Automatica 50 (2014) 3038–3053

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Multi-agent discrete-time graphical games and reinforcement learning solutions☆

Mohammed I. Abouheaf [a,1], Frank L. Lewis [b], Kyriakos G. Vamvoudakis [c], Sofie Haesaert [d], Robert Babuska [e]

[a] Systems Engineering Department, King Fahd University of Petroleum & Minerals, P. O. Box: 5067, Dhahran-3126, Saudi Arabia
[b] University of Texas at Arlington Research Institute, The University of Texas at Arlington, USA
[c] Center for Control, Dynamical-systems, and Computation (CCDC), University of California, Santa Barbara, USA
[d] Eindhoven University of Technology, Netherlands
[e] Delft Center for Systems and Control, Delft University of Technology, Netherlands

ARTICLE INFO

Article history:
Received 16 November 2011
Received in revised form
14 June 2014
Accepted 29 July 2014
Available online 4 November 2014

Keywords:
Dynamic graphical games
Optimal control
Nash equilibrium
Best response
Reinforcement learning

ABSTRACT

This paper introduces a new class of multi-agent discrete-time dynamic games, known in the literature as dynamic graphical games. For that reason a *local* performance index is defined for each agent that depends only *on the local information available to each agent*. Nash equilibrium policies and best-response policies are given in terms of the solutions to the discrete-time coupled Hamilton–Jacobi equations. Since in these games the interactions between the agents are prescribed by a communication graph structure we have to introduce a new notion of Nash equilibrium. It is proved that this notion holds if all agents are in Nash equilibrium and the graph is strongly connected. A novel reinforcement learning value iteration algorithm is given to solve the dynamic graphical games in an online manner along with its proof of convergence. The policies of the agents form a Nash equilibrium when all the agents in the neighborhood update their policies, and a best response outcome when the agents in the neighborhood are kept constant. The paper brings together discrete Hamiltonian mechanics, distributed multi-agent control, optimal control theory, and game theory to formulate and solve these multi-agent dynamic graphical games. A simulation example shows the effectiveness of the proposed approach in a leader-synchronization case along with optimality guarantees.

1. Introduction

Research on distributed multi-agent cooperative control systems has received extensive attention in the last two decades, mainly due to their applications in computer science, spacecraft, unmanned air vehicles, mobile robots, sensor networks, networked autonomous team, and so on (Beard & Stepanyan, 2003; Mu, Chu, & Wang, 2005). Synchronization (Jadbabaie, Lin, & Morse, 2003; Olfati-Saber, Fax, & Murray, 2007; Olfati-Saber & Murray, 2004; Qu, 2009; Ren & Beard, 2005, 2008; Ren, Beard, & Atkins, 2005; Tsitsiklis, 1984) allows each agent of a cooperative team to reach the same state by the proper selection and design of decision and control protocols.

Consensus has been studied for systems on communication graphs with fixed or varying topologies and communication delays. In the cooperative regulator consensus problem without leader, all agents converge to an uncontrollable common value depending on their initial conditions. In the cooperative tracking consensus problem, all agents synchronize to a leader or control agent state (Hong, Hu, & Gao, 2006; Ren, Moore, & Chen, 2007; Wang & Chen, 2002). In Fax and Murray (2004), the state of each agent is given by identical linear decoupled dynamics. A dynamical system is realized to supply each agent with common reference to be used in the collective behavior. In Ren et al. (2007), the authors generalize the first-order and second-order consensus problems to higher-order

☆ This work was supported by NSF grant ECCS-1128050, ONR grant N00014-13-1-0562, AFOSR EOARD Grant 13-3055, ARO grant W911NF-11-D-0001, China NNSF grant 61120106011, and China Education Ministry Project 111 (No. B08015). The material in this paper was partially presented at the 2013 American Control Conference (ACC), June 17–19, 2013, Washington, DC, USA. This paper was recommended for publication in revised form by Associate Editor Hideaki Ishii under the direction of Editor Ian R. Petersen.

E-mail addresses: abouheaf@kfupm.edu.sa (M.I. Abouheaf), lewis@uta.edu (F.L. Lewis), kyriakos@ece.ucsb.edu (K.G. Vamvoudakis), s.haesaert@tue.nl (S. Haesaert), r.babuska@tudelft.nl (R. Babuska).

[1] Tel.: +966 0138602968; fax: +966 0138602965.

http://dx.doi.org/10.1016/j.automatica.2014.10.047
0005-1098/

consensus problems. In Li, Duan, Chen, and Huang (2010), the synchronization among multi-agent systems with identical dynamics is achieved using a distributed observer consensus protocol based on relative output measurements. In Li, Wang, and Chen (2004), a complex dynamical network is controlled via pinning partially to a small number of nodes in the network using local feedback controllers. The placement of the local controllers is affected by the topology of the dynamical network. In the aforementioned consensus algorithm there are no optimality guarantees. For that reason one has to use a game-theoretic framework to overcome this issue.

Game theory provides an environment for formulating multi-player decision control problems for dynamical interacting systems (Başar & Olsder, 1999). Each agent optimizes its performance index independently to determine its optimal policy. The result is the Nash equilibrium solution. In order to find the Nash equilibrium one has to solve coupled Hamilton–Jacobi (HJ) equations, which in the linear quadratic case (LQR) reduce to the coupled game algebraic Riccati equations (Başar & Olsder, 1999; Freiling, Jank, & Abou-Kandil, 2002; Gajic & Li, 1988). Solution methods are generally offline and generate fixed control policies that are used to implement real-time online controllers. These coupled HJ equations are difficult or impossible to solve and depend on global information. As we shall see in the present paper, we will overcome this global information issue by designing dynamic graphical games. Static graphical games have been studied in the computational intelligence community by several researchers (e.g. Kakade, Kearns, Langford, & Ortiz, 2003; Shoham & Leyton-Brown, 2009).

In discrete mechanics, discrete versions of variational principles are used to derive the discrete equivalents of the continuous Euler–Lagrange and Hamiltonian equations (Marsden & West, 2001; Suris, 2003, 2004). The theory of discrete Lagrangian mechanics was introduced in Marsden and West (2001). A formulation for discrete Hamilton mechanics using direct approaches was developed in Gonzalez (1996) and McLachlan, Quispel, and Robidoux (1999). Later, Lall and West (2006), introduced the canonical form of the Hamiltonian theory that corresponds to the theory of discrete Lagrangian mechanics. The importance of discrete-time Hamiltonian relies on the Euclidean relationship between symplectic integrators, discrete time optimal control (Lewis, Vrabie, & Syrmos, 2012), and distributed network optimization (Lall & West, 2006). Moreover, it is often more direct to solve problems based on the discrete Hamiltonian rather than discrete Lagrangian (Lall & West, 2006).

Reinforcement Learning (RL) is an area of machine learning concerned with how an agent can pick its actions in a dynamic environment to transit to new states in such a way that optimizes the sum of cumulative reward (Sen & Weiss, 1999; Sutton & Barto, 1998). RL methods allow the development of algorithms to learn online the solutions to optimal control problems for dynamic systems that are described by difference equations (Sutton & Barto, 1998; Werbos, 1974, 1989, 1992). These involve two-step techniques known as Policy Iteration (PI) or Value Iteration (VI) (Bertsekas & Tsitsiklis, 1996). Policy iteration and value iteration algorithms have been developed for continuous time systems in Al-Tamimi, Lewis, and Abu-Khalaf (2008), Vamvoudakis and Lewis (2010), and Vrabie, Pastravanu, Lewis, and Abu-Khalaf (2009). RL algorithms are used to solve multi-player games for finite-state systems in Busoniu, Babuska, and De Schutter (2008), Littman (2001), and Vrancx, Verbeeck, and Nowe (2008), and to learn online in real-time the solutions for optimal control problems of dynamic systems and differential games in Dierks and Jagannathan (2010), Johnson, Hiramatsu, Fitz-Coy, and Dixon (2010), and Vamvoudakis and Lewis (2010, 2011). Wang, Liu, Wei, Zhao, and Jin (2012) used mathematical induction to prove convergence of an iterative dynamic programming algorithm in order to solve the optimal control problem for unknown non-affine nonlinear discrete-time systems. Actor–critic networks are one type of RL methods. The actor component applies actions or control policies to their environment, while the critic component assesses the values of these actions. Based on this assessment, the actor policy is updated at each learning step (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998).

Game theory is extensively used in multi-agent reinforcement learning. In such scenarios the agents do not have perfect knowledge about the game (e.g. Chang, Hsu, & Fu, 2007; Gopalakrishnan, Marden, & Wierman, 2011; Marden, Arslan, & Shamma, 2009; Young, 1998). In Bowling (2004), different methods are used to learn the optimal policy of every agent through repeated interactions among all agents. Interactions among a large number of players or complicated stochastic large systems are studied by population games and mean field games (Hofbauer & Sigmund, 1998; Sandholm, 2010). In Nourian, Caines, and Malhame (2011), a continuum based mean field control approach is used to solve the initial mean consensus problem. In that game, a set of coupled deterministic Hamilton–Jacobi Bellman and Fokker Planck Kolmogorov equations are used to approximate the stochastic system of agents in the continuum. The dynamics of population and mean field games covers a large class of game dynamics known in evolutionary game theory (Tembine, 2011).

This paper extends the results of continuous time graphical games that where introduced in Vamvoudakis, Lewis, and Hudas (2012) to discrete-time systems. It is shown that these dynamic graphical games are a special case of standard dynamic games (Başar & Olsder, 1999) and explicitly capture the structure of the communication graph topology. As such, they allow an analysis that shows the restrictions imposed on local control protocols by the graph topology.

The contributions of the paper are fourfold. The first involves the formulation of a graphical game for dynamic discrete-time multi-agent systems where information flow is restricted by a communication graph structure. A new notion of interactive Nash equilibrium is introduced which holds if the agents are all in Nash and the graph is strongly connected. The second contribution is in showing the relation between the discrete-time Bellman equation and the discrete-time Hamilton equation using discrete mechanics for dynamic graphical games. Coupled discrete-time HJ equations are formulated for the dynamic graphical games. The third contribution lies in providing policies in terms of those coupled HJ equations that converge to Nash equilibrium and best-response. Finally, a value iteration Heuristic Dynamic Programming (HDP) algorithm is given to solve the dynamic graphical games online in real-time by measuring the states along the system trajectories.

The paper is organized as follows. Section 2 provides a brief background on the synchronization problem in multi-agent systems. Section 3 formulates the dynamic graphical game and finds the relation between the discrete-time Hamilton–Jacobi Bellman equation and discrete-time Bellman optimality equation for multi-agent graphical games. Section 4 introduces the new notion of interactive Nash equilibrium. Moreover, this section provides existence solutions for policies forming an interactive Nash equilibrium and best response in terms of the solutions of the coupled HJ equations. Section 5 proposes a value iteration HDP algorithm to solve the coupled HJ equations along with its proof of convergence. Finally, Section 6 develops an online adaptive learning algorithm by using an actor–critic neural network framework to solve the graphical game, along with a simulation example to verify its effectiveness.

2. Graphs and synchronization of multi-agent dynamical systems

2.1. Graphs

The directed graph Gr is defined as the pair $Gr = (V, \mathcal{E})$ with a nonempty finite set of N vertices $V = \{v_1, \ldots, v_N\}$ and a set

of edges $\mathcal{E} \subseteq V \times V$. The connectivity matrix E is defined such that $E = [e_{ij}]$ with $e_{ij} > 0$ if $(v_j, v_i) \in \mathcal{E}$ and $e_{ij} = 0$ otherwise. The set of neighbors of every node v_i is $N_i = \{v_j : (v_j, v_i) \in \mathcal{E}\}$. Define the in-degree matrix D as a diagonal matrix $D = \text{diag}\{d_i\}$, with $d_i = \sum_{j \in N_i} e_{ij}$ the weighted in-degree of node i. The graph Laplacian matrix L is defined as $L = D - E$.

A directed path from node v_0 to node v_r is defined as a sequence of edges $v_0, v_1, \ldots, v_r$ such that $(v_i, v_{i+1}) \in \mathcal{E}$, $i \in \{0, 1, \ldots, r-1\}$. A directed graph is strongly connected if there is a directed path from v_i to v_j and vice versa for all distinct nodes $v_i, v_j \in V$.

2.2. Synchronization and tracking error dynamics

Consider the communication graph $Gr = (V, \mathcal{E})$ having N agents, each with local dynamics given by

$$x_i(k+1) = Ax_i(k) + B_i u_i(k) \tag{1}$$

where $x_i(k) \in R^n$ is the state vector of node i, and $u_i(k) \in R^{m_i}$ is the control input vector for node i. Also consider a control or leader node v_0 that has command generator dynamics (Lewis, 1992) $x_0(k) \in R^n$ given by

$$x_0(k+1) = Ax_0(k). \tag{2}$$

The leader is connected to a small percentage of the nodes in the graph.

The objective is to design the control inputs $u_i(k)$, using information only from neighbor nodes, so that all agent states synchronize to the leader state, that is $\lim_{k\to\infty} \|x_i(k) - x_0(k)\| = 0$, $\forall i$.

To study the synchronization problem on graphs, we define the local neighborhood tracking error (Khoo, Xie, & Man, 2009) $\varepsilon_i(k) \in R^n$ for each node i as

$$\varepsilon_i(k) = \sum_{j \in N_i} e_{ij}(x_j(k) - x_i(k)) + g_i(x_0(k) - x_i(k)) \tag{3}$$

where $g_i \geq 0$ is the pinning gain of node i, which is nonzero if node i is coupled to the control node x_0 (Li et al., 2004).

The overall tracking error vector for all nodes is given by

$$\varepsilon(k) = -((L+G) \otimes I_n)x(k) + ((L+G) \otimes I_n)\underline{x}_0(k) \tag{4}$$

where the global node state vector is $x = \begin{bmatrix} x_1^T & x_2^T & .. & x_N^T \end{bmatrix}^T$ and the global tracking error vector is $\varepsilon = \begin{bmatrix} \varepsilon_1^T & \varepsilon_2^T & .. & \varepsilon_N^T \end{bmatrix}^T$.

Eq. (4) can be written as

$$\varepsilon(k) = -((L+G) \otimes I_n)\eta(k) \tag{5}$$

where the global disagreement vector or the synchronization error vector (Olfati-Saber & Murray, 2004) is

$$\eta(k) = (x(k) - \underline{x}_0(k)) \in R^{nN} \tag{6}$$

with $\underline{x}_0 = \underline{I}x_0$, $\underline{I} = \underline{1} \otimes I_n$ and $\underline{1}$ the N-vector of ones. $G = \text{diag}\{g_i\} \in R^{N \times N}$ is a diagonal matrix of pinning gains.

If the graph contains a spanning tree and $g_i \neq 0$ for a root node, then $(L+G)$ is nonsingular (Khoo et al., 2009).

The next result shows that the disagreement vector can be made arbitrarily small by making the local neighborhood tracking errors small.

The maximum and minimum singular values of a matrix are denoted respectively as $\overline{\sigma}(.)$, $\underline{\sigma}(.)$.

Lemma 1. *Let $(L+G)$ be nonsingu lar. Then the synchronization error is bounded by*

$$\|\eta(k)\| \leq \|\varepsilon(k)\| / \underline{\sigma}(L+G). \tag{7}$$

Proof. Under the Hypothesis $(L+G)$ is nonsingular. Then $\underline{\sigma}(L+G) \neq 0$ and (5) implies (7), with $\varepsilon(k) \equiv 0$ if and only if the nodes synchronize, that is

$$x(k) = \underline{I}x_0(k). \quad \blacksquare \tag{8}$$

For ease of notation, we write x_{ik} for $x_i(k)$.

The dynamics of the local neighborhood tracking error for node i are given by

$$\begin{aligned} \varepsilon_{i(k+1)} &\equiv f_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) \\ &= A\varepsilon_{ik} - (d_i + g_i)B_i u_{ik} + \sum_{j \in N_i} e_{ij} B_j u_{jk}. \end{aligned} \tag{9}$$

These error dynamics are interacting dynamical systems driven by the control actions of agent i and all of its neighbors. Our objective is to minimize the local neighborhood tracking errors $\varepsilon_i(k)$, which in view of Lemma 1 will guarantee approximate synchronization.

3. Dynamic graphical games

In this section we define multi-player dynamic games on graphs. These dynamic graphical games are defined based on the error systems (9), which are locally coupled in the sense that they are driven by the agent's control actions and those of its neighbors. This structure arises from the nature of the synchronization problem for dynamic systems (1) on communication graphs. Therefore, in similar fashion, we define locally coupled performance indices that depend on the state of an agent, its control action, and the control actions of its neighbors.

Following that, principles of optimal control (Bellman, 1957; Bryson, 1996; Lewis et al., 2012) are used to develop the Hamiltonian functions and the Bellman equations for dynamic graphical games. The Discrete Hamilton–Jacobi theory is used to show the relation between the Hamiltonian equation and the Bellman equation (Lall & West, 2006). These results lay the foundation to solve the dynamic graphical games in subsequent sections.

3.1. Graphical games

Graphical games are based on the responses of each agent i to other players in graph. Define the control actions of the neighbors of agent i as

$$u_{-i} = \left\{u_j \mid j \in N_i\right\} \tag{10}$$

and the actions of all the other agents in the graph excluding i as

$$u_{\bar{i}} = \left\{u_j \mid j \in N, j \neq i\right\}. \tag{11}$$

The structure of the error dynamics (9) arises from the nature of the synchronization problem for dynamic systems on communication graphs. Therefore, in order to define the dynamic graphical game, in similar fashion, we write the local performance indices for each agent i as

$$\begin{aligned} J_i(\{\varepsilon_{ik}, u_{ik}, u_{-ik}\}_{k\geq 0}) &= \sum_{k=0}^{\infty} U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) \\ &= \frac{1}{2}\sum_{k=0}^{\infty}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + u_{ik}^T R_{ii} u_{ik} + \sum_{j \in N_i} u_{jk}^T R_{ij} u_{jk}\right) \end{aligned} \tag{12}$$

where $Q_{ii} > 0 \in R^{n_i \times n_i}$, $R_{ii} > 0 \in R^{m_i \times m_i}$, $R_{ij} > 0 \in R^{m_j \times m_j}$, are symmetric time-invariant weighting matrices.

The dynamics (9) and the performance indices (12) depend on the graph topology $Gr = (V, \mathcal{E})$.

Given fixed policies of agent i and its neighbors, the value function for each agent i is given as

$$V_i(\varepsilon_{ik}) = \sum_{l=k}^{\infty} U_i(\varepsilon_{il}, u_{il}, u_{-il}). \tag{13}$$

Definition 1. The policies u_{ik}, $\forall i$ are said to be admissible if they stabilize (9) and guarantee that $V_i(\varepsilon_{ik})$, $\forall i$ are finite (Zhang, Luo, & Liu, 2009). $\blacksquare$

Definition 2. The dynamic graphical game with local dynamics (9) and performance indices (12) is well-formed if $B_j \neq 0 \rightleftharpoons e_{ij} \in E$, $R_{ij} \neq 0 \rightleftharpoons e_{ij} \in E$. $\blacksquare$

3.2. Comparison of dynamic graphical games with standard dynamic games

Dynamic graphical games as introduced in this paper are a special case of the dynamic games normally discussed in the literature. The standard N-player dynamic game as defined in Başar and Olsder (1999), has centralized dynamics

$$\delta(k+1) = F\delta(k) + \sum_{i=1}^{N} B^i u_i(k) \tag{14}$$

where $\delta(k) \in R^{\tilde{N}}$ is the centralized state and $u_i(k) \in R^{m_i}$ is the control input for each player i. In these standard games, the performance index for each player i, is defined as

$$\tilde{J}_i(\{\delta(k), u_i(k)\}_{k\geq 0}) = \frac{1}{2}\sum_{k=0}^{\infty}\Big(\delta(k)^T \hat{Q}_{ii}\delta(k) + \sum_{j\in N} u_j^T(k)\hat{R}_{ij}u_j(k)\Big) \tag{15}$$

where $\hat{Q}_{ii} > 0 \in R^{\tilde{N}\times\tilde{N}}$, $\hat{R}_{ij} > 0 \in R^{m_j\times m_j}$.

Both the dynamics (14) and the performance indices (15) depend on the control actions of all other players.

To compare these standard games to the dynamic graphical games defined in this paper, write the global error dynamics for the dynamic graphical game as

$$\varepsilon(k+1) = (I_N \otimes A)\varepsilon(k) - ((L+G)\otimes I_n)\bar{B}u(k) \tag{16}$$

with $\bar{B} = \text{diag}\{B_1, \ldots, B_i, \ldots, B_N\}$ and $u(k) = \begin{bmatrix} u_1^T & u_2^T & \cdots & u_N^T\end{bmatrix}^T$ is the global vector of control inputs. To find the relation between the standard centralized game (14) and the dynamic graphical game (16), define $F = (I_N \otimes A)$ and write (14) as

$$\delta(k+1) = (I_N \otimes A)\delta(k) + \begin{bmatrix} B^1 & \cdots & B^N\end{bmatrix} u(k). \tag{17}$$

Now, define l_{ij} as the ijth element of $(L+G)$ and write (16) as

$$\varepsilon(k+1) = (I_N \otimes A)\varepsilon(k) - [l_{ij}B_j]u(k) \tag{18}$$

where $[l_{ij}B_j]$ is a matrix whose ijth block is $l_{ij}B_j$. Eqs. (17) and (18) are same if one defines the block column matrix B^j as $B^j = \begin{bmatrix} l_{1j}B_j^T & l_{2j}B_j^T & \cdots & l_{Nj}B_j^T\end{bmatrix}^T$.

It is easily observed that the graphical game local performance index (12) is a special case of the standard game performance index (15) with appropriate definition of $\hat{Q}_{ij}, \hat{R}_{ij}$. Therefore, the dynamic graphical game is a special case of the standard game which explicitly displays the graph topology through $(L+G)$ in (18).

The dynamic graphical game formulation explicitly captures the structure of the communication graph. Therefore, its analysis clearly reveals the interplay of individual node dynamics and the graph topology within a multi-player game. Moreover, as seen in Section 6, it allows the solution of the game in a distributed fashion. Note that the coupled game Riccati equations of each agent presented in Başar and Olsder (1999), depend on the policies of all other agents, and so provide a centralized solution for the game. Moreover, existence of solutions to those coupled game Riccati equations requires reachability conditions that are closely related to requirement (49) in our definition of interactive Nash equilibrium in Section 4.2. See Lemma 3. For these conditions to hold, the graph must be strongly connected.

3.3. Bellman equation for dynamic graphical games

Taking the first difference of (13) yields the graphical game Bellman equations

$$V_i(\varepsilon_{ik}) = U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) + V_i(\varepsilon_{i(k+1)}) \tag{19}$$

with initial conditions $V_i(0) = 0$.

The objective of the graphical games optimization problem is to find the optimal value

$$V_i^o(\varepsilon_{ik}) = \min_{\bar{u}_i}(V_i(\varepsilon_{ik})) = \min_{\bar{u}_i}\left(\sum_{l=k}^{\infty} U_i(\varepsilon_{il}, u_{il}, u_{-il})\right) \tag{20}$$

where $\bar{u}_i = \{u_{ik}\}_{k=0}^{\infty}$, $\forall i \in N$. According to the Bellman optimality principle

$$V_i^o(\varepsilon_{ik}) = \min_{u_{ik}}(U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) + V_i^o(\varepsilon_{i(k+1)})). \tag{21}$$

Consequently, the optimal control policy for each agent i is

$$u_{ik}^o = (d_i + g_i)R_{ii}^{-1}B_i^T\nabla V_i^o(\varepsilon_{i(k+1)}). \tag{22}$$

Substituting (22) into (21) yields the coupled graphical game Bellman optimality equations

$$\begin{aligned} V_i^o(\varepsilon_{ik}) = {} & V_i^o(\varepsilon_{i(k+1)}) + \frac{1}{2}\Big(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} \\ & + (d_i+g_i)^2\nabla V_i^o(\varepsilon_{i(k+1)})^T B_i R_{ii}^{-1}B_i^T\nabla V_i^o(\varepsilon_{i(k+1)}) \\ & + \sum_{j\in N_i}(d_j+g_j)^2\nabla V_j^o(\varepsilon_{j(k+1)})^T \\ & \times B_j R_{jj}^{-1}R_{ij}R_{jj}^{-1}B_j^T\nabla V_j^o(\varepsilon_{j(k+1)})\Big) \end{aligned} \tag{23}$$

with initial conditions given by $V_i^o(0) = 0$.

3.4. Hamiltonian function for dynamic graphical games

Consider the node error dynamics (9) and the performance indices (12). We can define the Hamiltonian function (Lewis et al., 2012) of each agent i as

$$\begin{aligned} & H_i(\varepsilon_{ik}, \lambda_{i(k+1)}, u_{ik}, u_{-ik}) \\ & = \lambda_{i(k+1)}^T\left(A\varepsilon_{ik} - (d_i+g_i)B_i u_{ik} + \sum_{j\in N_i} e_{ij}B_j u_{jk}\right) \\ & \quad + \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + u_{ik}^T R_{ii}u_{ik} + \sum_{j\in N_i} u_{jk}^T R_{ij}u_{jk}\right) \end{aligned} \tag{24}$$

where $\lambda_{ik} \equiv \lambda_i(k)$ is the costate or adjoint variable of each agent i. The necessary conditions for optimality allow us to find the costate equations as

$$\partial H_i/\partial\varepsilon_{ik} = \lambda_{ik} \Rightarrow \lambda_{ik} = A^T\lambda_{i(k+1)} + Q_{ii}\varepsilon_{ik}. \tag{25}$$

The optimal control policy based on the Hamiltonian (24) is given by the stationarity condition (Lewis et al., 2012) $\partial H_i/\partial u_{ik} = 0$, so that

$$u_{ik}^* = \arg\min_{u_{ik}}(H_i(\varepsilon_{ik}, \lambda_{i(k+1)}, u_{ik}, u_{-ik})) \tag{26}$$

or

$$u_{ik}^* = (d_i+g_i)R_{ii}^{-1}B_i^T\lambda_{i(k+1)}. \tag{27}$$

3.5. Discrete Hamilton–Jacobi theory: equivalence of Hamiltonian and Bellman optimality equations

For the following development, we will define the first difference of the value function $V_i(\varepsilon_{ik})$ as

$$\Delta V_i(\varepsilon_{ik}) = V_i(\varepsilon_{i(k+1)}) - V_i(\varepsilon_{ik}) \tag{28}$$

and its gradient as

$$\nabla V_i(\varepsilon_{i(k+1)}) = \partial V_i(\varepsilon_{i(k+1)})/\partial\varepsilon_{i(k+1)}. \tag{29}$$

The next result (cf. Lall & West, 2006) relates the Hamiltonian (24) to the value (13). It also introduces the discrete-time HJ equation, which relates (28) and (29).

Theorem 1 (*Discrete-Time Hamilton–Jacobi Equation*). *Consider the Hamiltonian equation (24) and define the value function* $V_i(\varepsilon_{ik})$

by (13). Then, $V_i(\varepsilon_{ik})$ satisfies the following discrete-time Hamilton–Jacobi (DTHJ) equation

$$\Delta V_i(\varepsilon_{ik}) - \nabla V_i(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)} + H_i(\varepsilon_{ik}, \nabla V_i(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) = 0. \tag{30}$$

Proof. The objective of optimal control is to minimize the performance index for each agent i (12). The optimization problem is subject to the equality constraint

$$\varepsilon_{i(k+1)} = f_i(\varepsilon_{ik}, u_{ik}, u_{-ik}). \tag{31}$$

Then one should define augmented value function as follows

$$V_i(\varepsilon_{ik}) = \sum_{l=k}^{\infty} \left\{ U_i(\varepsilon_{il}, u_{il}, u_{-il}) + \lambda_{i(l+1)}^T \left[f_i(\varepsilon_{il}, u_{il}, u_{-il}) - \varepsilon_{i(l+1)} \right] \right\} \tag{32}$$

with a corresponding Hamiltonian given by,

$$H_i(\varepsilon_{il}, \lambda_{i(l+1)}, u_{il}, u_{-il}) = U_i(\varepsilon_{il}, u_{il}, u_{-il}) + \lambda_{i(l+1)}^T f_i(\varepsilon_{il}, u_{il}, u_{-il}). \tag{33}$$

Eqs. (32) and (33) yield

$$V_i(\varepsilon_{ik}) = \sum_{l=k}^{\infty} \left\{ H_i(\varepsilon_{il}, \lambda_{i(l+1)}, u_{il}, u_{-il}) - \lambda_{i(l+1)}^T \varepsilon_{i(l+1)} \right\}$$

$$V_i(\varepsilon_{i(k+1)}) = \sum_{l=k+1}^{\infty} \left\{ H_i(\varepsilon_{il}, \lambda_{i(l+1)}, u_{il}, u_{-il}) - \lambda_{i(l+1)}^T \varepsilon_{i(l+1)} \right\}.$$

Substituting these equations into (19) yields

$$\Delta V_i(\varepsilon_{ik}) + H_i(\varepsilon_{ik}, \lambda_{i(k+1)}, u_{ik}, u_{-ik}) - \lambda_{i(k+1)}^T \varepsilon_{i(k+1)} = 0. \tag{34}$$

Taking the derivative of (34) with respect to $\varepsilon_{i(k+1)}$ yields

$$\nabla V_i(\varepsilon_{i(k+1)}) - \left[(\partial \lambda_{i(k+1)} / \partial \varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)} + \lambda_{i(k+1)} \right] + (\partial \lambda_{i(k+1)} / \partial \varepsilon_{i(k+1)})^T \partial H_i(\varepsilon_{ik}, \lambda_{i(k+1)}, u_{ik}, u_{-ik}) / \partial \lambda_{i(k+1)} = 0.$$

Rearranging this equation yields

$$\nabla V_i(\varepsilon_{i(k+1)}) = \lambda_{i(k+1)} + (\partial \lambda_{i(k+1)} / \partial \varepsilon_{i(k+1)})^T \times (\varepsilon_{i(k+1)} - \partial H_i(\varepsilon_{ik}, \lambda_{i(k+1)}, u_{ik}, u_{-ik}) / \partial \lambda_{i(k+1)}). \tag{35}$$

Eq. (35) yields $\lambda_{i(k+1)} = \nabla V_i(\varepsilon_{i(k+1)})$. Substituting this into (34) yields (30). ■

This proof motivates us to define the costate in terms of the value function as

$$\lambda_{i(k+1)} = \nabla V_i(\varepsilon_{i(k+1)}). \tag{36}$$

The optimal control policy based on the Bellman optimality equation (23) is given by (22). The next result shows the relation between the policies (22) and (27). It also relates the Hamiltonian (24) and Bellman optimality equation (23).

Theorem 2 (*Discrete-Time Hamilton–Jacobi Bellman Equation*)**.**

a. *Let $0 < V_i^*(\varepsilon_{ik}) \in C^2$, $\forall i$ satisfy the Discrete-Time Hamilton–Jacobi Bellman (DTHJB) equation*

$$H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}^*) = \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)} + \frac{1}{2} \left(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + u_{ik}^{*T} R_{ii} u_{ik}^* + \sum_{j \in N_i} u_{jk}^{*T} R_{ij} u_{jk}^* \right) = 0 \tag{37}$$

with initial condition given by $V_i^*(0) = 0$, where

$$u_{ik}^* = (d_i + g_i) R_{ii}^{-1} B_i^T \nabla V_i^*(\varepsilon_{i(k+1)}). \tag{38}$$

Then $V_i^*(\varepsilon_{ik})$ satisfies the Bellman optimality equation (23).

b. *Let (A, B_i) $\forall i$ be reachable. Let $0 < V_i^*(\varepsilon_{ik}) \in C^2$, $\forall i$ satisfy (23). Then $V_i^*(\varepsilon_{ik})$ satisfies (37).*

Proof. a. $V_i^*(\varepsilon_{ik})$ satisfies (37) and u_{ik}^* is given by (38) such that $H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}^*) = 0$. Then, by applying the results from Theorem 1 we have $\Delta V_i^*(\varepsilon_{ik}) = \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}$ from which the result follows.

b. Completing the squares on the Hamiltonian (24) for an arbitrary smooth function $V_i(\varepsilon_{ik})$ yields

$$\begin{aligned} &H_i(\varepsilon_{ik}, \nabla V_i(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) \\ &= H_i(\varepsilon_{ik}, \nabla V_i(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}^*) + \frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) \\ &\quad + \frac{1}{2} \sum_{j \in N_i} (u_{jk} - u_{jk}^*)^T R_{ij} (u_{jk} - u_{jk}^*) + \sum_{j \in N_i} u_{jk}^{*T} R_{ij} (u_{jk} - u_{jk}^*) \\ &\quad + \sum_{j \in N_i} e_{ij} \nabla V_i(\varepsilon_{i(k+1)})^T B_j (u_{jk} - u_{jk}^*) \end{aligned} \tag{39}$$

where $u_{ik}^* = (g_i + d_i) R_{ii}^{-1} B_i^T \nabla V_i(\varepsilon_{i(k+1)})$, $\forall i$. Now, let $V_i(\varepsilon_{ik}) \in C^2$, $\forall i$ satisfy the Bellman equation (19). The Hamiltonian with optimal value given by $V_i^*(\varepsilon_{ik})$ for arbitrary control policies yields,

$$\begin{aligned} &H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) \\ &= (\nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}) + U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) \\ &= \frac{1}{2} \sum_{j \in N_i} (u_{jk} - u_{jk}^*)^T R_{ij} (u_{jk} - u_{jk}^*) \\ &\quad + \frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) + \sum_{j \in N_i} u_{jk}^{*T} R_{ij} (u_{jk} - u_{jk}^*) \\ &\quad + \sum_{j \in N_i} e_{ij} \nabla V_i^*(\varepsilon_{i(k+1)})^T B_j (u_{jk} - u_{jk}^*). \end{aligned} \tag{40}$$

Bellman equation (19) can be written as follows

$$V_i(\varepsilon_{ik}) = U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) + (\nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}) - (\nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}) + V_i(\varepsilon_{i(k+1)})).$$

Substituting the Hamiltonian (40) into this equation yields

$$\begin{aligned} V_i(\varepsilon_{ik}) &= V_i(\varepsilon_{i(k+1)}) - (\nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}) \\ &\quad + \frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) \\ &\quad + \frac{1}{2} \sum_{j \in N_i} (u_{jk} - u_{jk}^*)^T R_{ij} (u_{jk} - u_{jk}^*) \\ &\quad + \sum_{j \in N_i} e_{ij} \nabla V_i^*(\varepsilon_{i(k+1)})^T B_j (u_{jk} - u_{jk}^*) \\ &\quad + \sum_{j \in N_i} u_{jk}^{*T} R_{ij} (u_{jk} - u_{jk}^*). \end{aligned}$$

Bellman's optimality principle, yields that $V_i^o(\varepsilon_{ik})$ has to satisfy the following equation

$$\begin{aligned} V_i^o(\varepsilon_{ik}) &= \min_{u_{ik}} (V_i^o(\varepsilon_{i(k+1)}) - (\nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}) \\ &\quad + \frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) + \sum_{j \in N_i} u_{jk}^{*T} R_{ij} (u_{jk} - u_{jk}^*) \\ &\quad + \frac{1}{2} \sum_{j \in N_i} (u_{jk} - u_{jk}^*)^T R_{ij} (u_{jk} - u_{jk}^*) \\ &\quad + \sum_{j \in N_i} e_{ij} \nabla V_i^*(\varepsilon_{i(k+1)})^T B_j (u_{jk} - u_{jk}^*)). \end{aligned} \tag{41}$$

Applying the stationarity conditions $\partial V_i^o(\varepsilon_{ik}) / \partial u_{ik} = 0$, the control policy u_{ik}^o is given by solving

$$-(g_i + d_i) B_i^T (\nabla V_i^o(\varepsilon_{i(k+1)}) - \nabla V_i^*(\varepsilon_{i(k+1)})) + R_{ii}(u_{ik}^o - u_{ik}^*) = 0$$

or,

$$(u_{ik}^o - u_{ik}^*) = (g_i + d_i)R_{ii}^{-1}B_i^T(\nabla V_i^o(\varepsilon_{i(k+1)}) - \nabla V_i^*(\varepsilon_{i(k+1)})). \quad (42)$$

The hessians of the Hamiltonian (24) and Bellman equations (19) with respect to all control policies are positive definite values since $\nabla^2_{u_{ik}}(H_i) = R_{ii}$ and $\nabla^2_{u_{ik}}(V_i) = R_{ii}$. Therefore the optimal control policy is unique $u_{ik}^* = u_{ik}^o,\ \forall k$.

Now, (42) and the costate (25) show that

$$(g_i + d_i)R_{ii}^{-1}B_i^T(A^T)^p(\nabla V_i^o(\varepsilon_{i(k+1)}) - \nabla V_i^*(\varepsilon_{i(k+1)})) = 0,$$
$$\forall k,\ p = 0, 1, \ldots, n-1. \quad (43)$$

The reachability matrix

$$\tilde{U}_i = \begin{bmatrix} B_i & AB_i & A^2B_i & \cdots & A^{n-1}B_i \end{bmatrix} \quad (44)$$

under the hypothesis has full rank. Therefore, since $V_i^*(0) = 0$ and $V_i^o(0) = 0$ then,

$$V_i^*(\varepsilon_{ik}) = V_i^o(\varepsilon_{ik}), \quad \forall k \quad (45)$$

from which the result follows. ■

The next lemma relates the first difference of the optimal value function and its gradient.

Lemma 2. *Let $V_i^*(\varepsilon_{ik})$ satisfy the Bellman optimality equation (23) or equivalently the coupled HJ (37) then,*

$$\nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)} = \Delta V_i^*(\varepsilon_{ik}) \quad (46)$$

or

$$(\partial V_i^*(\varepsilon_{i(k+1)})/\partial\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)} = V_i^*(\varepsilon_{i(k+1)}) - V_i^*(\varepsilon_{ik}).$$

Proof. The proof follows from the results of Theorems 1 and 2. ■

4. Nash solution for dynamic graphical games

The objective of a dynamic graphical game is to solve the non-cooperative minimization problem given by the coupled equations (20), which lead to the Bellman optimality equations (23). It is shown here that the concept of Nash equilibrium is incomplete for dynamic graphical games, since it does not take into account the graph connectivity properties.

4.1. Nash equilibrium and idiosyncrasies of dynamic graphical games

The next definition is given with respect to the actions of all the other players $u_{\bar{i}} = \{u_j \mid j \in N, j \neq i\}$, namely a complete graph.

Definition 3 (*Başar & Olsder, 1999*)**.** The N-player dynamic graphical game with N-tuple of optimal control policies $\{u_1^*, u_2^*, \ldots, u_N^*\}$ is said to have a global Nash equilibrium solution if for all $i \in N$

$$J_i^* \triangleq J_i(u_i^*, u_{\bar{i}}^*) \leq J_i(u_i, u_{\bar{i}}^*). \quad (47)$$

The N-tuple $\{J_1^*, J_2^*, \ldots, J_N^*\}$ is called the Nash equilibrium outcome of the N-player game. ■

4.2. Interactive Nash equilibrium

In the case of a disconnected graph, the agents can be in Nash equilibrium, yet have no influence on each other. In such situations, the definition of coalition-proof Nash equilibrium (Shinohara, 2010) may also hold, that is, no set of agents has an incentive to break away from the Nash equilibrium and seek a new Nash solution among them. To guarantee that all agents in a graph are involved in the same game, the stronger definition of interactive Nash equilibrium is introduced here (Vamvoudakis et al., 2012).

Definition 4 (*Shoham & Leyton-Brown, 2009*)**.** Agent i's best response to fixed control actions of his neighbors is the policy u_i^* such that

$$J_i(u_i^*, u_{-i}) \leq J_i(u_i, u_{-i}). \quad \blacksquare \quad (48)$$

It is noted that, as in the standard case of multi-player games with a centralized state (14), all agents are in best response to their neighbors (e.g. Definition 4) if and only if they are in global Nash equilibrium (e.g. Definition 3).

The next definition provides a strengthened notion of Nash equilibrium that is suitable for graphical games.

Definition 5 (*Interactive Global Nash Equilibrium*)**.** An N-*tuple* of policies $\{u_1^*, u_2^*, \ldots, u_N^*\}$ is said to constitute an interactive global Nash equilibrium solution for an N-player game if, for all $i \in N$, the Nash condition (47) holds and in addition there exists a policy u_z' such that

$$J_i(u_z^*, u_{-z}^*) \neq J_i(u_z', u_{-z}^*). \quad (49)$$

For all $i, z \in N$. That is, at the equilibrium point there exists a policy of every player z that influences the performance of all other players i. ■

Condition (49) means that the reaction curve (Başar & Olsder, 1999) of any player i is not constant with respect to all variations in the policy of any other player z. For that reason one has to find conditions under which the local best responses in Definition 4 imply the interactive global Nash of Definition 5.

Consider the systems given by (9) in closed-loop with admissible feedbacks given by (25) and (27). Denote a suboptimal policy $u_{zk} = K_{\lambda_z}\lambda_{zk} + K_{\varepsilon_z}\varepsilon_{zk} + v_{zk}$ for a single node z and optimal policies given by $u_{jk} = K_{\lambda_j}\lambda_{jk} + K_{\varepsilon_j}\varepsilon_{jk},\ \forall j \neq z$. That is, node z has an extra control input v_{zk}. Then we have

$$\varepsilon_{i(k+1)} = A\varepsilon_{ik} - (d_i + g_i)B_iu_{ik} + \sum_{j\in N_i} e_{ij}B_ju_{jk} + e_{iz}B_zv_{zk}, \quad (50)$$

and

$$K_{\varepsilon i} = -(d_i + g_i)R_{ii}^{-1}B_i^TA^{-T}Q_{ii}. \quad (51)$$

According to (25) and (50) the global closed-loop dynamics are given by the Hamiltonian system

$$[\varepsilon_{k+1}^T\ \lambda_{k+1}^T]^T = \overline{A}[\varepsilon_k^T\ \lambda_k^T]^T + \overline{B}\overline{v}_z \quad (52)$$

where $\overline{B} = \begin{bmatrix} ((L+G)\otimes I_n)\mathrm{diag}(B_i) \\ 0 \end{bmatrix}$, $\overline{v}_z = [0..v_{zk}..0]^T$, and $\overline{A} = \begin{bmatrix} (I_N\otimes A) + ((L+G)\otimes I_n)\mathrm{diag}(B_iK_{\varepsilon_i}) & ((L+G)\otimes I_n)\mathrm{diag}(B_iK_{\lambda_i}) \\ -\mathrm{diag}(A^{-T}Q_{ii}) & (I_N\otimes A^{-T}) \end{bmatrix}$. Consider node i and let $\overline{M} > 0$ be the first integer such that $[(L+G)^{\overline{M}}]_{iz} \neq 0$, where $[.]_{iz}$ denotes the element (i, z) of a matrix. That is, $\overline{M}$ is the length of the shortest directed path from node z to node i. Denote the nodes along this path by $z = z_0, z_1, \ldots, z_{\overline{M}-1}, z_{\overline{M}} = i$. Denote element (i, z) of $(L+G)$ by ℓ_{iz}. Then the $n \times m$ block element in block row i and block column z of matrix $\overline{A}^{(\overline{M}-1)}\overline{B} \in R^{2nN\times m_z}$ is equal to

$$[\overline{A}^{(\overline{M}-1)}\overline{B}]^{iz} = \sum_{z_{\overline{M}-1},\ldots,z_1} \ell_{i,z_{\overline{M}-1}}..\ell_{z_1,z}B_{z_{\overline{M}-1}}K_{\varepsilon z_{\overline{M}-1}}B_{z_{\overline{M}-2}}..B_{z_1}K_{\varepsilon z_1}B_z$$
$$\equiv \sum_{z_{\overline{M}-1}} B_{z_{\overline{M}-1}}\overline{B}_{z_{\overline{M}-1},z} \quad (53)$$

where $\overline{B}_{z_{\overline{M}-1},z} \in R^{m_{z_{\overline{M}-1}}\times m_z}$ and $[]^{iz}$ denotes the position of the block element in the block matrix.

Assumption 1. a. The matrix $(B_{z_j}^TA^{-T}Q_{jj}B_{z_{j-1}})$ has full row rank for $\overline{M} > j \geq 0$.

b. There is unique shortest path between every two nodes. ■

This assumption holds for a large class of systems and graphs. Note that condition (a) holds if $m_i = m_j,\ \forall i, j, B_i = B_j,\ \forall i, j$, with B_i of full column rank, $Q_{ii} > 0,\ \forall i$, and A is nonsingular.

Lemma 3. *Let Assumption 1 hold and let the control policies are given by (27). Then the ith performance index of closed-loop system (52) depends on the input v_z if and only if there exists a directed path from node z to node i.*

Proof. *Sufficiency*. If $z = i$ the result is obvious. Otherwise, the reachability matrix from node z to node i has the following composition

$$\left[[\bar{A}^{(\bar{M}-1)}\bar{B}]^{iz}\ [\bar{A}^{\bar{M}}\bar{B}]^{iz}\ [\bar{A}^{(\bar{M}+1)}\bar{B}]^{iz}\ \dots \right]$$

with

$$\begin{aligned}[\bar{A}^{(\bar{M}-1)}\bar{B}]^{iz} = &\sum_{z_{\bar{M}-1},..,z_1} \ell_{i,z_{\bar{M}-1}}..\ell_{z_1,z} B_{z_{\bar{M}-1}}(d_{z_{\bar{M}-1}} + g_{z_{\bar{M}-1}}) \\ &\times R^{-1}_{z_{\bar{M}-1}} B^T_{z_{\bar{M}-1}} A^{-T} Q_{z_{\bar{M}-1}} B_{z_{\bar{M}-2}}..B_{z_1} R^{-1}_{z_1} \\ &\times (d_{z_1} + g_{z_1}) B^T_{z_1} A^{-T} Q_{z_1} B_z. \end{aligned} \tag{54}$$

If there exists a path from node z to i, then the first block matrix has rank $m_{z_{\bar{M}-1}}$, therefore at least $m_{z_{\bar{M}-1}}$ modes of the states of node i are reachable. Thus the state will be a function of v_{zk}, because Q_{ii} is positive definite. Therefore, the ith performance index will depend on v_{zk}.

Necessity. If there is no path from node z to node i, then the control input of node z cannot influence the state or value of node i. ■

Theorem 3. *Let every node i be in best response to all its neighbors $j \in N_i$. Let Assumption 1 hold. Then all nodes in the graph are in interactive global Nash equilibrium if and only if the graph is strongly connected.*

Proof. Let every node i be in best response to all its neighbors $j \in N_i$. Then $J_i(u_i^*, u_{-i}) \le J_i(u_i, u_{-i}),\ \forall i$. Hence $u_j = u_j^*,\ \forall u_j \in u_{-i}$ and $J_i(u_i^*, u_{-i}^*) \le J_i(u_i, u_{-i}^*),\ \forall i$.

However, according to (12) $J_i(u_i^*, u_{-i}^*) = J_i(u_i^*, u_{-i}^*, u_z),\ \forall z \notin \{i\} \cup N_i$ so that $J_i(u_i^*, u_i^*) \le J_i(u_i, u_i^*),\ \forall i$ and the nodes are in Nash equilibrium.

Necessity. If the graph is not strongly connected, then there exist nodes z and i such that there is no path from node z to node i. Then, the control input of node z cannot influence the state or the value of node i. Therefore, the Nash equilibrium is not interactive.

Sufficiency. If there is a path from node z to node i, the performance index depends on u_z. Strong connectivity means there is a path from every node k to every node i and condition (49) holds for all $i, z \in N$. ■

According to the results just established, the following assumption is made.

Assumption 2. The graph is strongly connected and g_i is non-zero for at least one root node i. ■

Note that existence of a spanning tree is necessary and sufficient for synchronization of all states with dynamics given in (1). However, interactive Nash equilibrium requires the graph to be *strongly connected*.

4.3. Stability and Nash solution of the graphical games

We will now prove that the policies given in terms of the solutions to the coupled Bellman optimality equations (23) provide Nash equilibrium solution for the dynamic graphical game.

Theorem 4 (*Stability and Nash Equilibrium Solution*)**.** *Let $0 < V_i^*(\varepsilon_{ik}) \in C^2$ satisfy DTHJB (37), or equivalently the Bellman optimality equation (23). Let all agents use the control policies given by (38). Let the graph contain a spanning tree with at least one nonzero pinning gain. Then:*

a. *The error dynamics (9) are asymptotically stable and all agents synchronize to the target node dynamics (2).*
b. *The optimal performance index for each agent i is given by $J_i^*(\varepsilon_{il}, u_{il}^*, u_{il}^*) = V_i^*(\varepsilon_{il})$.*
c. *All agents are in Nash equilibrium.*

Proof. a. $V_i^*(\varepsilon_{ik})$ satisfies the Bellman optimality equation such that

$$V_i^*(\varepsilon_{i(k+1)}) - V_i^*(\varepsilon_{i(k)}) = -U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}^*) < 0. \tag{55}$$

Therefore, $V_i^*(\varepsilon_{ik})$ serves as Lyapunov function for (9), and the error system (9) is asymptotically stable. If there is a spanning tree, then according to Lemma 1, all agents synchronize to the targets node dynamics.

b. Using Theorem 2 and DTHJB (37), then the Hamiltonian (39) for arbitrary control policies is given by

$$\begin{aligned} &H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) \\ &= \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)} + U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) \\ &= \frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) + \frac{1}{2}\sum_{j\in N_i}(u_{jk} - u_{jk}^*)^T R_{ij}(u_{jk} - u_{jk}^*) \\ &\quad + \sum_{j\in N_i} u_{jk}^{*T} R_{ij}(u_{jk} - u_{jk}^*) + \sum_{j\in N_i} e_{ij}\nabla V_i^*(\varepsilon_{i(k+1)})^T B_j(u_{jk} - u_{jk}^*). \end{aligned} \tag{56}$$

Using the result in part a, $\varepsilon_i(\infty) \to 0$. Therefore $V_i^*(\varepsilon_i(\infty)) = 0$ and

$$J_i(\varepsilon_{il}, u_{il}, u_{-il}) = V_i^*(\varepsilon_i(\infty)) + \sum_{k=l}^{\infty} U_i(\varepsilon_{il}, u_{il}, u_{-il}). \tag{57}$$

Rearranging this equation yields,

$$\begin{aligned} J_i(\varepsilon_{il}, u_{il}, u_{-il}) = V_{il}^*(\varepsilon_{il}) + \sum_{k=l}^{\infty}(&U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) \\ &- U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}^*)). \end{aligned} \tag{58}$$

The Hamiltonian for arbitrary control inputs is given by

$$\begin{aligned} &H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) \\ &= \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik}, u_{-ik}} + U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}). \end{aligned} \tag{59}$$

The Hamiltonian for optimal control inputs is given by

$$\begin{aligned} &H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}^*) \\ &= \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik}^*, u_{-ik}^*} + U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}^*) = 0. \end{aligned} \tag{60}$$

It is further noted that,

$$\begin{aligned} &-\nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik}, u_{-ik}} + \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik}^*, u_{-ik}^*} \\ &= u_{ik}^{*T} R_{ii}(u_{ik} - u_{ik}^*) - \sum_{j\in N_i} e_{ij}\nabla V_i^*(\varepsilon_{i(k+1)})^T B_j(u_{jk} - u_{jk}^*). \end{aligned} \tag{61}$$

Eqs. (59), (60), and (61) yield

$$\begin{aligned} &U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) - U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}^*) \\ &= \frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) + u_{ik}^{*T} R_{ii}(u_{ik} - u_{ik}^*) \\ &\quad + \frac{1}{2}\sum_{j\in N_i}(u_{jk} - u_{jk}^*)^T R_{ij}(u_{jk} - u_{jk}^*) + \sum_{j\in N_i} u_{jk}^{*T} R_{ij}(u_{jk} - u_{jk}^*). \end{aligned} \tag{62}$$

Using (62) into (58) yields

$$J_i(\varepsilon_{il}, u_{il}, u_{-il}) = V_i^*(\varepsilon_{il}) + \sum_{k=l}^{\infty}\left(\sum_{j\in N_i}\frac{1}{2}(u_{ik}-u_{ik}^*)^T R_{ii}(u_{ik}-u_{ik}^*) + u_{ik}^{*T}R_{ii}(u_{ik}-u_{ik}^*) + \frac{1}{2}\sum_{j\in N_i}(u_{jk}-u_{jk}^*)^T R_{ij}(u_{jk}-u_{jk}^*) + \sum_{j\in N_i}u_{jk}^{*T}R_{ij}(u_{jk}-u_{jk}^*)\right). \quad (63)$$

Using (63) with the optimal policies (38) yields the optimal performance index J_i^* such that

$$J_i^*(\varepsilon_{il}, u_{il}^*, u_{-il}^*) = V_i^*(\varepsilon_{il}). \quad (64)$$

c. Given that the summation of the performance index (58) is positive for arbitrary control policies such that

$$\sum_{k=l}^{\infty} U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}^*) - U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}^*) > 0. \quad (65)$$

Eqs. (63), (64), and (65) yield

$$J_i^*(\varepsilon_{il}, u_{il}^*, u_{\bar{l}l}^*) \le J_i(\varepsilon_{il}, u_{il}, u_{\bar{l}l}^*) \quad (66)$$

from which the result follows, according to Definition 3. ■

The next result shows that the interactive Nash equilibrium requires the graph to be strongly.

Lemma 4. *Let the hypotheses of Theorem 4 and Assumptions 1 and 2 hold. Then* $\{u_1^*, u_2^*, \ldots, u_N^*\}$ *are in interactive Nash equilibrium and hence all agents synchronize to the target node dynamics.*

Proof. The proof is a consequence of Theorems 1–4. ■

4.4. Best response solution of dynamic graphical games

Theorem 4 and Lemma 4 reveal that the agents are in interactive Nash equilibrium if, for all $i \in N$, agent i selects its best response to its neighbors policies and the graph is strongly connected.

Now by considering fixed neighbor policies $u_{-i} = \{u_j : j \in N_i\}$ we can define the best response Bellman equation for each agent i as

$$V_i^o(\varepsilon_{ik}) = V_i^o(\varepsilon_{i(k+1)}) + \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + \sum_{j\in N_i}u_{jk}^T R_{ij}u_{jk} + (d_i+g_i)^2\nabla V_i^o(\varepsilon_{i(k+1)})^T B_i R_{ii}^{-1}B_i^T\nabla V_i^o(\varepsilon_{i(k+1)})\right) \quad (67)$$

with initial condition given by $V_i^o(0) = 0$ and $u_{ik} = u_{ik}^o$ where u_{ik}^o is given by (22) in terms of the solution (67).

Define the best response Hamilton–Jacobi (HJ) equation as

$$H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}) = \nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)} + \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + u_{ik}^{*T}R_{ii}u_{ik}^* + \sum_{j\in N_i}u_{jk}^T R_{ij}u_{jk}\right) = 0 \quad (68)$$

with $V_i^*(0) = 0$ and $u_{ik} = u_{ik}^*$ where u_{ik}^* is given by (38) in terms of the solution to (68).

The next lemma shows the relation between the best response Bellman equation (67) and the best response Hamilton–Jacobi equation (68).

Lemma 5. a. *Let* $0 < V_i^*(\varepsilon_{ik}) \in C^2$, $\forall i$ *satisfy the best response (DTHJ) equation (68) with initial condition given by* $V_i^*(0) = 0$, *and optimal control policy* u_{ik}^* *given by (38). Then,* $V_i^*(\varepsilon_{ik})$ *satisfies the best response Bellman equation (67).*

b. *Let* (A, B_i) $\forall i$ *be reachable. Let* $0 < V_i^*(\varepsilon_{ik}) \in C^2$, $\forall i$ *satisfy (67). Then* $V_i^*(\varepsilon_{ik})$ *satisfies (68).*

Proof. a. Let $V_i^*(\varepsilon_{ik})$ satisfy (68) and u_{ik}^* be given by (38), then $H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}) = 0$. Then by using Theorem 1, we have $\Delta V_i^*(\varepsilon_{ik}) = \nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)}$. Therefore $V_i^*(\varepsilon_{ik})$ satisfies (67).

b. The best response Hamiltonian with arbitrary smooth value $V_i^*(\varepsilon_{ik})$ for arbitrary control policy u_{ik} is given by,

$$\begin{aligned} &H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) \\ &= \nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)} + U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) \\ &= H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}) \\ &\quad + \frac{1}{2}(u_{ik}-u_{ik}^*)^T R_{ii}(u_{ik}-u_{ik}^*). \end{aligned} \quad (69)$$

Now, let $V_i(\varepsilon_{ik}) \in C^2$, $\forall i$ satisfy (19) such that

$$V_i(\varepsilon_{ik}) = U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) + (\nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)}) - (\nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)}) + V_i(\varepsilon_{i(k+1)}). \quad (70)$$

Substituting (69) into (70) yields

$$V_i(\varepsilon_{ik}) = V_i(\varepsilon_{i(k+1)}) - (\nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)}) + \frac{1}{2}(u_{ik}-u_{ik}^*)^T R_{ii}(u_{ik}-u_{ik}^*).$$

By applying Bellman's optimality principle, yields that $V_i^o(\varepsilon_{ik})$ has to satisfy the following equation

$$V_i^o(\varepsilon_{ik}) = \min_{u_{ik}}(V_i^o(\varepsilon_{i(k+1)}) - (\nabla V_i^*(\varepsilon_{i(k+1)})^T\varepsilon_{i(k+1)}) + \frac{1}{2}(u_{ik}-u_{ik}^*)^T R_{ii}(u_{ik}-u_{ik}^*)).$$

By applying the stationarity condition $\partial V_i(\varepsilon_{ik})/\partial u_{ik} = 0$ one has to solve the following equation for the control policy u_{ik}^o.

$$u_{ik}^o - u_{ik}^* = (g_i+d_i)R_{ii}^{-1}B_i^T(\nabla V_i^o(\varepsilon_{i(k+1)}) - \nabla V_i^*(\varepsilon_{i(k+1)})). \quad (71)$$

The remainder of the proof involves the reachability result from Theorem 2. ■

Theorem 5 (*Best Response Solution*). *Given fixed neighboring policies* $u_{-i} = \{u_j : j \in N_i\}$, *assume there exists an admissible policy* u_i. *Let* $0 < V_i^*(\varepsilon_{ik}) \in C^2$ *satisfy the best response (HJ) equation (68), or equivalently the best response Bellman equation (67). Let each agent* i *use control policy (38). Let the graph contain a spanning tree with at least one pinning gain nonzero. Then:*

a. *The error dynamics (9) are asymptotically stable and all the agents synchronize to the target node dynamics (2).*
b. *The optimal performance index for each agent* i *is given by* $J_i^*(\varepsilon_{il}, u_{il}^*, u_{\bar{l}l}) = V_i^*(\varepsilon_{il})$.
c. *All agents are in Nash equilibrium.*

Proof. a. Suppose that $V_i^*(\varepsilon_{ik})$ satisfies (67) such that,

$$V_i^*(\varepsilon_{i(k+1)}) - V_i^*(\varepsilon_{ik}) = -U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}) < 0. \quad (72)$$

Therefore, $V_i^*(\varepsilon_{ik})$ serves as Lyapunov function for the error system (9), and the error system (9) is asymptotically stable. Hence, according to Lemma 1, all agents synchronize to the targets node dynamics (2).

b. Using Lemma 5 and the best response (HJ) equation (68), then the best response Hamiltonian for arbitrary control policies is given by (69).

Using the result in part a, $\varepsilon_i(\infty) \to 0$. Therefore $V_i^*(\varepsilon_i(\infty)) = 0$ and the best response performance index for each agent i is given by

$$J_i(\varepsilon_{il}, u_{il}, u_{-il}) = V_{il}^*(\varepsilon_{il}, u_{il}^*, u_{-il}) + \sum_{k=l}^{\infty}(U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) - U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik})). \quad (73)$$

The best response Hamiltonian with arbitrary control input u_{ik} is given by

$$H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) = \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik},u_{-ik}} + U_i^*(\varepsilon_{ik}, u_{ik}, u_{-ik}). \quad (74)$$

The best response (HJ) equation for agent i is given by

$$H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}^*, u_{-ik}) = \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik}^*,u_{-ik}} + U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}) = 0. \quad (75)$$

It is straightforward to see that,

$$U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) - U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}) = H_i(\varepsilon_{ik}, \nabla V_i^*(\varepsilon_{i(k+1)}), u_{ik}, u_{-ik}) - \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik},u_{-ik}} + \nabla V_i^*(\varepsilon_{i(k+1)})^T \varepsilon_{i(k+1)}|_{u_{ik}^*,u_{-ik}}.$$

Finally one has,

$$U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) - U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}) = \frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) + u_{ik}^{*T} R_{ii}(u_{ik} - u_{ik}^*). \quad (76)$$

Using (76) into (73) yields

$$J_i(\varepsilon_{il}, u_{il}, u_{-il}) = V_i^*(\varepsilon_{il}, u_{il}^*, u_{-il}) + \sum_{k=l}^{\infty}\left(\frac{1}{2}(u_{ik} - u_{ik}^*)^T R_{ii}(u_{ik} - u_{ik}^*) + u_{ik}^{*T} R_{ii}(u_{ik} - u_{ik}^*)\right). \quad (77)$$

The best response performance index (77) with the optimal control policy (38) is given by the unique value $V_i^*(\varepsilon_{il})$.

$$J_i^*(\varepsilon_{il}, u_{il}^*, u_{-il}) = V_i^*(\varepsilon_{il}). \quad (78)$$

c. The summation of the best response performance index (77) is positive for arbitrary control policies u_{ik} such that

$$\sum_{k=l}^{\infty} U_i(\varepsilon_{ik}, u_{ik}, u_{-ik}) - U_i^*(\varepsilon_{ik}, u_{ik}^*, u_{-ik}) > 0. \quad (79)$$

Eqs. (77), (78), and (79) yield

$$J_i^*(\varepsilon_{il}, u_{il}^*, u_{\bar{i}l}) \le J_i(\varepsilon_{il}, u_{il}, u_{\bar{i}l}) \quad (80)$$

which proves that the policies u_{il}^*, $\forall i$ form a Nash equilibrium according to Definitions 3 and 4. ■

5. Value iteration algorithm for graphical games

In this section, a value iteration HDP algorithm is proposed for solving the discrete-time dynamic graphical games. This is a cooperative version of adaptive dynamic programming (Werbos, 1974, 1992). Specifically, the single-agent HDP algorithm is extended to the multi-player graphical game.

Algorithm 1 (*HDP Algorithm for Graphical Games*). *Step* 1: Start with arbitrary initial policies u_{ik}^0 and values $V_i^0(\varepsilon_{ik})$.

Step 2: Solve for V_i^{l+1} using Bellman equations

$$V_i^{l+1}(\varepsilon_{ik}) = U_i(\varepsilon_{ik}, u_{ik}^l, u_{-ik}^l) + V_i^l(\varepsilon_{i(k+1)}) \quad (81)$$

where l is the iteration index.

Step 3: Update the control policies using

$$u_{ik}^{l+1} = (d_i + g_i)R_{ii}^{-1} B_i^T \nabla V_i(\varepsilon_{i(k+1)})^{l+1} \quad (82)$$

Step 4: On convergence of $\left\| V_i(\varepsilon_{ik})^{l+1} - V_i(\varepsilon_{ik})^l \right\|$ *End.* ■

Theorems 6 and 7 that follow provide the convergence proof for Value Iteration Algorithm 1 for two separate cases. In the first case, every agent i performs the Algorithm 1 while its neighboring agents *hold their policies fixed*. In second case, all agents update their policies *simultaneously* using Algorithm 1.

5.1. Best response solution using HDP algorithm

For the first case, all the neighbors of agent i retain fixed policies u_j, $j \in N_i$. Then the optimal control policy sequence $\{L_i^l\}_{l=0}^{\infty} \in R^{m_i}$ for each agent i at each iteration step in Algorithm 1 is given by

$$L_i^l = \arg\min_{u_{ik}} \left(\frac{1}{2} \left(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + u_{ik}^T R_{ii} u_{ik} + \sum_{j \in N_i} u_{jk}^T R_{ij} u_{jk} \right) + V_i^l(\varepsilon_{i(k+1)}) \right). \quad (83)$$

The notation has been streamlined to simplify the presentation of the proofs, and l is the iteration index. The associated value function sequence is

$$V_i^{l+1}(\varepsilon_{ik}) \equiv F_i(V_i^l, L_i^l) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \overline{V}_i^l(R_{ij}, u_{-i}^l)) + V_i^l(\varepsilon_{i(k+1)}) \quad (84)$$

where $\overline{V}_i^l(R_{ij}, u_{-i}^l) = \sum_{j \in N_i} u_j^{lT} R_{ij} u_j^l$ has a fixed value.

Now consider arbitrary stabilizing control policy sequence $\{M_i^l\}_{l=0}^{\infty} \in R^{m_i}$. The associated value function sequence is

$$Z_i^{l+1}(\varepsilon_{ik}) \equiv F_i(Z_i^l, M_i^l) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + M_i^{lT} R_{ii} M_i^l + \overline{Z}_i^l(R_{ij}, u_{-i}^l)) + Z_i^l(\varepsilon_{i(k+1)}) \quad (85)$$

where $\overline{Z}_i^l(R_{ij}, u_{-i}^l) = \sum_{j \in N_i} u_j^{lT} R_{ij} u_j^l$ has a fixed value. Note that $\overline{V}_i^l(R_{ij}, u_{-i}^l) = \overline{Z}_i^l(R_{ij}, u_{-i}^l)$.

The following theorem proves convergence of the HDP algorithm with fixed neighboring policies by using an induction proof following the results of Wang et al. (2012) and Zhang et al. (2009).

Theorem 6 (*Convergence of HDP Algorithm*). *Let the neighbors of each agent i have fixed policies u_j. Assume there exists an admissible policy u_i. Let agent i perform Algorithm 1. Then the solution sequence $\{V_i^l\}_{l=0}^{\infty} \in R^1$ converges to the best response solution $V_i^*(\varepsilon_{ik})$ $\forall i$ of (67).*

Proof. According to Lemmas 6, 7 in the Appendix, one has

$$0 \le V_i^l \le Z_i^l \le \overline{U}. \quad (86)$$

Using the value function sequences (84) and (85) we can write,

$$V_i^{l+1}(\varepsilon_{ik}) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \overline{V}_i^l(R_{ij}, u_{-i}^l)) + V_i^l(\varepsilon_{i(k+1)}) \quad (87)$$

$$Z_i^l(\varepsilon_{ik}) = \frac{1}{2}(\varepsilon_i^T Q_{ii} \varepsilon_i + L_i^{lT} R_{ii} L_i^l + \overline{V}_i^{l-1}(R_{ij}, u_{-i}^{l-1})) + Z_i^{l-1}(\varepsilon_{i(k+1)}). \quad (88)$$

Starting with $l = 0$, we will show that the hypothesis $Z_i^l(\varepsilon_{ik}) \le V_i^{l+1}(\varepsilon_{ik})$ holds. By setting $V_i^0 = Z_i^0 = 0$ and $V_i^1(\varepsilon_{ik}) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \overline{V}_i^l(R_{ij}, u_{-i}^l)) \ge 0$, then (87) and (88) for $\forall L_i^l$ yield

$$V_i^1(\varepsilon_{ik}) - Z_i^0(\varepsilon_{ik}) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \overline{V}_i^l(R_{ij}, u_{-i}^l)) \ge 0 \quad (89)$$

or

$$Z_i^0(\varepsilon_{ik}) \le V_i^1(\varepsilon_{ik}). \tag{90}$$

Now we need to assume that $Z_i^l(\varepsilon_{ik}) \le V_i^{l+1}(\varepsilon_{ik})$ holds for $l-1$ such that $Z_i^{l-1}(\varepsilon_{ik}) \le V_i^l(\varepsilon_{ik}),\ \forall L_i^l$. Then at (*step l*) (87) and (88) yield

$$V_i^{l+1}(\varepsilon_{ik}) - Z_i^l(\varepsilon_{ik}) = V_i^l(\varepsilon_{ik}) - Z_i^{l-1}(\varepsilon_{ik}) \ge 0. \tag{91}$$

Therefore the mathematical induction for $\forall l$ gives

$$Z_i^l(\varepsilon_{ik}) \le V_i^{l+1}(\varepsilon_{ik}), \quad \forall l. \tag{92}$$

Consequently, for all policies L_i^l, $\forall l$ the following monotonic sequence holds,

$$V_i^{l+1} > Z_i^l > V_i^l > \cdots \ge 0. \tag{93}$$

Since $Z_i^l(\varepsilon_{ik})$ is a lower bound on $V_i^{l+1}(\varepsilon_{ik})$ and Lemma 7 sets $\overline{U}$ as an upper bound on V_i^{l+1}, then (86) and (93) yield

$$0 \le V_i^l \le Z_i^l \le V_i^{l+1} \le \overline{U}. \tag{94}$$

From (94) the sequence V_i^l is increasing and it has an upper bound $\overline{U}$ which means that V_i^l converges to the best response solution V_i^* that satisfies the following equation

$$V_i^*(\varepsilon_{ik}) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{*T} R_{ii} L_i^* + \overline{V}_i(R_{ij}, u_{-i})) + V_i^*(\varepsilon_{i(k+1)}) \tag{95}$$

where

$$L_i^* = (d_i + g_i)R_{ii}^{-1} B_i^T \nabla V_i^*(\varepsilon_{i(k+1)}). \quad \blacksquare \tag{96}$$

Remark 1. In this theorem, the existence of an admissible best response policy is required. Admissibility is a standard assumption and is required in order for an agent to have a best response towards its neighborhood. If this assumption does not hold, there is no solution to the game for the selected policies in the neighborhood. ■

5.2. Nash solution using HDP algorithm

For the second case, all the agents i perform Algorithm 1 simultaneously at each iteration step. Then the control policy sequences $\{L_i^l\}_{l=0}^{\infty} \in R^{m_i}$ for every agent i are given by (83) and the associated value sequences are given by

$$V_i^{l+1}(\varepsilon_{ik}) = \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} L_j^{lT} R_{ij} L_j^l\right) + V_i^l(\varepsilon_{i(k+1)}). \tag{97}$$

For arbitrary admissible policies for the agents $\{M_{il}\}_{l=0}^{\infty} \in R^{m_i}$, their associated value sequences are given by

$$Z_i^{l+1}(\varepsilon_{ik}) = \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + M_i^{lT} R_{ii} M_i^l + \sum_{j\in N_i} M_j^{lT} R_{ij} M_j^l\right) + Z_i^l(\varepsilon_{i(k+1)}). \tag{98}$$

The following theorem proves convergence of the HDP algorithm when all agents update their policies simultaneously. The proof follows the induction proof of Theorem 6.

Theorem 7 (*Convergence of HDP Algorithm*)**.** *Assume there exist admissible policies u_i $\forall i$. Let all agents update their policies simultaneously using Algorithm 1. Suppose that $\overline{\sigma}(R_{jj}^{-1}R_{ij})$ is small. Then the solution sequences $\{V_i^l\}_{l=0}^{\infty} \in R^1$ converge monotonically to the optimal solution $\tilde{V}_i^*(\varepsilon_{ik})$ $\forall i$ of (23).*

Proof. According to Lemmas 8, 9 in the Appendix, one has the following inequality for every agent

$$0 \le V_i^l \le Z_i^l \le \overline{\overline{U}}. \tag{99}$$

The value function sequences (97) and (98), with control policies given by (83) yield

$$V_i^{l+1}(\varepsilon_{ik}) = \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} L_j^{lT} R_{ij} L_j^l\right) + V_i^l(\varepsilon_{i(k+1)}) \tag{100}$$

and

$$Z_i^l(\varepsilon_{ik}) = \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} L_j^{lT} R_{ij} L_j^l\right) + Z_i^{l-1}(\varepsilon_{i(k+1)}). \tag{101}$$

Starting with $l = 0$, we will show that the hypothesis $Z_i^l(\varepsilon_{ik}) \le V_i^{l+1}(\varepsilon_{ik})$ holds. By setting $V_i^0 = Z_i^0 = 0$ and $V_i^1(\varepsilon_{ik}) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} L_j^{lT} R_{ij} L_j^l) \ge 0$, then (100) and (101) for $\forall L_i^l$, L_j^l yield

$$V_i^1(\varepsilon_{ik}) - Z_i^0(\varepsilon_{ik}) = \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} L_j^{lT} R_{ij} L_j^l\right) \ge 0 \tag{102}$$

or

$$Z_i^0(\varepsilon_{ik}) \le V_i^1(\varepsilon_{ik}). \tag{103}$$

Now we need to assume that $Z_i^l(\varepsilon_{ik}) \le V_i^{l+1}(\varepsilon_{ik})$ holds for $l-1$ such that $Z_i^{l-1}(\varepsilon_{ik}) \le V_i^l(\varepsilon_{ik}),\ \forall L_i^l$. Then at (*step l*) (100) and (101) yield

$$V_i^{l+1}(\varepsilon_{ik}) - Z_i^l(\varepsilon_{ik}) = V_i^l(\varepsilon_{ik}) - Z_i^{l-1}(\varepsilon_{ik}) \ge 0. \tag{104}$$

Therefore the mathematical induction for $\forall l$ yields

$$Z_i^l(\varepsilon_{ik}) \le V_i^{l+1}(\varepsilon_{ik}), \quad \forall l. \tag{105}$$

Consequently, for all the policies L_i^l, L_j^l, $\forall l$ the following monotonic sequence holds

$$V_i^{l+1} > Z_i^l > V_i^l > \cdots \ge 0. \tag{106}$$

Since $Z_i^l(\varepsilon_{ik})$ is the lower bound of $V_i^{l+1}(\varepsilon_{ik})$ and Lemma 9 sets $\overline{\overline{U}}$ as an upper value of V_i^{l+1}, then (99) and (106) yield

$$0 \le V_i^l \le Z_i^l \le V_i^{l+1} \le \overline{\overline{U}}. \tag{107}$$

From (107) the sequence of V_i^l is increasing and it has an upper bound $\overline{\overline{U}}$, which means that V_i^l will converge to the optimal solution $\tilde{V}_i^*$ monotonically by satisfying

$$\tilde{V}_i^*(\varepsilon_{ik}) = \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + \tilde{L}_i^{*T} R_{ii} \tilde{L}_i^* + \sum_{j\in N_i} \tilde{L}_j^{*T} R_{ij} \tilde{L}_j^*\right) + \tilde{V}_i^*(\varepsilon_{i(k+1)}) \tag{108}$$

where

$$\tilde{L}_i^* = (d_i + g_i)R_{ii}^{-1} B_i^T \nabla \tilde{V}_i^*(\varepsilon_{i(k+1)}). \quad \blacksquare \tag{109}$$

Remark 2. The condition $\overline{\sigma}(R_{jj}^{-1}R_{ij})$ small means that agent j weights his own control effort in his value function higher than his out-neighbors weight his control effort in their value functions. This can be guaranteed for any choice of R_{ij} of agent i by selecting R_{jj} large enough. This is related to the weakly coupling condition defined in Chapter 6 of Başar and Olsder (1999). ■

6. Graphical game solutions by actor–critic learning

This section develops an actor–critic framework based on value function approximation which will be used to solve the dynamic graphical game online. This framework is motivated by Algorithm 1. Each agent i has its own critic network to perform the value update (81) and an actor network to perform the policy improvement (82). The actor–critic network structures depend only on local information.

6.1. Actor–critic networks and tuning

The value function $V_i(\varepsilon_{i(k+1)})$ for each agent i is approximated by a critic network $\hat{V}_i(.|\tilde{W}_{ic})$, and the control policy is approximated by an actor network $\hat{u}_i(.|\tilde{W}_{ia})$ so that

$$\hat{V}_{ik}(\tilde{W}_{ic}) = Z_{ik}^T \tilde{W}_{ic}^T Z_{ik} \tag{110}$$

$$\hat{u}_{ik}(\tilde{W}_{ia}) = \tilde{W}_{ia}^T Z_{ik} \tag{111}$$

where $\tilde{W}_{ic} \in R^{nN_{i,j} \times nN_{i,j}}$ and $\tilde{W}_{ia} \in R^{nN_{i,j} \times m_i}$ are the critic and actor weights respectively. $N_{i,j}$ is the total number of each agent i and its neighbors. $Z_{ik} \in R^{nN_{i,j}}$ is a vector of the state ε_{ik} of agent i and the states of its neighbors.

Let $\xi_{u_{ik}}^{V_i(\varepsilon_{ik})}$ be the approximation error of the actor network so that

$$\xi_{u_{ik}}^{V_i(\varepsilon_{ik})} = \hat{u}_{ik}(\tilde{W}_{ai}) - \tilde{u}_{ik} = \tilde{W}_{ai}^T Z_{ik} - \tilde{u}_{ik} \tag{112}$$

where, based on (82), the target control policy $\tilde{u}_{ik}$ is given in terms of the critic network such that

$$\tilde{u}_{ik} = (g_i + d_i) R_{ii}^{-1} B_i^T \nabla \hat{V}_{i(k+1)}. \tag{113}$$

This target control policy $\tilde{u}_{ik}$ can be expressed in terms of the critic network weights $\tilde{W}_{ic}$ such that

$$\tilde{u}_{ik} = (g_i + d_i) R_{ii}^{-1} B_i^T O_i \tilde{W}_{ic}^T Z_{ik} \tag{114}$$

where $O_i = 2 \times [0 \dots [I]_{ii} \dots 0] \in R^{n \times nN_{i,j}}$.

The squared approximation error is

$$\mathrm{err}_{\mathrm{actor}} = \frac{1}{2} (\xi_{u_{ik}}^{V_i(\varepsilon_{ik})})^T \xi_{u_{ik}}^{V_i(\varepsilon_{ik})}. \tag{115}$$

The change in the actor weights is given by the gradient descent. The update rule for the actor weights is therefore given by

$$\tilde{W}_{ia}^{(l+1)T} = \tilde{W}_{ia}^{lT} - \tilde{\mu}_{ia} ((\tilde{W}_{ia}^{lT} Z_{ik} - \tilde{u}_{ik}^l)(Z_{ik})^T) \tag{116}$$

where $0 < \tilde{\mu}_{ia} < 1$ is the actor network learning rate.

The value update equation is given by (81). Let $\Im_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})}$ be the target value of the critic network at step l such that

$$\Im_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})} = \frac{1}{2} \left(\varepsilon_{ik}^T Q_{ii} \varepsilon_{ik} + \hat{u}_{ik}^{lT} R_{ii} \hat{u}_{ik}^l + \sum_{j \in N_i} \hat{u}_{jk}^{lT} R_{ij} \hat{u}_{jk}^l \right) + \hat{V}_{i(k+1)}^l. \tag{117}$$

The critic network approximation error at step l is given by

$$\xi_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})} = \Im_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})} - \hat{V}_{ik}(\tilde{W}_{ic}). \tag{118}$$

Similarly, we define squared approximation error for the critic as

$$\mathrm{err}_{\mathrm{critic}} = \frac{1}{2} (\xi_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})})^T \xi_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})} = \frac{1}{2} \left\| \Im_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})} - Z_{ik}^T \tilde{W}_{ic}^T Z_{ik} \right\|_2^2. \tag{119}$$

By employing gradient descent, the update rule for the critic weights is given by

$$\tilde{W}_{ic}^{(l+1)T} = \tilde{W}_{ic}^{lT} - \tilde{\mu}_{ic} \left(\Im_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})} - Z_{ik}^T \tilde{W}_{ic}^{lT} Z_{ik} \right) Z_{ik} Z_{ik}^T \tag{120}$$

where $0 < \tilde{\mu}_{ic} < 1$ is the critic network learning rate.

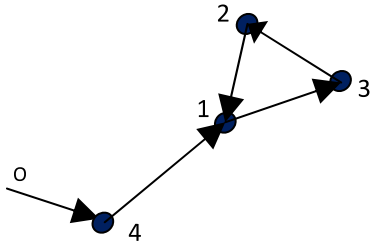

Fig. 1. Graphical game example.

6.2. Actor–critic online tuning in real-time

The following algorithm is used for online tuning of the actor–critic networks in real-time using data measured along the system trajectories.

Algorithm 2 (*Actor–Critic Network Online Tuning*)**.**

1. Initialize the actor weights $\tilde{W}_{ia}^0$ randomly and initialize the critic weights $\tilde{W}_{ic}^0$ with zero values.
2. Do Loop (l iterations) {
   - 2.1 Start with initial state ε_{i0} on the system trajectory.
   - 2.2 Calculate $\hat{u}_{ik}^l$ using (111).
   - 2.3 Calculate the dynamics $\varepsilon_{i(k+1)}^l$ using (9).
   - 2.4 Calculate the value $\hat{V}_{i(k+1)}^l$ using (110).
   - 2.5 Critic update rule
     $\tilde{W}_{ic}^{(l+1)T} = \tilde{W}_{ic}^{lT} - \mu_{ic} (\Im_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})} - Z_{ik}^T \tilde{W}_{ic}^{lT} Z_{ik}) Z_{ik} Z_{ik}^T$
     where $\Im_{\varepsilon_{ik}}^{V_i(\varepsilon_{ik})}$ is given by (117).
   - 2.6 Actor update rule
     $\tilde{W}_{ia}^{(l+1)T} = \tilde{W}_{ia}^{lT} - \mu_{ia} ((\tilde{W}_{ia}^{lT} Z_{ik} - \tilde{u}_{ik}^l)(Z_{ik})^T)$
     where $\tilde{u}_{ik}^l = (g_i + d_i) R_{ii}^{-1} B_i^T O_i \tilde{W}_{ic}^T Z_{ik}$
   - 2.7 On convergence of $\left\| \hat{V}_i(\varepsilon_{ik})^{l+1} - \hat{V}_i(\varepsilon_{ik})^l \right\|$ End Loop}. ■

Remark 3. Algorithm 2 uses gradient descent to tune the weights of the critic and actor networks at each iteration. Assuming that the gradient descent algorithms converge exactly at each iteration, then Algorithm 2 at each step solves the Bellman equation (81). Then, Theorem 7 proves convergence of Algorithm 1. Unfortunately, gradient descent cannot always be guaranteed to converge to the exact solutions in approximation structures. However, simulations have shown the effectiveness of this algorithm. ■

6.3. Graphical game example and simulation results

The graphical game can be solved online in real-time by using Algorithm 2. In this section we perform simulations to verify the theoretical developments. Consider the directed graph with four agents shown in Fig. 1.

The plant and input matrices for every agent are given as,

$$A = \begin{bmatrix} 0.995 & 0.09983 \\ -0.09983 & 0.995 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 0.2047 \\ 0.08984 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0.2147 \\ 0.2895 \end{bmatrix},$$

$$B_3 = \begin{bmatrix} 0.2097 \\ 0.1897 \end{bmatrix}, \quad B_4 = \begin{bmatrix} 0.2 \\ 0.1 \end{bmatrix}$$

with pinning gains given by $g_1 = g_2 = g_3 = 0, g_4 = 1$, edge weights given by $e_{12} = 0.8, e_{14} = 0.7, e_{23} = 0.6, e_{31} = 0.8$ and the learning rates are selected as ($\tilde{\mu}_{ic} = 0.1$, $\tilde{\mu}_{ia} = 0.1$, $\forall i$).

The user defined matrices in the performance indices are selected to be,

$$Q_{11} = Q_{22} = Q_{33} = Q_{44} = I_{2\times 2}, \quad R_{11} = R_{22} = R_{33} = R_{44} = 1,$$
$$R_{13} = R_{21} = R_{32} = R_{41} = 0, \quad R_{12} = R_{14} = R_{23} = R_{31} = 1.$$

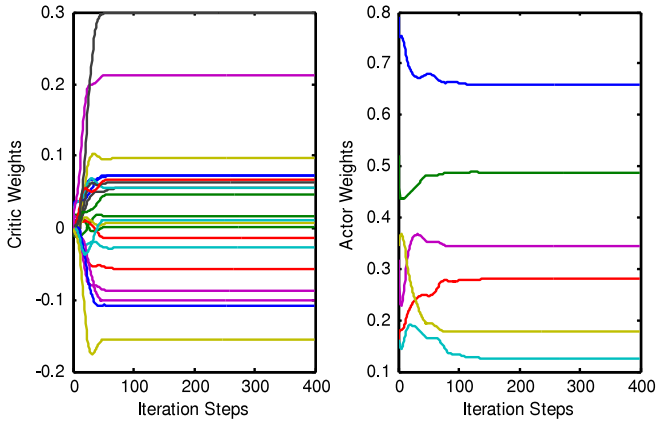

Fig. 2. Critic–actor weights update of Agent (1) versus iteration steps.

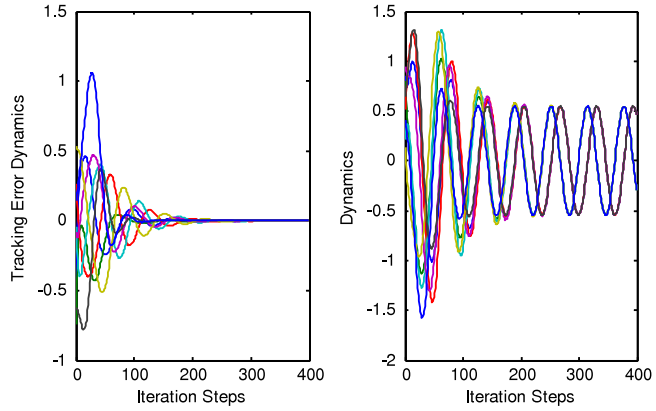

Fig. 3. Tracking error and agents' states versus iteration steps.

Fig. 2 shows the critic and actor weights of agent 1. Fig. 3 shows the neighborhood tracking error dynamics and the dynamics of all four agents. In this figure it is shown that the tracking error dynamics vanish, and all the agents synchronize to the leader while preserving their optimality. Fig. 4 shows the phase plane plots of all four agents. It is shown that starting with random initial values, a synchronization to the leader is achieved. These figures show that Algorithm 2 yields stability and synchronization to the leader's state.

7. Conclusion

This paper introduces a new class of discrete-time dynamic games known as dynamic graphical games. It brings together discrete Hamiltonian mechanics, distributed multi-agent control, optimal control theory, and game theory to formulate and solve these dynamic graphical games. The relation between the Bellman optimality equation and the discrete-time HJB equation for the graphical game is shown. It is shown that standard notions of Nash equilibrium are insufficient to guarantee that all agents are involved in the same game and for that reason a new notion of interactive Nash equilibrium is introduced which holds if the agents are all in Nash and the graph is strongly connected. A value iteration algorithm is proposed to solve the dynamic graphical games. Based on this algorithm, a real adaptive learning structure is developed to solve the dynamic graphical game in real-time. Simulation results show the effectiveness of the proposed structure.

Appendix

The next two technical lemmas are required to prove Theorem 6. They are motivated by Al-Tamimi et al. (2008), and Lancaster and Rodman (1995).

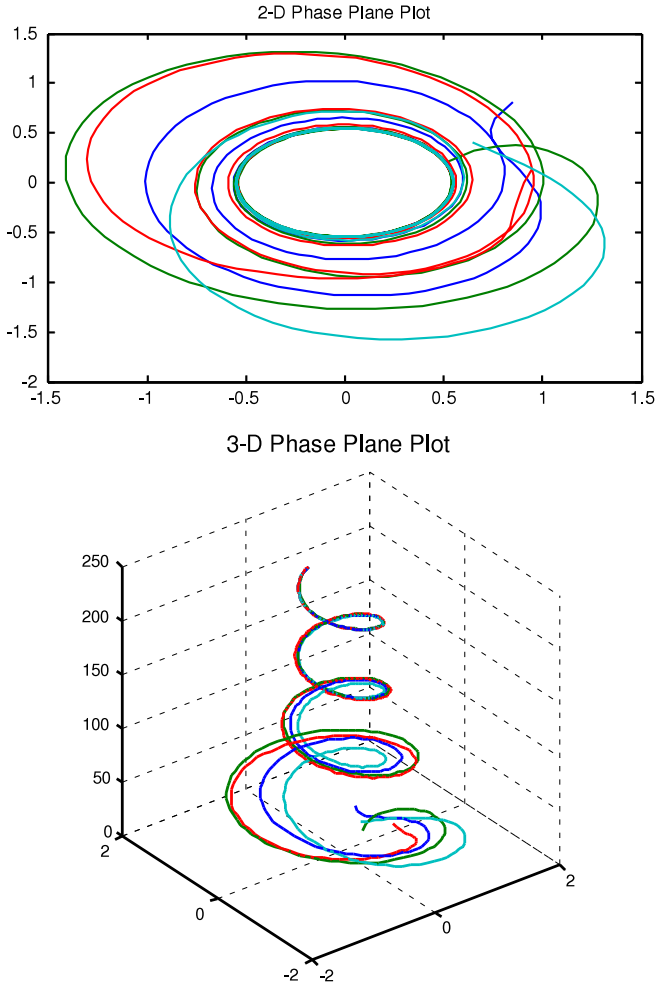

Fig. 4. Phase plane plots of the agents that show that the agents are being synchronized.

Lemma 6. *Let the neighbors of agent i have fixed policies u_j. Given arbitrary stabilizing control policies $\{M_i^l\}_{l=0}^{\infty}$ for every agent i, let the associated value sequence be $\{Z_i^l\}_{l=0}^{\infty}$, $Z_i^0 \geq 0$. Define the sequence of control policies generated by Algorithm 1 as $\{L_i^l\}_{l=0}^{\infty}$ with associated value sequence $\{V_i^l\}_{l=0}^{\infty}$. Then starting with $0 \leq V_i^0 \leq Z_i^0$ one has*

$$0 \leq V_i^l \leq Z_i^l. \tag{121}$$

Proof. The value function sequences for each agent i are given by (84) and (85), the arbitrary stabilizing control sequence for each agent i can be written as $M_i^l = (L_i^l + (M_i^l - L_i^l))$. The arbitrary value sequence $\{Z_i^l\}_{l=0}^{\infty}$ for each agent i is given as

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &\equiv F_i(Z_i^l, M_i^l) \\ &= Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l}) + \frac{1}{2}(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + \overline{Z}_i^l(R_{ij}, u_{-i}^l)) \\ &\quad + \frac{1}{2}(L_i^l + (M_i^l - L_i^l))^T R_{ii}(L_i^l + (M_i^l - L_i^l)) \end{aligned}$$

where $\overline{V}_i^l(R_{ij}, u_{-i}^l) = \overline{Z}_i^l(R_{ij}, u_{-i}^l) = \sum_{j\in N_i} u_j^{lT} R_{ij} u_j^l$.

Rearranging this equation yields

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &\equiv F_i(Z_i^l, L_i^l) + \frac{1}{2}(M_i^l - L_i^l)^T R_{ii}(M_i^l - L_i^l) \\ &\quad + (Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l}) + M_i^{lT} R_{ii} L_i^l) - (Z_i^l(\varepsilon_{i(k+1)}|_{L_i^l}) + L_i^{lT} R_{ii} L_i^l) \end{aligned} \tag{122}$$

where,

$$F_i(Z_i^l, L_i^l) = \frac{1}{2}(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \overline{Z}_i^l(R_{ij}, u_{-i}^l)) + Z_i^l(\varepsilon_{i(k+1)}|_{L_i^l}).$$

By induction,

$$\frac{1}{2}(M_i^l - L_i^l)^T R_{ii}(M_i^l - L_i^l) + (Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l}) + M_i^{lT} R_{ii} L_i^l) - (Z_i^l(\varepsilon_{i(k+1)}|_{L_i^l}) + L_i^{lT} R_{ii} L_i^l) > 0. \quad (123)$$

Then (122) and the induction result (123) yield,

$$Z_i^{l+1}(\varepsilon_{ik}) = F_i(Z_i^l, M_i^l) \geq \hat{Z}_i^{l+1} = F_i(Z_i^l, L_i^l). \quad (124)$$

Similarly,

$$V_i^{l+1}(\varepsilon_{ik}) = F_i(V_i^l, M_i^l) \geq V_i^{l+1}(\varepsilon_{ik}) = F_i(V_i^l, L_i^l). \quad (125)$$

Using the initial sequence $0 \leq V_i^0 \leq Z_i^0$, which by applying induction becomes $0 \leq V_i^l \leq Z_i^l$. Inequality (124) gives the lower bound on the arbitrary value sequence $Z_i^{l+1}(\varepsilon_{ik})$ such that

$$0 \leq V_i^{l+1}(\varepsilon_{ik}) = F_i(V_i^l, M_i^l) \leq \hat{Z}_i^{l+1} = F_i(Z_i^l, L_i^l) \leq Z_i^{l+1}(\varepsilon_{ik}) = F_i(Z_i^l, M_i^l). \quad (126)$$

Then,

$$0 \leq V_i^l \leq \hat{Z}_i^l \leq Z_i^l. \quad (127)$$

Eq. (127) yields (121). ■

Lemma 7. *Let the neighbors of agent i have fixed policies u_j. Define the sequence of control policies generated by Algorithm 1 for each agent i as $\{L_i^l\}_{l=0}^{\infty}$ with associated value sequence $\{V_i^l\}_{l=0}^{\infty}$. Then, there exists an upper bound $\overline{U}$ such that*

$$0 \leq V_i^l \leq \overline{U}. \quad (128)$$

Proof. M_i is stabilizing control policy for each agent i. Using (98) and the sequence $V_i^0 = Z_i^0 = 0$ yields

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) - Z_i^l(\varepsilon_{ik}) &= F_i(Z_i^l, M_i^l) - F_i(Z_i^{l-1}, M_i^l) \\ &= Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l}) + \frac{1}{2}(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + M_i^{lT} R_{ii} M_i^l \\ &\quad + \overline{Z}_i^l(R_{ij}, u_{-i}^l)) - Z_i^{l-1}(\varepsilon_{i(k+1)}|_{M_i^l}) \\ &\quad - \frac{1}{2}(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + M_i^{lT} R_{ii} M_i^l + \overline{Z}_i^{l-1}(R_{ij}, u_{-i}^{l-1})) \end{aligned} \quad (129)$$

where $\overline{Z}_i^l(R_{ij}, u_{-i}^l) = \overline{Z}_i^{l-1}(R_{ij}, u_{-i}^{l-1})$.

Rearranging (129) yields,

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) - Z_i^l(\varepsilon_{ik}) &= Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l}) - Z_i^{l-1}(\varepsilon_{i(k+1)}|_{M_i^l}) \\ &= Z_i^{l-1}(\varepsilon_{i(k+2)}|_{M_i^l}) - Z_i^{l-2}(\varepsilon_{i(k+2)}|_{M_i^l}) \\ &\vdots \\ &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l}) - Z_i^0(\varepsilon_{i(k+l)}|_{M_i^l}) \end{aligned} \quad (130)$$

with $Z_i^0(\varepsilon_{i(k+l)}|_{M_i^l}) = 0$ Eq. (130) yields,

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l}) + Z_i^l(\varepsilon_{ik}) \\ &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l}) + Z_i^1(\varepsilon_{i(k+l-1)}|_{M_i^l}) + Z_i^{l-1}(\varepsilon_{ik}) \\ &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l}) + Z_i^1(\varepsilon_{i(k+l-1)}|_{M_i^l}) + .. + Z_i^1(\varepsilon_{ik}). \end{aligned} \quad (131)$$

Then (131) is rearranged so that

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &= \sum_{n=0}^{l} Z_i^1(\varepsilon_{i(k+n)}|_{M_i^l}) \\ &= \sum_{n=0}^{l} \frac{1}{2}(\varepsilon_{i(k+n)}^T|_{M_i^l} Q_{ii}\varepsilon_{i(k+n)}|_{M_i^l} \\ &\quad + M_i^{lT} R_{ii} M_i^l + \overline{Z}_i^l(R_{ij}, u_{-i}^l)). \end{aligned} \quad (132)$$

Since the used policies are stabilizable policies, then

$$Z_i^{l+1}(\varepsilon_{ik}) \leq \sum_{n=0}^{\infty} Z_i^1(\varepsilon_{i(k+n)}|_{M_i^l}) = \overline{\overline{U}}. \quad (133)$$

This inequality satisfies (128). ■

The next two technical lemmas are required for the proof of Theorem 7.

Lemma 8. *Given arbitrary stabilizing control policies $\{M_i^l\}_{l=0}^{\infty}$ and $\{M_j^l\}_{l=0}^{\infty}$ for agent i and its neighbor j, let the associated value sequence be $\{Z_i^l\}_{l=0}^{\infty}, Z_i^0 \geq 0$. Define the sequences of control policies generated by Algorithm 1 for agent i and its neighbor j as $\{L_i^l\}_{l=0}^{\infty}$ and $\{L_j^l\}_{l=0}^{\infty}$ respectively, with associated value sequence $\{V_i^l\}_{l=0}^{\infty}$. Suppose that $\overline{\sigma}(R_{jj}^{-1}R_{ij})$ is small. Then, starting with $0 \leq V_i^0 \leq Z_i^0$, one has*

$$0 \leq V_i^l \leq Z_i^l. \quad (134)$$

Proof. The value function sequences for each agent i are given by (97) and (98), the arbitrary stabilizing control sequences for each agent i and its neighbor j can be written as $M_i^l = (L_i^l + (M_i^l - L_i^l))$ and $M_j^l = (L_j^l + (M_j^l - L_j^l))$ respectively. The arbitrary value sequence $\{Z_i^l\}_{l=0}^{\infty}$ for each agent i is given as

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &\equiv F_i(Z_i^l, L_i^l, L_j^l) + \frac{1}{2}(M_i^l - L_i^l)^T R_{ii}(M_i^l - L_i^l) \\ &\quad + \frac{1}{2}\sum_{j\in N_i}(M_j^l - L_j^l)^T R_{ij}(M_j^l - L_j^l) \\ &\quad + \left(Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l, M_j^l}) + M_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} M_j^{lT} R_{ij} L_j^l\right) \\ &\quad - \left(Z_i^l(\varepsilon_{i(k+1)}|_{L_i^l, L_j^l}) + L_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} L_j^{lT} R_{ij} L_j^l\right) \end{aligned} \quad (135)$$

where,

$$\begin{aligned} F_i(Z_i^l, L_i^l, L_j^l) &= Z_i^l(\varepsilon_{i(k+1)}|_{L_i^l, L_j^l}) \\ &\quad + \frac{1}{2}\left(\varepsilon_{ik}^T Q_{ii}\varepsilon_{ik} + L_i^{lT} R_{ii} L_i^l + \sum_{j\in N_i} L_j^{lT} R_{ij} L_j^l\right). \end{aligned}$$

Eq. (135) yields

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &\equiv F_i(Z_i^l, L_i^l, L_j^l) + \frac{1}{2}\sum_{j\in N_i}(M_j^l - L_j^l)^T R_{ij}(M_j^l - L_j^l) \\ &\quad + \frac{1}{2}(M_i^l - L_i^l)^T R_{ii}(M_i^l - L_i^l) + \sum_{j\in N_i}(M_j^l - L_j^l)^T R_{ij} L_j^l \\ &\quad + (Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l, M_j^l}) + M_i^{lT} R_{ii} L_i^l) \\ &\quad - (Z_i^l(\varepsilon_{i(k+1)}|_{L_i^l, L_j^l}) + L_i^{lT} R_{ii} L_i^l). \end{aligned} \quad (136)$$

Since the following inequality is true,

$$(Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l,M_j^l}) + M_i^{lT}R_{ii}L_i^l) - (Z_i^l(\varepsilon_{i(k+1)}|_{L_i^l,L_j^l}) + L_i^{lT}R_{ii}L_i^l) + \frac{1}{2}(M_i^l - L_i^l)^T R_{ii}(M_i^l - L_i^l) > 0 \quad (137)$$

and $\frac{1}{2}\sum_{j\in N_i}(M_j^l - L_j^l)^T R_{ij}(M_j^l - L_j^l) + \sum_{j\in N_i}(M_j^l - L_j^l)^T R_{ij}L_j^l > 0$ by using the condition

$$\frac{1}{2}(M_j^l - L_j^l)^T R_{ij}(M_j^l - L_j^l) > (L_j^l - M_j^l)^T R_{ij}L_j^l, \quad \forall j. \quad (138)$$

Considering the optimal policy given by

$$L_j^l = (d_j + g_j)R_{jj}^{-1}B_j^T\nabla V_j(\varepsilon_{j(k+1)})^l.$$

The inequality (138) becomes

$$\sum_{j\in N_i}\frac{1}{2}\underline{\sigma}(R_{ij})\left\|\Delta E_{jk}\right\| > \sum_{j\in N_i}(g_j + d_j)\overline{\sigma}(R_{jj}^{-1}R_{ij})\left\|\nabla\tilde{V}_i^l(\varepsilon_{i(k+1)})\right\|\left\|B_{jk}\right\|$$

where $\left\|\Delta E_{jk}\right\| = \left\|L_j^l - M_j^l\right\|$.

By considering (137) and under the assumption that $\overline{\sigma}(R_{jj}^{-1}R_{ij})$ is small then (135) yields

$$Z_i^{l+1}(\varepsilon_{ik}) = F_i(Z_i^l, M_i^l, M_j^l) \geq \hat{Z}_i^{l+1} = F_i(Z_i^l, L_i^l, L_j^l). \quad (139)$$

Similarly we have,

$$V_i^{l+1}(\varepsilon_{ik}) = F_i(V_i^l, M_i^l, M_j^l) \geq V_i^{l+1}(\varepsilon_{ik}) = F_i(V_i^l, L_i^l, L_j^l). \quad (140)$$

By setting the initial sequence $0 \leq V_i^0 \leq Z_i^0$, and after applying induction we have $0 \leq V_i^l \leq Z_i^l$. The inequality (139) gives a lower bound on the arbitrary value sequence $Z_i^{l+1}(\varepsilon_{ik})$ such that

$$0 \leq V_i^{l+1}(\varepsilon_{ik}) = F_i(V_i^l, M_i^l) \leq \hat{Z}_i^{l+1} = F_i(Z_i^l, L_i^l) \leq Z_i^{l+1}(\varepsilon_{ik}) = F_i(Z_i^l M_i^l). \quad (141)$$

Then finally we have that,

$$0 \leq V_i^l \leq \hat{Z}_i^l \leq Z_i^l. \quad (142)$$

Eq. (142) yields (134). ■

Lemma 9. *Define the sequences of control policies generated by Algorithm 1 for agent i and its neighbor j as $\{L_i^l\}_{l=0}^{\infty}$ and $\{L_j^l\}_{l=0}^{\infty}$ respectively, with associated value sequences $\{V_i^l\}_{l=0}^{\infty}$. Suppose that $\overline{\sigma}(R_{jj}^{-1}R_{ij})$ is small. Then there exists a finite upper bound $\overline{\overline{U}}$ such that*

$$0 \leq V_i^l \leq \overline{\overline{U}}. \quad (143)$$

Proof. The policies M_i and M_j are stabilizing control policies for each agent i and its neighbor j. Using (98) and the sequence $V_i^0 = Z_i^0 = 0$ we have

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) - Z_i^l(\varepsilon_{ik}) &= Z_i^l(\varepsilon_{i(k+1)}|_{M_i^l,M_j^l}) - Z_i^{l-1}(\varepsilon_{i(k+1)}|_{M_i^l,M_j^l}) \\ &= Z_i^{l-1}(\varepsilon_{i(k+2)}|_{M_i^l,M_j^l}) - Z_i^{l-2}(\varepsilon_{i(k+2)}|_{M_i^l,M_j^l}) \\ &\vdots \\ &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l,M_j^l}) - Z_i^0(\varepsilon_{i(k+l)}|_{M_i^l,M_j^l}) \end{aligned} \quad (144)$$

with $Z_i^0(\varepsilon_{i(k+l)}|_{M_i^l,M_j^l}) = 0$.

By doing some manipulations in (144) we have

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l,M_j^l}) + Z_i^l(\varepsilon_{ik}) \\ &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l,M_j^l}) + Z_i^1(\varepsilon_{i(k+l-1)}|_{M_i^l,M_j^l}) + Z_i^{l-1}(\varepsilon_{ik}) \\ &= Z_i^1(\varepsilon_{i(k+l)}|_{M_i^l,M_j^l}) + Z_i^1(\varepsilon_{i(k+l-1)}|_{M_i^l,M_j^l}) + .. + Z_i^1(\varepsilon_{ik}). \end{aligned} \quad (145)$$

Finally (145) can be written as

$$\begin{aligned} Z_i^{l+1}(\varepsilon_{ik}) &= \sum_{n=0}^{l} Z_i^1(\varepsilon_{i(k+n)}|_{M_i^l,M_j^l}) \\ &= \sum_{n=0}^{l}\frac{1}{2}\left(\varepsilon_{i(k+n)}^T|_{M_i^l,M_j^l}Q_{ii}\varepsilon_{i(k+n)}|_{M_i^l,M_j^l} + M_i^{lT}R_{ii}M_i^l + \sum_{j\in N_i}M_j^{lT}R_{ij}M_j^l\right) \end{aligned} \quad (146)$$

and since the used policies are stabilizable we have

$$Z_i^{l+1}(\varepsilon_{ik}) \leq \sum_{n=0}^{\infty} Z_i^1(\varepsilon_{i(k+n)}|_{M_i^l,M_j^l}) = \overline{\overline{U}}. \quad (147)$$

This inequality satisfies (143) and hence the result follows. ■

References

Al-Tamimi, A., Lewis, F. L., & Abu-Khalaf, M. (2008). Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, *38*(4), 943–949.

Başar, T., & Olsder, G. J. (1999). *Classics in applied mathematics, Dynamic noncooperative game theory* (2nd ed.). Philadelphia: SIAM.

Beard, R. W., & Stepanyan, V. (2003). Synchronization of information in distributed multiple vehicle coordination control. In *Proc. of the IEEE conference on decision and control, Maui, HI* (pp. 2029–2034).

Bellman, R. 1957. *Dynamic programming*. Princeton.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. MA: Athena Scientific.

Bowling, M. 2004. Convergence and no-regret in multiagent learning. In *NIPS*.

Bryson, A. E. (1996). Optimal control-1950 to 1985. *IEEE Control Systems*, *16*(3), 26–33.

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multi-agent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, *38*(2), 156–172.

Chang, C. F., Hsu, J. Y., & Fu, L. (2007). Dynamic game-based analysis for a hybrid multi-agent robotic system. *International Journal of Electronic Business Management (IJEBM)*.

Dierks, T., & Jagannathan, S. (2010). Optimal control of affine nonlinear continuous-time systems using an online Hamilton–Jacobi–Isaacs formulation1. In *Proc. IEEE conf. decision and control, Atlanta* (pp. 3048–3053).

Fax, J., & Murray, R. (2004). Information flow and cooperative control of vehicle formations. *IEEE Transactions on Automatic Control*, *49*(9), 1465–1476.

Freiling, G., Jank, G., & Abou-Kandil, H. (2002). On global existence of solutions to coupled matrix Riccati equations in closed loop Nash games. *IEEE Transactions on Automatic Control*, *41*(2), 264–269.

Gajic, Z., & Li, T.-Y. (1988). Simulation results for two new algorithms for solving coupled algebraic Riccati equations. In *Third int. symp. on differential games, Sophia, Antipolis, France*.

Gonzalez, O. (1996). Time integration and discrete Hamiltonian systems. *Journal of Nonlinear Science*, *6*(5), 449–467.

Gopalakrishnan, R., Marden, J. R., & Wierman, A. (2011). An architectural view of game theoretic control. *ACM SIGMETRICS Performance Evaluation Review*, *38*(3), 31–36.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press.

Hong, Y., Hu, J., & Gao, L. (2006). Tracking control for multi-agent consensus with an active leader and variable topology. *Automatica*, *42*(7), 1177–1182.

Jadbabaie, A., Lin, J., & Morse, A. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, *48*(6), 988–1001.

Johnson, M., Hiramatsu, T., Fitz-Coy, N., & Dixon, W.E. (2010). Asymptotic stackelberg optimal control design for an uncertain Euler Lagrange system. In *IEEE conference on decision and control* (pp. 6686–6691).

Kakade, S., Kearns, M., Langford, J., & Ortiz, L. (2003). Correlated equilibria in graphical games. In *Proc. 4th ACM conference on electronic commerce* (pp. 42–47).

Khoo, S., Xie, L., & Man, Z. (2009). Robust finite-time consensus tracking algorithm for multirobot systems. *IEEE Transactions on Mechatronics*, *14*, 219–228.

Lall, S., & West, M. (2006). Discrete variational Hamiltonian mechanics. *Journal of Physics A: Mathematical and General*, *39*(19), 5509–5519.

Lancaster, P., & Rodman, L. (1995). *Algebraic Riccati equations*. Oxford: Clarendon Press.
Lewis, F. (1992). *Applied optimal control and estimation: digital design and implementation*. New Jersey: Prentice-Hall.
Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). *Optimal control* (3rd ed.). John Wiley.
Li, Z., Duan, Z., Chen, G., & Huang, L. (2010). Consensus of multi-agent systems and synchronization of complex networks: Aunified viewpoint. *IEEE Transactions on Circuits and Systems. I. Regular Papers*, *57*(1), 213–224.
Li, X., Wang, X., & Chen, G. (2004). Pinning a complex dynamical network to its equilibrium. *IEEE Transactions on Circuits and Systems. I. Regular Papers*, *51*(10), 2074–2087.
Littman, M. L. (2001). Value-function reinforcement learning in Markov games. *Journal of Cognitive Systems Research*, *2*(1), 55–66.
Marden, J. R., Arslan, G., & Shamma, J. S. (2009). Joint strategy fictitious play with inertia for potential games. *IEEE Transactions on Automatic Control*, *54*(2), 208–220.
Marsden, J. E., & West, M. (2001). Discrete mechanics and variational integrators. *Acta Numerica*, *10*(5), 357–514.
McLachlan, R. I., Quispel, G. R. W., & Robidoux, N. (1999). Geometric integration using discrete gradients. *Philosophical Transactions of the Royal Society A*, *357*(1754), 1021–1045.
Mu, S., Chu, T., & Wang, L. (2005). Coordinated collective motion in a motile particle group with a leader. *Physica A*, *351*, 211–226.
Nourian, M., Caines, P. E., & Malhame, R. P. (2011). A solution to the initial mean consensus problem via a continuum based mean field control approach. In *Conference on decision and control, CDC* (pp. 5708–5713).
Olfati-Saber, R., Fax, J., & Murray, R. (2007). Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, *95*(1), 215–233.
Olfati-Saber, R., & Murray, R. M. (2004). Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, *49*(9), 1520–1533.
Qu, Z. (2009). *Cooperative control of dynamical systems: applications to autonomous vehicles*. New York: Springer-Verlag.
Ren, W., & Beard, R. (2005). Consensus seeking in multi-agent systems under dynamically changing interaction topologies. *IEEE Transactions on Automatic Control*, *50*(5), 655–661.
Ren, W., & Beard, R. W. (2008). *Distributed consensus in multi-vehicle cooperative control*. Berlin: Springer.
Ren, W., Beard, R., & Atkins, E. (2005). A survey of consensus problems in multi-agent coordination. In *Proc. Amer. control conf.* (pp. 1859–1864).
Ren, W., Moore, K., & Chen, Y. (2007). High-order and model reference consensus algorithms in cooperative control of multivehicle systems. *Journal of Dynamic Systems, Measurement and Control*, *129*(5), 678–688.
Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.
Sen, S., & Weiss, G. (1999). *Learning in multi-agent systems, in multi-agent systems: a modern approach to distributed artificial intelligence*. (pp. 259–298). Cambridge, MA: MIT Press.
Shinohara, R. (2010). Coalition proof equilibria in a voluntary participation game. *International Journal of Game Theory*, *39*(4), 603–615.
Shoham, Y., & Leyton-Brown, K. (2009). *Multi-agent systems: algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.
Suris, Y. B. (2003). *The problem of integrable discretization: Hamiltonian approach*. Basel: Birkhäuser.
Suris, Y. B. (2004). Discrete Lagrangian models. In *Lecture notes in physics*: *Vol. 644. Discrete integrable systems* (pp. 111–184). Springer.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning—an introduction*. Cambridge, MA: MIT Press.
Tembine, H. (2011). Hybrid mean field game dynamics in large populations. In *American control conference, ACC, San Francisco, California, US*.
Tsitsiklis, J. (1984). *Problems in decentralized decision making and computation* (Ph.D. dissertation), Cambridge, MA: Dept. Elect. Eng. and Comput. Sci., MIT.
Vamvoudakis, K. G., & Lewis, F. L. (2010). Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, *46*(5), 878–888.
Vamvoudakis, K. G., & Lewis, F. L. (2011). Multi-player non-zero sum games: online adaptive learning solution of coupled Hamilton–Jacobi equations. *Automatica*, *47*(8), 1556–1569.
Vamvoudakis, K. G., Lewis, F. L., & Hudas, G. R. (2012). Multi-agent differential graphical games: online adaptive learning solution for synchronization with optimality. *Automatica*, *48*(8), 1598–1611.
Vrabie, D., Pastravanu, O., Lewis, F. L., & Abu-Khalaf, M. (2009). Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, *45*(2), 477–484.
Vrancx, P., Verbeeck, K., & Nowe, A. (2008). Decentralized learning in Markov games. *IEEE Transactions on Systems, Man and Cybernetics*, *38*(4), 976–981.
Wang, X., & Chen, G. (2002). Pinning control of scale-free dynamical networks. *Physica A*, *310*(3–4), 521–531.
Wang, D., Liu, D., Wei, Q., Zhao, D., & Jin, N. (2012). Optimal control of unknown non-affine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica*, *48*(8), 1825–1832.
Werbos, P. J. (1974). *Beyond regression: new tools for prediction and analysis in the behavior sciences*. (Ph.D. Thesis)
Werbos, P.J. (1989). Neural-networks for control and system identification. In *IEEE Proc. CDC89 (1)* (pp. 260–265).
Werbos, P. J. (1992). 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.
Young, H. P. (1998). *Individual strategy and social structure*. Princeton, NJ: Princeton Univ. Press.
Zhang, H., Luo, Y., & Liu, D. (2009). Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Transactions on Neural Networks*, *20*(9), 1490–1503.

Mohammed I. Abouheaf was born in Smanoud, Egypt. He received his B.Sc. and M.Sc. degrees in Electronics and Communication Engineering, Mansoura College of Engineering, Mansoura, Egypt (2000, 2006). He worked as an assistant lecturer with the Air Defense College, Alexandria, Egypt (2001–2002). He worked as a Planning Engineer for the maintenance department, Suez Oil Company (SUCO), South Sinai, Egypt, (2002–2004). He worked as an Assistant Lecturer with the Electrical Engineering Department, Aswan College of Energy Engineering, Aswan, Egypt (2004–2008). He received his Ph.D. degree in Electrical Engineering, University of Texas at Arlington (UTA), Arlington, Texas, USA (2012). He worked as a Postdoctoral Fellow with the University of Texas at Arlington Research Institute (UTARI), Fort Worth, Texas, USA (2012–2013). He worked as Adjunct Faculty with the Electrical Engineering Department, University of Texas at Arlington (UTA), Arlington, Texas, USA (2012–2013). He was a member of the Advanced Controls and Sensor Group (ACS) and the Energy Systems Research Center (ESRC), University of Texas at Arlington, Arlington, Texas, USA (2008–2012). Currently, he is Assistant Professor with the Systems Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. His research interests include optimal control, adaptive control, reinforcement learning, fuzzy systems, game theory, microgrids, and economic dispatch.

Frank L. Lewis, Member, National Academy of Inventors. Fellow IEEE, Fellow IFAC, Fellow UK Institute of Measurement & Control, PE Texas, UK Chartered Engineer. UTA Distinguished Scholar Professor, UTA Distinguished Teaching Professor, and Moncrief–O'Donnell Chair at the University of Texas at Arlington Research Institute. Qian Ren Thousand Talents Professor, Northeastern University, Shenyang, China. He obtained the Bachelor's Degree in Physics/EE and the MSEE at Rice University, the M.S. in Aeronautical Engineering from Univ. W. Florida, and the Ph.D. at Ga. Tech. He works in feedback control, intelligent systems, cooperative control systems, and nonlinear systems. He is author of 6 US patents, numerous journal special issues, journal papers, and 14 books, including *Optimal Control, Aircraft Control, Optimal Estimation, and Robot Manipulator Control* which are used as university textbooks worldwide. He received the Fulbright Research Award, NSF Research Initiation Grant, ASEE *Terman Award*, Int. Neural Network Soc. *Gabor Award*, UK Inst Measurement & Control *Honeywell Field Engineering Medal*, and IEEE Computational Intelligence Society *Neural Networks Pioneer Award*; received Outstanding Service Award from Dallas IEEE Section, selected as Engineer of the year by Ft. Worth IEEE Section. He was listed in Ft. Worth Business Press Top 200 Leaders in Manufacturing. Texas Regents Outstanding Teaching Award 2013. He is Distinguished Visiting Professor at Nanjing University of Science & Technology and Project 111 Professor at Northeastern University in Shenyang, China, and Founding Member of the Board of Governors of the Mediterranean Control Association.

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 at Arlington with Professor Frank L. Lewis as his advisor and he received his M.S. and Ph.D. in Electrical Engineering in 2008 and 2011 respectively. From May 2011 to January 2012, he was working as an Adjunct Professor and Faculty Research Associate at the University of Texas at Arlington and at the Automation and Robotics Research Institute. He currently serves as a Project Research Scientist at the Center for Control, Dynamical systems and Computation (CCDC) at the University of California, Santa Barbara. His research interests include approximate dynamic programming, game theory, neural network feedback control, and optimal control. Recently, his research has focused on network security and multi-agent optimization. Dr. Vamvoudakis is the recipient of several international awards including 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 member of Tau Beta Pi, Eta Kappa Nu and Golden Key honor societies and is listed in Who is Who in the World, Who is Who in Science and Engineering, and Who is Who in America. He has also served on various international program committees and has organized special sessions for several international conferences. He currently is a member of the Technical Committee on Intelligent Control of the IEEE Control Systems Society (TCIC), a member of the Technical Committee on Adaptive Dynamic Programming and Reinforcement

Learning of the IEEE Computational Intelligence Society (ADPRLTC), an Associate Editor on the IEEE Control Systems Society Conference Editorial Board, a registered Electrical/Computer engineer (PE) and a member of the Technical Chamber of Greece.

Sofie Haesaert was born in Leuven, Belgium. She obtained the Bachelor's Degree in Mechanical Engineering in 2010 with distinction and the Master of Science Degree in Systems and Control in 2012 with distinction in combination with the Masters Honors Program at Delft University of Technology. She is a recipient of a Ph.D. grant from the Dutch Institute of Systems and Control (DISC) in the scope of the Graduate Program of The Netherlands Organization for Scientific Research (NOW). Currently, she is working towards the Ph.D. degree in the Control Systems group of the Electrical Engineering Department at Eindhoven University of Technology. Her current research interests include identification for control, Bayesian identification, machine learning, formal verification and correct-by-design controller synthesis.

Robert Babuska received his M.Sc. degree (with honors) in Electrical Engineering from the Czech Technical University Prague (1990), and his Ph.D. degree (cum laude) in Control from the Delft University of Technology (1997). He has had faculty appointments at the Technical Cybernetics Department of the Czech Technical University Prague (1991–1993) and at the Faculty of Information Technology and Systems of the Delft University of Technology (1993–2003). Currently he is professor of Intelligent Control and Robotics at the Delft Center for Systems and Control, Delft University of Technology. He is the founder and scientific director of the TU Delft Robotics Institute. His research interests include adaptive and learning control, nonlinear identification and state-estimation, dynamic multi-agent systems, predictive control, neural networks, fuzzy systems and machine learning. Robert Babuska has co-authored over 390 publications, including three research monographs, three edited books and 85 journal papers. He has been serving as Associate Editor of several archived journals including Automatica, Engineering Applications of Artificial Intelligence, and IEEE Transactions on Fuzzy Systems. He served the chairman of IFAC (International Federation of Automatic Control) technical committee on Cognition and Control.