Automatica 145 (2022) 110566

Contents lists available at ScienceDirect

Automatica

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

Decentralized inertial best-response with voluntary and limited communication in random communication networks☆

Sarper Aydın, Ceyhun Eksin *

Industrial and Systems Engineering Department, Texas A&M University, College Station, TX 77843, United States of America

ARTICLE INFO

Article history:
Received 3 June 2021
Received in revised form 28 April 2022
Accepted 27 May 2022
Available online 7 September 2022

Keywords:
Multi-agent systems
Game theory
Decentralized algorithms

ABSTRACT

Multiple autonomous agents interact over a random communication network to maximize their individual utility functions which depend on the actions of other agents. We consider decentralized best-response with inertia type algorithms in which agents form beliefs about the future actions of other players based on local information, and take actions that maximize their expected utilities computed with respect to these beliefs or continue to take their previous actions. We show convergence of these types of algorithms to a Nash equilibrium in weakly acyclic games. The result depends on the condition that the belief update and information exchange protocols successfully learn the actions of other players with positive probability in finite time given a static environment, i.e., when other agents' actions do not change. We design a decentralized fictitious play algorithm with voluntary and limited communication (DFP-VL) protocols that satisfy this condition. In the voluntary communication protocol, each agent decides whom to exchange information with by assessing the novelty of its information and the potential effect of its information on others' beliefs. The limited communication protocol entails agents sending only their most frequent action to agents that they decide to communicate with. Numerical experiments on a target assignment game demonstrate that the voluntary and limited communication protocol can more than halve the number of communication attempts while retaining the same convergence rate as DFP in which agents constantly attempt to communicate.

1. Introduction

Multi-agent systems comprise of interlinked decision-makers (agents) aiming to maximize objectives that depend on the actions of other agents in the system. In epidemics, the preemptive measures taken by individuals affect the risks associated with socialization (Bauch & Earn, 2004; Eksin, Shamma, & Weitz, 2017). In a smart grid, multiple devices determine generation and consumption levels to reach a balance while minimizing costs (Kar, Hug, Mohammadi, & Moura, 2014; Zhang, Gatsis, & Giannakis, 2012). In autonomous teams of mobile robots, each robot decides its direction of movement and position to maximize a team objective that depends on the movements and positions of other robots (Aydın & Eksin, 2020a; Kantaros, Guo, & Zavlanos, 2019; Kantaros & Zavlanos, 2016). In all of these settings, agents have to reason about the motives of other agents based on local information. Game theoretic equilibrium concepts, i.e., Nash equilibrium (NE), provide a benchmark for rational reasoning where agents assume other agents are also trying to achieve their individual objectives. However, computation of NE is not feasible given limited computation capabilities and local information. Here, we develop decentralized game-theoretic learning algorithms for settings in which agents do not know the incentives of other agents, and need to communicate over a random network that is subject to failures in order to reason about other agents' actions.

Success of a communication attempt is often subject to random failures in social and technological settings. Moreover, in social settings communication is often voluntary, i.e., agents attempt to communicate upon the need for information exchange. In technological settings, communication is costly to the agents. Because of this, persistent communication attempts are neither realistic in social settings, nor practical in technological settings. Here, we propose decentralized learning algorithms in which agents consider the effect of their information on a potentially receiving agent's beliefs before attempting to communicate.

In the decentralized algorithms considered in this paper, agents use best-response with inertia to determine their next

☆ This work was supported by National Science Foundation (NSF), United States of America CCF-2008855. The material in this paper was partially presented at the 59th IEEE Conference on Decision and Control, December 14–18, 2020, Jeju Island, Republic of Korea. 4th IEEE Conference on Control Technology and Applications (CCTA), August 24–26, 2020, Montreal, Canada. This paper was recommended for publication in revised form by Associate Editor Kostas Margellos under the direction of Editor Christos G. Cassandras.

* Corresponding author.
E-mail addresses: sarper.aydin@tamu.edu (S. Aydın), eksinc@tamu.edu (C. Eksin).

https://doi.org/10.1016/j.automatica.2022.110566
0005-1098/

actions. In best-response with inertia, each agent forms beliefs about the actions of other agents, and takes an action that either maximizes its expected utility computed with respect to its beliefs (best-responds) or continues to take its former action (shows inertia). Whether an agent best-responds or shows inertia in a given step is random. Agents form beliefs about other agents' behavior via information exchanges over a random communication network. The randomness of communication means that agents cannot receive information from every other agent at each step. Given this setting and learning updates, we show convergence of the best-response with inertia behavior to a NE of any weakly acyclic game in finite time almost surely. This result holds as long as the information exchange and belief update protocols ensure that agents learn another agent's action if that agent repeats the same action long enough (Theorem 1).

We call this sufficient condition for convergence (Condition 1) as *prediction under static actions*. Based on this condition, we design voluntary communication protocols in which agents attempt to send information to an agent if they see the need to communicate (Section 4). Agents determine the need to communicate based upon the novelty of their information to the potential receiving agent. For such an assessment, agents form second order beliefs, i.e., reason about the beliefs that other agents have about their behavior. In this voluntary communication protocol, agents assume other agents act according to a stationary distribution determined by the past empirical frequencies of their actions similar to standard fictitious play (FP) (Brown, 1951; Marden, Arslan, & Shamma, 2009b; Young, 2004). Unlike FP, agents cannot keep track of the empirical frequencies of all the agents when the communication is random and voluntary. We show that the voluntary communication protocol satisfies the prediction under static actions condition. (Theorem 2). Via numerical experiments in a target assignment problem, we show that the proposed DFP algorithm with voluntary communication and limited information exchange (DFP-VL) more than halves the number of communication attempts per link (Section 5). In addition DFP-VL retains a convergence rate comparable to the standard DFP with constant communication attempts.

1.1. Related literature

FP converges to rational behavior in various games including potential (Monderer & Shapley, 1996), weakly acyclic (Marden, Arslan, & Shamma, 2009a; Young, 2004), zero-sum (Robinson, 1951), and stochastic games (Sayin, Parise, & Ozdaglar, 2020). Applications of best-response type algorithms and FP include, but are not limited to, traffic routing (Garcia, Reaume, & Smith, 2000), target assignment (Arslan, Marden, & Shamma, 2007), scheduling problems (Al Sheikh, Brun, Hladik, & Prabhu, 2011; Bell, 1996), target tracking (Williams, Goldfain, Drews, Rehg, & Theodorou, 2018) and network formation (Chen & Zhu, 2019) for autonomous teams. In FP, each agent takes an action that maximizes its expected utility (best responds) assuming other agents select their actions randomly from a stationary distribution. Agents assume this stationary distribution is equal to the empirical frequency of past actions. FP is not a decentralized algorithm, since agents need to observe past actions of everyone to be able to keep track of empirical frequencies. Recent works (Arefizadeh & Eksin, 2019; Eksin & Ribeiro, 2017; Swenson, Eksin, Kar, & Ribeiro, 2018; Swenson, Kar, & Xavier, 2015) consider a decentralized form of the fictitious play, in which agents form estimates on empirical frequencies of other agents' actions by averaging the estimates received from their neighbors in a communication network. These algorithms are shown to converge to a NE in weakly acyclic games, i.e., games that admit finite best-response improvement paths. However, they rely on communication with neighbors after every decision-making step. This assumption ignores the randomness of communication attempts, e.g., in wireless communication settings, and the energy costs of communication. Preliminary versions of this paper either consider a specific setting for the voluntary communication protocol design, namely the target assignment game in Aydın and Eksin (2020a), or focus on the convergence of a specific communication protocol for DFP in Aydin and Eksin (2020b). Theorem 1 generalizes prior convergence results in DFP by showing that a generic inertial best-response type behavior will converge to a rational action profile as long as there exists a belief update and information exchange protocol in which agents are able to learn the actions of other agents when the environment is static. We then leverage this result to design an intuitive and novel class of communication efficient belief update and information exchange protocols.

In the voluntary information exchange protocols, the assessment of the novelty of information is based on two metrics: *(i)* novelty of local information and *(ii)* its potential effect on the belief of the receiving agent. Such metrics that are based on second order beliefs (estimating the estimates of the receiving agents) has the potential to improve communication efficiency in other decentralized game-theoretic learning algorithms based on, e.g., gradient descent (Alpcan & Başar, 2005; De Persis & Grammatico, 2019; Koshal, Nedić, & Shanbhag, 2016; Shamma & Arslan, 2005), best-response (Scutari & Pang, 2013), ADMM (Salehisadaghiani, Shi, & Pavel, 2019), and other adaptive strategies (Ye & Hu, 2021). Indeed, communication-censoring protocols that rely on some form of novelty of information metrics proved viable in reducing communication attempts in distributed stochastic gradient descent (Chen, Giannakis, Sun and Yin, 2018; Chen, Sadler and Blum, 2018) and ADMM (Li, Liu, Tian, & Ling, 2019) in the context of optimization. In the class of information exchange protocols considered here, while the novelty of information metric is sender specific, the metric on potential effect of information on other's assessment is receiving agent specific. Thus, agents manage their local information by deciding whom to communicate with. This is a novel communication protocol that relies on agents keeping track of second order beliefs, i.e., forming beliefs on beliefs, in order to estimate the novelty of their information to the candidate receiving agent.

2. Learning Nash equilibria in time-varying random networks

2.1. Notation

We use $\|.\|$ to denote Euclidean norm. We use $|.|$ to denote both the absolute value of a scalar and the cardinality of a set. The notation $\Delta(.)$ defines the space of probability distributions over a given set. $\mathbf{1}_{(.)}$ is an indicator function. We denote the set of N agents with $\mathcal{N} = \{1, 2, \ldots, N\}$. We implement the standard index notation $(i, -i)$ to differentiate agent $i \in \mathcal{N}$ from all of the other agents $-i := \{j \in \mathcal{N} : j \neq i\}$. For any set or an element of a set $\mathcal{X}$, if an index subscript is used, e.g., $\mathcal{X}_i$, $\mathcal{X}_j$, or $\mathcal{X}_{-i}$, it indicates the ownership of the set by the given agent(s).

2.2. Problem statement

We consider a non-cooperative game Γ among a set of N agents. Each agent i chooses an action a_i from a common action set $\mathcal{A}$ with finitely many actions, i.e., $|\mathcal{A}| = K$. We represent each action with an unit vector $\mathbf{e}_k \in \mathbb{R}^K$ so that $\mathcal{A} := \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$. Agent i obtains a bounded payoff $U_i(a) \in \mathbb{R}$ from the joint action profile $a \in \mathcal{A}^N$.

A mixed action (strategy) σ_i is a probability distribution over the action space $\mathcal{A}$. We define the space of probability distributions over the action space as $\Delta(\mathcal{A})$. A strategy profile $\sigma =$

(σ_i, σ_{-i}) is a joint mixed action profile belonging to the set of independent probability distributions over the space of action profiles, i.e., $\Delta^N(\mathcal{A}) := \prod_{i\in\mathcal{N}} \Delta(\mathcal{A})$. Each agent has an utility function $u_i : \Delta^N(\mathcal{A}) \to \mathbb{R}$. Given a strategy profile $\sigma \in \Delta^N(\mathcal{A})$ and payoff values $U_i(a)$ for $a \in \mathcal{A}^N$, the (expected) utility of agent i is given as,

$$u_i(\sigma) := \sum_{a\in\mathcal{A}^N} U_i(a)\sigma(a) \tag{1}$$

where $\sigma(a)$ is the probability of action profile $a \in \mathcal{A}^N$. Note that a pure action $a_i = \mathbf{e}_k$ can be interpreted as a degenerate probability distribution that selects action k with probability 1. In the rest of the paper, we will use $u_i(a) = U_i(a)$ to indicate the (expected) payoff value obtained from the degenerate distribution $a \in \Delta^N(\mathcal{A})$. Then, the game can be defined by the tuple $\Gamma := (\mathcal{N}, \mathcal{A}^N, \{u_i\}_{i\in\mathcal{N}})$.

We assume point-to-point communication between each pair of agents, but communication is subject to random failures. The probability of the existence of a communication link between agent $i \in \mathcal{N}$ and agent $j \in \mathcal{N} \setminus \{i\}$ at time $t \in \mathbb{N}^+$ is distributed with a Bernoulli random variable,

$$c_{ij}(t) \sim \text{Bernoulli}(p_{ij}(t)), \tag{2}$$

where the probability of success is $0 \le p_{ij}(t) \le 1$. We denote the random communication network at time t with $G(t) = (\mathcal{N}, \mathcal{E}(t))$ where $\mathcal{E}(t) := \{(i, j) \in \mathcal{N} \times \mathcal{N} : c_{ij}(t) = 1\}$ is the set of edges realized according to (2). The random communication network $G(t)$ belongs to the space of all possible networks $\mathcal{G}$ given the set of nodes $\mathcal{N}$.

Given the scenario above, agents need to determine their actions given locally available information. In addition to the randomness of communication, the communication resources can be limited or costly further deterring agents from transmitting their information continuously. The objective of this paper to develop a decentralized algorithm that is aware of the potential limits and costs of communication attempts, and that reaches an optimal action profile, defined as a pure NE, given a random communication channel.

Next, we describe the standard FP and then introduce a generalization of FP for random communication networks.

2.3. Fictitious play with inertia

FP is a distributed game-theoretic learning algorithm in which agents repeatedly take actions in discrete time steps that maximize their expected utilities. Agents form estimates of other agents' strategies assuming that other agents are taking actions drawn from a stationary probability distribution determined by the *empirical frequency* of past actions. The empirical frequency $f_i \in \Delta(\mathcal{A})$ of agent i is computed as follows,

$$f_i(t) = (1 - \rho)f_i(t - 1) + \rho a_i(t), \tag{3}$$

where $a_i(t) \in \mathcal{A}$ is the action of agent i at time $t \in \mathbb{N}^+$ and $\rho \in (0, 1)$ is a fading memory constant.

Given the empirical frequencies of other agents $f_{-i}(t) = \{f_j(t) \in \Delta(\mathcal{A})\}_{j\in\mathcal{N}\setminus i}$, agent i's expected utility from taking action a_i is given as,

$$u_i(a_i, f_{-i}(t)) = \sum_{a_{-i}\in\mathcal{A}^{N-1}} u_i(a_i, a_{-i})f_{-i}(t)(a_{-i}), \tag{4}$$

$f_{-i}(t)(a_{-i})$ represents the probability of action profile a_{-i} occurring.

In FP with inertia, each agent best-responds with inertia, i.e., either takes an action that maximizes its expected utility, or follows its previous action with a small probability $\epsilon \in (0, 1)$. Agent i needs to observe the past actions of *all* agents in order to compute the empirical frequencies as per (3) so that it can compute the best response action.

2.4. Decentralized fictitious play (DFP) in random networks

When communication between agents is subject to failures, agents do not have immediate and permanent access to others' actions. One way to address this problem is by agents keeping local estimates of empirical frequencies of past actions. We denote the estimate of agent i on agent j's empirical frequency in (3) with $f_j^i(t) \in \Delta(\mathcal{A})$. As in standard FP with inertia, agents best-respond with inertia, i.e., maximize their expected utility or continue taking the previous action with probability ϵ,

$$a_i(t) = \begin{cases} \text{argmax}_{a_i\in\mathcal{A}}\, u_i(a_i, f_{-i}^i(t-1)) & \text{w.pr. } 1-\epsilon, \\ a_i(t-1) & \text{w.pr. } \epsilon. \end{cases} \tag{5}$$

Note that we replaced the empirical frequencies $f_{-i}(t)$ in (4) with the estimates $f_{-i}^i(t) := \{f_j^i(t)\}_{j\in\mathcal{N}\setminus\{i\}}$ to get the expected utility of agent i from taking action $a_i \in \mathcal{A}$ $(u_i(a_i, f_{-i}^i(t)))$ in (5).

In DFP, agents update their local estimates based on information they receive from their neighbors in the network. We denote the information available to agent i at time t with $H_i(t)$. The information exchange protocol of agent i, denoted with $\Omega_i : H_i(t) \to -i \times I_{-i}^i(t)$, determines the information I_{-i}^i agent i shares with other agents $(-i)$ using its local information $H_i(t)$. If $I_j^i(t) \neq \emptyset$, this implies that agent i would like to communicate with agent j. We define the set of agents that agent i is willing to communicate with as $\mathcal{N}_i^{out}(t) := \{j \in \mathcal{N} : I_j^i(t) \neq \emptyset\} \subseteq -i$. Upon receiving information from its neighbors, $\mathcal{N}_i^{in}(t) := \{j \in -i : i \in \mathcal{N}_j^{out}(t) \cap \{c_{ji}(t) = 1\}\}$, agent i updates its estimates $\{f_j^i(t)\}_{j\in\mathcal{N}}$ according to a function $\Phi_{i,j} : H_i(t) \to \Delta(\mathcal{A})$. Given the exchange protocol of all the agents $\{\Omega_j\}_{j\in\mathcal{N}}$, we define the information available to agent i at time t as $H_i(t) := \{\{a_i(s)\}_{s=1}^{t-1}, \prod_{s=1}^{t-1}\prod_{j\in\mathcal{N}_i^{in}(s)} I_i^j(s)\}$. Equivalently, the information available to agent i at time $t+1$ is a concatenation of the information available at time t with the new information revealed at time t, i.e., $H_i(t+1) = \{H_i(t), a_i(t), \prod_{j\in\mathcal{N}_i^{in}(t)} I_i^j(t)\}$.

For the convergence analysis, we will be agnostic to the specifics of the estimate updates $(\Phi_i := \{\Phi_{i,j},\ j \in -i\})$ and the information exchange process (Ω_i), as long as they ensure that agents are able to learn others' actions under a static action profile. We state the condition formally next.

Condition 1 (*Prediction Under Static Actions*)**.** *There exists a positive probability $\hat{\epsilon} > 0$ and a finite time $\hat{T}$ such that if an agent $j \in \mathcal{N}$ repeats the same action for at least $T > \hat{T}$ times starting from time $t > 0$, i.e., conditioned on the event $\hat{E}_j(t) = \{a_j(s) = \mathbf{e}_k$ for $s = t, t+1, \ldots, t+T-1\}$ and $\mathbf{e}_k \in \mathcal{A}$, agent $i \in \mathcal{N}$ learns agent j's action with positive probability $\hat{\epsilon} > 0$, i.e., $\mathbb{P}(\|a_j(t+T) - f_j^i(t+T)\| \le \xi | H(t), \hat{E}_j(t)) \ge \hat{\epsilon}$ for any $\xi > 0$.*

Any estimate update and information exchange process that satisfies Condition 1 makes sure that agent i's estimate of agent j's action $(f_j^i(t))$ gets close to agent j's action $(a_j(t))$ whenever agent j repeats its action long enough.

We summarize the key steps of the generic DFP in Algorithm 1.

Algorithm 1 Generic DFP for Agent i

1: **Input:** Inertia probability ϵ and fading constant ρ.
2: **Given:** $f_{-i}^i(0)$ and $a(0)$ for all $i \in \mathcal{N}$.
3: **for** $t = 1, 2, \cdots$ **do**
4: *Best-respond*: Use $f^i(t-1) := \{f_j^i(t-1)\}_{j\in-i}$ in (5)
5: *Share information*: Use Ω_i to determine $\mathcal{N}_i^{out}(t)$ and the information to be exchanged
6: *Observe*: Receive information from $\mathcal{N}_i^{in}(t)$
7: *Update estimates*: $f_j^i(t+1) = \Phi_{i,j}(H_i(t+1))$ for $j \in -i$.
8: **end for**

3. DFP convergence for weakly acyclic games

We consider convergence of the DFP in the class of weakly acyclic games. A weakly acyclic game has finite best-response paths, i.e., starting from any action profile there exists a (finite) sequence of best-response updates that reach a pure NE (Milchtaich, 1996; Young, 1993). A best-response path is a sequence of action profiles obtained by a single agent best-responding to the current action profile at each step of the sequence. Next we provide formal definitions for a NE strategy and weakly acyclic games.

Definition 1 (*Nash Equilibrium*)**.** A strategy profile $\sigma^* = (\sigma_i^*, \sigma_{-i}^*) \in \Delta^N(\mathcal{A})$ is a Nash equilibrium of the game Γ if and only if

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(\sigma_i, \sigma_{-i}^*), \quad \text{for all } \sigma_i \in \Delta(\mathcal{A}), i \in \mathcal{N}. \tag{6}$$

A pure NE strategy profile σ^* is a NE that selects an action profile $a^* = (a_i^*, a_{-i}^*) \in \mathcal{A}^N$ with probability 1.

A NE strategy is an (mixed) action profile in which no individual agent can benefit by unilaterally switching to another action.

Definition 2 (*Weakly Acyclic Game*)**.** A game Γ is weakly acyclic if from any joint action profile $a = (a_i, a_{-i}) \in \mathcal{A}^N$, there exists a best-response path ending at a pure NE $a^* = (a_i^*, a_{-i}^*)$.

The existence of a finite best-response path ensures that no agent can improve its utility after some finite number of iterations. Weakly acyclic games are a broad class of games that include potential games and its several variants such as best-response potential and pseudo-potential games.

We consider weakly acyclic games in which optimal action is unique when other agents take NE actions. Specifically, we make the following assumption.

Assumption 1. For any pure NE action profile $a^* \in \mathcal{A}^N$ of the game Γ, it holds that,

$$\{a_i^*\} = \operatorname*{argmax}_{a_i \in \mathcal{A}} u_i(a_i, a_{-i}^*). \tag{7}$$

This assumption makes sure that agents are not indifferent between multiple actions at a pure NE.

3.1. Convergence to a pure Nash equilibrium

We show almost sure convergence of joint action profile $a(t)$ to a pure NE a^* (Theorem 1). The convergence result relies on the fact that the DFP dynamics stays at a pure NE once it reaches that NE (Lemma 2), and there is a positive probability to reach a pure NE from any action profile (Lemma 3). Before showing these lemmas, we show that the best response action of an agent computed with respect to the estimated empirical frequencies $\{f_j^i(t)\}_{j \in \mathcal{N}}$ belongs to the best response action set computed with respect to the actual actions of others $a_{-i}(t)$, whenever the estimates are close enough to $a_{-i}(t)$–see Appendix A.1 for the proof.

Lemma 1. *There exists a small enough $\xi > 0$ such that if $\|a_j(t) - f_j^i(t)\| \leq \xi$ for agents $j \in -i$ at time step t, then* $\operatorname{argmax}_{a_i \in \mathcal{A}} u_i(a_i, f_{-i}^i(t)) \subseteq \operatorname{argmax}_{a_i \in \mathcal{A}} u_i(a_i, a_{-i})$ *for all $i \in \mathcal{N}$ and $a_{-i} \in \mathcal{A}^{N-1}$.*

Next, we prove that when agents play a pure NE and are aware of others' actions, agents are going to stay in this pure NE indefinitely.

Lemma 2 (*Absorption Property*)**.** *Suppose Assumption 1 holds. Assume $\|a_j(t+T) - f_j^i(t+T)\| \leq \xi$ where $\xi > 0$ satisfies Lemma 1 for all pairs of agents $(i, j) \in \mathcal{N} \times \mathcal{N} \setminus \{i\}$ at time step $t+T$. Further, let $a^* \in \mathcal{A}^N$ be a pure NE action profile and $a(t+T) = a^*$. Then, pure NE are the absorbing states such that $a(s) = a^*$ holds for all $s \geq t+T$.*

Proof. By Assumption 1, the set of optimal actions given others' actions $a_{-i}(t+T) = a_{-i}^*$ is a singleton given by $\operatorname{argmax}_{a_i \in \mathcal{A}} u_i(a_i, f_{-i}^i(t+T)) = \operatorname{argmax}_{a_i \in \mathcal{A}} u_i(a_i, a_{-i}^*) = \{a_i^*\}$. Otherwise, by inertia agent i takes the same action a_i^*. Thus, the joint action profile remains at the pure NE, i.e., $a(s) = a^*$, for all $s \geq t+T$. □

In the above proof, we use the fact that a NE is a fixed point of a best-response mapping. By the fixed point definition of NE, and the fact that agents best respond as in (5), only Nash equilibria can be the absorbing joint action profiles.

The next lemma states that there is a positive probability that agents can reach a NE action profile given Condition 1.

Lemma 3 (*Positive Probability of Absorption*)**.** *Suppose Assumption 1 and Condition 1 hold. Let $a(t) \in \mathcal{A}^N$ be the joint action profile at time t. We define the following event starting from time t,*

$$\begin{aligned} E(t) = \{ & a(s) = a^*, \ \|a_j(\bar{s}+T) - f_j^i(\bar{s}+T)\| \leq \xi \\ & \text{for all } (i,j) \in \mathcal{N} \times \mathcal{N} \setminus \{i\} \\ & \text{for all } s \in \{\bar{s}, \bar{s}+1, \ldots, \bar{s}+T-1\} \\ & \text{for some } \bar{s} \in \{t, t+1, \ldots, t+K^N T\}\} \end{aligned}$$

where $a^* \in \mathcal{A}^N$ is a pure NE and $\xi > 0$ is small enough such that the condition in Lemma 1 is satisfied. Then the transition probability $\mathbb{P}(E(t)|H(t))$, is bounded below by some positive constant $\bar{\epsilon}(T) > 0$ for all $t \in \mathbb{N}^+$.

Proof. To show the result, we are going to use the fact that in weakly acyclic games, there exists a finite path from any action profile to a pure NE. The action set of each agent has cardinality K. There exists K^N different joint action profiles in total. Hence, K^N is an upper bound on the length of any finite path to a pure NE.

If $a(t) = a^*$, the pure NE is reached, and there is no improvement step, and $E(t)$ is realized by the fact that beliefs are close enough to the actual actions so that they satisfy Lemma 1. If $a(t) \neq a^*$, we can exploit the fact that there is a finite best-response path to a pure NE. In each improvement step, only one agent improves its utility by changing its action. First, we define the set of best-response actions for agent i against the current action profile of other agents using local empirical frequencies,

$$\begin{aligned} \hat{\mathcal{A}}_i(t) = \{a_i \in \mathcal{A} | & \operatorname*{argmax}_{a_i \in \mathcal{A}} u_i(a_i, f_{-i}^i(t+T+1)) \\ & \subseteq \operatorname*{argmax}_{a_i \in \mathcal{A}} u_i(a_i, a_{-i}(t+T+1))\}. \end{aligned}$$

Then, the following event can be defined accordingly,

$$E_1(t) = \{a_i(t+T+1) \in \hat{\mathcal{A}}_i(t)\}.$$

We aim to find a lower-bound for the probability of the event $E_1(T)$. For this purpose, we define the following events below,

$$E_2(t) = \{\|a_j(t+T) - f_j^i(t+T)\| \leq \xi \text{ for all } j \in -i\}$$

$$E_3(t) = \{a_i(t+T+1) \in \operatorname*{argmax}_{a_i \in \mathcal{A}} u_i(a_i, f_{-i}^i(t+T+1))\}$$

$$E_4(t) = \{a_j(t+T+1) = a_j(t+T) \text{ for all } j \in -i\}$$

For a small enough selected ξ, it holds by Lemma 1,

$$\mathbb{P}(E_1(t)|H(t)) \geq \mathbb{P}(E_2(t), E_3(t), E_4(t)|H(t)). \tag{8}$$

Using the chain rule, RHS of (8) is equal to,

$$\mathbb{P}(E_2(t), E_3(t), E_4(t)|H(t)) = \mathbb{P}(E_3(t), E_4(t)|E_2(t), H(t))\mathbb{P}(E_2(t)|H(t)). \quad (9)$$

By Condition 1 and inertia probability, the second term in (9) can again be conditioned, and be lower bounded,

$$\mathbb{P}(E_2(t)|H(t)) \geq \mathbb{P}(E_2(t)|H(t), \hat{E}_j(t) \text{ for all } j \in -i) \times \mathbb{P}(\hat{E}_j(t) \text{ for all } j \in -i|H(t)) \quad (10)$$

$$\geq \hat{\epsilon}^N \epsilon^{NT}. \quad (11)$$

Then, the remaining part of (9) has positive probability under the condition of prediction under static actions. This probability is equal to the probability that agent i best-responds, while other agents take the same action, i.e.,

$$\mathbb{P}(E_3(t), E_4(t)|E_2(t), H(t)) \geq (1-\epsilon)\epsilon^{(N-1)} \quad (12)$$

Thus, the finite improvement step has positive lower bound,

$$\mathbb{P}(E_1(t)|H(t)) \geq \epsilon_1 := \hat{\epsilon}^N \epsilon^{NT}(1-\epsilon)\epsilon^{(N-1)} > 0. \quad (13)$$

After the completion of an improvement step, the event of another improvement step until a^* is reached has at least the same positive probability. As stated before, total number of improvement paths cannot exceed K^N times. Once $a(\bar{s}) = a^*$, the probability of repeating the same action profile by all agents and learning other's actions is again $\hat{\epsilon}^N \epsilon^{NT}$. Using this, the probability to reach a pure NE is bounded below as $\mathbb{P}(E(t)|H(t)) \geq \bar{\epsilon} = \epsilon_1^{K^N} \hat{\epsilon}^N \epsilon^{NT}$. □

Lemma 3 relies on showing that the DFP dynamics can follow a best-response path with positive probability if agents can obtain accurate enough information on other agents' actions in finite time.

Remark 1. We show that a finite improvement path has a positive probability by considering the worst possible case where all agents have to repeat the same action NT times so that Condition 1 is satisfied for a small enough ξ such that Lemma 1 holds for agent i. Moreover, we assume the longest possible improvement path where all K^N actions have to be visited. Given that our derivation relies on the worst case scenario, the positive probability of absorption does not inform the rate of convergence as we demonstrate in numerical experiments (Section 5). The studies (Arslan & Yüksel, 2016; Gao, Ma, Başar, & Birge, 2021; Marden et al., 2009b; Marden, Young, Arslan, & Shamma, 2009; Swenson et al., 2018) on weakly acyclic games indicate similar lower bounds in terms of the best-response path length on the value of positive probability to reach NE. The numerical results in these studies also corroborate our observation that worst-case-type analysis does not reflect the average convergence rates in practice for different kinds of games including, but not limited to, target assignment and congestion games.

Next, we state the main convergence theorem.

Theorem 1. *Suppose Assumption 1 and Condition 1 hold. Let $\{a(t) = (a_1(t), a_2(t), \ldots, a_N(t))\}_{t\geq 1}$ be a sequence of actions by the DFP Algorithm (Algorithm 1) and random time-varying communication networks $\{G(t)\}_{t\geq 1}$. The action sequence $\{a(t)\}_{t\geq 1}$ converges to a pure NE a^* of the game Γ, almost surely.*

Proof. By Lemma 2, pure Nash equilibria are the absorbing states of the DFP dynamics among all joint action profiles. By Lemma 3, there exists a positive probability to reach a pure NE. Further, Lemmas 2 and 3 together imply that there are no recurrent sets (infinitely visited) of action profiles other than absorbing states (pure NE). This is because the sequence of actions a_t reaches a pure NE with a positive probability (Lemma 3), which means all other actions besides the pure Nash equilibria are transient states. Therefore, in finite time with probability 1, a pure NE is reached and action profile stays the same once reached. Thus, the action sequence $\{a(t)\}_{t\geq 1}$ converges to a pure NE a^* of the game Γ, almost surely. □

The convergence result relies on the idea of absorbing Markov chains in which pure Nash equilibria are the only absorbing states among all joint action profiles (states) and there is a positive probability of reaching a NE action profile starting from any action.

4. Information exchange and belief update protocols for random communication networks

We introduce information exchange $\Omega_i(\cdot)$ and belief update $\Phi_i(\cdot)$ protocols that aim to reduce the number of communication attempts while at the same time guaranteeing that prediction under static actions condition (Condition 1) holds.

4.1. Voluntary communication protocols

We use two metrics, novelty and belief similarity, to determine whether agent i attempts to send information to agent j or not. The novelty metric is the distance between the empirical frequency of agent i and its current action denoted with $h_{ii}(t) := \|a_i(t) - f_i(t)\|$. The belief similarity metric, defined as $h_{ij}(t) := \|f_i(t) - f_i^{j(i)}(t)\|$, is the distance between agent i's empirical frequency $f_i(t)$ and the second order belief of agent i, i.e., agent i's belief on agent j's belief on $f_i(t)$ denoted with $f_i^{j(i)}(t)$. Based on these metrics, agent i decides to communicate its empirical frequency $f_i(t)$ to agent j if the following logical condition is satisfied,

$$\mathbf{1}(\eta_1 \leq h_{ii}(t) \leq \eta_2) \vee \mathbf{1}(h_{ij}(t) \geq \eta_3) \quad (14)$$

where $\eta_2 > \eta_1 \geq 0$ and $\eta_3 \geq 0$, $\mathbf{1}(\cdot)$ is the indicator function, and $\vee$ is the logical OR operator. Condition (14) determines the set of agents agent that i is willing to communicate with at time step t, i.e., $\mathcal{N}_i^{out}(t)$.

The intuition for the condition in (14) is as follows. The novelty metric $h_{ii}(t)$ is likely to be small when agent i takes the same action for several steps indicating that it may have converged on an action. If $h_{ii}(t)$ is large, it means agent i is undecided, taking a different action from its past set of actions. When h_{ii} is neither too small or too large, agent i attempts to communicate with all the other agents. Agent i attempts to send its empirical frequency specifically to agent j, if it believes agent j does not have an accurate estimate of its empirical frequency, i.e., if h_{ij} is large enough. When neither of these conditions holds, that is if agent i's novelty is small or large, and agent $j \in \mathcal{N} \setminus i$ has an accurate belief about agent i's actions, then agent i ceases to communicate.

Given the communication scheme, agent i updates its belief $f_j^i \in \Delta(\mathcal{A})$ about agent j's empirical frequency f_j at each time step as follows,

$$f_j^i(t) = \begin{cases} f_j(t), \text{ if } c_{ji}(t) = 1, \\ f_j^i(t-1), \text{ otherwise.} \end{cases} \quad (15)$$

That is, agent i replaces its estimate on agent j's empirical frequency with the empirical frequency received from agent j upon a successful communication attempt. Otherwise, its estimate remains the same.

In computing the belief similarity $h_{ij}(t)$, agent i has to form and update beliefs about agent j's belief on its own empirical frequency $f_i^j(t)$. This can be done via an acknowledgment scheme

where each time agent i makes a successful communication attempt to agent j, agent j sends back 1-bit acknowledgment signal. We allow the acknowledgment signal to be subject to failures with a Bernoulli variable $b_{ij}(t) \sim \text{Bernoulli}(\beta_{ij}(t))$ with success rate $0 \leq \beta_{ij}(t) \leq 1$. We note that the acknowledgment procedure is executed if and only if agent i receives information from agent j. Thus, we have $\mathbb{P}(b_{ij}(t) = 0 | c_{ji}(t) = 0) = 1$. Otherwise, we have $\mathbb{P}(b_{ij}(t) = 1 | c_{ji}(t) = 1) > \beta_{ij}(t)$. Given the acknowledgment scheme, agent i's second order belief $f_i^{j(i)}(t) \in \Delta(\mathcal{A})$ is updated as follows,

$$f_i^{j(i)}(t) = \begin{cases} f_i(t), \text{ if } b_{ji}(t) = 1, \\ f_i^{j(i)}(t-1), \text{ otherwise.} \end{cases} \tag{16}$$

Upon receiving the acknowledgment, agent i knows that its empirical frequency is transmitted to agent j, and agent j has updated its belief as per (15). In a scenario where $c_{ij}(t) = 1$ and $b_{ji}(t) = 1$, empirical frequencies and estimates align, i.e., $f_i^j(t) = f_i^{j(i)}(t) = f_i(t)$.

Remark 2. In the information exchange and belief update protocols described above, each agent keeps an estimate of the empirical frequencies of all agents $\{f_j^i(t)\}_{j\in\mathcal{N}}$, an $N \times K$ real-valued matrix, and second order beliefs about other agents' estimates about its empirical frequency $\{f_i^{j(i)}(t)\}_{j\in\mathcal{N}}$, an $N \times K$ real-valued matrix. Agent i attempts to send its empirical frequency $f_i(t)$, a real-valued vector of length K, to a subset of agents in $\mathcal{N}$ according to the condition in (14). In prior works that consider DFP (Swenson et al., 2018), each agent shares their estimates of all the other agents, $\{f_j^i(t)\}_{j\in\mathcal{N}}$, an $N \times K$ real-valued matrix, to all of their neighbors at every step.

4.2. Limited information communication

Agents share the maximum value and the index of their empirical frequency, i.e.,

$$\upsilon_i(t) = \max_{k\in\mathcal{K}} f_{ik}^i(t), \tag{17}$$

$$\kappa_i(t) = \operatorname*{argmax}_{k\in\mathcal{K}} f_{ik}^i(t), \tag{18}$$

instead of their empirical frequencies, where $f_{ik}^i \in [0, 1]$ is the frequency of action $k \in \mathcal{K}$ in agent i's past actions. When an agent j successfully sends the maximum value $\upsilon_j(t)$ and its index $\kappa_j(t)$ (18) to agent i, agent i needs to reconstruct a well-defined empirical frequency and update its belief $f_j^i(t)$ accordingly. Upon successful communication of $\upsilon_j(t)$ and $\kappa_j(t)$, the reconstructed belief $f_j^i(t)$ has to satisfy

$$\sum_{k\in\mathcal{K}} f_{jk}^i(t) = 1,\ f_{jk}^i(t) \geq 0,\ f_{j\kappa_i(t)}^i(t) \geq \upsilon_i(t), \tag{19}$$

where f_{jk}^i denotes the kth index. While the first two constraints above define a proper distribution over the space of actions, the third constraint makes sure that the receiving agent uses the information received. There could multiple update rules $\Phi_i(\kappa_j(t), \upsilon_j(t))$ that satisfy the conditions in (19). For instance, one update rule can assume full support on the most frequent action of agent j, i.e., $f_{j\kappa_j(t)}^i(t) = 1$ and $f_{jk}^i(t) = 0$ for $k \in \mathcal{K} \setminus \kappa_j(t)$. Another update rule can assume actions other than the most common are equally likely, i.e., $f_{j\kappa_i(t)}^i(t) = \upsilon_j(t)$ and $f_{jk}^i(t) = (1 - \upsilon_j(t))/(|\mathcal{K}| - 1)$ for $k \in \mathcal{K} \setminus \kappa_j(t)$.

Remark 3. The limited communication protocol described further reduces the information sent per communication attempt to a single real value $\upsilon_i(t)$ and an integer $\kappa_i(t)$.

4.3. Convergence of communication and belief update protocols

We describe the specific steps of the DFP with voluntary and limited communication protocols (DFP-VL) in Algorithm 2. Step 4 corresponds to the *best response* step in Algorithm 1. Steps 5–7 correspond to the information sharing and observation steps in Algorithm 1. Steps 8–9 update the empirical frequency estimates and second order beliefs, respectively.

Algorithm 2 DFP-VL for Agent i

1: **Input:** The parameters $\rho, \epsilon, \eta_1, \eta_2, \eta_3$.
2: **Given:** $f_{-i}^i(0) = \{f_j^i(0)\}_{j\in\mathcal{N}\setminus\{i\}}$, $f_i^{j(i)}(0)$ for all $j \in \mathcal{N} \setminus i$ and $a(0)$ for all $i \in \mathcal{N}$.
3: **for** $t = 1, 2, \cdots$ **do**
4: Agent i takes action $a_i(t)$ using (5).
5: Determine $\mathcal{N}_i^{out}(t)$ by checking (14) for all $j \in \mathcal{N} \setminus \{i\}$.
6: Transmit $\upsilon_i(t)$ and $\kappa_i(t)$ to agent $j \in \mathcal{N}_i^{out}$.
7: Send an acknowledgment signal to $j \in \mathcal{N}_i^{in}(t)$
8: Update $f_j^i(t) = \Phi_{i,j}(\kappa_j(t), \nu_j(t))$ for $j \in \mathcal{N}_i^{in}(t)$.
9: Update $f_i^{j(i)}(t) = f_i(t)$ for agent $\{j \in \mathcal{N}_i^{out} \cap \{j : b_{ji}(t) = 1\}\}$.
10: **end for**

Theorem 2. *Suppose the communication and acknowledgment success probabilities are lower bounded by a positive value, i.e., $p_{ij}(t) > \nu > 0$ and $b_{ji}(t) > \nu > 0$ for all $t \in \mathbb{N}^+$ and $i \in \mathcal{N}, j \in \mathcal{N}$. Let $\{a(t) = (a_1(t), a_2(t), \ldots, a_N(t))\}_{t\geq 1}$ be a sequence of actions generated by the DFP-VL (Algorithm 2). Then, Condition 1 is satisfied for any $\xi > 0$ given small enough $0 \leq \eta_1 < \xi/2$, large enough $\xi/2 < \eta_2$, and small enough $0 \leq \eta_3 \leq \xi/2$ given the repetition of the same action by agent $j \in \mathcal{N}$ as stated in Condition 1.*

Proof. See the Appendix. □

Theorem 2 implies that DFP-VL converges to a pure NE of any weakly acyclic game via Theorem 1.

5. Numerical experiments

We investigate the performance of different communication protocols in terms of convergence rate and cost of communication in the target assignment game.

5.1. Target assignment game

A team of N agents aim to cover N targets with minimum effort. We can represent the problem as a game with the following payoff values for agent i,

$$U_i(a_i, a_{-i}) = \frac{a_i^T \mathbf{1}_{a_{-ik}=0}}{a_i^T d_i}, \tag{20}$$

where $a_i = \mathbf{e}_k \in \mathbb{R}^K$ is an unit vector and $\mathbf{1}_{a_{-ik}=0} \in \{0, 1\}^K$ is a binary vector whose kth index is 1 if none of the other agents $j \in \mathcal{N} \setminus \{i\}$ selects target k, and otherwise the kth index is equal to 0. The distance vector $d_i = [d_{i1}, \ldots, d_{ik}, \ldots, d_{iK}] \in \mathbb{R}_+^K$ measures the distance between agent i and each target k in the 2-dimensional plane, where $d_{ik} = \|\theta_i - \theta_k\|$. Agent i obtains a positive utility that is inversely proportional to the distance of the agent to the selected target if the target is not selected by another agent j. Otherwise, agent i receives zero utility. Given the utility function (20), any joint action that is an one-to-one assignment between agents and targets is a pure NE.

In the numerical experiments, we consider a target assignment problem with $N = 20$ agents and $K = 20$ targets. Positions of agents and targets are randomly generated on the plane. Target positions are generated using polar coordinates with radii and

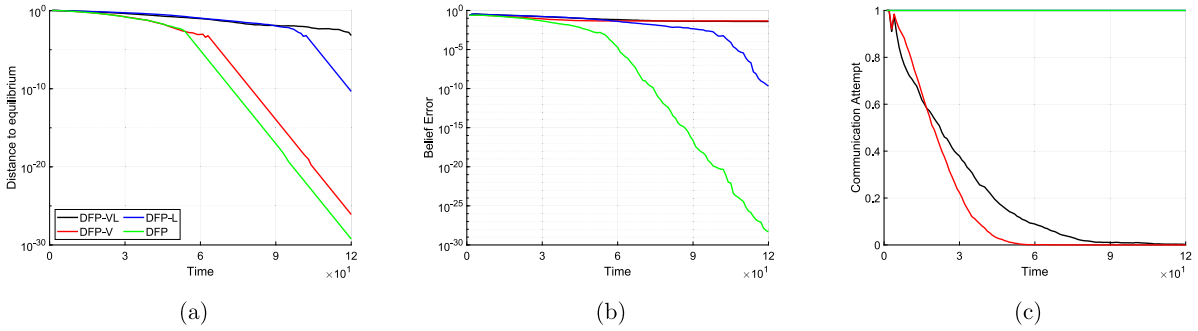

Fig. 1. Convergence results over 100 replications for DFP, DFP-V, DFP-L, and DFP-VL. See Table 1 for parameter values. (a) Convergence of empirical frequencies to pure NE $\frac{1}{N}\sum_{i\in\mathcal{N}}\|f_i(t)-a_i^*\|$ on average. We obtain the nearest pure NE by solving a linear assignment problem. (b) Convergence of beliefs $\frac{1}{N(N-1)}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}\setminus\{i\}}\|f_i(t)-f_i^j(t)\|$. (c) Average attempt per communication link over time.

Table 1
Parameter values of algorithms.

	Parameters			
	DFP-VL	DFP-V	DFP-L	DFP
η_1	0.2	0.2	–	–
η_2	0.6	0.6	–	–
η_3	0.1	0.1	–	–
ϵ	0.3	0.3	0.3	0.3
ρ	0.6	0.6	0.6	0.6
1-bit	Yes	No	Yes	No

angular coordinates uniformly sampled from 15 to 20, and from 0 to 2π, respectively. Similarly, the positions of agents on the 2-dimensional plane are determined by sampling from a normal distribution with mean 0 and standard deviation 1 independently for each dimension. The pairwise distances between agents and targets are computed based on the positions of agents and targets.

The communication and acknowledgment probability for each link are given as $p_{ij}(t)=0.6$ and $\beta_{ij}(t)=0.9$ for all $t\geq 1$ and $i\in\mathcal{N}, j\in\mathcal{N}\setminus\{i\}$. Initial empirical frequencies of agents $f_i(0)$ are set to uniform discrete distribution, i.e., $f_{ik}=1/K$ for $k=\{1,\ldots,K\}$. We run each simulation for $T_f=120$ steps.

5.2. Effects of the communication protocol

We compare the effects of different communication protocols to the standard DFP in which agents attempt to communicate with all the agents after each decision with DFP-V, DFP-L, and DFP-VL (Algorithm 2). DFP-V uses only the voluntary communication protocol (Section 4.1). In DFP-L, agents attempt to communicate at every step but use the limited communication protocol (Section 4.2). The parameter values are given in Table 1.

In all communication protocols, the final action profile a_{T_f} is a pure NE for all 100 realizations. That is, the action profile a time $T_f=120$ is a one-to-one assignment of agents to targets. The empirical frequencies converge to the pure NE the fastest on average in DFP followed by the second fastest DFP-V (Fig. 1(a)). In protocols that use limited communication, the decrease in distance of empirical frequencies to a NE tends to be slower. The average time to reach a final pure NE action profile is the fastest for DFP ($t=21$) and slowest for DFP-VL ($t=46$).

We observe that constant communication in DFP achieves a faster convergence of beliefs, while voluntary and limited communication protocols slow down the convergence in beliefs as shown in Fig. 1(b). Together, Fig. 1(a)–(b) signify that communication protocols increase belief error but preserve convergence to an equilibrium. Indeed, for voluntary communication protocols (DFP-V and DFP-VL), the belief errors stays constant around 10^{-2} as agents cease communication attempts. In contrast, the belief error converges to 0 in DFP and DFP-L as communication attempts continue. For DFP-L, the reduced error in beliefs does not translate to smaller distance of empirical frequencies to a NE compared to DFP-V (compare DFP-L and DFP-V in Fig. 1(a)). The possible reason for this is the difference in actions selected based on the real empirical frequencies and estimates based on 1-bit signals. That is, limited communication reduces total cost by $O(K)$ which leads to a loss of information valuable to convergence rate of empirical frequencies.

DFP-V and DFP-VL start at full usage of links and then cease the communication attempts almost entirely toward the end of the simulation horizon (Fig. 1(c)). DFP-V and DFP-VL utilize 17% and 22% of the communication links on average. Further, note that even though DFP-V uses less communication links on average, the total communication cost for DFP-VL is an $O(K)$ less than DFP-V since DFP-VL sends only 1-bit information. This also implies that DFP-VL has communication cost less than 1 percent of DFP given the number of actions $K=20$. Thus, voluntary and limited communication protocols effectively reduce communication cost, while converging to a pure NE in all cases within $T_f=120$ with the given set of parameters.

5.3. Parameter sensitivity

We assess the performance of DFP-VL under different communication thresholds in Fig. 2. Here we consider smaller fading $\rho=0.4$ and inertia $\epsilon=0.1$ values compared to Fig. 1. We select the set of threshold values starting from the tightest case with $(\eta_1,\eta_2,\eta_3)=(0.4,0.5,0.4)$, and we relax each threshold value by 0.1 until $(\eta_1,\eta_2,\eta_3)=(0.1,0.8,0.1)$. Compared with the baseline case (DFP-VL shown with black line in Fig. 1), we observe that DFP-VL converges faster on average with smaller fading and smaller inertia—observe all the lines in Fig. 2(a) reach below 10^{-5}.

As expected, Fig. 2(a–c) show that as threshold values are relaxed, the time to reach the final pure NE is faster on an average game at the expense of increased communication cost. The average time to reach the final NE action profile is between $t=33$ (green line) and $t=41$ (black line). Average communication attempts are between 14% (black line) and 21% (green line). In all parameter values, communication attempts are reduced at least by 95% after the half time $T_f/2=60$. As the threshold values are relaxed, we see smaller belief error in the final time step $T_f=120$ (Fig. 2(b)).

We note that the final action profile a_{T_f} is a pure NE in all of the replications for each parameter set. Fig. 3 shows an instance of the time evolution of the action profile a_t for each parameter set. Overall, we do not see abrupt changes in the overall performance of DFP-VL with different parameter values. With different

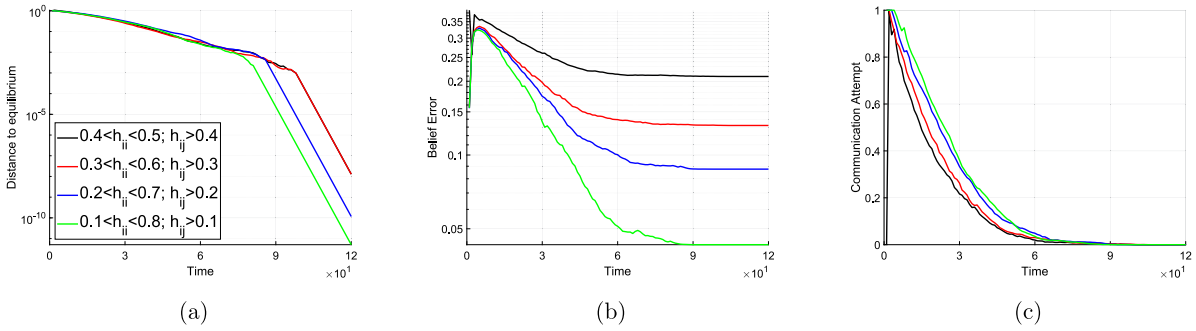

Fig. 2. Convergence results of DFP-VL with different parameter values over 100 replications. Fading rate $\rho = 0.8$ and inertia probability $\epsilon = 0.1$. (a) Convergence of empirical frequencies to pure NE $\frac{1}{N}\sum_{i\in\mathcal{N}} \|f_i(t) - a_i^*\|$ on average. (b) Convergence of Beliefs $\frac{1}{N(N-1)}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}\setminus\{i\}} \|f_i(t) - f_i^j(t)\|$. (c) Average attempt per communication link over time.

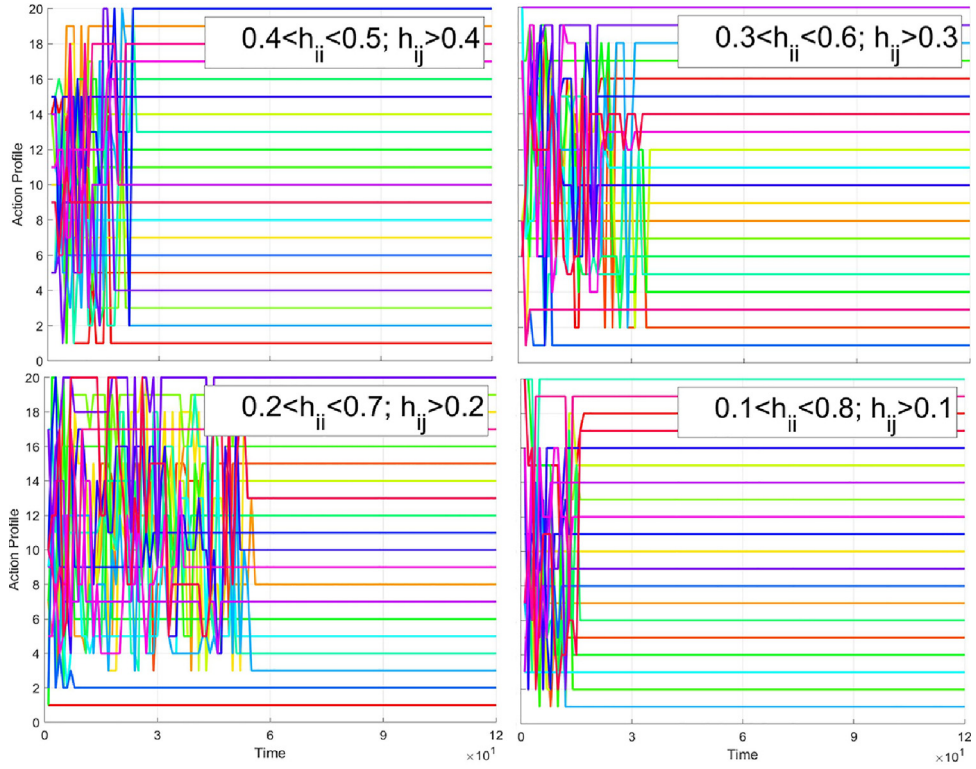

Fig. 3. Instances of joint actions profile a_t over time generated by DFP-VL with the given parameters. Joint action profiles converges to one-to-one assignments which are pure NE.

parameter values, agents can still converge to a pure NE in all cases, albeit smaller fading rate and inertia values appear to be more preferable in terms of convergence rate.

6. Conclusion

We considered inertial best-response type algorithms given random communication networks for learning Nash equilibria in weakly acyclic games. We showed that the actions generated from inertial best-response type algorithms converge to a pure Nash equilibrium almost surely under the condition that agents learn to predict the actions of other agents when those agents repeat the same action. We then proposed voluntary and limited communication protocols for DFP. Using this protocol, agents choose the subset of agents they want to send information to. We further showed that the proposed communication protocols satisfy the prediction under static actions condition, and thus are guaranteed to converge to a pure NE. Compared to standard DFP with constant communication attempts, numerical experiments showed that the proposed communication protocols significantly reduce communication attempts while achieving comparable convergence rates.

Appendix

A.1. Proof of Lemma 1

The expected utility $u_i : \Delta\mathcal{A}^N \to \mathbb{R}$ in (1) is a linear combination of bounded payoff values $U_i(a)$. Thus, there exists a

Lipschitz constant $L > 0$ such that the following holds,

$$|u_i(a_i, f^i_{-i}(t)) - u_i(a_i, a_{-i})| \leq L\|a_{-i} - f^i_{-i}(t)\|, \tag{A.1}$$

$$\leq L \sum_{j\in\mathcal{N}\setminus\{i\}} \|a_j - f^i_j(t)\|,$$

$$\leq L(N-1)\xi < \frac{\mu}{2}, \tag{A.2}$$

for some $\mu > 0$. Next, we define the following mutually exclusive subsets of action space $\mathcal{A}$ for all $i \in \mathcal{N}$,

$$\mathcal{A}_1(i) = \{\mathbf{e}_{k_1} \in \mathcal{A} \,|a_i = \mathbf{e}_{k_1} \in \operatorname{argmax} u_i(a_i, a_{-i})\}, \tag{A.3}$$

$$\mathcal{A}_2(i) = \{\mathbf{e}_{k_2} \in \mathcal{A} \,|a_i = \mathbf{e}_{k_2} \notin \operatorname{argmax} u_i(a_i, a_{-i})\}. \tag{A.4}$$

Since they are mutually exclusive subsets, it holds $\mathcal{A}_i = \mathcal{A}_1(i) \cup \mathcal{A}_2(i)$ and $\mathcal{A}_1(i) \cap \mathcal{A}_2(i) = \emptyset$. Then, optimal set over a finite feasible set of utility functions cannot be an empty set $\mathcal{A}_1(i) \neq \emptyset$, while it is possible that $\mathcal{A}_2(i) = \emptyset$. Firstly, suppose that $\mathcal{A}_2(i) \neq \emptyset$. Hence, there exist actions $a'_i \in \mathcal{A}_1(i)$ and $a''_i \in \mathcal{A}_2(i)$ such that,

$$u_i(a'_i, a_{-i}) - u_i(a''_i, a_{-i}) > \mu \tag{A.5}$$

for some $\mu > 0$ satisfying (A.2) where we note that μ can be made small enough by selecting a small enough ξ.

Note that (A.2) holds for both actions $a'_i \in \mathcal{A}_1(i)$ and $a''_i \in \mathcal{A}_2(i)$,

$$|u_i(a'_i, f^i_{-i}(t)) - u_i(a'_i, a_{-i})| < \frac{\mu}{2}, \tag{A.6}$$

$$|u_i(a''_i, f^i_{-i}(t)) - u_i(a''_i, a_{-i})| < \frac{\mu}{2}. \tag{A.7}$$

Next, we add the terms in (A.6) and (A.7) to the left and right hand sides of (A.5), respectively. Since the difference between the terms in (A.6) and (A.7) must be no worse than $-\mu/2$, it yields,

$$\begin{aligned} u_i(a'_i, f^i_{-i}(t)) &- u_i(a''_i, f^i_{-i}(t)) \\ &= u_i(a'_i, f^i_{-i}(t)) + u_i(a'_i, a_{-i}) - u_i(a''_i, a_{-i}) \\ &\quad - u_i(a'_i, a_{-i}) + u_i(a''_i, a_{-i}) - u_i(a''_i, f^i_{-i}(t)) \\ &> \mu - \frac{\mu}{2} - \frac{\mu}{2} = 0. \end{aligned} \tag{A.8}$$

Further, for any two best-response actions, $a'_i \in \mathcal{A}_1(i)$ and $\tilde{a}'_i \in \mathcal{A}_1(i)$, it can be shown that

$$|u_i(a'_i, f^i_{-i}(t)) - u_i(\tilde{a}'_i, f^i_{-i}(t))| < \mu. \tag{A.9}$$

As a result, using its estimates $f^i_{-i}(t)$, agent i only chooses an action from its optimal action set $\mathcal{A}_1(i)$ for the both cases $\mathcal{A}_2(i) = \emptyset$ and $\mathcal{A}_2(i) \neq \emptyset$. Thus, it holds for all $i \in \mathcal{N}$,

$$\operatorname*{argmax}_{a_i\in\mathcal{A}} u_i(a_i, f^i_{-i}(t)) \subseteq \operatorname*{argmax}_{a_i\in\mathcal{A}} u_i(a_i, a_{-i}). \tag{A.10}$$

A.2. Proof of Theorem 2

We note that the randomness stems from inertia, and communication and acknowledgment failures. The probability of given events in the following part, only depends on these random variables. Thus, showing that the event $\{\|a_j(t+T) - f^i_j(t+T)\| \leq \xi\}$ has a positive probability follows from the positive probability of successful communication and acknowledgment, and the positive probability of agent j repeating the same action via inertia. Consider the following events:

$$E_5(t) = \{\|a_j(t+T) - f^i_j(t+T)\| \leq \|a_j(t+T) - f_j(t+T)\| + \|f_j(t+T) - f^{i(j)}_j(t+T)\|\}$$

$$E_6(t) = \{\|a_j(t+T) - f_j(t+T)\| \leq \xi/2\}$$

$$E_7(t) = \{\|f_j(t+T) - f^{i(j)}_j(t+T)\| \leq \xi/2\}$$

By triangle equality we have,

$$\|a_j(t+T) - f^i_j(t+T)\| \leq \|a_j(t+T) - f_j(t+T)\| + \|f_j(t+T) - f^i_j(t+T)\|. \tag{A.11}$$

Then, via triangle inequality, showing that $E_5(t)$ happens with positive probability reduces to showing the positive probability of the following event,

$$E_8(t) = \{\| f_j(t+T) - f^i_j(t+T) \| \leq \|f_j(t+T) - f^{i(j)}_j(t+T)\|\}.$$

Given the assumptions on η_1, η_2 and η_3, condition in (14) is satisfied, i.e., agent j attempts to communicate with agent i, until the events E_6 and E_7 happen together.

In the event that agent j successfully communicates with agent i and receives an acknowledgment, we have $f^i_j(t+T) = f^{i(j)}_j(t+T)$. Hence, it follows,

$$\begin{aligned} &\mathbb{P}(E_5(t)|H(t), \hat{E}_j(t)) = \mathbb{P}(E_8(t)|H(t), \hat{E}_j(t)) \\ &= \mathbb{P}(c_{ji}(t+T) = 1, b_{ji}(t+T) = 1) \geq \nu^2, \end{aligned} \tag{A.12}$$

where the event $\hat{E}_j(t)$ is as defined in Condition 1, and the inequality follows via the lower bound on communication and acknowledgment success probabilities. Next, the event $E_6(t)$ is certain given the repetition of the same action by agent j, and by Lemma 4*(a)* there exists a long enough T such that,

$$\mathbb{P}(E_6(t)|H(t), \hat{E}_j(t)) = 1. \tag{A.13}$$

Now, let $\phi_j(t+T)$ be the estimate of empirical frequency of agent j constructed using limited information $\upsilon_j(t+T)$ (17) and $\kappa_j(t+T)$ (18) at time $t+T$. By triangle equality, we have

$$\begin{aligned} \|f_j(t+T) - f^{i(j)}_j(t+T)\| &\leq \|f_j(t+T) - \phi_j(t+T)\| \\ &+ \|\phi_j(t+T) - f^i_j(t+T)\| + \|f^i_j(t+T) - f^{i(j)}_j(t+T)\|. \end{aligned} \tag{A.14}$$

Now, consider the following events,

$$E_9(t) = \{\|f_j(t+T) - \phi_j(t+T)\| \leq \xi/2\}$$

$$E_{10}(t) = \{\|\phi_j(t+T) - f^i_j(t+T)\| = 0\}$$

$$E_{11}(t) = \{\|f^i_j(t+T) - f^{i(j)}_j(t+T)\| = 0\}$$

Given the repetition of the same actions by agents $j \in \mathcal{N} \setminus \{i\}$ and Lemma 4*(b)*, there exists a long enough T such that $\mathbb{P}(E_9(t)|H(t), \hat{E}(t)) = 1$ similar to (A.13). Further, see the remaining events have also positive probability as a result of the lower bound on the chance of successful communication and acknowledgment,

$$\mathbb{P}(E_{10}(t)|H(t), \hat{E}_j(t)) = \mathbb{P}(c_{ji}(t+T) = 1) \geq \nu > 0, \tag{A.15}$$

$$\begin{aligned} &\mathbb{P}(E_{11}(t)|H(t), \hat{E}_j(t)) = \mathbb{P}(c_{ji}(t+T) = 1, b_{ji}(t+T) = 1) \\ &\geq \nu^2 > 0. \end{aligned} \tag{A.16}$$

The equality in (A.15) follows by the fact that the estimate based on limited information from j, i.e., $\phi_j(t+T)$, can only be constructed at node i if there is a successful communication from node j to i. Similarly, j's belief about i's belief is correct only when both the communication and acknowledgment attempts are successful. From (A.14) and the bounds above, we have

$$\begin{aligned} &\mathbb{P}(E_7(t)|H(t), \hat{E}_j(t)) \geq \\ &\qquad \mathbb{P}(E_9(t), E_{10}(t), E_{11}(t)|H(t), \hat{E}_j(t)) \geq \nu^2 > 0. \end{aligned} \tag{A.17}$$

Thus, there exists a positive real number $\hat{\epsilon} > 0$ such that,

$$\begin{aligned} &\mathbb{P}(\|a_j(t+T) - f^i_j(t+T)\| \leq \xi \,|H(t), \hat{E}_j(t)) \geq \\ &\mathbb{P}(E_5(t), E_6(t), E_7(t)|H(t), \hat{E}_j(t)) \geq \nu^2 = \hat{\epsilon} > 0. \end{aligned} \tag{A.18}$$

A.3. Technical result

Lemma 4. *Let $\{a(t) = (a_1(t), a_2(t), \ldots, a_N(t))\}_{t\geq 1}$ be a sequence of actions generated by the DFP-VL (Algorithm 2). For a given $0 < \xi_1$ and $0 < \xi_2$, there exists a long enough T such that if agent $j \in \mathcal{N} \setminus \{i\}$ repeats the same action $a_j(s) = \mathbf{e}_k$ at least $T > 0$ times for $s = t, t+1, \ldots, t+T-1$, then*

(a) $\|a_j(t+T) - f_j(t+T)\| \leq \xi_1$ *for all* $j \in \mathcal{N} \setminus \{i\}$,
(b) $\|\phi_j(t+T) - f_j(t+T)\| \leq \xi_2$ *for all* $j \in \mathcal{N} \setminus \{i\}$,

where $\phi_j(t)$ is the reconstructed belief of agent j's empirical frequency using $\upsilon_j(t)$ and $\kappa_j(t)$ defined in (17) and (18), respectively.

Proof.

(a) From (3), it holds that if $\mathbf{e}_k$ is repeated for any $\tau \in \{0, 1, 2, \ldots\}$ starting from time t by a player $j \in \mathcal{N} \setminus \{i\}$,

$$f_j(t+\tau) = (1-\rho)^\tau f_j(t) + (1-(1-\rho)^\tau)\mathbf{e}_k, \tag{A.19}$$

Subtracting $\mathbf{e}_k$ from both sides and taking the norm we obtain the following,

$$\begin{aligned}\|f_j(t+\tau) - \mathbf{e}_k\| &= \|(1-\rho)^\tau (f_j(t) - \mathbf{e}_k)\|,\\ &= O((1-\rho)^\tau).\end{aligned} \tag{A.20}$$

Therefore, if agent $j \in \mathcal{N} \setminus \{i\}$ repeat the same action $a_j(s) = \mathbf{e}_k$ for long enough $T > 0$ times for $s = t, t+1, \ldots, t+T-1$, we have the inequality in (a).

(b) We use triangle inequality to get,

$$\begin{aligned}\|\phi_j(t+T) - f_j(t+T)\| &\leq \|\phi_j(t+T) - a_j(t+T)\|\\ &\quad + \|a_j(t+T) - f_j(t+T)\|.\end{aligned} \tag{A.21}$$

Then, notice that $\|a_j(t+T) - f_j(t+T)\| = O((1-\rho)^{T+1})$ implies $|\upsilon_j(t) - 1| = O((1-\rho)^{T+1})$. Since $\phi_{j\kappa_i(t)}(t+T) \geq \upsilon_j(t)$ via (19), it also holds $\|\phi_j(t+T) - a_j(t+T)\| = O((1-\rho)^{T+1})$. Thus, given the repetition of the same action for T times, we have

$$\begin{aligned}&\|\phi_j(t+T) - f_j(t+T)\|\\ &\leq \|\phi_j(t+T) - a_j(t+T)\| + \|a_j(t+T) - f_j(t+T)\|\\ &= O((1-\rho)^{T+1}) \leq \xi_2. \quad \square\end{aligned} \tag{A.22}$$

References

Al Sheikh, A., Brun, O., Hladik, P.-E., & Prabhu, B. J. (2011). A best-response algorithm for multiprocessor periodic scheduling. In *2011 23rd Euromicro conference on real-time systems* (pp. 228–237). IEEE.

Alpcan, T., & Başar, T. (2005). Distributed algorithms for Nash equilibria of flow control games. In *Advances in dynamic games* (pp. 473–498). Springer.

Arefizadeh, S., & Eksin, C. (2019). Distributed fictitious play in potential games with time varying communication networks. In *2019 53rd Asilomar conference on signals, systems, and computers* (pp. 1755–1759). IEEE.

Arslan, G., Marden, J. R., & Shamma, J. S. (2007). Autonomous vehicle-target assignment: A game-theoretical formulation. *Journal of Dynamic Systems, Measurement, and Control*, *129*(5), 584–596.

Arslan, G., & Yüksel, S. (2016). Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, *62*(4), 1545–1558.

Aydın, S., & Eksin, C. (2020a). Communication censoring in decentralized fictitious play for the target assignment problem. In *2020 IEEE conference on control technology and applications (CCTA)* (pp. 334–339). IEEE.

Aydin, S., & Eksin, C. (2020b). Decentralized fictitious play with voluntary communication in random communication networks. In *2020 59th IEEE conference on decision and control (CDC)* (pp. 337–342). IEEE.

Bauch, C. T., & Earn, D. J. (2004). Vaccination and the theory of games. *Proceedings of the National Academy of Sciences*, *101*(36), 13391–13394.

Bell, C. E. (1996). Finding improving directions in Lagrangian relaxation by fictitious play: A NASA scheduling application. *European Journal of Operational Research*, *88*(3), 550–562.

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, *13*(1), 374–376.

Chen, T., Giannakis, G., Sun, T., & Yin, W. (2018). LAG: Lazily aggregated gradient for communication-efficient distributed learning. In *Advances in neural information processing systems* (pp. 5050–5060).

Chen, Y., Sadler, B. M., & Blum, R. S. (2018). Ordered transmission for efficient wireless autonomy. In *2018 52nd asilomar conference on signals, systems, and computers* (pp. 1299–1303). IEEE.

Chen, J., & Zhu, Q. (2019). Control of multilayer mobile autonomous systems in adversarial environments: A games-in-games approach. *IEEE Transactions on Control of Network Systems*, *7*(3), 1056–1068.

De Persis, C., & Grammatico, S. (2019). Distributed averaging integral Nash equilibrium seeking on networks. *Automatica*, *110*, Article 108548.

Eksin, C., & Ribeiro, A. (2017). Distributed fictitious play for multiagent systems in uncertain environments. *IEEE Transactions on Automatic Control*, *63*(4), 1177–1184.

Eksin, C., Shamma, J. S., & Weitz, J. S. (2017). Disease dynamics in a stochastic network game: a little empathy goes a long way in averting outbreaks. *Scientific Reports*, *7*, 44122.

Gao, Z., Ma, Q., Başar, T., & Birge, J. R. (2021). Finite-sample analysis of decentralized Q-learning for stochastic games. arXiv preprint arXiv:2112.07859.

Garcia, A., Reaume, D., & Smith, R. L. (2000). Fictitious play for finding system optimal routings in dynamic traffic networks. *Transportation Research, Part B (Methodological)*, *34*(2), 147–156.

Kantaros, Y., Guo, M., & Zavlanos, M. M. (2019). Temporal logic task planning and intermittent connectivity control of mobile robot networks. *IEEE Transactions on Automatic Control*.

Kantaros, Y., & Zavlanos, M. M. (2016). Distributed communication-aware coverage control by mobile sensor networks. *Automatica*, *63*, 209–220.

Kar, S., Hug, G., Mohammadi, J., & Moura, J. M. (2014). Distributed state estimation and energy management in smart grids: A consensus + innovations approach. *IEEE Journal of Selected Topics in Signal Processing*, *8*(6), 1022–1038.

Koshal, J., Nedić, A., & Shanbhag, U. V. (2016). Distributed algorithms for aggregative games on graphs. *Operations Research*, *64*(3), 680–704.

Li, W., Liu, Y., Tian, Z., & Ling, Q. (2019). Communication-censored linearized ADMM for decentralized consensus optimization. *IEEE Transactions on Signal and Information Processing over Networks*, *6*, 18–34.

Marden, J., Arslan, G., & Shamma, J. (2009a). Cooperative control and potential games. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, *39*(6), 1393–1407.

Marden, J. R., Arslan, G., & Shamma, J. S. (2009b). Joint strategy fictitious play with inertia for potential games. *IEEE Transactions on Automatic Control*, *54*(2), 208–220.

Marden, J. R., Young, H. P., Arslan, G., & Shamma, J. S. (2009). Payoff-based dynamics for multiplayer weakly acyclic games. *SIAM Journal on Control and Optimization*, *48*(1), 373–396.

Milchtaich, I. (1996). Congestion games with player-specific payoff functions. *Games and Economic Behavior*, *13*(1), 111–124.

Monderer, D., & Shapley, L. S. (1996). Fictitious play property for games with identical interests. *Journal of Economic Theory*, *68*(1), 258–265.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, 296–301.

Salehisadaghiani, F., Shi, W., & Pavel, L. (2019). Distributed Nash equilibrium seeking under partial-decision information via the alternating direction method of multipliers. *Automatica*, *103*, 27–35.

Sayin, M. O., Parise, F., & Ozdaglar, A. (2020). Fictitious play in zero-sum stochastic games. arXiv preprint arXiv:2010.04223.

Scutari, G., & Pang, J.-S. (2013). Joint sensing and power allocation in nonconvex cognitive radio games: Nash equilibria and distributed algorithms. *IEEE Transactions on Information Theory*, *59*(7), 4626–4661.

Shamma, J., & Arslan, G. (2005). Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria. *IEEE Transactions on Automatic Control*, *50*(3), 312–327.

Swenson, B., Eksin, C., Kar, S., & Ribeiro, A. (2018). Distributed inertial best-response dynamics. *IEEE Transactions on Automatic Control*, *63*(12), 4294–4300.

Swenson, B., Kar, S., & Xavier, J. (2015). Empirical centroid fictitious play: An approach for distributed learning in multi-agent games. *IEEE Transactions on Signal Processing*, *63*(15), 3888–3901.

Williams, G., Goldfain, B., Drews, P., Rehg, J. M., & Theodorou, E. A. (2018). Best response model predictive control for agile interactions between autonomous ground vehicles. In *2018 IEEE international conference on robotics and automation (ICRA)* (pp. 2403–2410). IEEE.

Ye, M., & Hu, G. (2021). Adaptive approaches for fully distributed Nash equilibrium seeking in networked games. *Automatica*, *129*, Article 109661.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, 57–84.

Young, H. P. (2004). *Strategic learning and its limits*. OUP Oxford.
Zhang, Y., Gatsis, N., & Giannakis, G. B. (2012). Robust distributed energy management for microgrids with renewables. In *2012 IEEE third international conference on smart grid communications (SmartGridComm)* (pp. 510–515). IEEE.

Sarper Aydın received the B.Sc. degree in industrial engineering from Bilkent University, Ankara, Turkey in 2017. During this period, he spent one semester as an exchange student at Queensland University of Technology, Brisbane, QLD, Australia. From 2017 to 2019, he was with Lehigh University, Bethlehem, PA, USA as a Ph.D. student. He joined Texas A&M University, College Station, TX, USA in 2019. Currently, he is working toward the Ph.D. degree in the Department of Industrial and Systems Engineering, Texas A&M University. His current research interests include decentralized control of multi-agent systems and game theoretic learning with applications to assignment problems in autonomous robot teams.

Ceyhun Eksin received the B.S. degree in control engineering from Istanbul Technical University, in 2005, an M.S. degree in industrial engineering from Boğaziçi University, Istanbul, Turkey in 2008, an M.A. degree in statistics from Wharton School in 2015, and the Ph.D. degree in Electrical and Systems Engineering from the University of Pennsylvania in 2015. He was a postdoctoral researcher in Georgia Institute of Technology from 2015 to 2017. He is currently an assistant professor with the Industrial and Systems Engineering Department, Texas A&M University. His research interests focus on modeling and design of networked multi-agent systems using game theory, control theory, and distributed optimization.