theorem 1 hold. Assume that the agents are strongly connected in the networked system $\mathcal{G}$. Consider the neighborhood tracking error dynamics given by* (3)*, the critic approximator for each agent $i \in \mathcal{N}$ given by* (31)*, the optimal control for each agent $i \in \mathcal{N}$ given by* (32)*, the worst case adversarial input for each agent $i \in \mathcal{N}$ given by* (33)*, the tuning law for the weights of the critic given by* (39)*, for the weights of the control actor given by* (40)*, and for the weights of the adversarial actor given by* (41)*. Assume that the tuning laws for the critic are faster than the tuning laws for the control and adversarial actors, and*

$$\alpha_{ia} > 1;\ \alpha_{id} > 1, \tag{50}$$

$$\alpha_{ia}\bar{\lambda}(R_{ii}^{-1}) + \frac{\alpha_{id}}{\gamma_{ii}^2} < \frac{2}{\delta_i}\Bigg(\underline{\lambda}\big(H_i + Q^i_{\mathrm{e_i u_i}}R_{ii}^{-1}(Q^i_{\mathrm{e_i u_i}})^T\big) - \frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_i u_i}}(Q^i_{\mathrm{e_i u_i}})^T + \frac{2}{\gamma_{ii}^2}Q^i_{\mathrm{e_i v_i}}(Q^i_{\mathrm{e_i v_i}})^T\big)\Bigg), \tag{51}$$

$$\alpha_{ja}\bar{\lambda}(R_{jj}^{-1}) + \frac{\alpha_{jd}}{\gamma_{jj}^2} < \frac{2}{\delta_j}\Bigg(\underline{\lambda}\big(Q^j_{\mathrm{e_j u_j}}R_{jj}^{-1}R_{ij}R_{jj}^{-1}(Q^j_{\mathrm{e_j u_j}})^T\big) - \frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_i u_j}}(Q^i_{\mathrm{e_i u_j}})^T + \frac{\gamma_{ij}^2}{\gamma_{jj}^4}Q^i_{\mathrm{e_i v_j}}(Q^i_{\mathrm{e_i v_j}})^T\big)\Bigg),\ \forall j \in \mathcal{N}_i, \tag{52}$$

where $\delta_i \in (0,1)$. Then the equilibrium point of $\psi_i,\ \forall i \in \mathcal{N}$ and for all initial conditions $\psi_i(0)$ is globally asymptotically stable. □

Remark 10. In order to satisfy the conditions (50), (51), (52) we have to pick the tuning gains $\alpha_{ia}, \alpha_{id}, \forall i \in \mathcal{N}$ for each control and adversarial actor such that all three inequalities are met simultaneously. □

Remark 11. The necessary conditions given in (51) and (52) can be satisfied by picking appropriately the user defined matrices $H_i,\ R_{ii}, R_{ij}, \gamma_{ii}, \gamma_{ij}\ \ \forall i,j \in \mathcal{N}$ and the tuning gains $\alpha_{ic} \gg \alpha_{ia}$, $\alpha_{ic} \gg \alpha_{id}$, $\forall i \in \mathcal{N}$ without needing to know the exact values of the submatrices $Q^i_{(\cdot)}$. Moreover by following Property 1, from the actor approximation errors for the control and the worst case adversarial inputs $\tilde{W}_{ia} := -Q^i_{\mathrm{e_i u_i}}R_{ii}^{-1} - \hat{W}_{ia}$ and $\tilde{W}_{id} := -Q^i_{\mathrm{e_i v_i}}\gamma_{ii}^{-2} - \hat{W}_{id}$ respectively one also has known bounds on $Q^i_{\mathrm{e_i u_i}}R_{ii}^{-1}$, $Q^j_{\mathrm{e_j u_j}}R_{jj}^{-1}$ and $\gamma_{ii}^{-2}Q^i_{\mathrm{e_i v_i}}$, $\gamma_{ij}^2\gamma_{jj}^{-4}Q^j_{\mathrm{e_j v_j}}$. Regarding the tuning gains conditions, as noted in adaptive control [11], large adaptive gains can cause high frequency oscillations in the control signal and reduced tolerance to time delays that will destabilize the system. There are not any systematic approaches to pick a satisfactory adaptation gain, hence trial and error, intuition or Monte Carlo simulations can serve as guidelines. □

Remark 12. In our algorithm there are no offline computations, neither requirements for storage of any data in the memory, but everything happens as in a plug-n-play framework for every agent by measuring the system trajectories in the neighborhood. The size of the network N increases the complexity since we will have more distributed optimization criteria to optimize but on the other side the position of leader(s) does not affect the convergence as long as the graph is strongly connected. The involved computation is as in many adaptive control algorithms and is dominated by the training algorithm for $\hat{W}_{ic}$, $\hat{W}_{ia}$ and $\hat{W}_{id}$ in order to approximate $\hat{Q}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i})$, $\hat{u}_i(e_i)$ and $\hat{v}_i(e_i)$. The complexity of the scheme for every agent increases with the number of states n and control inputs and worst case adversarial inputs in the neighborhood $(m_i+l_i)+\sum_{j\in\mathcal{N}_i}(m_j+l_j)$. Thus, for every agent the complexity is $\mathcal{O}\big(\frac{1}{2}(n+(m_i+l_i)+\sum_{j\in\mathcal{N}_i}(m_j+l_j))(n+(m_i+l_i)+\sum_{j\in\mathcal{N}_i}(m_j+l_j)+1)+n(m_i+l_i)\big)$. It has been shown in [12] (c.f. Theorem 4.3.2) and [32] (cf. Chapter 7) that the convergence rate of gradient descent algorithms of the form (39) (and of course (40) and (41) that copy the weights of (39)) as shown in Lemma 2 converge exponentially fast with rate that depends on the level of excitation β_i, σ_i, the size of the time interval T_{PE}, $m_{0i} = \sup_{t\geqslant 0}\Big\{1+\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)^T\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)\Big\}$ and α_{ic}. □

The following Corollary shows that convergence to the graphical Nash equilibrium and synchronization of all the agents to the leader is achieved asymptotically despite the presence of persistent adversarial inputs.

Corollary 1. *Suppose that the hypotheses of Theorem 2 hold. Then, the policies $\hat{u}_i, \hat{v}_i\ \forall i \in \mathcal{N}$ given by* (32) *and* (33) *asymptotically form a graphical Nash equilibrium and all the agents are synchronized to the leader state.*

Proof: According to Theorem 2, we have $\left\|W_{ic} - \hat{W}_{ic}\right\| \to$

$0,\ \forall i \in \mathcal{N}$, $\left\|W_{ia}-\hat{W}_{ia}\right\| \rightarrow 0,\ \forall i \in \mathcal{N}$ and $\left\|W_{id}-\hat{W}_{id}\right\| \rightarrow 0,\ \forall i \in \mathcal{N}$ and after taking into consideration (27), (32) and (28), (33) we have that $\|u_i^*-\hat{u}_i\| \rightarrow 0,\ \forall i \in \mathcal{N}$ and $\|v_i^*-\hat{v}_i\| \rightarrow 0,\ \forall i \in \mathcal{N}$ from which the results follow according to (5). Moreover we also conclude that from Theorem 2 that $\|e_i\| \rightarrow 0,\ \forall i \in \mathcal{N}$, which means that all the agents synchronize to the leader state. ■

C. Proof of Theorem 2

In order to prove stability we shall start with the following Lyapunov function $\mathcal{V}: \mathbb{R}^n \times \mathbb{R}^{\frac{1}{2}(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j))(n+m_i+l_i+\sum_{j\in\mathcal{N}_i}(m_j+l_j)+1)} \times \mathbb{R}^{n\times m_i} \times \mathbb{R}^{n\times\sum_{j\in\mathcal{N}_i} m_j} \times \mathbb{R}^{n\times l_i} \times \mathbb{R}^{n\times\sum_{j\in\mathcal{N}_i} l_j} \rightarrow \mathbb{R}$,

$$\begin{aligned}\mathcal{V} = \sum_{i=1}^{N}\Bigg(&V_i^*(e_i)+\frac{1}{2}\left\|\tilde{W}_{ic}\right\|^2 \\ &+\frac{1}{2}\operatorname{trace}\{\tilde{W}_{ia}^T\tilde{W}_{ia}\}+\frac{1}{2}\sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{ja}^T\tilde{W}_{ja}\} \\ &+\frac{1}{2}\operatorname{trace}\{\tilde{W}_{id}^T\tilde{W}_{id}\}+\frac{1}{2}\sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{jd}^T\tilde{W}_{jd}\}\Bigg),\ t \geqslant 0.\end{aligned} \tag{53}$$

The time derivative of (53) after computing the derivative of $V_i^*(e_i) := \frac{1}{2}e_i^T P_i e_i$ along the closed-loop trajectories with controllers $\hat{u}_i,\ \forall i \in \mathcal{N}$ and adversarial inputs $\hat{v}_i,\ \forall i \in \mathcal{N}$ yields,

$$\begin{aligned}\dot{\mathcal{V}} = \sum_{i=1}^{N}\Bigg(&\frac{\partial V_i^*(e_i)}{\partial e_i}^T\Big(Ae_i+(d_i+g_i)(B_i\hat{u}_i+D_i\hat{v}_i) \\ &-\sum_{j\in\mathcal{N}}a_{ij}(B_j\hat{u}_j+D_j\hat{v}_j)\Big) \\ &+\tilde{W}_{ic}^T\dot{\tilde{W}}_{ic}+\operatorname{trace}\{\tilde{W}_{ia}^T\dot{\tilde{W}}_{ia}\}+\sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{ja}^T\dot{\tilde{W}}_{ja}\} \\ &+\operatorname{trace}\{\tilde{W}_{id}^T\dot{\tilde{W}}_{id}\}+\sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{jd}^T\dot{\tilde{W}}_{jd}\}\Bigg),\end{aligned} \tag{54}$$

and after substituting (42), (45), and (46) in (54) we have (55) (next page).

We can write (55) as four separate terms,

$$\dot{\mathcal{V}} = \sum_{i=1}^{N}\big(T_{i1}+T_{i2}+T_{i3}+T_{i4}\big),$$

where,

$$\begin{aligned}T_{i1} := &\frac{\partial V_i^*(e_i)}{\partial e_i}^T\Big(Ae_i-(d_i+g_i)B_i\tilde{W}_{ia}^Te_i-(d_i+g_i)D_i\tilde{W}_{id}^Te_i \\ &+(d_i+g_i)B_iu_i^*+(d_i+g_i)D_iv_i^*+\sum_{j\in\mathcal{N}}a_{ij}(B_j\tilde{W}_{ja}^Te_j-B_ju_j^*) \\ &+\sum_{j\in\mathcal{N}}a_{ij}(D_j\tilde{W}_{jd}^Te_j-D_jv_j^*)\Big),\end{aligned} \tag{56}$$

T_{i2} is given as (57) (next page), T_{i3} is given as,

$$\begin{aligned}T_{i3} := &-\alpha_{ia}\operatorname{trace}\{\tilde{W}_{ia}^Te_ie_i^T\tilde{W}_{ia}+\tilde{W}_{ia}^Te_ie_i^T\tilde{Q}^i_{\mathrm{e_iu_i}}R_{ii}^{-1}\} \\ &-\sum_{j\in\mathcal{N}_i}\alpha_{ja}\operatorname{trace}\{\tilde{W}_{ja}^Te_je_j^T\tilde{W}_{ja}+\tilde{W}_{ja}^Te_je_j^T\tilde{Q}^j_{\mathrm{e_ju_j}}R_{jj}^{-1}\},\end{aligned} \tag{58}$$

and finally, T_{i4} is given as,

$$\begin{aligned}T_{i4} := &-\alpha_{id}\operatorname{trace}\{\tilde{W}_{id}^Te_ie_i^T\tilde{W}_{id}+\tilde{W}_{ia}^Te_ie_i^T\tilde{Q}^i_{\mathrm{e_iv_i}}\gamma_{ii}^{-2}\} \\ &-\sum_{j\in\mathcal{N}_i}\alpha_{jd}\operatorname{trace}\{\tilde{W}_{jd}^Te_je_j^T\tilde{W}_{jd}+\tilde{W}_{jd}^Te_je_j^T\tilde{Q}^j_{\mathrm{e_jv_j}}\gamma_{jj}^{-2}\}.\end{aligned} \tag{59}$$

Using the coupled Hamilton-Jacobi equations (10), (12), i.e. subtracting zero from (56) we have,

$$\begin{aligned}T_{i1} = &\frac{\partial V_i^*(e_i)}{\partial e_i}^T\Big(Ae_i-(d_i+g_i)B_i\tilde{W}_{ia}^Te_i-(d_i+g_i)D_i\tilde{W}_{id}^Te_i \\ &+(d_i+g_i)B_iu_i^*+(d_i+g_i)D_iv_i^*+\sum_{j\in\mathcal{N}}a_{ij}(B_j\tilde{W}_{ja}^Te_j-B_ju_j^*) \\ &+\sum_{j\in\mathcal{N}}a_{ij}(D_j\tilde{W}_{jd}^Te_j-D_jv_j^*)\Big) \\ &-\Big(e_i^TP_i\big(Ae_i-(d_i+g_i)^2(B_iR_{ii}^{-1}B_i^T-\frac{1}{\gamma_{ii}^2}D_iD_i^T)P_ie_i \\ &+\sum_{j\in\mathcal{N}}a_{ij}(d_j+g_j)(B_jR_{jj}^{-1}B_j^T-\frac{1}{\gamma_{jj}^2}D_jD_j^T)P_je_j\big) \\ &+\frac{1}{2}\sum_{j\in\mathcal{N}_i}(d_j+g_j)^2e_j^TP_j(B_jR_{jj}^{-T}R_{ij}R_{jj}^{-1}B_j^T-\frac{\gamma_{ij}^2}{\gamma_{jj}^4}D_jD_j^T)P_je_j \\ &+\frac{1}{2}(d_i+g_i)^2e_i^TP_i(B_iR_{ii}B_i^T-\frac{1}{\gamma_{ii}^2}D_iD_i^T)P_ie_i+\frac{1}{2}e_i^TH_ie_i\Big),\end{aligned}$$

or

$$\begin{aligned}T_{i1} = &-\frac{1}{2}\sum_{j\in\mathcal{N}_i}(d_j+g_j)^2e_j^TP_j(B_jR_{jj}^{-T}R_{ij}R_{jj}^{-1}B_j^T \\ &-\frac{\gamma_{ij}^2}{\gamma_{jj}^4}D_jD_j^T)P_je_j-\frac{1}{2}(d_i+g_i)^2e_i^TP_i(B_iR_{ii}B_i^T \\ &-\frac{1}{\gamma_{ii}^2}D_iD_i^T)P_ie_i-\frac{1}{2}e_i^TH_ie_i \\ &+\frac{\partial V_i^*(e_i)}{\partial e_i}^T\big(\sum_{j\in\mathcal{N}}a_{ij}(B_j\tilde{W}_{ja}^T+D_j\tilde{W}_{jd}^T)e_j \\ &-(d_i+g_i)(B_i\tilde{W}_{ia}^T+D_i\tilde{W}_{id}^T)e_i\big).\end{aligned} \tag{60}$$

The term (60) can be upper bounded after using Young's inequality by,

$$\begin{aligned}T_{i1} \leqslant &-\Big(\underline{\lambda}\big(H_i+Q^i_{\mathrm{e_iu_i}}R_{ii}^{-1}(Q^i_{\mathrm{e_iu_i}})^T\big) \\ &-\frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_iu_i}}(Q^i_{\mathrm{e_iu_i}})^T+\frac{2}{\gamma_{ii}^2}Q^i_{\mathrm{e_iv_i}}(Q^i_{\mathrm{e_iv_i}})^T\big)\Big)\|e_i\|^2 \\ &-\sum_{j\in\mathcal{N}_i}\Big(\underline{\lambda}\big(Q^j_{\mathrm{e_ju_j}}R_{jj}^{-1}R_{ij}R_{jj}^{-1}(Q^j_{\mathrm{e_ju_j}})^T\big) \\ &-\frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_iu_j}}(Q^i_{\mathrm{e_iu_j}})^T+\frac{\gamma_{ij}^2}{\gamma_{jj}^4}Q^i_{\mathrm{e_iv_j}}(Q^i_{\mathrm{e_iv_j}})^T\big)\Big)\|e_j\|^2 \\ &+\frac{1}{2}\left\|e_i^T\tilde{W}_{ia}\right\|^2+\frac{1}{2}\sum_{j\in\mathcal{N}_i}\left\|e_j^T\tilde{W}_{ja}\right\|^2 \\ &+\frac{1}{2}\left\|e_i^T\tilde{W}_{id}\right\|^2+\frac{1}{2}\sum_{j\in\mathcal{N}_i}\left\|e_j^T\tilde{W}_{jd}\right\|^2,\end{aligned} \tag{61}$$

$$\dot{\mathcal{V}} = \sum_{i=1}^{N}\Bigg(\frac{\partial V_i^*(e_i)}{\partial e_i}^T\Bigg(Ae_i - (d_i+g_i)B_i\tilde{W}_{ia}^T e_i - (d_i+g_i)D_i\tilde{W}_{id}^T e_i$$
$$+ (d_i+g_i)B_i u_i^* + (d_i+g_i)D_i v_i^* + \sum_{j\in\mathcal{N}} a_{ij}(B_j\tilde{W}_{ja}^T e_j - B_j u_j^*) + \sum_{j\in\mathcal{N}} a_{ij}(D_j\tilde{W}_{jd}^T e_j - D_j v_j^*)\Bigg)$$
$$- \alpha_{ic}\tilde{W}_{ic}^T\frac{\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)^T}{\Big(1+\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)^T\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)\Big)^2}\tilde{W}_{ic}$$
$$- \alpha_{ia}\operatorname{trace}\{\tilde{W}_{ia}^T e_i e_i^T \tilde{W}_{ia} + \tilde{W}_{ia}^T e_i e_i^T \tilde{Q}^i_{\mathrm{e_iu_i}} R_{ii}^{-1}\} - \sum_{j\in\mathcal{N}_i}\alpha_{ja}\operatorname{trace}\{\tilde{W}_{ja}^T e_j e_j^T\tilde{W}_{ja} + \tilde{W}_{ja}^T e_j e_j^T\tilde{Q}^j_{\mathrm{e_ju_j}}R_{jj}^{-1}\}$$
$$- \alpha_{id}\operatorname{trace}\{\tilde{W}_{id}^T e_i e_i^T\tilde{W}_{id} + \frac{1}{\gamma_{ii}^2}\tilde{W}_{id}^T e_i e_i^T\tilde{Q}^i_{\mathrm{e_iv_i}}\} - \sum_{j\in\mathcal{N}_i}\alpha_{jd}\operatorname{trace}\{\tilde{W}_{jd}^T e_j e_j^T\tilde{W}_{jd} + \frac{1}{\gamma_{jj}^2}\tilde{W}_{jd}^T e_j e_j^T\tilde{Q}^j_{\mathrm{e_jv_j}}\}\Bigg),\ t\geqslant 0. \quad (55)$$

$$T_{i2} := -\alpha_{ic}\tilde{W}_{ic}^T\frac{\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)^T}{\Big(1+\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)^T\Big(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\Big)\Big)^2}\tilde{W}_{ic}, \quad (57)$$

Now the term (57) can be upper bounded to write,

$$T_{i2} \leqslant -\frac{\alpha_{ic}}{4}\left\|\tilde{W}_{ic}\right\|^2. \quad (62)$$

According to Lemma 2 and (62) we have proved that the critic weights $\left\|\tilde{W}_{ic}\right\| \to 0,\ \forall i\in\mathcal{N}$. Since the actor error dynamics is based on the critic error dynamics of each agent in a form of a cascaded system, one needs to establish convergence of the critic way faster than the control actor and adversarial input actor approximators. This means that we need to pick $\alpha_{ic} \gg \alpha_{ia}, \alpha_{ic} \gg \alpha_{id}\ \forall i\in\mathcal{N}$. This can also be tackled similarly in view of Lemma 4.7 from [15] and treat the whole system of (42), (45), and (46) as input-to-state stable systems with $\tilde{W}_{ic}$ as an input in (45) and (46). Following the result of these developments we can upper bound (58), after using Young's inequality and also that the entry-wise norm or Frobenius norm of each submatrix of a matrix is less than the entry-wise norm of the matrix it is part of, i.e. $\left\|\tilde{Q}^i_{(\cdot)}\right\| \leqslant \left\|\tilde{W}_{ic}\right\|$, as follows,

$$\begin{aligned}T_{i3} &= -\alpha_{ia}\operatorname{trace}\{\tilde{W}_{ia}^T e_i e_i^T\tilde{W}_{ia}\} - \alpha_{ia}\operatorname{trace}\{\tilde{W}_{ia}^T e_i e_i^T\tilde{Q}^i_{\mathrm{e_iu_i}}R_{ii}^{-1}\}\\ &\quad - \sum_{j\in\mathcal{N}_i}\alpha_{ja}\operatorname{trace}\{\tilde{W}_{ja}^T e_j e_j^T\tilde{W}_{ja} + \tilde{W}_{ja}^T e_j e_j^T\tilde{Q}^j_{\mathrm{e_ju_j}}R_{jj}^{-1}\}\\ &\leqslant -\frac{\alpha_{ia}}{2}\left\|e_i^T\tilde{W}_{ia}\right\|^2 + \frac{\alpha_{ia}\delta_i}{2}\bar{\lambda}(R_{ii}^{-1})\left\|e_i\right\|^2\\ &\quad - \sum_{j\in\mathcal{N}_i}\frac{\alpha_{ja}}{2}\left\|e_j^T\tilde{W}_{ja}\right\|^2 + \sum_{j\in\mathcal{N}_i}\frac{\alpha_{ja}\delta_j}{2}\bar{\lambda}(R_{jj}^{-1})\left\|e_j\right\|^2, \quad (63)\end{aligned}$$

with $\delta_i(0,1),\ \forall i\in\mathcal{N}$. Similarly for the adversarial actor approximator we have for (59),

$$\begin{aligned}T_{i4} &= -\alpha_{id}\operatorname{trace}\{\tilde{W}_{id}^T e_i e_i^T\tilde{W}_{id}\} - \alpha_{id}\operatorname{trace}\{\tilde{W}_{id}^T e_i e_i^T\tilde{Q}^i_{\mathrm{e_iv_i}}\gamma_{ii}^{-2}\}\\ &\quad - \sum_{j\in\mathcal{N}_i}\alpha_{jd}\operatorname{trace}\{\tilde{W}_{jd}^T e_j e_j^T\tilde{W}_{jd} + \tilde{W}_{jd}^T e_j e_j^T\tilde{Q}^j_{\mathrm{e_jv_j}}\gamma_{jj}^{-2}\}\\ &\leqslant -\frac{\alpha_{id}}{2}\left\|e_i^T\tilde{W}_{id}\right\|^2 + \frac{\alpha_{id}\delta_i}{2\gamma_{ii}^2}\left\|e_i\right\|^2 - \sum_{j\in\mathcal{N}_i}\frac{\alpha_{jd}}{2}\left\|e_j^T\tilde{W}_{jd}\right\|^2\\ &\quad + \sum_{j\in\mathcal{N}_i}\frac{\alpha_{jd}\delta_j}{2\gamma_{jj}^2}\left\|e_j\right\|^2. \quad (64)\end{aligned}$$

Finally, by combining the terms (61), (62), (63) and (64) we have that,

$$\begin{aligned}\dot{\mathcal{L}} \leqslant \sum_{i=1}^{N}\Bigg(&-\Bigg(\underline{\lambda}\big(H_i + Q^i_{\mathrm{e_iu_i}}R_{ii}^{-1}(Q^i_{\mathrm{e_iu_i}})^T\big)\\ &-\frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_iu_i}}(Q^i_{\mathrm{e_iu_i}})^T + \frac{2}{\gamma_{ii}^2}Q^i_{\mathrm{e_iv_i}}(Q^i_{\mathrm{e_iv_i}})^T\big)\\ &\quad -\frac{\alpha_{ia}\delta_i}{2}\bar{\lambda}(R_{ii}^{-1}) - \frac{\alpha_{id}\delta_i}{2\gamma_{ii}^2}\Bigg)\left\|e_i\right\|^2\\ &-\sum_{j\in\mathcal{N}_i}\Bigg(\underline{\lambda}\big(Q^j_{\mathrm{e_ju_j}}R_{jj}^{-1}R_{ij}R_{jj}^{-1}(Q^j_{\mathrm{e_ju_j}})^T\big)\\ &-\frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_iu_j}}(Q^i_{\mathrm{e_iu_j}})^T + \frac{\gamma_{ij}^2}{\gamma_{jj}^4}Q^i_{\mathrm{e_iv_j}}(Q^i_{\mathrm{e_iv_j}})^T\big)\\ &\quad -\frac{\alpha_{ja}\delta_j}{2}\bar{\lambda}(R_{jj}^{-1}) - \frac{\alpha_{jd}\delta_j}{2\gamma_{jj}^2}\Bigg)\left\|e_j\right\|^2\\ &-\frac{\alpha_{ic}}{4}\left\|\tilde{W}_{ic}\right\|^2 - \frac{1}{2}(\alpha_{ia}-1)\left\|e_i^T\tilde{W}_{ia}\right\|^2\\ &-\frac{1}{2}\sum_{j\in\mathcal{N}_i}(\alpha_{ja}-1)\left\|e_j^T\tilde{W}_{ja}\right\|^2\\ &-\frac{1}{2}(\alpha_{id}-1)\left\|e_i^T\tilde{W}_{id}\right\|^2 - \frac{1}{2}\sum_{j\in\mathcal{N}_i}(\alpha_{jd}-1)\left\|e_j^T\tilde{W}_{jd}\right\|^2\Bigg),\end{aligned}$$

and after taking into account the conditions (50), (51), and (52) the result follows. ■

V. Numerical Example

In order to show the effectiveness of the proposed distributed Q-learning model-free approach we use a connected graph $\mathcal{G}$ with 10 agents as shown in Figure 1. The system dynamics are given in the form of (1) with a marginally stable matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, and $B_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $B_2 = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$, $B_3 = \begin{bmatrix} 0.6 \\ -4 \end{bmatrix}$, $B_4 = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$, $B_5 = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$, $B_6 = \begin{bmatrix} 4 \\ -3 \end{bmatrix}$, $B_7 = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$, $B_8 = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$, $B_9 = \begin{bmatrix} 0.5 \\ -2 \end{bmatrix}$, $B_{10} = \begin{bmatrix} 5 \\ -0.5 \end{bmatrix}$ and $D_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $D_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $D_3 = \begin{bmatrix} 1.2 \\ -2 \end{bmatrix}$, $D_4 = \begin{bmatrix} 1 \\ -3.7 \end{bmatrix}$, $D_5 = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$, $D_6 = \begin{bmatrix} 0.8 \\ -4.2 \end{bmatrix}$, $D_7 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $D_8 = \begin{bmatrix} 2.2 \\ -0.4 \end{bmatrix}$, $D_9 = \begin{bmatrix} 3.5 \\ 1.12 \end{bmatrix}$, $D_{10} = \begin{bmatrix} 1 \\ 0.9 \end{bmatrix}$ and a leader given as $\dot{x}_L = A x_L$.

The user-defined matrices in the distributed performances are given as $H_i = 10 I_2, \forall i \in \{1, 2, \ldots, 10\}$, $R_{ii} = 5, \forall i \in \{1, 2, \ldots, 10\}$, $R_{ij} = 1, \forall i \neq j \in \{1, 2, \ldots, 10\}$, $\gamma_{ii} = 6, \forall i \in \{1, 2, \ldots, 10\}$, $\gamma_{ij} = 2, \forall i \neq j \in \{1, 2, \ldots, 10\}$ and the tuning gains of the algorithm in Theorem 2 are picked as $\alpha_{ic} = 50, \alpha_{ia} = \alpha_{id} = 5, \forall i \in \{1, 2, \ldots, 10\}$ and $T = 0.01$sec. The initial actor and critic weights are picked randomly in $[0, 5]$ and the initial positions of the agents are shown with circles in Figure 2. For $0 \leqslant t \leqslant 0.5$ second we add exponentially decreasing exploration noise in the control inputs to ensure persistence of excitation and state exploration (see Remark 9).

Figure 2 shows the phase plane plot of the synchronization of the agents.

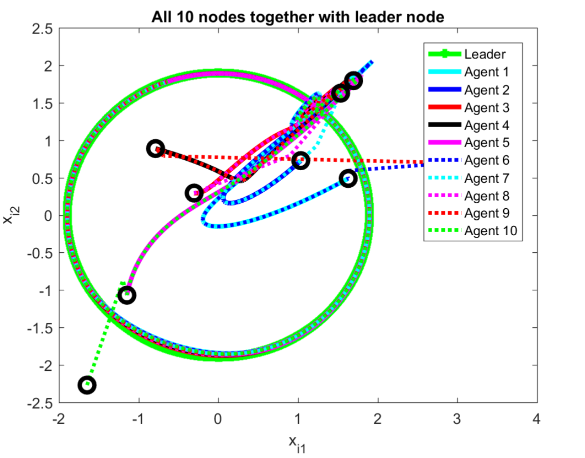

Fig. 2. Phase-plane plot of the states of agents that shows synchronization to the Lissajous curve of the leader.

VI. Conclusion and Future Work

This work proposes a model-free distributed Q-learning algorithm to learn the graphical Nash equilibrium while attenuating worst case adversarial inputs. In order to achieve this, we write a distributed Q-function for each agent as a parametrization of the neighborhood tracking error, and the control and adversarial inputs in the neighborhood. It is shown that all the agents pick their control and adversarial actions in order to form a graphical Nash equilibrium by solving N-coupled distributed optimization problems. In general, Q-learning functions need to store a lot of data throughout learning, we use an approximate dynamic programming framework, which is based on a $3N$ approximators structure that are tuned based on a distributed information available to every agent. The approximators are tuned in an adaptive manner such that they can guarantee closed-loop stability and the control and adversarial policies form a graphical Nash equilibrium, without requiring any information of the system dynamics while their states are being synchronized to the leader state. A numerical example is provided to show the effectiveness of the proposed method.

Future research efforts will focus on extending the results to nonlinear networked uncertain/unknown multi-agent systems with output feedback.

References

[1] T. Alpcan and T. Basar, "An intrusion detection game with limited observations," *Proc. 12th Int. Symp. on Dynamic Games and Applications*, 2006

[2] T. Basar, G. J. Olsder, *Dynamic noncooperative game theory* (2nd ed.), Philadelphia, PA: SIAM, 1999

[3] T. Basar, P. Bernhard, "H_∞ Optimal Control and Related Minimax Design Problems," Birkhauser, Boston, MA, 1991

[4] D. P. Bertsekas, J. N. Tsitsiklis, *Neuro-dynamic programming* MA: Athena Scientific, 1996

[5] L. Busoniu, R. Babuska, B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 2, pp. 156-172, 2008

[6] Y. Cao, W. Yu, W. Ren and G. Chen, "An overview of recent progress in the study of distributed multiagent coordination," *IEEE Transactions on Industrial informatics*, vol. 9, no. 1, pp. 427-438, 2013

[7] S. N. Dorogovtsev, J. F. Mendes, "Evolution of networks: From biological nets to the Internet and WWW," *Oxford University Press*, 2013

[8] Q. Duong, M. P. Wellman, S. Singh, M. Kearns, "Learning and predicting dynamic networked behavior with graphical multiagent models," *Proc. of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pp. 441-448, 2012

[9] A. Geramifard, J. Redding, J. P. How, "Intelligent Cooperative Control Architecture: A framework for performance improvement using safe learning," *Journal of Intelligent and Robotic Systems*, vol. 72, no. 1, pp. 83-103, 2013

[10] C. Guestrin, S. Venkataraman, D. Koller, "Context-specific multiagent coordination and planning with factored MDPs, *In Proc. AAAI/IAAI*, pp. 253-259), 2002

[11] P. Ioannou, B. Fidan, *Adaptive control tutorial*, Advances in design and control, PA: SIAM, 2006

[12] P. A. Ioannou, J. Sun, *Robust adaptive control*, Courier Corporation, 2012

[13] D. Jin, S. Cho, Y. Sung, K. Cho, K. Um, "Reactive virtual agent learning for NUI-based HRI applications," *Multimedia Tools and Applications*, pp. 1-14, 2014

[14] M. Kearns, M. Littman, and S. Singh, "Graphical Models for Game Theory," *In Proceedings of UAI*, 2001

[15] H. K. Khalil, *Nonlinear Systems*, Prentice Hall, 2002

[16] J. R. Kok, N. Vlassis, "Sparse cooperative Q-learning," *In Proceedings of the twenty-first international conference on Machine learning*, p. 61, 2004

[17] F.L. Lewis, S. Jagannathan, and A. Yesildirek, *Neural Network Control of Robot Manipulators and Nonlinear Systems*, Taylor and Francis, London, 1999

[18] Y. Y. Liu, J. J. Slotine, A. L. Barabsi, "Controllability of complex networks," *Nature*, vol. 473, no. 7346, pp. 167-173, 2011

[19] F. S. Melo, M. Veloso, "Learning of coordination: Exploiting sparse interactions in multiagent systems," *In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, vol. 2, pp. 773-780, 2009

[20] M. Mesbahi, M. Egerstedt, *Graph theoretic methods in multiagent networks*, Princeton University Press, 2010
[21] S. Phelps, "Emergence of social networks via direct and indirect reciprocity," *Autonomous agents and multi-agent systems*, vol. 27, no. 3, pp. 355-374, 2014
[22] A. J. Van Der Schaft, "L_2-gain analysis of nonlinear systems and nonlinear state-feedback H_∞ control, "IEEE transactions on automatic control," vol. 37, no. 6, pp. 770-784, 1992
[23] A. Galindo-Serrano, L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823-1834, 2010
[24] S. Shahaboddin, A. Patel, N. B. Anuar, M. L. M. Kiah, and A. Abraham, "Cooperative game theoretic approach using fuzzy Q-learning for detecting and preventing intrusions in wireless sensor networks," *Engineering Applications of Artificial Intelligence*, vol. 32, pp. 228-241, 2014
[25] R. Sharma, M. Gopal, "Synergizing reinforcement learning and game theory-A new direction for control," *Applied Soft Computing*, vol. 10, no. 3, pp. 675-688, 2010
[26] D. F. Sittig, H. Singh, "A new sociotechnical model for studying health information technology in complex adaptive healthcare systems," *Quality and Safety in Health Care*, vol. 19, no. 3, pp. i68-i74, 2010
[27] S.H. Strogatz, "Exploring Complex Networks," *Nature*, vol. 410, pp. 268-276, 2001
[28] R. S. Sutton, A. G. Barto, *Reinforcement Learning- An Introduction*, MIT Press, Cambridge, Massachusetts, 1998
[29] K. G. Vamvoudakis, F. L. Lewis, "Multi-player non-zero-sum games: online adaptive learning solution of coupled Hamilton-Jacobi equations," *Automatica*, vol. 47, no. 8, pp. 1556-1569, 2011
[30] K. G. Vamvoudakis, F. L. Lewis, G. R. Hudas, "Multi-Agent Differential Graphical Games: Online Adaptive Learning Solution for Synchronization with Optimality," *Automatica*, vol. 48, no. 8, pp. 1598-1611, 2012
[31] K. G. Vamvoudakis, L.R. Garcia-Carrillo, J. P. Hespanha, "Learning consensus in adversarial environments," SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, pp.87410K-87410K, 2013
[32] D. Vrabie, K. G. Vamvoudakis, F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*, Control Engineering Series, IET Press, 2012
[33] F. Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications, *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 630-638, 2010
[34] S. Wang, M. S. Szalay, C. Zhang, P. Csermely, "Learning and innovative elements of strategy adoption rules expand cooperative network topologies," *PLoS One*, vol. 3, no. 4, e1917, 2008
[35] C. J. C. H. Watkins, P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279-292, 1992
[36] G. Weiss, *Multiagent systems: a modern approach to distributed artificial intelligence*, MIT press, 1999
[37] P. J. Werbos, *Approximate dynamic programming for real-time control and neural modeling* In D. A. White, D. A. Sofge (Eds.), Handbook of intelligent control. New York: Van Nostrand Reinhold, 1992

Kyriakos G. Vamvoudakis was born in Athens, Greece. He received the Diploma (a 5 year degree, equivalent to a Master of Science) in Electronic and Computer Engineering from Technical University of Crete, Greece in 2006 with highest honors. After moving to the United States of America, he studied at The University of Texas and he received his M.S. and Ph.D. in Electrical Engineering in 2008 and 2011 respectively. During the period from 2012 to 2016 he was a project research scientist at the Center for Control, Dynamical Systems and Computation at the University of California, Santa Barbara. He is now an assistant professor at the Kevin T. Crofton Department of Aerospace and Ocean Engineering at Virginia Tech. His research interests include approximate dynamic programming, game theory, and optimal control. Recently, his research has focused on network security, smart grid and multi-agent optimization. Dr. Vamvoudakis is the recipient of several international awards including the 2016 International Neural Network Society Young Investigator (INNS) Award, the Best Paper Award for Autonomous/Unmanned Vehicles at the 27th Army Science Conference in 2010, the Best Presentation Award at the World Congress of Computational Intelligence in 2010, and the Best Researcher Award from the Automation and Robotics Research Institute in 2011. He is a Senior Member of IEEE.

João P. Hespanha was born in Coimbra, Portugal, in 1968. He received the Licenciatura in electrical and computer engineering from the Instituto Superior Tcnico, Lisbon, Portugal in 1991 and the Ph.D. degree in electrical engineering and applied science from Yale University, New Haven, Connecticut in 1998. From 1999 to 2001, he was Assistant Professor at the University of Southern California, Los Angeles. He moved to the University of California, Santa Barbara in 2002, where he currently holds a Professor position with the Department of Electrical and Computer Engineering. Prof. Hespanha is the Chair of the Department of Electrical and Computer Engineering; Director for the Center for Control, Dynamical-systems, and Computation (CCDC); and a member of the Executive Committee for the Institute for Collaborative Biotechnologies (ICB). From 2004-2007 he was an associate editor for the IEEE Transactions on Automatic Control. His current research interests include hybrid and switched systems; multi-agent control systems; distributed control over communication networks (also known as networked control systems); the use of vision in feedback control; stochastic modeling in biology; and network security. Dr. Hespanha is the recipient of the Yale University's Henry Prentiss Becton Graduate Prize for exceptional achievement in research in Engineering and Applied Science, a National Science Foundation CAREER Award, the 2005 best paper award at the 2nd Int. Conf. on Intelligent Sensing and Information Processing, the 2005 Automatica Theory/Methodology best paper prize, the 2006 George S. Axelby Outstanding Paper Award, and the 2009 Ruberti Young Researcher Prize. He is a Fellow of the IEEE and IFAC.