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

Citation for published version (APA):

DOI:
10.1016/j.automatica.2014.10.047

Document status and date:
Published: 01/01/2014

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 09. Mar. 2020
Multi-agent discrete-time graphical games and reinforcement learning solutions

Mohammed I. Abouheaf\textsuperscript{a,1}, Frank L. Lewis\textsuperscript{b}, Kyriakos G. Vamvoudakis\textsuperscript{c}, Sofie Haesaert\textsuperscript{d}, Robert Babuska\textsuperscript{e}

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

A R T I C L E   I N F O

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

A B S T R A C T

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.

© 2014 Elsevier Ltd. All rights reserved.

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 Fox 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...
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, Pasravanu, 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 Vanx, 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 $G = (V, E)$ is defined as the pair $G = (V, E)$ 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 $L = E - D$.

A directed path from node $v_0$ to node $v_i$ is defined as a sequence of edges $v_0, v_1, \ldots, v_l$ such that $(v_j, v_{j+1}) \in \mathcal{E}, \ i \in \{0, 1, \ldots, l - 1\}$. 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 $G = (V, \mathcal{E})$ having $N$ agents, each with local dynamics given by

$$x_i(k + 1) = Ax_i(k) + Bu_i(k)$$

(1)

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

$$x_0(k) \in \mathbb{R}^n$$

given by

$$x_0(k + 1) = Ax_0(k).$$

(2)

The local neighborhood tracking error (Khoo, Xie, & Man, 2009)

$$e_i(k) \in \mathbb{R}^n$$

for each node $i$ as

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

(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

$$e(k) = -((L + G) \otimes I_n)\text{vec}x(k) + ((L + G) \otimes I_n)x_0$$

(4)

where the global node state vector is $x = \begin{bmatrix} x_1^T & x_2^T & \ldots & x_n^T \end{bmatrix}^T$ and the tracking error vector is $e = \begin{bmatrix} e_1^T & e_2^T & \ldots & e_n^T \end{bmatrix}^T$. Eq. (4) can be written as

$$e(k) = -((L + G) \otimes I_n)\text{vec}e(k)$$

(5)

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

$$\eta(k) = (x(k) - x_0(k)) \in \mathbb{R}^{nN}$$

(6)

with $x_0 = [x_0, L = 1 \otimes I_n$ and $1$ the $N$-vector of ones. $G = \text{diag}(g_i) \in \mathbb{R}^{nN \times nN}$ 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 $\sigma_{\text{max}}(.)$, $\sigma_{\text{min}}(.)$.

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

$$\|\eta(k)\| \leq \|e(k)\| / \sigma_{\text{min}}(L + G).$$

(7)

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

$x(k) = Lx_0(k).$  

(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

$$e_{i(k+1)} = f_i(e_{i, k}, u_{ik}, u_{ijk})$$

$$= A_i e_i - (d_i + g_i)B_i u_{ik} + \sum_{j \in N_i} e_{ij} B_{ij} u_{ijk}.$$  

(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 $e_i(k)$, 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

$$u_{-i} = \{u_j | j \in N_i\}$$

(10)

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

$$u_i = \{u_j | j \in N, i \neq j\}.$$  

(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

$$J_i([e_{ik}, u_{ik}, u_{-ik}, k \geq 0]) = \sum_{k=0}^{\infty} U_i(e_{ik}, u_{ik}, u_{-ik})$$

$$= \frac{1}{2} \sum_{k=0}^{\infty} \left( e_{ik}^T Q_k e_{ik} + u_{ik}^T R_k u_{ik} + \sum_{j \in N_i} u_{ijk}^T S_{ijk} u_{ijk} \right)$$

(12)

where $Q_k > 0 \in \mathbb{R}^{n_i \times n_i}$, $R_k > 0 \in \mathbb{R}^{m \times m}$, and $S_{ijk} > 0 \in \mathbb{R}^{m \times m_{ij}}$, are symmetric time-invariant weighting matrices.

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

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

$$V_i(e_{ik}) = \sum_{k=0}^{\infty} U_i(e_{ik}, u_{ik}, u_{-ik}).$$  

(13)

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

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

...
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

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

(14)

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

$$J_i(\delta(k), u_i(k)) = \frac{1}{2} \sum_{k=0}^{\infty} \delta(k)^T Q_i \delta(k) + \sum_{j \in N} u_i^T(k) R_{ij} u_j(k)$$

(15)

where $Q_i > 0 \in \mathbb{R}^{N_i \times N_i}$, $R_{ij} > 0 \in \mathbb{R}^{n_i \times n_j}$.

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

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

(16)

with $\tilde{B} = \text{diag}[B_1, \ldots, B_i, \ldots, B_N]$ and $u(k) = [u_1^T, u_2^T, \ldots, u_N^T]^T$ 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

$$\delta(k+1) = (I_N \otimes A) \delta(k) + \tilde{B}^1 \cdots \tilde{B}^N u(k).$$

(17)

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

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

(18)

where $[l_{ij}]$ is a matrix whose $ij$th block is $l_{ij} B_j$. Eqs. (17) and (18) are same if one defines the block column matrix $\tilde{B}'$ as $\tilde{B}' = \{l_{ij} B_i, l_{ij} B_j, \ldots, l_{ij} B_N\}^T$. 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 $\tilde{Q}_{ij}, \tilde{R}_{ij}$. 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

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

(19)

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

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

$$V^*_i(\varepsilon_{ik}) = \min_{\varepsilon_{ik}} \{V_i(\varepsilon_{ik})\} = \min_{\varepsilon_{ik}} \left\{ \sum_{l \in N} U_l(\varepsilon_{lk}, u_{lk}, u_{\ldots l}) \right\}$$

(20)

where $U_l = \{u_{lk}\}^\infty_{l=0}, \forall l \in N$. According to the Bellman optimality principle

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

(21)

Consequently, the optimal control policy for each agent $i$ is

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

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

$$V_i^*(\varepsilon_{ik}) = V_i^*(\varepsilon_{i(k+1)}) + \frac{1}{2} \left(\epsilon_{ik}^T Q_{i} \epsilon_{ik} + \sum_{j \in N} (d_j + g_j)^2 \nabla V_j^*(\varepsilon_{j(k+1)})^T B_j R_j^{-1} B_j^T \nabla V_j^*(\varepsilon_{j(k+1)}) \right)$$

(23)

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

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

$$H_i(\varepsilon_{ik}, \lambda_{i(k+1)}, u_{ik}, u_{\ldots ik})$$

$$= \lambda_{i(k+1)}^T A \varepsilon_{ik} - (d_i + g_i) B_i u_{ik} + \sum_{j \in N} c_{ij} B_j u_{jk}$$

$$+ \frac{1}{2} \left(\epsilon_{ik}^T Q_{i} \epsilon_{ik} + \sum_{j \in N} u_{ik}^T R_{ij} u_{ik} + \sum_{j \in N} u_{jk}^T R_{ij} u_{jk} \right)$$

(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

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

(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

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

(26)

or

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

(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

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

(28)

and its gradient as

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

(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

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

(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

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

(31)

Then one should define augmented value function as follows

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

(32)

with a corresponding Hamiltonian given by,

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

(33)

Eqs. (32) and (33) yield

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

Substituting these equations into (19) yields

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

(34)

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

$$
\nabla V_i(\varepsilon_{i(k+1)}) - \left[ (\partial \lambda_{i(k+1)})/\partial \varepsilon_{i(k+1)} \right] \varepsilon_{i(k+1)} + \lambda_{i(k+1)}
$$

$$
+ \left[ (\partial \lambda_{i(k+1)})/\partial \varepsilon_{i(k+1)} \right] \varepsilon_{i(k+1)} + \lambda_{i(k+1)} = 0.
$$

Rearranging this equation yields

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

(35)

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

Eq. (35) yields $\lambda_{i(k+1)}$. Substituting this into (34) yields (30). \hfill \Box

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

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

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

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

$$
+ \frac{1}{2} \left( \varepsilon_{ik}^T Q_{ik} + u_{ik}^T R_{ik} u_{ik} + \sum_{j \in N_i} u_{ikj}^T R_{ikj} u_{ikj} \right) = 0
$$

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

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

(37)

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

b. Let $(A, B)$ be reachable. Let $0 < V_i^*(\varepsilon_{ik}) \in \mathbb{C}^2$, $i$ satisfies (23). Then $V_i^*(\varepsilon_{ik})$ satisfies (37).
4.2. Interactive Nash equilibrium

The hessians of the Hamiltonian (24) and Bellman equations (19) with respect to all control policies are positive definite values since

\[ V''_{ik}(e_{ik+1}) = \nabla^2 V''_{ik}(e_{ik+1}) - \nabla^2 V''_{ik}(e_{ik+1}) = 0. \]

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

\[ \nabla^3 V''_{ik}(e_{ik+1}) = 0, \quad \forall k, \quad p = 0, 1, \ldots, n = 1. \]

The reachability matrix

\[ \tilde{U}_i = [B_i, AB_i, A^2 B_i, \ldots, A^{n-1} B_i] \]

under the hypothesis has full rank. Therefore, since \( V''_{ik}(0) = 0 \) and \( V''_{ik}(0) = 0 \) then,

\[ V''_{ik}(e_{ik}) = V''_{ik}(e_{ik}), \quad \forall k \]

from which the result follows. ■

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

**Lemma 2.** Let \( V''_{ik}(e_{ik}) \) satisfy the Bellman optimality equation (23) or equivalently the coupled HJ (37) then,

\[ \nabla V''_{ik}(e_{ik+1})^T e_{ik+1} = \Delta V''_{ik}(e_{ik}) \]

or

\[ (\partial V''_{ik}(e_{ik+1})/\partial e_{ik+1})^T e_{ik+1} = V''_{ik}(e_{ik+1}) - V''_{ik}(e_{ik}) \].

**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 noncooperative 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 \( u_j = \{ u_j \mid j \in N, j \neq i \} \), 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 \( i \in N \)

\[ J''_{i}(u^{*}_i, u^{*}_j) \leq J''(u_i, u_j). \]  

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 (Shimohara, 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

\[ J''(u^{*}_i, u_{-i}) \leq J''(u_i, u_{-i}). \]  

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 \( u^{*}_i \) such that

\[ J''(u^{*}_i, u_{-i}') \neq J''(u'_i, u_{-i}'). \]

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 \( u_k = K, \lambda_k + K, \lambda_k + \mu_k \) for a single node \( z \) and optimal policies given by \( u_k = K, \lambda_k + K, \lambda_k, \forall z \neq i \). That is, node \( z \) has an extra control input \( \mu_k \). Then we have

\[ e_{ik+1} = Ae_{ik} - \sum_{j \in N} e_{jk} B_j u_j + \sum_{j \in N} e_{jk} B_j u_j - e_{zk} B_j u_{zk}, \]

and

\[ K, = -(d_i + g_i) R_{ik}^{-1} B_i A^T Q_i. \]

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

\[ e_{ik}^{T} \lambda_{ik}^{T} = \lambda_{ik}^{T} \lambda_{ik}^{T} + B_{ik} \]

where \( \lambda_{ik} \) is the length of the shortest directed path from node \( z \) to node \( i \). Denote the nodes along this path by \( z = z_0, z_1, \ldots, z_{\lambda_{ik}} \). Denote element \( (i, z) \) of \((L) \) by \( \lambda_{ik} \). Then the \( n \times m \) block element in block row \( i \) and block column \( z \) of matrix \( \lambda_{ik}^{T} \lambda_{ik}^{T} \) is equal to

\[ \lambda_{ik}^{T} \lambda_{ik}^{T} = \sum_{z_{\lambda_{ik}}} \lambda_{iz_{\lambda_{ik}}} \lambda_{iz_{\lambda_{ik}}} \]

where \( \lambda_{iz_{\lambda_{ik}}} \) is the position of the block element in the block matrix.

**Assumption 1.** a. The matrix \( B_{ik}^{T} A^T Q_i B_{ik} \) has full row rank for \( M > j > 0 \).

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 $m_i = m_j$, $i, j, B_i = B_j$, $i, j$, with $B_i$ of full column rank, $Q_i > 0$, $i, j$, and $A$ is nonsingular.

**Lemma 3.** Let Assumption 1 hold and let the control policies are given by (27). Then the $i$th performance index of closed-loop system (52) depends on the input $v_i$ 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:

$$
[A^{\Delta T - 1} B]^{iz} [A^{\Delta T} B]^{iz} [A^{\Delta T + 1} B]^{iz} \ldots
$$

with

$$
[A^{\Delta T - 1} B]^{iz} = \sum_{\Delta T - 1, \ldots, \Delta T} e_i \xi_{\Delta T - 1} \xi_{\Delta T - 2} (d_{\Delta T - 1} + g_{\Delta T - 1}) \times R_{\Delta T - 1} B_{\Delta T - 1} A^{-T} Q_{\Delta T - 1} B_{\Delta T - 2} R_{\Delta T - 1}^{-1} \times (d_z + g_z) B_z A^{-T} Q_z B_z.
$$

(54)

If there exists a path from node $z$ to $i$, then the first block matrix has rank $m_{\Delta T - 1}$, therefore at least $m_{\Delta T - 1}$, modes of the nodes of node $i$ are reachable. Thus the state will be a function of $v_{ik}$, because $Q_i$ is positive definite. Therefore, the $i$th performance index will depend on $v_{ik}$.

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 $j_i(u_i, u_{-i}) \leq j_i(u_i, u_{-i})$, $\forall i$. Hence $u_i = u_i^*$, $\forall u_i \in u_i$ and $j_i(u_i^*, u_{-i}^*) = j_i(u_i^*, u_{-i}^*)$, $\forall i$. According to (12) $j_i(u_i^*, u_{-i}^*) = j_i(u_i^*, u_{-i}^*)$, $\forall u_i \not\in N_i$ so that $j_i(u_i^*, u_{-i}^*) = j_i(u_i^*, u_{-i}^*)$, $\forall i$ 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 $i$, $z \in N_i$. ■

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

**Assumption 2.** The graph is strongly connected and $g_j$ 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 $0 < V_i^*(\epsilon_{ik}) \in C^2$ 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 $I_i^*(\epsilon_{ii}, u_i^*, u_{-i}^*) = V_i^*(\epsilon_{ii})$.

c. All agents are in Nash equilibrium.

**Proof.** a. $V_i^*(\epsilon_{ik+1})$ satisfies the Bellman optimality equation such that

$$
V_i^*(\epsilon_{ik+1}) - V_i^*(\epsilon_{ik}) = -U_i^*(\epsilon_{ik}, u_{ik}, u_{-ik}^*) < 0.
$$

(55)

Therefore, $V_i^*(\epsilon_{ik})$ 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

$$
H_i(\epsilon_{ik}, \nabla V_i^*(\epsilon_{ik+1}), u_{ik}, u_{-ik}) = \nabla V_i^*(\epsilon_{ik+1})^T \epsilon_{ik+1} + \sum_{j \neq i} U_i(\epsilon_{ik}, u_{ik}, u_{-ik}).
$$

(56)

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

$$
J_i(\epsilon_{ik}, u_{ik}, u_{-ik}) = V_i^*(\epsilon_i(\infty)) + \sum_{k=1} U_i(\epsilon_{ik}, u_{ik}, u_{ik+1}).
$$

(57)

Rearranging this equation yields,

$$
J_i(\epsilon_{ik}, u_{ik}, u_{-ik}) = V_i^*(\epsilon_i(\infty)) + \sum_{k=1} U_i(\epsilon_{ik}, u_{ik}, u_{-ik})
$$

$$
- U_i^*(\epsilon_{ik}, u_{ik}^*, u_{-ik}^*).
$$

(58)

The Hamiltonian for arbitrary control inputs is given by

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

(59)

The Hamiltonian for optimal control inputs is given by

$$
H_i(\epsilon_{ik}, \nabla V_i^*(\epsilon_{ik+1}), u_{ik}^*, u_{-ik}^*) = \nabla V_i^*(\epsilon_{ik+1})^T \epsilon_{ik+1} + U_i^*(\epsilon_{ik}, u_{ik}^*, u_{-ik}^*) = 0.
$$

(60)

It is further noted that,

$$
- \nabla V_i^*(\epsilon_{ik+1})^T \epsilon_{ik+1} + U_i^*(\epsilon_{ik}, u_{ik}^*, u_{-ik}^*)
$$

$$
u_{ik}^* R_{ik} (u_{ik} - u_{ik}^*) - \sum_{j \neq i} \epsilon_j \nabla V_i^*(\epsilon_{ik+1}) B_j (u_{ik} - u_{ik}^*).
$$

(61)

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

$$
U_i(\epsilon_{ik}, u_{ik}, u_{-ik}) = -U_i^*(\epsilon_{ik}, u_{ik}^*, u_{-ik}^*) = \frac{1}{2} (u_{ik} - u_{ik}^*)^T R_{ik} (u_{ik} - u_{ik}^*) + u_{ik}^* R_{ik} (u_{ik} - u_{ik}^*)
$$

$$
+ \frac{1}{2} \sum_{j \neq i} (u_{ik} - u_{ik}^*)^T R_{ik} (u_{ik} - u_{ik}^*) + \sum_{j \neq i} u_{ik}^* R_{ik} (u_{ik} - u_{ik}^*).
$$

(62)
Using (62) into (58) yields
\[
J_i(\varepsilon_i, u_i, u_{-i}) = V^*_i(\varepsilon_i) + \sum_{k \neq i} \left( \sum_{j \in N} \frac{1}{2} (u_{ik} - u^*_k)^T R_k (u_{ik} - u^*_k) + u^*_k R_k (u_{ik} - u^*_k) + \frac{1}{2} \sum_{j \in N} u^*_j R_j (u_{jk} - u^*_j) \right). 
\]
(63)

Using (63) with the optimal policies (38) yields the optimal performance index \( J^*_i(\varepsilon_i, u^*_i, u^*_{-i}) = V^*_i(\varepsilon_i) \).

Given that the summation of the performance index (58) is positive for arbitrary control policies such that
\[
\sum_{k \neq i} U_i(\varepsilon_i, u_i, u_{-i}) - U^*_i(\varepsilon_i, u^*_i, u^*_{-i}) > 0. 
\]
(65)

Eqs. (63), (64), and (65) yield
\[
J^*_i(\varepsilon_i, u^*_i, u^*_{-i}) \leq J_i(\varepsilon_i, u_i, u_i^*) 
\]
(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 \( \{u_i^1, u_i^2, \ldots, u_i^N\} \) 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 \( u_{-i} = \{ u_j : j \in N \} \) we can define the best response Bellman equation for each agent \( i \) as
\[
V^*_i(\varepsilon_i) = V^*_i(\varepsilon_{i(i+1)}) + \frac{1}{2} \left( \varepsilon^T Q_i \varepsilon_i + \sum_{j \in N} u^*_i R_i u_j + (d_i + g_i)^2 \nabla V^*_i(\varepsilon_{i(i+1)})^T B_i R_i^{-1} B_i^T \nabla V^*_i(\varepsilon_{i(i+1)}) \right) 
\]
(67)

with initial condition given by \( V^*_i(0) = 0 \) and \( u^*_i = u^*_0 \), where \( u^*_0 \) is given by (22) in terms of the solution (67).

Define the best response Hamilton-Jacobi (HJ) equation as
\[
H_i(\varepsilon_i, \nabla V^*_i(\varepsilon_{i(i+1)}), u^*_i, u_{-i}) = \nabla V^*_i(\varepsilon_{i(i+1)})^T \varepsilon_i + \frac{1}{2} \left( \varepsilon^T Q_i \varepsilon_i + u^*_i R_i u_i + \sum_{j \in N} u^*_j R_j u_j \right) = 0 
\]
(68)

with \( V^*_i(0) = 0 \) and \( u^*_i = u^*_0 \) where \( u^*_0 \) 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 \( 0 < V^*_i(\varepsilon_i) \in C^2 \), \( \forall i \) satisfy the best response (DTJH) equation (68) with initial condition given by \( V^*_i(0) = 0 \), and optimal control policy \( u^*_i \) given by (38). Then, \( V^*_i(\varepsilon_i) \) satisfies the best response Bellman equation (67).

b. Let \( (A, B, \varepsilon) \) be reachable. Let \( 0 < V^*_i(\varepsilon_i) \in C^2 \), \( \forall i \) satisfy (67). Then \( V^*_i(\varepsilon_i) \) satisfies (68).

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

b. The best response Hamiltonian with arbitrary smooth value \( V^*_i(\varepsilon_i) \) for arbitrary control policy \( u_i \) is given by,
\[
H_i(\varepsilon_i, \nabla V^*_i(\varepsilon_{i(i+1)}), u_i, u_{-i}) = \nabla V^*_i(\varepsilon_{i(i+1)})^T \varepsilon_{i(i+1)} + u_i \varepsilon_i + u_{-i} \varepsilon_{-i} \]
\[
+ \frac{1}{2} (u_i - u^*_i)^T R_i (u_i - u^*_i). 
\]
(69)

Now, let \( V_i(\varepsilon_i) \in C^2 \), \( \forall i \) satisfy (19) such that
\[
V_i(\varepsilon_i) = U_i(\varepsilon_i, u_i, u_{-i}) + (\nabla V^*_i(\varepsilon_{i(i+1)})^T \varepsilon_{i(i+1)}) + V_i(\varepsilon_{i(i+1)}). 
\]
(70)

Substituting (69) into (70) yields
\[
V_i(\varepsilon_i) = V_i(\varepsilon_{i(i+1)}) - (\nabla V^*_i(\varepsilon_{i(i+1)})^T \varepsilon_{i(i+1)}) + \frac{1}{2} (u_i - u^*_i)^T R_i (u_i - u^*_i). 
\]

By applying Bellman’s optimality principle, yields that \( V^*_i(\varepsilon_i) \) has to satisfy the following equation
\[
V^*_i(\varepsilon_i) = \min_{u_i} [V^*_i(\varepsilon_{i(i+1)}) - (\nabla V^*_i(\varepsilon_{i(i+1)})^T \varepsilon_{i(i+1)}) + \frac{1}{2} (u_i - u^*_i)^T R_i (u_i - u^*_i)]. 
\]
(71)

By applying the stationarity condition \( \partial V_i(\varepsilon_i)/\partial u_i = 0 \) one has to solve the following equation for the control policy \( u^*_i \):
\[
u^*_i - u^*_0 = (g_i + d_i) R_i^{-1} B_i^T (\nabla V^*_i(\varepsilon_{i(i+1)}) - \nabla V^*_i(\varepsilon_{i(i+1)})). 
\]

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

Theorem 5 (Best Response Solution). Given fixed neighboring policies \( u_{-i} = \{ u_j : j \in N \} \), assume there exists an admissible policy \( u_i \). Let \( 0 < V^*_i(\varepsilon_i) \in C^2 \) satisfy the best response (HJ) equation (68), or equivalently the best response Bellman equation (67). Then 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 \( J^*_i(\varepsilon_i, u^*_i, u^*_i, u_{-i}) = V^*_i(\varepsilon_i) \).

c. All agents are in Nash equilibrium.

Proof. a. Suppose that \( V^*_i(\varepsilon_i) \) satisfies (67) such that,
\[
V^*_i(\varepsilon_{i(i+1)}) - V^*_i(\varepsilon_i) = -U_i^T(\varepsilon_i, u^*_i, u_{-i}) < 0. 
\]
(72)

Therefore, \( V^*_i(\varepsilon_i) \) 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, \( \varepsilon_i(\infty) \rightarrow 0 \). Therefore \( V_0^*(\varepsilon_i(\infty)) = 0 \) and the best response performance index for each agent \( i \) is given by
\[
J_i(\varepsilon_i, u_i^* , u_{-i} - u_i) = V_i^*(\varepsilon_i, u_i^*, u_{-i} - u_i) = U_i^*(\varepsilon_i, u_i^*, u_{-i} - u_i).
\]
(73)

The best response Hamiltonian with arbitrary control input \( u_i \) is given by
\[
H_i(\varepsilon_i, \nabla V_i^*(\varepsilon_i(k+1)), u_i, u_{-i} - u_i)
= \nabla V_i^*(\varepsilon_i(k+1))T \nabla V_i^*(\varepsilon_i(k+1)) u_i, u_{-i} - u_i + U_i^*(\varepsilon_i, u_i, u_{-i} - u_i).
\]
(74)

The best response (HJ) equation for agent \( i \) is given by
\[
H_i(\varepsilon_i, \nabla V_i^*(\varepsilon_i(k+1)), u_i, u_{-i} - u_i)
= \nabla V_i^*(\varepsilon_i(k+1))T \nabla V_i^*(\varepsilon_i(k+1)) u_i, u_{-i} - u_i + U_i^*(\varepsilon_i, u_i, u_{-i} - u_i) = 0.
\]
(75)

It is straightforward to see that,
\[
U_i(\varepsilon_i, u_i, u_{-i} - u_i) - U_i^*(\varepsilon_i, u_i^*, u_{-i} - u_i)
= H_i(\varepsilon_i, \nabla V_i^*(\varepsilon_i(k+1)), u_i, u_{-i} - u_i)
- \nabla V_i^*(\varepsilon_i(k+1))T \nabla V_i^*(\varepsilon_i(k+1)) u_i, u_{-i} - u_i + U_i^*(\varepsilon_i, u_i, u_{-i} - u_i) + U_i^0(\varepsilon_i).
\]
(76)

Using (76) into (73) yields
\[
J_i(\varepsilon_i, u_i, u_{-i} - u_i) = V_i^*(\varepsilon_i, u_i^*, u_{-i} - u_i)
+ \sum_{k=1}^{\infty} \left( \frac{1}{2} (u_i - u_i^0)^T R_i (u_i - u_i^0) + u_i^T R_i (u_i - u_i^0) \right).
\]
(77)

The best response performance index (77) with the optimal control policy (38) is given by the unique value \( V_i^*(\varepsilon_i) \).
\[ J_i^*(\varepsilon_i, u_i^*, u_{-i} - u_i) = V_i^*(\varepsilon_i). \]
(78)

c. The summation of the best response performance index (77) is positive for arbitrary control policies \( u_i \) such that
\[
\sum_{k=1}^{\infty} U_i(\varepsilon_i, u_i, u_{-i} - u_i) - U_i^*(\varepsilon_i, u_i^*, u_{-i} - u_i) > 0.
\]
(79)

Eqs. (77), (78), and (79) yield
\[
J_i^*(\varepsilon_i, u_i^*, u_{-i} - u_i) \leq J_i(\varepsilon_i, u_i, u_{-i} - u_i)
\]
(80)

which proves that the policies \( u_i^0 \) vi 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 \( u_i^0 \) and values \( V_i^0(\varepsilon_i) \).

Step 2: Solve for \( V_i^{l+1} \) using Bellman equations
\[
V_i^{l+1}(\varepsilon_i) = U_i(\varepsilon_i, u_i^l, u_{-i} - u_i^l) + V_i^l(\varepsilon_i(k+1)) \tag{81}
\]

where \( l \) is the iteration index.

Step 3: Update the control policies using
\[
u_i^{l+1} = (d_i + g_i)R_i^{-1}B_i^T \nabla V_i(\varepsilon_i(k+1)) \tag{82}
\]

Step 4: On convergence of \( |V_i(\varepsilon_i(k+1)) - V_i(\varepsilon_i)\| \) 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 \( \{u_i^l\}_{l=0}^{\infty} \in R^{n_i} \) for each agent \( i \) at each iteration step in Algorithm 1 is given by
\[
l_i = \arg \min_{u_i} \left( \frac{1}{2} \left( \epsilon_i Q_i \epsilon_i u_i + u_i^T R_i u_i + \sum_{j \in N_i} u_j^T R_j u_j \right) + V_i^l(\varepsilon_i(k+1)) \right). \tag{83}
\]

The notation has been streamlined to simplify the presentation of the proofs, and \( l \) is the iteration index. The associated value function sequence
\[
V_i^{l+1}(\varepsilon_i) \equiv F_i(V_i^l, l_i)
= \frac{1}{2} \epsilon_i^T Q_i \epsilon_i u_i + u_i^T R_i u_i + \sum_{j \in N_i} u_j^T R_j u_j \tag{84}
\]
\[
= \frac{1}{2} \epsilon_i^T Q_i \epsilon_i u_i + u_i^T R_i u_i + \sum_{j \in N_i} u_j^T R_j u_j + \epsilon_i^T V_i(\varepsilon_i(k+1)) \tag{85}
\]
\[
\text{where } V_i(\varepsilon_i, u_{-i} - u_i) = \sum_{j \in N_i} u_j^T R_j u_j \text{ has a fixed value.}
\]

Now consider arbitrary stabilizing control policy sequence \( \{M_i^l\}_{l=0}^{\infty} \in R^{n_i} \). The associated value function sequence
\[
Z_i^{l+1}(\varepsilon_i) \equiv F_i(Z_i, M_i^l)
= \frac{1}{2} \epsilon_i^T Q_i \epsilon_i u_i + M_i^T R_i M_i^l
+ \frac{1}{2} \epsilon_i^T V_i(\varepsilon_i(k+1)) \tag{86}
\]
\[
\text{where } Z_i(\varepsilon_i, u_{-i} - u_i) \text{ has a fixed value. Note that }
\]
\[
\epsilon_i^T V_i(\varepsilon_i, u_{-i} - u_i) \neq \epsilon_i^T V_i(\varepsilon_i, u_{-i} - u_i).
\]

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 \( u_i \). Assume there exists an admissible policy \( u_i \). Let agent \( i \) perform Algorithm 1. Then the solution sequence \( \{V_i^l\}_{l=0}^{\infty} \in R^l \) converges to the best response solution \( V_i^*(\varepsilon_i) \text{ } \forall i \).

**Proof.** According to Lemmas 6, 7 in the Appendix, one has
\[
0 \leq V_i^l \leq Z_i^{l} \leq U_i \tag{86}
\]

Using the value function sequences (84) and (85) we can write,
\[
V_i^{l+1}(\varepsilon_i) = \frac{1}{2} \epsilon_i^T Q_i \epsilon_i u_i + l_i^T R_i l_i + V_i(\varepsilon_i, u_{-i} - u_i) + V_i(\varepsilon_i(k+1)) \tag{87}
\]
\[
Z_i^{l+1}(\varepsilon_i) = \frac{1}{2} \epsilon_i^T Q_i \epsilon_i u_i + l_i^T R_i l_i + V_i(\varepsilon_i, u_{-i} - u_i) + Z_i^{l-1}(\varepsilon_i(k+1)) \tag{88}
\]

Starting with \( l = 0 \), we will show that the hypothesis \( Z_i^{l+1}(\varepsilon_i) \leq V_i^{l+1}(\varepsilon_i) \) holds. By setting \( V_i^l = Z_i^{l} = 0 \) and \( V_i^l(\varepsilon_i) = \frac{1}{2} \epsilon_i^T Q_i \epsilon_i u_i + l_i^T R_i l_i + V_i(\varepsilon_i, u_{-i} - u_i) \geq 0 \), then (87) and (88) for \( \forall l \) yield
\[
V_i^{l}(\varepsilon_i) - Z_i^{0}(\varepsilon_i) = \frac{1}{2} \epsilon_i^T Q_i \epsilon_i u_i + l_i^T R_i l_i + V_i(\varepsilon_i, u_{-i} - u_i) \geq 0 \tag{89}
\]
or
\[ Z^0_i(e_i) \leq V^1_i(e_i). \tag{90} \]

Now we need to assume that \( Z^l_i(e_i) \leq V^{l+1}_i(e_i) \) holds for \( l \) such that \( Z^{l-1}_i(e_i) \leq V^l_i(e_i) \), \( \forall l \). Then at (step l) (87) and (88) yield
\[ V^{l+1}_i(e_i) - Z^l_i(e_i) = V^l_i(e_i) - Z^{l-1}_i(e_i) \geq 0. \tag{91} \]

Therefore the mathematical induction for \( \forall l \) gives
\[ Z^l_i(e_i) \leq V^{l+1}_i(e_i), \quad \forall l. \tag{92} \]

Consequently, for all policies \( \bar{l}_i \), \( \forall l \) the following monotonic sequence holds,
\[ V^{l+1}_i > V^l_i > \cdots \geq 0. \tag{93} \]

Since \( Z^l_i(e_i) \) is a lower bound on \( V^{l+1}_i(e_i) \) and Lemma 7 sets \( \bar{U} \) as an upper bound on \( V^i \), then (86) and (93) yield
\[ 0 \leq V^l_i \leq \bar{Z}^l_i \leq V^{l+1}_i \leq \bar{U}. \tag{94} \]

From (94) the sequence \( V^l_i \) is increasing and it has an upper bound \( \bar{U} \) which means that \( V^* \) converges to the best response solution \( V^*_i \) that satisfies the following equation
\[ V^*_i(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{R}_i^j \bar{L}_i^j + \bar{V}_i(R_i, u_{-i}) \right) + V^*_i(e_i+1) \tag{95} \]

where
\[ \bar{L}_i^* = (d_i + g_i)R_i^{-1}B_i \nabla V^*_i(e_i+1). \tag{96} \]

**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. \( \blacksquare \)

### 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 \( \{l_i^j\}_{j=0}^\infty \in \mathbb{R}^{m_i} \) for every agent \( i \) are given by (83) and the associated value sequences are given by
\[ V^{l+1}_i(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{R}_i^j \bar{L}_i^j + \sum_{j \in N_i} \bar{R}_i^j \bar{L}_i^j \right) + V^*_i(e_i+1). \tag{97} \]

For arbitrary admissible policies for the agents \( \{M_i\}_{i=0}^\infty \in \mathbb{R}^m_i \), their associated value sequences are given by
\[ Z^{l+1}_i(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{M}_i^j \bar{M}_i^j + \sum_{j \in N_i} M_i^j \bar{R}_i^j M_i^j \right) + Z^*_i(e_i+1). \tag{98} \]

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 \( u, v \). Let all agents update their policies simultaneously using Algorithm 1. Suppose that \( \sigma(R_i^{-1} R_j) \) is small. Then the solution sequences \( \{V_i^l\}_{l=0}^\infty \in \mathbb{R}^l \) converge monotonically to the optimal solution \( \bar{V}_i^* \) \( \forall i \) of (23).

**Proof.** According to Lemmas 8, 9 in the Appendix, one has the following inequality for every agent
\[ 0 \leq V^l_i \leq \bar{Z}^l_i \leq \bar{U}. \tag{99} \]

The value function sequences (97) and (98), with control policies given by (83) yield
\[ V^{l+1}_i(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{Q}_i^j \bar{Q}_i^j + \sum_{j \in N_i} \bar{R}_i^j \bar{R}_i^j \right) + V^*_i(e_i+1) \tag{100} \]

and
\[ Z^{l+1}_i(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{Q}_i^j \bar{Q}_i^j + \sum_{j \in N_i} \bar{R}_i^j \bar{R}_i^j \right) + Z^*_i(e_i+1). \tag{101} \]

Starting with \( l = 0 \), we will show that the hypothesis \( Z^l_i(e_i) \leq V^{l+1}_i(e_i) \) holds. By setting \( V^0_i = \bar{Z}^0_i = 0 \) and \( V^1_i(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{Q}_i^j \bar{Q}_i^j + \sum_{j \in N_i} \bar{R}_i^j \bar{R}_i^j \right) \geq 0 \), then (100) and (101) for \( V^0_i, V^1_i \) yield
\[ V^1_i(e_i) - Z^0_i(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{Q}_i^j \bar{Q}_i^j + \sum_{j \in N_i} \bar{R}_i^j \bar{R}_i^j \right) \geq 0 \tag{102} \]

or
\[ Z^0_i(e_i) \leq V^1_i(e_i). \tag{103} \]

Now we need to assume that \( Z^l_i(e_i) \leq V^{l+1}_i(e_i) \) holds for \( l \) such that \( Z^{l-1}_i(e_i) \leq V^l_i(e_i) \), \( \forall l \). Then at (step l) (100) and (101) yield
\[ V^{l+1}_i(e_i) - Z^l_i(e_i) = V^l_i(e_i) - Z^{l-1}_i(e_i) \geq 0. \tag{104} \]

Therefore the mathematical induction for \( \forall l \) yields
\[ Z^l_i(e_i) \leq V^{l+1}_i(e_i), \quad \forall l. \tag{105} \]

Consequently, for all the policies \( l_i^k, l_j^k \), \( \forall l \) the following monotonic sequence holds
\[ V^{l+1}_i > V^l_i > \cdots \geq 0. \tag{106} \]

Since \( Z^l_i(e_i) \) is the lower bound of \( V^{l+1}_i(e_i) \) and Lemma 9 sets \( \bar{U} \) as an upper value of \( V^{l+1}_i \), then (99) and (106) yield
\[ 0 \leq V^l_i \leq \bar{Z}^l_i \leq V^{l+1}_i \leq \bar{U}. \tag{107} \]

From (107) the sequence of \( V^l_i \) is increasing and it has an upper bound \( \bar{U} \), which means that \( V^l_i \) will converge to the optimal solution \( \bar{V}_i^* \) monotonically by satisfying
\[ \bar{V}_i^*(e_i) = \frac{1}{2} \left( \sum_{j \in N_i} \bar{Q}_i^j \bar{Q}_i^j + \sum_{j \in N_i} \bar{R}_i^j \bar{R}_i^j \right) + \bar{V}_i^*(e_i+1) \tag{108} \]

where
\[ \bar{L}_i^* = (d_i + g_i)R_i^{-1}B_i \nabla \bar{V}_i^*(e_i+1). \tag{109} \]

**Remark 2.** The condition \( \sigma(R_i^{-1} R_j) \) 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 \( R_i \) of agent \( i \) by selecting \( R_j \) large enough. This is related to the weakly coupling condition defined in Chapter 6 of Başar and Olsder (1999). \( \blacksquare \)
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 $V_i(ε_t|k+1)$ for each agent $i$ is approximated by a critic network $\hat{V}_i(\cdot|W_{ic})$, and the control policy is approximated by an actor network $\hat{u}_i(\cdot|W_{ia})$ so that

$$\hat{V}_{ic}(W_{ic}) = Z_{ik}^T W_{ic}^T Z_{ik}$$

(110)

$$\hat{u}_i(W_{ia}) = \hat{W}_{ia} Z_{ik}$$

(111)

where $W_{ic} \in \mathbb{R}^{n_i \times n_{i,j}}$ and $W_{ia} \in \mathbb{R}^{n_i \times m_i}$ are the critic and actor weights respectively. $N_{ij}$ is the total number of each agent $i$ and its neighbors. $Z_{ik} \in \mathbb{R}^{n_i}$ is a vector of the state $ε_t$ of agent $i$ and the states of its neighbors.

Let $\tilde{e}_{\text{critic}}(ε_t)$ be the approximation error of the actor network so that

$$\tilde{e}_{\text{critic}}(ε_t) = \hat{u}_i(W_{ia}) - \tilde{u}_i = \hat{W}_{ia} Z_{ik} - \tilde{u}_i$$

(112)

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

$$\tilde{u}_i = (g_i + d_i R_{ii}^{-1} B_i^T \nabla \hat{V}_i(k+1))$$

(113)

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

$$\tilde{u}_i = (g_i + d_i) R_{ii}^{-1} B_i^T \nabla \hat{V}_i(k+1)$$

(114)

where, $O_i = 2 \times [0 \ldots I_n \ldots 0] \in \mathbb{R}^{p_n \times n_{i,j}}$.

The squared approximation error is

$$\text{err}_{\text{actor}} = \frac{1}{2} \tilde{e}_{\text{critic}}(ε_t)^T \tilde{e}_{\text{critic}}(ε_t)$$

(115)

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

$$W_{ia}^{(l+1)T} = W_{ia}^{lT} - \mu_{ia} \left( \nabla V_{ic}(ε_t) - \tilde{u}_i(\hat{W}_{ic}(k)) \right)$$

(116)

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

The value update equation is given by (81). Let $\gamma_{\text{critic}}(ε_t)$ be the target value of the critic network at step $l$ such that

$$\gamma_{\text{critic}}(ε_t) = \frac{1}{2} \left( \tilde{e}_{\text{critic}}^T(ε_t) Q_{ie} \tilde{e}_{\text{critic}}(ε_t) + \tilde{u}_i^T R_{ii} \tilde{u}_i + \sum_{j \in N_i} \tilde{u}_j^T R_{ij} \tilde{u}_j \right) + V_{ic}(ε_t)^T$$

(117)

The critic network approximation error at step $l$ is given by

$$\tilde{e}_{\text{critic}}(ε_t) = \gamma_{\text{critic}}(ε_t) - \hat{V}_i(W_{ic})$$

(118)

Similarly, we define squared approximation error for the critic as

$$\text{err}_{\text{critic}} = \frac{1}{2} \tilde{e}_{\text{critic}}(ε_t)^T \tilde{e}_{\text{critic}}(ε_t) = \frac{1}{2} \left\| \gamma_{\text{critic}}(ε_t) - Z_{ik}^T W_{ic} Z_{ik} \right\|_2^2$$

(119)

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

$$W_{ic}^{(l+1)T} = W_{ic}^{lT} - \mu_{ic} \left( \gamma_{\text{critic}}(ε_t) - Z_{ik}^T W_{ic} Z_{ik} \right)$$

(120)

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

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 $W_{ia}$ randomly and initialize the critic weights $W_{ic}$ with zero values.
2. Do Loop ($l$ iterations)
   2.1 Start with initial state $ε_t$ on the system trajectory.
   2.2 Calculate $\hat{u}_i$ using (111).
   2.3 Calculate the dynamics $ε_t^{l+1}$ using (9).
   2.4 Calculate the value $\hat{V}_i(k+1)$ using (110).
   2.5 Critic update rule
      $$W_{ic}^{(l+1)T} = W_{ic}^{lT} - \mu_{ic} \left( \gamma_{\text{critic}}(ε_t) - Z_{ik}^T W_{ic} Z_{ik} \right)$$
      where $\gamma_{\text{critic}}(ε_t)$ is given by (117).
   2.6 Actor update rule
      $$W_{ia}^{(l+1)T} = W_{ia}^{lT} - \mu_{ia} (\nabla V_{ic}(ε_t) - \tilde{u}_i(\hat{W}_{ic}(k)) \right)$$
      where $\tilde{u}_i = (g_i + d_i) R_{ii}^{-1} B_i^T \nabla \hat{V}_i(k)$
   2.7 On convergence of $\gamma_{\text{critic}}(ε_t) - \hat{V}_i(W_{ic})$ 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 $(\mu_{ia} = 0.1, \mu_{ia} = 0.1, \forall i)$.

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

$$Q_{31} = Q_{23} = Q_{43} = Q_{44} = I_2 \times 2, \quad R_{11} = R_{22} = R_{33} = R_{44} = 1, \quad R_{13} = R_{21} = R_{32} = R_{41} = 0, \quad R_{12} = R_{14} = R_{23} = R_{31} = 1.$$
and Rodman (1995). They are motivated by Al-Tamimi et al. (2008), and Lancaster.

Appendix results show the effectiveness of the proposed structure. 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.

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).
where,

$$F_l(Z_l^i, L_l^i) = \frac{1}{2} \left( e_{ik}^T Q_{ij} e_{ik} + L_{ij}^T R_{ij} L_{ij} + Z_l^i(R_{ij}, u_{ij}) \right) + Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}).$$

By induction,

$$\frac{1}{2} (M_l^i - L_l^i)^T R_{ij} (M_l^i - L_l^i) + (Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik})) + M_l^T R_l L_l^i$$

$$- (Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik})) + L_{ij}^T R_{ij} L_{ij} > 0.$$  \hspace{1cm} (123)

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

$$Z_l^{i+1}(\epsilon_{ik}) = F_l(Z_l^i, M_l^i) \geq Z_l^{i+1} = F_l(Z_l^i, L_l^i).$$ \hspace{1cm} (124)

Similarly,

$$V_l^{i+1}(\epsilon_{ik}) = F_l(V_l^i, M_l^i) \geq V_l^{i+1}(\epsilon_{ik}) = F_l(V_l^i, L_l^i).$$ \hspace{1cm} (125)

Using the initial sequence 0 ≤ V_l^0 ≤ Z_l^0, which by applying induction becomes 0 ≤ V_l^i ≤ Z_l^i. Inequality (124) gives the lower bound on the average value sequence Z_l^{i+1}(\epsilon_{ik}) such that

$$0 \leq V_l^{i+1}(\epsilon_{ik}) = F_l(V_l^i, M_l^i) \leq Z_l^{i+1}(\epsilon_{ik}) = F_l(Z_l^i, L_l^i) \leq Z_l^{i+1}(\epsilon_{ik}) = F_l(Z_l^i, M_l^i).$$ \hspace{1cm} (126)

Then,

$$0 \leq V_l^i \leq \hat{Z}_l^i \leq Z_l^i.$$ \hspace{1cm} (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 \([L_l^i]_0^\infty\) with associated average value sequence \([V_l^i]_0^\infty\). Then, there exists an upper bound \(\overline{U}\) such that

$$0 \leq V_l^i \leq \overline{U}.$$ \hspace{1cm} (128)

**Proof.** M_l is stabilizing control policy for each agent i. Using (98) the sequence V_l^0 = Z_l^0 = 0 yields

$$Z_l^{i+1}(\epsilon_{ik}) - Z_l^i(\epsilon_{ik}) = F_l(Z_l^i, M_l^i) - F_l(Z_l^{i-1}, M_l^i)$$

$$= Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) + \frac{1}{2} (e_{ik}^T Q_{ij} e_{ik} + M_l^T R_l M_l^i)$$

$$+ Z_l^i(R_{ij}, u_{ij}) - Z_l^{i-1}(\epsilon_{ik+l+1} | \epsilon_{ik})$$

$$- \frac{1}{2} (e_{ik}^T Q_{ij} e_{ik} + M_l^T R_l M_l^i)$$ \hspace{1cm} (129)

where Z_l^i(R_{ij}, u_{ij}) = Z_l^{i-1}(R_{ij}, u_{ij}).

Rearranging (129) yields,

$$Z_l^{i+1}(\epsilon_{ik}) - Z_l^i(\epsilon_{ik}) = Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) - Z_l^{i-1}(\epsilon_{ik+l+1} | \epsilon_{ik})$$

$$= Z_l^{i-1}(-\epsilon_{ik+l+2} | \epsilon_{ik}) - Z_l^{i-2}(-\epsilon_{ik+l+2} | \epsilon_{ik})$$

$$\vdots$$

$$= Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) - Z_l^0(\epsilon_{ik+l} | \epsilon_{ik}).$$ \hspace{1cm} (130)

with Z_l^0(\epsilon_{ik+l} | \epsilon_{ik}) = 0 Eq. (130) yields,

$$Z_l^{i+1}(-\epsilon_{ik}) = Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) + Z_l^i(-\epsilon_{ik})$$

$$= Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) + Z_l^i(\epsilon_{ik+l+1} | | \epsilon_{ik}) + Z_l^{i-1}(\epsilon_{ik})$$

$$= Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) + Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) + \ldots + Z_l^i(\epsilon_{ik}).$$ \hspace{1cm} (131)

Then (131) is rearranged so that

$$Z_l^{i+1}(\epsilon_{ik}) = \sum_{n=0}^{l} Z_l^i(\epsilon_{ik+n} | \epsilon_{ik})$$

$$= \sum_{n=0}^{l} \frac{1}{2} (z_{ik+n}^T M_l^i Q_{ij} z_{ik+n} + M_l^T R_l M_l^i)$$

$$+ M_l^T R_l M_l^i + Z_l^i(R_{ij}, u_{ij}).$$ \hspace{1cm} (132)

Since the used policies are stabilizable policies, then

$$Z_l^{i+1}(\epsilon_{ik}) \leq Z_l^i(\epsilon_{ik+n} | \epsilon_{ik}) = \overline{U}.$$ \hspace{1cm} (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 \([M_l^i]_0^\infty\) and \([M_l^j]_0^\infty\) for agent i and its neighbor j, let the associated value sequence be \([Z_l^i]_0^\infty\), \(Z_l^0 \geq 0\). Define the sequence of control policies generated by Algorithm 1 for agent i and its neighbor j as \([L_l^i]_0^\infty\) and \([L_l^j]_0^\infty\) respectively, with associated value sequence \([V_l^i]_0^\infty\). Suppose that \(\sigma(R_j^{-1} R_i)\) is small. Then, starting with 0 ≤ V_l^0 ≤ Z_l^0, one has

0 ≤ V_l^i ≤ \(\overline{U}\). \hspace{1cm} (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 \(M_l^i = (L_l^i + (M_l^i - L_l^i))\) and \(M_l^j = (L_l^j + (M_l^j - L_l^j))\) respectively. The arbitrary average control sequence \([Z_l^i]_0^\infty\) for each agent i is given as

$$Z_l^{i+1}(\epsilon_{ik}) = F_l(Z_l^i, L_l^i, L_l^j) + \frac{1}{2} (M_l^j - L_l^j)^T R_{ij} (M_l^i - L_l^i)$$

$$+ \sum_{j \in N_l} (M_l^j - L_l^j)^T R_{ij} (M_l^i - L_l^i)$$

$$+ Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) + \sum_{j \in N_l} Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik})$$ \hspace{1cm} (135)

where

$$F_l(Z_l^i, L_l^i, L_l^j) = Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik})$$

$$+ \frac{1}{2} \left( e_{ik}^T Q_{ij} e_{ik} + L_l^T R_{ij} L_l^i + \sum_{j \in N_l} L_l^j R_{ij} L_l^j \right).$$ \hspace{1cm} (136)

Eq. (135) yields

$$Z_l^{i+1}(\epsilon_{ik}) = F_l(Z_l^i, L_l^i, L_l^j) + \frac{1}{2} \sum_{j \in N_l} (M_l^j - L_l^j)^T R_{ij} (M_l^i - L_l^i)$$

$$+ \frac{1}{2} (M_l^j - L_l^j)^T R_{ij} (M_l^i - L_l^i) + \sum_{j \in N_l} (M_l^j - L_l^j)^T R_{ij} L_l^j$$

$$+ Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik}) + \sum_{j \in N_l} Z_l^i(\epsilon_{ik+l+1} | \epsilon_{ik})$$ \hspace{1cm} (137)
Since the following inequality is true,
\[
Z_i^j(\varepsilon_{i(k+1)}|\lambda_i^j, M_j^i) + M_i^j R_i(L_j^i) - (Z_i^j(\varepsilon_{i(k+1)}|L_j^i, L_j^i) + L_i^j R_i(L_j^i))
+ \frac{1}{2} (M_i^j - L_i^j)^T R_i (M_i^j - L_i^j) > 0
\]
and \(\frac{1}{2} \sum_{i \in N} (M_i^j - L_i^j)^T R_i (M_i^j - L_i^j) + \sum_{j \in N} (M_j^i - L_j^i)^T R_j L_j^i > 0\) by using the condition
\[
\frac{1}{2} (M_i^j - L_i^j)^T R_i (M_i^j - L_i^j) > (L_i^j - M_i^j)^T R_i L_j^i, \quad \forall j.
\]
Considering the optimal policy given by
\[
L_j^i = (d_i + g_i) R_i^{-1} \nabla V_j(\varepsilon_{i(k+1)}),
\]
The inequality (138) becomes
\[
\sum_{i \in N} \frac{1}{2} \sigma(R_i) \| \Delta E_{i,k} \| > \sum_{i \in N} (g_i + d_i) \sigma(R_j^i) R_i \nabla V_j(\varepsilon_{i(k+1)}) \| B_{i,k} \|
\]
where \(\| \Delta E_{i,k} \| = \| L_j^i - M_j^i \| \).

By considering (137) and under the assumption that \(\sigma(R_j^i) R_i \) is small then (135) yields
\[
Z_i^{j+1}(\varepsilon_{i,k}) = F_i(Z_i^j, M_i^j, M_j^i) \geq Z_i^j(\varepsilon_{i,k}) \geq Z_i^{j+1}(\varepsilon_{i,k})
\]
Similarly we have,
\[
V_i^{j+1}(\varepsilon_{i,k}) = F_i(V_i^j, M_i^j, M_j^i) \geq Z_i^{j+1}(\varepsilon_{i,k}) \geq V_i^{j+1}(\varepsilon_{i,k}) = F_i(V_i^j, L_i^j, L_j^i).
\]
By setting the initial sequence \(0 \leq V_i^0 \leq Z_i^0\), and after applying induction we have \(0 \leq V_i^j \leq Z_i^j\). The inequality (139) gives a lower bound on the arbitrary value sequence \(Z_i^{j+1}(\varepsilon_{i,k})\) such that
\[
0 \leq V_i^{j+1}(\varepsilon_{i,k}) = F_i(V_i^j, M_i^j, M_j^i) \leq Z_i^{j+1}(\varepsilon_{i,k}) = F_i(Z_i^j, M_j^i).
\]
Then finally we have that,
\[
0 \leq V_i^j \leq Z_i^j \leq Z_i^j.
\]
Eq. (142) yields (134).

**Lemma 9.** Define the sequences of control policies generated by Algorithm 1 for agent \(i\) and its neighbor \(j\) as \(V_i^j\) and \(Z_i^j\) respectively, with associated value sequences \(V_i^j\) and \(Z_i^j\). Suppose that \(\sigma(R_j^i) R_i \) is small. Then there exists a finite upper bound \(\overline{T}\) such that
\[
0 \leq V_i^j \leq \overline{T}.
\]

**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
\[
Z_i^{j+1}(\varepsilon_{i,k}) - Z_i^j(\varepsilon_{i,k}) = Z_i^j(\varepsilon_{i(k+1)}|M_j^i, M_i^j) - Z_i^j(\varepsilon_{i(k+1)}|L_j^i, L_j^i)
+ \frac{1}{2} (M_i^j - L_i^j)^T R_i (M_i^j - L_i^j) > 0.
\]
By doing some manipulations in (144) we have
\[
Z_i^{j+1}(\varepsilon_{i,k}) = Z_i^j(\varepsilon_{i(k+1)}|M_j^i, M_i^j) + Z_i^j(\varepsilon_{i,k})
= Z_i^j(\varepsilon_{i(k+1)}|M_j^i, M_i^j) + Z_i^j(\varepsilon_{i(k+1)}|M_j^i, M_i^j) + Z_i^j(\varepsilon_{i,k})
= Z_i^j(\varepsilon_{i(k+1)}|M_j^i, M_i^j) + Z_i^j(\varepsilon_{i(k+1)}|M_j^i, M_i^j) + ... + Z_i^j(\varepsilon_{i,k}).
\]
Finally (145) can be written as
\[
Z_i^{j+1}(\varepsilon_{i,k}) = \sum_{n=0}^{j} Z_i^j(\varepsilon_{i(n+1)}|M_j^i, M_i^j)
= \sum_{n=0}^{j} \left( \varepsilon_{i(n+1)}|M_j^i, M_i^j \right) Q_{\varepsilon_{i(n+1)}}(\varepsilon_{i,n+1})|M_j^i, M_i^j)
+ M_i^j R_i M_j^i + \sum_{n=0}^{j} M_i^j R_i M_j^i
\]
and since the used policies are stabilizable we have
\[
Z_i^{j+1}(\varepsilon_{i,k}) \leq \sum_{n=0}^{\infty} Z_i^j(\varepsilon_{i(n+1)}|M_j^i, M_i^j) = \overline{T}.
\]
This inequality satisfies (143) and hence the result follows.

**References**


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.