

# Decentralized Learning for Optimality in Stochastic Dynamic Teams and Games with Local Control and Global State Information

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel

**Abstract**—Stochastic dynamic teams and games are rich models for decentralized systems and challenging testing grounds for multi-agent learning. Previous work that guaranteed team optimality assumed stateless dynamics, or an explicit coordination mechanism, or joint-control sharing. In this paper, we present an algorithm with guarantees of convergence to team optimal policies in teams and common interest games. The algorithm is a two-timescale method that uses a variant of Q-learning on the finer timescale to perform policy evaluation while exploring the policy space on the coarser timescale. Agents following this algorithm are “independent learners”: they use only local controls, local cost realizations, and global state information, without access to controls of other agents. The results presented here are the first, to our knowledge, to give formal guarantees of convergence to team optimality using independent learners in stochastic dynamic teams and common interest games.

**Index Terms**—Stochastic games; Stochastic optimal control; Cooperative control; Game Theory; Machine learning.

## I. INTRODUCTION

In modern control engineering applications, two challenges are becoming increasingly common: online problems and decentralization. In online problems, the system to be controlled is not initially known by the agent and must be learned. In decentralized systems, several autonomous decision-makers act in a shared environment. This paper is concerned with multi-agent reinforcement learning (MARL), which is at the intersection of these two challenges. We use stochastic games to model the shared environment, and we present algorithms suitable for stochastic dynamic teams under a particular decentralized information structure.

In online problems, important knowledge of the system to be controlled is initially unavailable to the controller. Classical methods for solving control problems, such as linear programming, dynamic programming, and convex analytic methods, cannot be implemented without access to the system model. Instead, the control agent must use observed feedback to learn control policies. Reinforcement learning has had considerable success in single-agent control problems, both in applications and in theory, where methods such as Q-learning [2], [3], [4] recover optimal policies when used in a stationary environment.

A conference version [1] was presented at the 2019 Conference on Decision and Control and serves as an announcement of the partial results presented here without details. The conference version [1] does not contain the results on weakly acyclic games or any of the proofs presented here.

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

B. Yongacoglu and S. Yüksel are with the Department of Mathematics and Statistics, Queen’s University, Kingston, ON, Canada, email: {lbmy,yuksel@queensu.ca}. G. Arslan is with the University of Hawaii. Email: {gurdal@hawaii.edu}.

A second challenge comes from decentralization. Decentralized systems are characterized by multiple agents acting in a common environment with some local information available to each agent. The costs incurred by one agent in a decentralized system depend, in general, on its own actions, the actions of other agents, and the history of the system. Such coupled interactions are common in complex, real-world engineering applications. Some examples of systems that are inherently decentralized are sensor networks, stochastic networked control systems, the Internet of Things, and energy systems.

Compared to the success of reinforcement learning in stationary single-agent problems, there are relatively few formal results on MARL. This is partly explained by the loss of stationarity: when multiple learning agents interact, a given agent will change its behaviour to exploit learned information. From the point-of-view of the remaining agents, this agent is a part of the environment, and so the environment is non-stationary [5]. Consequently, one of the fundamental assumptions made for single-agent theory does not hold in MARL, and theoretical guarantees do not carry over.

Stochastic games [6], [7], [8], [9] generalize both repeated games [10] and Markov decision problems (MDPs). Like repeated games, players in stochastic games must be strategic and respond to the policies used by other agents. Unlike repeated games, in which the same stage game is played at every time step, the stage game played at a given time in a stochastic game depends on the history of play, which is summarized by a state. As in MDPs, agents in stochastic games must select actions with the state process and its long-term cost implications in mind. As stochastic games provide a rich model for dynamic, strategic decision making, they are a popular framework for studying MARL [11].

Stochastic dynamic teams [12], [13] and common-interest games [14], [15] model cooperative systems and so are of special interest to decentralized control. In teams, all players incur the same costs and interests are perfectly aligned. Common interest games generalize teams in a natural way: in common interest games, agents do not necessarily incur identical costs, but there are a subset of joint policies which each agent strictly prefers to all other policies. Despite the incentive to coordinate behavior in common interest games, coordination is generally difficult in online problems when information is decentralized.

As we will outline in detail in Section II, there are relatively few theoretical results for stochastic games without control-sharing. Even when assuming full state observability at each agent rather than the more general assumption of partial state observability, there are no rigorous results that guarantee team optimality in truly stochastic teams without relying on control-sharing.

## A. Contributions

In this paper, we present a decentralized learning algorithm for playing stochastic common interest games, a class of games which model decentralized control problems and contain stochastic teams as a special case. We give formal guarantees of convergence to team optimal policies without use of control sharing among agents.

- (i) In Theorem 1, we consider stochastic common interest games and introduce an algorithm (that only uses local cost and local action history and the common state of the system) that provably converges to a team optimal policy in a probabilistic sense that is made precise in the theorem. What makes this algorithm different from our prior work [16], which guaranteed convergence to equilibrium but not team optimal policies, is the utilization of a finite window of the most recent (noisy) aggregate cost scores to adaptively estimate the lowest possible cost for each decision maker.
- (ii) Theorem 2 considers a specific implementation of our main algorithm in the context of weakly acyclic games. We show this algorithm leads to equilibrium policies in weakly acyclic games and, furthermore, if the game is also a common interest game then play will settle to a team optimal policy. This theorem strengthens one of the main results from [16].
- (iii) In Theorem 3, we obtain convergence to team optimality in the stronger sense of almost sure convergence by using constant, preset aspiration levels. This result requires a stronger assumption that the preset aspiration levels separate the team optimal policies from the other policies. Theorem 3 also describes the long-run behaviour of the algorithm with constant aspirations when used in a general stochastic game.

These contributions are the first formal guarantees of achieving team optimality in stochastic common interest games under full state observability but no action sharing.

The remainder of the paper is organized as follows: Section II surveys related literature. In Section III, we specify the stochastic game model and provide relevant background. In Section IV, we present our main algorithm and state Theorem 1. Section V presents Theorem 2, which strengthens a result from [16]. Section VI considers a variant of the main algorithm and presents Theorem 3, which studies the variant algorithm's long-term behaviour in general sum games. Section VII contains numerical results from a simulation study, and Section VIII concludes the paper. The proofs of our main technical results are contained in the appendices.

## B. Notation

$\mathbb{R}$  denotes the real numbers,  $\mathbb{N}$  and  $\mathbb{N}_+$  denote the nonnegative and positive integers, respectively.  $\Pr(\cdot)$  and  $E(\cdot)$  denote the probability and the expectation, respectively. For a finite set  $S$ ,  $\mathcal{P}(S)$  denotes the set of probability distributions over  $S$ . For finite sets  $S, S'$ , we let  $\mathcal{P}(S'|S)$  denote the set of stochastic kernels on  $S'$  given  $S$ . An element  $\mathcal{T} \in \mathcal{P}(S'|S)$  is a collection of probabilities distributions on  $S'$ , with one distribution for each  $s \in S$ , and we write  $\mathcal{T}(\cdot|s)$  for  $s \in S$  to make this distributional dependence on  $s$  explicit. We write  $Y \sim f$  to denote that the random variable  $Y$  has distribution  $f$ . If the distribution of  $Y$  is a mixture of other distributions, say with

mixture components  $f_i$  and weights  $p_i$  for  $1 \leq i \leq n$ , we write  $Y \sim \sum_{i=1}^n p_i f_i$ . The Dirac distribution concentrated at  $x \in \mathbb{R}$  is denoted  $\mathbb{I}_x$ . For a finite set  $S$ ,  $\text{Unif}(S)$  denotes the uniform distribution over  $S$  and  $2^S$  denotes the set of subsets of  $S$ .  $(x)^+ := \max\{x, 0\}$ , for  $x \in \mathbb{R}$ .

## II. LITERATURE REVIEW

Interest in using single-agent reinforcement learning in multi-agent environments dates at least as far back as [17], in which Q-learning is studied in a cooperative predator-prey simulation. In [18], multiple agents run Q-learning in a block-pushing task without sharing actions with one another, and the authors suggest that cooperative behaviour may emerge even without explicit communication between agents.

In addition to presenting empirical results and formal conjectures, an important terminological distinction was popularized in [19], where the authors distinguish between joint action learners and independent learners: joint action learners use the past actions of all agents in their learning, while independent learners use only local action histories.

Early rigorous work on MARL in games was concerned mostly with joint-action learners. In [11], Littman proposed stochastic games as a framework for studying MARL and presented the *Minimax Q-learning* algorithm, a joint-action learner designed for two-player zero-sum games. Convergence results for this method were proved in [20]. The main idea from [11] was extended in [21] and [22], which present *Nash Q-learning*, another joint-action learner with convergence guarantees under certain restrictive assumptions. Further contributions in this line include *Friend-or-Foe Q-learning* [23], *Team Q-Learning* [24], and several others, e.g. [25], [26]. A considerably different approach is taken in [27], which presents *Optimal Adaptive Play* (OAP), a joint-action learner based on adaptive play [28] rather than on Q-learning. OAP is shown to converge to a team optimal policy when used in a stochastic team.

Though early rigorous work focused on joint action learners, there has also been persistent interest in independent learners. As the number of joint actions is exponential in the number of agents, the computational burden of a joint action learner at any one agent becomes intractable for problems of even a moderate size. Scalability, robustness, and faster convergence are potential advantages of independent learners over joint action learners [29], [30]. The applicability of the set-up considered here and other advantages are covered in greater detail in [29] and [31]. For a recent survey of MARL that discusses other decentralized set-ups, see [32].

*Distributed Q-learning*, an independent learner designed for teams, was presented in [33], along with a guarantee of convergence to team optimality in teams with deterministic state dynamics and costs. When using this algorithm, an agent only updates its Q-factors when an improvement is observed, attributing unfavourable feedback to its teammates' experimenting with other actions. This optimistic approach leads to poor performance in problems with random state transitions or cost readings [29].

An algorithm called *Win or Learn Fast Policy Hill Climbing* (WoLF-PHC) was introduced in [34]. An agent using WoLF-PHC selects actions according to an exploration policy and iteratively improves its exploration policy using its learned Q-factors by updating toward a best-response. Although no

formal results are presented for stochastic games, the key innovation of [34] is its policy update: the agent compares the performance of the current exploration policy to that of a distinguished “average policy.” When the current policy outperforms the average policy, the agent changes its policy relatively slowly; when the current policy is underperforming, the agent changes its policy more rapidly.

Following [33] and [34], a number of algorithms based on Q-learning were proposed for stochastic games. Some of these algorithms, such as *Hysteretic Q-learning* [35], modify the Q-factor update. Other methods, including the *Frequency Maximum Q* heuristic presented in [36] and its extensions to stochastic games [37], modify action selection. Still other methods, such as lenient learning [38], [39], modify both the Q-factor update as well as the action selection mechanism in an attempt to achieve optimality in cooperative games. With the exception of [33] described above, these works offer only empirical support for their algorithms, rather than formal guarantees of convergence to team optimality. For a survey on this line of research and a description of obstacles in MARL, see [29].

While researchers in the machine learning community sought empirically successful algorithms for stochastic games, a parallel line of research in the control and operations research communities sought rigorous results in the more restricted class of stateless repeated games.

Among the literature on MARL for repeated games, [40], [41] and [42] are most relevant to this paper. Although the algorithms and analysis presented in these works differ from one another, each operates using the principle of exploring an agent’s set of actions more aggressively when the agent perceives it is underperforming.

Reference [41] presents three algorithms, including *Safe Experimentation Dynamics*, which is shown to lead to team optimality in repeated teams with high probability. Using this method, an agent maintains a baseline action and baseline cost while experimenting with other actions. Each time an action is taken, its immediate cost is compared with the baseline cost; the baseline cost is adjusted when a lower cost is observed, and the action achieving this lower cost becomes the new baseline.

Agents using the algorithm from [42] maintain a binary “mood” variable, which is meant to capture whether the agent is content with its current performance. It is shown that all stochastically stable outcomes maximize the sum of joint payoffs across all agents.

*Aspiration learning* for repeated coordination games is presented in [40], along with formal results on the stochastic stability of efficient outcomes. An agent using this algorithm iteratively sets its aspiration level, a scalar threshold value that represents the highest cost (or lowest reward) that the agent finds acceptable. When receiving costs higher than its aspiration level, the agent is unsatisfied and explores alternative actions more aggressively.

Other work in this area includes [43], [44] and [45]. Variants of log-linear learning for repeated games were studied in [43], [44] and come with guarantees on the stochastic stability of efficient outcomes. The stochastic imitation dynamics introduced in [45] assign probability one to efficient outcomes in large class of repeated games.

One explanation for the greater number of rigorous results on independent learners in a repeated game setting is the lack

of state dynamics. In repeated games, the same stage game is played in each period and there is no tradeoff between short- and long-term costs. As such, the scalar cost realizations can be used directly when setting aspiration levels (as in [40]) or baseline costs (as in [41] and [42]). In contrast, policy evaluation is inherently slow (due to delayed rewards), noisy, and algorithm dependent in games with random state dynamics, and this is only exacerbated by the presence of other learning agents. Consequently, extending the preceding methods is a significant challenge.

In this paper, we study stochastic teams and common interest games with full state information at each agent but no action sharing between agents. This set-up arises naturally in problems where the state can be sensed by a global sensor and broadcast to agents. In [46], a (physically) distributed array of micro-electro-valves producing controlled and directed micro-air-jets is used to steer the motion of a small object on a smart surface. The state of this system is the current and previous positions of the object which is sensed by an overhead camera and accessed by all control units each controlling a separate valve. Each control unit implements a standard Q-learning algorithm based on the global state and its own control observations (by ignoring the other control units) for reasons stated as follows: “A fully centralized control architecture is not suitable due to processing complexity and the number of communication channels required”. In [31], robotics problems involving multi-dimensional action spaces are considered. The authors observe that centralized approaches in problems with multiple actuators are often intractable due to a combinatorial explosion of the joint state-action space. Among other decentralization schemes, the authors consider the case with full state but only local actions, wherein the actuators are able to sense the global state variable (e.g. two dimensional position and velocity in a vehicle navigation problem; three dimensional position in a joint manipulation task) but do not attempt to sense one another’s actions for computational tractability. Other applications for which this information structure is appropriate include problems where the state variable is a commonly observed price as well as problems in traffic networks, where link latencies can be broadcast using a mobile application.

Another motivation for studying this set-up is that algorithms designed for problems with full state information but no action-sharing have been successful even when used in problems possessing a different information structure, such as partial state observability. Examples of studies that use partial state observations as a surrogate for complete state observations and then use methods designed for our information structure include interference control in wireless networks [47], [48] and cache placement in wireless networks [49]. Many further examples can be found in the area of cognitive radio; see [50] and the references therein.

In [16], we introduced an independent learner that provably leads to equilibrium in weakly acyclic stochastic games in general and in teams in particular. However, stochastic teams generally have both team optimal equilibrium policies and suboptimal equilibrium policies, and suboptimal equilibria can perform arbitrarily worse than an optimal equilibrium. A simple but illustrative example is offered in Section III. Thus, guarantees of finding an equilibrium joint policy are not satisfactory in the context of decentralized control when

cost minimization is a design goal. In this paper, we modify the main algorithm from [16] to guarantee convergence to team optimality when possible.

### III. BACKGROUND

#### A. Stationary Markov Decision Problems and Q-Learning

A stationary Markov Decision Problem (MDP) with a discounted cost criterion is a discrete time process characterized by the following:

- 1) A finite set of states  $\mathbb{X}$
- 2) A random initial state  $x_0 \in \mathbb{X}$
- 3) A finite set of control actions  $\mathbb{U}$
- 4) A discount factor  $\beta \in (0, 1)$
- 5) A cost function  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$
- 6) A transition probability kernel  $P \in \mathcal{P}(\mathbb{X}|\mathbb{X} \times \mathbb{U})$  for determining the next state given the current state-action.

At time  $t \in \mathbb{N}$ , the system is in state  $x_t \in \mathbb{X}$  and the decision maker<sup>1</sup> (DM) selects a control action  $u_t \in \mathbb{U}$ . The DM then incurs a stage cost  $c(x_t, u_t)$ , and the system randomly transitions to the next state,  $x_{t+1}$ , according to the probability distribution  $P(\cdot|x_t, u_t)$ . We assume that, prior to selecting  $u_t$  at time  $t \in \mathbb{N}$ , the DM has access to the information  $I_t$  defined by

$$I_0 = \{x_0\}, \quad I_{t+1} = I_t \cup \{x_{t+1}, u_t, c(x_t, u_t)\}, \quad t \in \mathbb{N}.$$

A policy is a rule for selecting control actions based on the information available. In principle, the DM may use any function of  $I_t$  to choose  $u_t$ , possibly with randomization. Fixing a policy  $\theta$  induces a probability distribution on the sequence of state-actions  $\{(x_t, u_t)\}_{t \in \mathbb{N}}$ . This induced probability measure is used to define the cost criterion:

$$J_x(\theta) := E^\theta \left( \sum_{t \in \mathbb{N}} \beta^t c(x_t, u_t) \middle| x_0 = x \right), \quad \forall x \in \mathbb{X}$$

where  $E^\theta$  denotes that the stochastic process  $\{(x_t, u_t)\}_{t \in \mathbb{N}}$  is determined by the policy  $\theta$ .

The DM's goal is to select a policy that minimizes the cost functional  $J_x$  in every initial state  $x \in \mathbb{X}$ . Although the agent can use an arbitrarily complicated, history dependent policy, it is well-known (see, for example, [51]) that this minimum can be achieved within the simpler set of stationary randomized policies, which we identify with the set  $\Delta = \mathcal{P}(\mathbb{U}|\mathbb{X})$ . A stationary randomized policy  $\theta \in \Delta$  uses only the most recent state  $x_t$  to (randomly) select an action  $u_t$  in a time-invariant manner; that is, when the agent follows a policy  $\theta \in \Delta$ , we have  $u_t \sim \theta(\cdot|x_t)$ . Within  $\Delta$ , we can further restrict our attention (without loss of optimality [51]) to the set of stationary deterministic policies  $\Pi$ , which we identify as  $\Pi = \{\pi : \mathbb{X} \rightarrow \mathbb{U}\}$ . An agent following a policy  $\pi \in \Pi$  selects its action as a deterministic function of the state, and we write  $u_t = \pi(x_t)$  or  $u_t \sim \mathbb{I}_{\pi(x_t)}$ .

When the cost function and transition kernel are known, iterative methods such as value iteration can be used to obtain an optimal policy. Otherwise, model-free reinforcement learning techniques such as Q-learning [3] can be used to recover an optimal policy. In standard online Q-learning, the

DM begins with arbitrary Q-factors  $Q_0 \in \mathbb{R}^{\mathbb{X} \times \mathbb{U}}$  and updates its Q-factors as follows:

$$Q_{t+1}(x_t, u_t) = (1 - \alpha_t(x_t, u_t))Q_t(x_t, u_t) + \alpha_t(x_t, u_t)(c(x_t, u_t) + \beta \min_{v \in \mathbb{U}} Q_t(x_{t+1}, v))$$

$$Q_{t+1}(x, u) = Q_t(x, u), \quad \forall (x, u) \neq (x_t, u_t)$$

where  $\alpha_t(x_t, u_t) \in [0, 1]$  is the step-size at time  $t \in \mathbb{N}$ . If all state-action pairs are visited infinitely often and the step-sizes vanish properly, then  $\Pr(Q_t \rightarrow Q^*) = 1$ , where  $Q^*$  is the vector of optimal Q-factors, the unique solution of a Bellman fixed point equation [2], [4].

Once  $Q^*$  is attained, one can recover the value function  $V^*$ , using  $V^*(x) = \min_{u \in \mathbb{U}} Q^*(x, u)$ , or an optimal policy  $\pi^*$ , using  $\pi^*(x) \in \arg \min_{u \in \mathbb{U}} Q^*(x, u)$ . Moreover, learned Q-factors can be exploited during play: [52] presents a Q-learning algorithm in which the DM's action selection converges to that of an optimal policy.

The popularity of Q-learning in stationary MDPs is justified: it is easy to implement and asymptotically recovers an optimal policy. However, this theoretical guarantee is predicated on the stationarity of the system. When a state-action  $(x_t, u_t)$  is visited, the feedback received (in the form of a cost  $c(x_t, u_t)$  and next state  $x_{t+1}$ ) is always generated by the same Markovian source. If the system is not stationary, then convergence to the Q-factors  $Q^*$  is not guaranteed.

#### B. Stochastic Games and Decentralized Q-learning

A finite (discounted) stochastic game is a multi-agent generalization of a stationary MDP, and is characterized by

- 1)  $N \in \mathbb{N}_+$  decision makers, the  $i^{\text{th}}$  denoted by  $\text{DM}^i$
- 2) A finite set of states  $\mathbb{X}$
- 3) A random initial state  $x_0 \in \mathbb{X}$
- 4) For each  $\text{DM}^i$ :
  - A finite set of control actions  $\mathbb{U}^i$
  - A discount factor  $\beta^i \in (0, 1)$
  - A cost function  $c^i : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$ , where  $\mathbf{U} := \times_{i=1}^N \mathbb{U}^i$
- 5) A transition probability kernel  $P \in \mathcal{P}(\mathbb{X}|\mathbb{X} \times \mathbf{U})$  for determining the next state given the current state and joint action.

At time  $t \in \mathbb{N}$ , the system is in state  $x_t$ , and each  $\text{DM}^i$  chooses a control action  $u_t^i$ . While  $\text{DM}^i$  only selects  $u_t^i$ , its incurred cost is given by  $c^i(x_t, \mathbf{u}_t)$ , where  $\mathbf{u}_t := (u_t^1, \dots, u_t^N)$ . Following the play of this stage game, the system randomly transitions to state  $x_{t+1}$  according to  $P(\cdot|x_t, \mathbf{u}_t)$ . We consider the situation in which  $\text{DM}^i$  observes only the state variable, its own actions, and its own cost realizations ( $\text{DM}^i$  need not know the functional form of its cost). More precisely, prior to selecting  $u_t^i$  at time  $t \in \mathbb{N}$ ,  $\text{DM}^i$  has access to the information  $I_t^i$  defined by

$$I_0^i = \{x_0\}, \quad I_{t+1}^i = I_t^i \cup \{x_{t+1}, u_t^i, c^i(x_t, \mathbf{u}_t)\}, \quad t \in \mathbb{N}.$$

In particular,  $\text{DM}^i$  cannot see the past actions of the other DMs,  $u_s^j$ , for any  $j \neq i$ ,  $s \in \mathbb{N}$ . This is in contrast to previous works such as [21], [23], [27] and [53].

A policy for  $\text{DM}^i$  is a rule for selecting the sequence of local actions given the information available to  $\text{DM}^i$ . As in MDPs,  $\text{DM}^i$ 's goal is to minimize its long-term expected discounted cost. Unlike MDPs, however,  $\text{DM}^i$ 's cost is affected by the control actions of the other agents. We again restrict

<sup>1</sup>We use the terms agent, decision maker, and player interchangeably.

our attention to stationary randomized policies, which will be justified below. We denote the set of stationary randomized policies for  $DM^i$  by  $\Delta^i := \mathcal{P}(\mathbb{U}^i|\mathbb{X})$ , and similarly we use  $\Pi^i = \{\pi^i : \mathbb{X} \rightarrow \mathbb{U}^i\}$  to denote the set of stationary deterministic policies for  $DM^i$ .

We use boldface symbols to denote joint objects, i.e. lists of objects with one entry per agent, and we omit the agent superscript. The set of stationary joint policies is thus denoted by  $\Delta := \times_{i=1}^N \Delta^i$  and the set of stationary deterministic joint policies is denoted  $\Pi := \times_{i=1}^N \Pi^i$ .

For notational convenience, we will use the agent superscript  $-i$  to refer to a joint quantity for which  $DM^i$ 's position has been removed. Using this standard convention, the set of stationary joint policies for all agents except  $DM^i$  is denoted  $\Delta^{-i} := \times_{j \neq i} \Delta^j$ . Similarly,  $\Pi^{-i} = \times_{j \neq i} \Pi^j$  and  $\mathbb{U}^{-i} = \times_{j \neq i} \mathbb{U}^j$ . By convention, we may write  $\mathbf{U} = \mathbb{U}^i \times \mathbb{U}^{-i}$  for any  $DM^i$ , and similarly for the sets  $\Delta$  and  $\Pi$ . This allows us to re-write joint objects while isolating  $DM^i$ 's role: for instance, a joint action  $\mathbf{u} \in \mathbf{U}$  can be re-written as  $\mathbf{u} = (u^i, \mathbf{u}^{-i})$  and a joint policy  $\theta \in \Delta$  can be re-written as  $\theta = (\theta^i, \theta^{-i})$ .

A joint policy  $\theta \in \Delta$  induces a probability measure on sequences of states and joint actions, which we use in defining  $DM^i$ 's cost

$$J_x^i(\theta) := E^\theta \left( \sum_{t \in \mathbb{N}} (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = x \right), \quad \forall x \in \mathbb{X}$$

where  $E^\theta$  denotes that the stochastic process  $\{(x_t, \mathbf{u}_t)\}_{t \in \mathbb{N}}$  is determined by the policy  $\theta$ . Then, each  $DM^i$ 's goal is to select a policy  $\pi^i \in \Delta^i$  to minimize this cost.

**Definition 1.** A policy  $\pi^{*i} \in \Delta^i$  is called a best reply to  $\theta^{-i} \in \Delta^{-i}$  (for  $DM^i$ ) if

$$J_x^i(\pi^{*i}, \theta^{-i}) = \min_{\pi^i \in \Delta^i} J_x^i(\pi^i, \theta^{-i}), \quad \forall x \in \mathbb{X}.$$

Any best reply  $\pi^{*i} \in \Delta^i$  to  $\theta^{-i} \in \Delta^{-i}$  is called a strict best reply with respect to  $(\pi^i, \theta^{-i})$  if

$$J_x^i(\pi^{*i}, \theta^{-i}) < J_x^i(\pi^i, \theta^{-i}), \quad \text{for some } x \in \mathbb{X}.$$

For any fixed  $\theta^{-i} \in \Delta^{-i}$ ,  $DM^i$  faces a stationary MDP; hence,  $DM^i$  always has a deterministic best reply to any  $\theta^{-i} \in \Delta^{-i}$ . We denote the set of deterministic best replies by

$$BR^i(\theta^{-i}) := \{\pi^{*i} \in \Pi^i : \pi^{*i} \text{ is a best reply to } \theta^{-i}\}.$$

We can describe the set  $BR^i(\theta^{-i})$  using the optimal Q-factors for the MDP faced by  $DM^i$  when playing against the policy  $\theta^{-i}$ . The vector of optimal Q-factors for this environment is denoted  $Q_{\theta^{-i}}^{*i} \in \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i}$ . We include the policy  $\theta^{-i}$  in this notation as a reminder that the MDP and optimal Q-factors both depend on the policy used by all other players. Then,  $BR^i(\theta^{-i})$  can be expressed as

$$\begin{aligned} BR^i(\theta^{-i}) \\ = \{\pi^i \in \Pi^i : Q_{\theta^{-i}}^{*i}(x, \pi^i(x)) = \min_{v^i \in \mathbb{U}^i} Q_{\theta^{-i}}^{*i}(x, v^i), \forall x \in \mathbb{X}\}. \end{aligned}$$

**Definition 2.** A joint policy  $\theta^* \in \Delta$  is called a (Markov perfect) equilibrium if  $\theta^{*i}$  is a best reply to  $\theta^{*-i}$ , for all  $i$ .

We denote the set of all Markov perfect equilibrium policies by  $\Delta_{\text{eq}}$  and we denote the set of stationary deterministic

equilibrium policies by  $\Pi_{\text{eq}} := \Delta_{\text{eq}} \cap \Pi$ . In any finite discounted stochastic game, the set  $\Delta_{\text{eq}}$  is non-empty [10]. Note, however, that the set  $\Pi_{\text{eq}}$  may be empty in general stochastic games.

**Definition 3.** A stochastic game is called a stochastic team (or simply a team) if there exists  $c : \mathbb{X} \times \mathbf{U} \rightarrow \mathbb{R}$  and  $\beta \in (0, 1)$  such that

$$c^i = c, \quad \beta^i = \beta, \quad \forall DM^i.$$

**Definition 4.** A joint policy  $\pi^* \in \Pi$  is called team-optimal if

$$J_x^i(\pi^*) = \inf_{\pi \in \Pi} J_x^i(\pi) \quad \forall i, x \in \mathbb{X}. \quad (1)$$

We use  $\Pi_{\text{opt}}$  to denote the set of team optimal policies, which are stationary deterministic policies by definition. It is easy to see that  $\Pi_{\text{opt}}$  may be empty in a general stochastic game but that  $\Pi_{\text{opt}}$  is non-empty in any stochastic team.

**Definition 5.** A stochastic game is called a common interest game if (i)  $\Pi_{\text{opt}}$  is non-empty, and (ii) for any  $\tilde{\pi} \in \Pi \setminus \Pi_{\text{opt}}$ , we have

$$\inf_{\pi \in \Pi} \sum_{x \in \mathbb{X}} J_x^i(\pi) < \sum_{x \in \mathbb{X}} J_x^i(\tilde{\pi}), \quad \forall DM^i.$$

This definition is consistent with the definition of a common interest game introduced in [14] and used in other literature, e.g., [15]. Teams are a proper subclass of common interest games. The repeated game ( $|\mathbb{X}| = 1$ ) with the stage cost functions shown in Figure 1 is a common interest game for  $a, b > 0$  but not a team unless  $a = b$  and  $\beta^1 = \beta^2$ .

		$u_2^2$ :	
		1	2
$u_1^2$ :	1	$a, b$	$a+1, b+1$
2	1	$a+1, b+1$	$-a, -b$

Fig. 1. Stage cost for a two-DM game where  $DM^1$  ( $DM^2$ ) chooses a row (a column) and its cost is the first (the second) entry in the chosen cell.

It is immediate that a team-optimal policy is an equilibrium; however, the converse need not be true. For an illustration of how poorly an equilibrium policy can perform with respect to team-optimality, consider again the repeated game presented in Figure 1 with  $a = b > 0$  and  $\beta^1 = \beta^2 = \beta \in (0, 1)$ . Clearly, the joint policy  $\pi_{\text{sub}} := (1, 1)$  is an equilibrium policy, and so is the team-optimal policy  $\pi^* := (2, 2)$ . We have that  $J^i(\pi_{\text{sub}}) - J^i(\pi^*) = \frac{2a}{1-\beta}$ , for each agent  $i \in \{1, 2\}$ , which shows that the performance gap between an equilibrium policy and a team-optimal policy can be arbitrarily large. This provides the motivation for designing decentralized algorithms that allow agents to learn team-optimal policies, when they exist.

Our objective is the following: given a common interest game, we wish to provide each DM with a decentralized learning algorithm that does not use control sharing and that provably leads, in some appropriate sense, to a team optimal policy.

In [16], we presented an algorithm that leads to equilibrium policies in weakly acyclic games, another class of games (different from common interest games) that generalizes teams. These algorithms instruct DMs to use the same stationary policy, called baseline policies, for a large number of consecutive stages, the collection of which is called an exploration phase. At the end of an exploration phase, DMs update their baseline

policies in a synchronized manner. In this way, the system is stationary for long enough for Q-learning to return meaningful Q-factors. The Q-factors acquired during an exploration phase are used to construct best replies; Q-factors are then reset for the next exploration phase. The DMs use inertial best-responding to update their baseline policies, and it is shown that this process leads to equilibrium policies in weakly acyclic games.

In the next section, we present a decentralized learning algorithm that leads to team-optimal policies, when they exist. The algorithm here uses the exploration phase technique from [16], but modifies the baseline policy update in order to exploit the following structural result on Q-factors in teams and common interest games.

**Lemma 1.** *In a common interest game, for all  $i$ ,  $\pi^* \in \Pi_{\text{opt}}$ ,  $\tilde{\pi} \in \Pi \setminus \Pi_{\text{opt}}$ , we have*

$$\sum_{x \in \mathbb{X}} Q_{\pi^{*-i}}^i(x, \pi^{*i}(x)) < \sum_{x \in \mathbb{X}} Q_{\tilde{\pi}^{-i}}^i(x, \tilde{\pi}^i(x)).$$

This fact provides for us an avenue for separating team-optimal policies from the other policies by focusing on Q-factors.

*Proof.* For all  $i$ ,  $\pi^* \in \Pi_{\text{opt}}$ ,  $\tilde{\pi} \in \Pi \setminus \Pi_{\text{opt}}$ , we have

$$\sum_{x \in \mathbb{X}} Q_{\pi^{*-i}}^i(x, \pi^{*i}(x)) = \sum_{x \in \mathbb{X}} J_x^i(\pi^*) < \sum_{x \in \mathbb{X}} J_x^i(\tilde{\pi})$$

If  $\tilde{\pi}^i \in \text{BR}^i(\tilde{\pi}^{-i})$ , then  $J_x^i(\tilde{\pi}) = Q_{\tilde{\pi}^{-i}}^i(x, \tilde{\pi}^i(x))$ ; otherwise,

$$\sum_{x \in \mathbb{X}} J_x^i(\pi^*) \leq \sum_{x \in \mathbb{X}} \min_{u^i \in \mathbb{U}^i} Q_{\tilde{\pi}^{-i}}^i(x, u^i) < \sum_{x \in \mathbb{X}} Q_{\tilde{\pi}^{-i}}^i(x, \tilde{\pi}^i(x)).$$

□

#### IV. LEARNING TEAM OPTIMALITY

In this section, we introduce a learning algorithm for achieving team optimality in teams and common interest games. To motivate our algorithm, we first study a time-homogenous Markov chain  $\{\pi_k\}_{k \geq 0}$ , taking values in the set of joint stationary deterministic policies  $\Pi$ . The dynamics of this Markov chain will be determined by the Idealized Update Procedure (IUP), detailed in Algorithm 1. While the IUP cannot be implemented in a stochastic common interest game under the information structure of interest, the resulting Markov chain will be used in approximation arguments in the proofs of our main results.

Under *inertial best-responding* with inertia parameter  $\lambda^i \in (0, 1)$ , at time  $k \in \mathbb{N}$ ,  $\text{DM}^i$  checks whether its current policy  $\pi_k^i$  is a best-reply to the policy being used by other players, i.e. it checks if  $\pi_k^i \in \text{BR}^i(\pi_k^{-i})$ ; if it is, then  $\pi_{k+1}^i = \pi_k^i$ . Otherwise  $\text{DM}^i$  is not best-replying and selects

$$\pi_{k+1}^i \sim (1 - \lambda^i) \text{Unif}(\text{BR}^i(\pi_k^{-i})) + \lambda^i \mathbb{I}_{\pi_k^i},$$

that is, switches to a random best-reply with probability  $1 - \lambda^i$  or is inert (does not change away from  $\pi_k^i$ ) with probability  $\lambda^i$ . Including inertia in one's policy update can be used to avoid cycling in best-reply dynamics. For example, in the game in Figure 1, if play starts at either joint policy (1, 2) or at (2, 1) and both players switch to a best-reply at each step, the joint policy will cycle between (1, 2) and (2, 1) perpetually. Such cycling can be avoided by using explicit

coordination mechanisms for determining which DM should change its policy and at what time, but such mechanisms may not be feasible in decentralized settings. Simple decentralized mechanisms such as inertia can be used with the same effect [43], [41].

The condition-dependent nature of inertial best-responding can be captured using a stochastic kernel  $R^{i, \lambda^i} \in \mathcal{P}(\Pi^i | \Pi^i \times 2^{\Pi^i})$ , where  $R^{i, \lambda^i}$  selects a successor policy randomly, conditioning on the current policy and the current (perhaps estimated) best-reply set. To allow for uncertainty of  $\text{BR}^i(\pi_k^{-i})$ , we define  $R^{i, \lambda^i}$  as follows:

$$R^{i, \lambda^i}(\tilde{\pi}^i | \pi^i, B^i) := \begin{cases} 1, & \text{if } \pi^i \in B^i, \tilde{\pi}^i = \pi^i \\ \lambda^i, & \text{if } \pi^i \notin B^i, \tilde{\pi}^i = \pi^i \\ \frac{1 - \lambda^i}{|B^i|}, & \text{if } \pi^i \notin B^i, \tilde{\pi}^i \in B^i \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

for any  $\pi^i \in \Pi^i$ ,  $B^i \in 2^{\Pi^i}$  and  $\tilde{\pi}^i \in \Pi^i$ .

Note that selecting  $\pi_{k+1}^i \sim R^{i, \lambda^i}(\cdot | \pi^i, \text{BR}^i(\pi_k^{-i}))$  is equivalent to selecting  $\pi_{k+1}^i$  according to inertial best-responding with parameter  $\lambda^i$ .

Under the *Idealized Update Procedure (IUP)*, presented in Algorithm 1,  $\text{DM}^i$  chooses  $\pi_{k+1}^i$  according to a mixture of uniform random experimentation and inertial best-responding when the joint policy is team optimal, i.e.  $\pi_k \in \Pi_{\text{opt}}$ . When  $\pi_k \notin \Pi_{\text{opt}}$ ,  $\text{DM}^i$  uses a mixture of uniform random experimenting and a player selected stochastic kernel  $h^i \in \mathcal{P}(\Pi^i | \Pi^i \times 2^{\Pi^i})$  to choose  $\pi_{k+1}^i$ .

#### Algorithm 1: Idealized Update Procedure (IUP) for $\text{DM}^i$

1	<b>Set Parameters</b>
2	$\lambda^i \in [0, 1]$ : inertia probability
3	$h^i \in \mathcal{P}(\Pi^i   \Pi^i \times 2^{\Pi^i})$ , a policy update kernel
4	$\gamma^i, \kappa^i \in (0, 1)$ : exploration probabilities
5	<b>for</b> $k \geq 0$
6	<b>if</b> $\pi_k \in \Pi_{\text{opt}}$ <b>then</b>
	$\pi_{k+1}^i \sim (1 - \gamma^i) R^{i, \lambda^i}(\cdot   \pi_k^i, \text{BR}^i(\pi_k^{-i})) + \gamma^i \text{Unif}(\Pi^i)$
7	<b>else</b> ( $\pi_k \notin \Pi_{\text{opt}}$ )
8	$\pi_{k+1}^i \sim (1 - \kappa^i) h^i(\cdot   \pi_k^i, \text{BR}^i(\pi_k^{-i})) + \kappa^i \text{Unif}(\Pi^i)$
9	<b>end</b>
10	<b>end</b>

We will require that  $\text{DM}^i$  randomly explores  $\Pi^i$  more when the joint policy is not team optimal, i.e.  $\kappa^i \gg \gamma^i$ . Qualitatively, this results in shifting away from suboptimal joint policies more quickly than team optimal policies, and as a result the process spends a large fraction of time in  $\Pi_{\text{opt}}$ . We formalize this intuition below, and note that the guarantee of Lemma 2, on attaining team-optimality in common interest games, holds for arbitrary  $\{h^i\}_{i=1}^N$ . That is,  $\text{DM}^i$  has some flexibility in how it updates its policies when not experimenting and when the current joint policy is not team optimal.

**Lemma 2.** *Consider a common interest game, and suppose each  $\text{DM}^i$  updates its policies according to the IUP in Algorithm 1. Let  $A_{\gamma, \kappa, h}$  denote the matrix of the transition probabilities for the induced time-homogenous Markov chain on  $\Pi$ , where  $\gamma := \{\gamma^i\}_{i=1}^N$ ,  $\kappa := \{\kappa^i\}_{i=1}^N$ ,  $h = \{h^i\}_{i=1}^N$ . We*

denote the associated unique stationary distribution by  $\mu_{\gamma, \kappa, h}^*$ . For any  $\epsilon \in (0, 1)$ ,  $\kappa \in (0, 1)^N$ , there exists  $\bar{\gamma}_\epsilon(\kappa) > 0$  such that if  $\gamma^i \in (0, \bar{\gamma}_\epsilon(\kappa))$  for all  $i$ , then

$$\mu_{\gamma, \kappa, h}^*(\Pi_{\text{opt}}) \geq 1 - \epsilon/2. \quad (3)$$

Moreover, for all  $\mu_0 \in \mathcal{P}(\Pi)$ , we have

$$\lim_{n \rightarrow \infty} \mu_0 A_{\gamma, \kappa, h}^n = \mu_{\gamma, \kappa, h}^*.$$

*Proof.* Since  $\gamma^i, \kappa^i > 0$  for all  $i$ , the induced Markov chain is irreducible, hence there exists unique  $\mu_{\gamma, \kappa, h}^*$  such that  $\mu_{\gamma, \kappa, h}^* = \mu_{\gamma, \kappa, h}^* A_{\gamma, \kappa, h}$ . We have

$$\begin{aligned} & \sum_{\pi^* \in \Pi_{\text{opt}}} \mu_{\gamma, \kappa, h}^*(\pi^*) \\ &= \sum_{\pi^* \in \Pi_{\text{opt}}} \sum_{\pi \in \Pi_{\text{opt}}} \mu_{\gamma, \kappa, h}^*(\pi) A_{\gamma, \kappa, h}(\pi, \pi^*) \\ &+ \sum_{\pi^* \in \Pi_{\text{opt}}} \sum_{\pi \notin \Pi_{\text{opt}}} \mu_{\gamma, \kappa, h}^*(\pi) A_{\gamma, \kappa, h}(\pi, \pi^*) \\ &\geq \sum_{\pi \in \Pi_{\text{opt}}} \mu_{\gamma, \kappa, h}^*(\pi) \prod_i (1 - \gamma^i) \\ &+ \sum_{\pi \notin \Pi_{\text{opt}}} \mu_{\gamma, \kappa, h}^*(\pi) \prod_i (\kappa^i / |\Pi^i|) \end{aligned}$$

This leads to

$$\sum_{\pi^* \in \Pi_{\text{opt}}} \mu_{\gamma, \kappa, h}^*(\pi^*) \geq 1 - \frac{\sum_i \gamma^i}{\sum_i \gamma^i + \prod_i (\kappa^i / |\Pi^i|)}$$

which implies (3). The last part follows from the aperiodicity of the Markov chain.  $\square$

Lemma 2 shows that if DMs follow the IUP, then they would choose a team-optimal policy in the long run with arbitrarily high probability provided the experimentation probabilities of  $\gamma$  are positive but sufficiently small relative to  $\kappa$ .

It is clear that the IUP cannot be directly implemented in our study of decentralized, online teams. The first issue relates to decentralization:  $\text{DM}^i$  cannot observe the policy  $\pi_k^{-i}$ . The second issue relates to the online nature of the problem: even if  $\pi_k^{-i}$  were known,  $\text{DM}^i$  may not know its best-reply set or the set of team optimal policies. Nevertheless, the IUP motivates our decentralized learning algorithm, Algorithm 2, which can be viewed as a two timescale approximation of the IUP. We expand on this point below, after presenting the main result of this section.

We emphasize that Algorithm 2 is decentralized in the sense that it can be implemented by ‘‘independent learners,’’ in the terminology of [29], [32]. That is, each  $\text{DM}^i$  can run a separate copy of this algorithm without reference to the joint actions or policies of the remaining players. We recall that each  $\text{DM}^i$ 's interaction with its environment at any time  $t$  consists of sending its control decision  $u_t^i$  and receiving its cost realization  $c^i(x_t, u_t^1, \dots, u_t^N)$  as well as the next state  $x_{t+1}$  without observing any information about the other DMs, in particular, without observing the control decisions  $\mathbf{u}_t^{-i}$  of the other DMs. In fact, each  $\text{DM}^i$  need not even be aware of the presence of the other DMs or the fact it is engaged in learning in a multi-player game. Simply, each DM is running a single-agent algorithm similar to standard Q-learning (that is re-initialized after its baseline policy is updated at the end of

### Algorithm 2: Independent Team Q-Learning for $\text{DM}^i$

```

1 Set Parameters
2  $\mathbb{Q}^i \subset \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i}$ : a compact set
3  $\{T_k\}_{k \geq 0}$ : a sequence in  $\mathbb{N}_+$  of exploration phase
   lengths (common to all DMs)
4 Set  $t_0 = 0$  and  $t_{k+1} = t_k + T_k$  for all  $k \geq 0$ .
5  $\rho^i \in (0, 1)$ : action experimentation probability
6  $\gamma^i, \kappa^i \in (0, 1)$ : policy experimentation probabilities
7  $\lambda^i \in [0, 1]$ : probability of inertia when updating
   baseline policy
8  $\delta^i > 0$ : tolerance for sub-optimality when
   constructing best-reply sets
9  $d^i > 0$ : a tolerance for sub-optimality when setting
   the aspiration level
10  $W^i \in \mathbb{N}_+$ : a window for setting aspiration levels
11  $h^i \in \mathcal{P}(\Pi^i | \Pi^i \times 2^{\Pi^i})$ , a policy update kernel
12  $\{\alpha_n^i\}_{n \geq 0}$ : step sizes such that  $\alpha_n^i \in [0, 1]$ ,
    $\sum_n \alpha_n^i = \infty$ ,  $\sum_n (\alpha_n^i)^2 < \infty$ 
13 Initialize (arbitrary)  $\pi_0^i \in \Pi^i$ ,  $Q_0^i \in \mathbb{Q}^i$ 
14 Receive  $x_0$ 
15 for  $k \geq 0$  ( $k^{\text{th}}$  exploration phase)
16   for  $t = t_k, t_k + 1, \dots, t_{k+1} - 1$  // Learn
   best-replies for  $k^{\text{th}}$  EP
17     Select  $u_t^i \sim (1 - \rho^i) \mathbb{I}_{\pi_k^i(x_t)} + \rho^i \text{Unif}(\mathbb{U}^i)$ 
18     Receive cost  $c^i(x_t, u_t^i, \mathbf{u}_t^{-i})$ 
19     Receive state  $x_{t+1} \sim P(\cdot | x_t, \mathbf{u}_t)$ 
20     Set  $n_t^i =$  number of visits to  $(x_t, u_t^i)$  in  $[t_k, t]$ 
21      $Q_{t+1}^i(x_t, u_t^i) = (1 - \alpha_{n_t^i}^i) Q_t^i(x_t, u_t^i) +$ 
        $\alpha_{n_t^i}^i [c^i(x_t, u_t^i, \mathbf{u}_t^{-i}) + \beta^i \min_{v^i} Q_t^i(x_{t+1}, v^i)]$ 
22      $Q_{t+1}^i(x, u^i) = Q_t^i(x, u^i), \forall (x, u^i) \neq (x_t, u_t^i)$ 
23   end
24    $\text{BR}_k^i = \{\pi^i \in \Pi^i : Q_{t_{k+1}}^i(x, \pi^i(x)) \leq$ 
        $\min_{v^i} Q_{t_{k+1}}^i(x, v^i) + \delta^i, \forall x \in \mathbb{X}\}$ 
25    $S_k^i = \sum_{x \in \mathbb{X}} Q_{t_{k+1}}^i(x, \pi_k^i(x))$ 
26    $\Lambda_k^i = \min\{S_{k-1}^i, \dots, S_{(k-W^i)^+}^i\} + d^i$ 
27   if  $S_k^i \leq \Lambda_k^i$  then
        $\pi_{k+1}^i \sim (1 - \gamma^i) R^{i, \lambda^i}(\cdot | \pi_k^i, \text{BR}_k^i) + \gamma^i \text{Unif}(\Pi^i)$ 
28   else ( $S_k^i > \Lambda_k^i$ , not achieving aspiration)
        $\pi_{k+1}^i \sim (1 - \kappa^i) h^i(\cdot | \pi_k^i, \text{BR}_k^i) + \kappa^i \text{Unif}(\Pi^i)$ 
29   end
30   end
31   Reset  $Q_{t_{k+1}}^i$  to any  $Q^i \in \mathbb{Q}^i$  (e.g., project onto  $\mathbb{Q}^i$ )
32 end

```

each exploration phase). As such, all quantities computed by  $\text{DM}^i$ 's copy of Algorithm 2 are indexed by  $i$ . These remarks also apply verbatim to Algorithm 3 introduced in Section VI.

**Assumption 1.** For all  $x, x' \in \mathbb{X}$ , there exists  $H \in \mathbb{N}$  and  $\tilde{u}_0, \dots, \tilde{u}_H \in \mathbb{U}$  such that

$$\Pr(x_{H+1} = x' | x_0 = x, \mathbf{u}_j = \tilde{u}_j, \forall j \in \{0, 1, \dots, H\}) > 0.$$

**Assumption 2.** Assume, for all  $i$ ,  $\delta^i \in (0, \bar{\delta})$ ,  $d^i \in (0, \bar{d})$ ,  $\rho^i \in (0, \bar{\rho})$ , where  $\bar{\delta}, \bar{d}, \bar{\rho}$  are constants defined in Appendix A that depend only on the game.

**Theorem 1.** Consider a common interest game in which each  $DM^i$  uses Algorithm 2, and let Assumptions 1-2 hold. For any  $\epsilon > 0$ , there exist

$$\bar{\gamma}_\epsilon(\boldsymbol{\kappa}) \in (0, 1), \quad \bar{W}_\epsilon(\boldsymbol{\gamma}, \boldsymbol{\kappa}) \in \mathbb{N}_+, \quad \bar{T}_\epsilon(\boldsymbol{\gamma}, \boldsymbol{\kappa}, W_{\max}) \in \mathbb{N}_+$$

where  $W_{\max} := \max_i W^i$  such that if, for all  $i, k \in \mathbb{N}$

$$\gamma^i \in (0, \bar{\gamma}_\epsilon(\boldsymbol{\kappa})), \quad W^i \geq \bar{W}_\epsilon(\boldsymbol{\gamma}, \boldsymbol{\kappa}), \quad T_k \geq \bar{T}_\epsilon(\boldsymbol{\gamma}, \boldsymbol{\kappa}, W_{\max})$$

then

$$\liminf_{k \in \mathbb{N}} Pr(\boldsymbol{\pi}_k \in \boldsymbol{\Pi}_{\text{opt}}) \geq 1 - \epsilon$$

*Proof.* See Appendix A.  $\square$

### Discussion

Algorithm 2 can be viewed as a two timescale<sup>2</sup> approximation to the IUP in Algorithm 1. The faster timescale is that where time is indexed by the stage games, comprising lines 16-23 of Algorithm 2. The selection of actions, observation of costs and state transitions, and Q-factor updates all occur on this faster timescale. In contrast, the slower timescale is where time is indexed by the exploration phase. Decisions on the slower timescale involve processing learned Q-factors to estimate one's best-reply set (line 24), computing a "cost score" and comparing it to historical cost scores (lines 25-27), and updating one's baseline policy (lines 27-30).

We note that  $DM^i$  only uses its learned Q-factors at times  $\{t_{k+1}^i\}_{k \geq 0}$  in the selection of the policy  $\pi_{k+1}^i$ . As such, we are only interested in the sequence  $\{Q_t^i\}_{t \geq 0}$  sampled at times  $\{t_{k+1}^i\}_{k \geq 0}$ . In particular, we are concerned with the approximation of  $Q_{\pi_k^{-i}}^*$  by  $Q_{t_{k+1}^i}^i$ . Crucially, the baseline policies are fixed within an exploration phase and only change between exploration phases. This means that for any  $k \geq 0$ , from the point of view of any  $DM^i$ , the environment is stationary within the  $k^{\text{th}}$  EP and equivalent, during the interval  $[t_k, t_k + T_k - 1]$ , to an MDP determined by  $\pi_k^{-i}$ . It was shown in [16] that under certain conditions—satisfied here by Assumptions 1 and 2—that Q-learning within an EP leads to informative Q-factors that can be used, among other things, to recover one's best-reply set  $BR^i(\pi_k^{-i})$  with high probability. After the policy update suite (lines 27-30),  $DM^i$  resets its Q-factors and its counters ahead of the  $(k+1)^{\text{th}}$  EP, and the Q-learning process restarts. We make no claims about the asymptotic behaviour of the entire sequence  $\{Q_t^i\}_{t \geq 0}$ , as this is not needed for the analysis of policy updating.

The analogy between Algorithm 2 and the IUP can be seen by comparing the if-suite (lines 6-9) in Algorithm 1 with its counterpart (lines 27-30) in Algorithm 2. The unobservable condition  $\pi_k \in \boldsymbol{\Pi}_{\text{opt}}$  of the IUP has been replaced by a surrogate condition  $S_k^i \leq \Lambda_k^i$ . Here,  $S_k^i$  is a "cost score," which aggregates  $DM^i$ 's policy performance across all states for the  $k^{\text{th}}$  exploration phase, and  $\Lambda_k^i$  is a measure of  $DM^i$ 's best performance during the preceding  $W^i$  exploration phases.

<sup>2</sup>In two timescale algorithms in the literature (e.g. [54]), both the Q-factors and the policies would be updated incrementally at each time  $t = 1, 2, \dots$ . The step size sequences for Q-learning and policy updating would be selected so that policies are effectively fixed while Q-factors are learned. In our algorithms, the policies are updated without using any step sizes but only only at  $t = t_1 - 1, t_2 - 1, \dots$  whereas the Q-factors are updated at each time  $t = 1, 2, \dots$  using step sizes that are re-initialized at  $t = t_1 - 1, t_2 - 1, \dots$  and are reduced during  $t \in [t_k, t_{k+1} - 1]$  at a rate satisfying the assumptions of the standard (i.e., one time scale) stochastic approximation theory.

Importantly, the condition  $S_k^i \leq \Lambda_k^i$  can be verified by independent learners.

Algorithm 2 is in the spirit of aspiration learning algorithms [40], where  $\Lambda_k^i$  plays the role of  $DM^i$ 's aspiration level, a scalar quantity against which  $DM^i$  compares the performance of its policy  $\pi_k^i$  during the  $k^{\text{th}}$  exploration phase. Each  $DM^i$  aspires to perform at least as well as its aspiration level, which is updated at the end of each exploration phase and may be thought of as a maximum tolerable cost, i.e., if the baseline policy yields higher cost, then it is viewed as unsatisfactory.

Unlike the aspiration learning methods in the literature, which focus on repeated games with no state dynamics and players with no look ahead, Algorithm 2 is designed for stochastic dynamic games with nontrivial state dynamics and far-sighted players. Due to the long-run cost considerations in dynamic stochastic games, evaluating of the cost of a policy is a slow and noisy process, which leads to additional difficulties in setting the aspiration levels.

In light of Lemma 1, a viable approach is to use the learned Q-factors to produce cost scores and to set the aspiration levels to the minimum cost score over some window of the past. However, scores obtained from the (random) Q-factors are noisy estimates of the scores corresponding to the true cost of the policies. In particular, setting the aspiration levels to the minimum of the cost scores over the entire past based on the learned Q-factors can result in unattainable aspiration levels. Hence, to mitigate the effects of the noise present in the learned Q-factors, we set the aspiration levels of each  $DM^i$  to the minimum cost score obtained over a finite window of the most recent past within some tolerance. This allows DMs to discard unattainable cost scores in finite time.

Another aspect of Algorithm 2 is the persistent experimentation in the policy space. Experimentation when DMs feel that they meet their aspiration levels ( $S_k^i \leq \Lambda_k^i$ ) is required to prevent DMs settling in a policy that is not team-optimal. This is due to the finite window approach used for setting the aspiration levels and the possibility of setting suboptimal aspiration levels. Experimentation when  $S_k^i > \Lambda_k^i$  is also necessary to aid DMs in searching for team-optimal policies.

Finally, we note that the set of approximate best responses  $BR_k^i$  computed by each  $DM^i$  within each exploration phase  $k$  is a subset of  $\Pi^i$ , the set of stationary and deterministic policies of  $DM^i$ . Therefore,  $|BR_k^i| \leq |\Pi^i| = |\mathbb{U}^i|^{|X^i|}$ . We note that  $BR_k^i$  is computed via the Q-factors  $Q_{t_{k+1}^i}^i \in \mathbb{R}^{X^i \times U^i}$ , which is of size  $|X^i| |U^i|$ .

## V. BEYOND TEAM OPTIMALITY: APPLICATION TO WEAKLY ACYCLIC GAMES

In this section, we consider a special case of Algorithm 2 that has desirable convergence properties in weakly acyclic games, in addition to providing team-optimality in the sense of Theorem 1.

**Definition 6.** A (possibly finite) sequence  $\pi_0, \pi_1, \dots$  in  $\boldsymbol{\Pi}$  is called a multi-DM strict best reply path if, for each  $k$ ,  $\pi_k$  and  $\pi_{k+1}$  differ for at least one DM and, for each deviating  $DM^i$ ,  $\pi_{k+1}^i$  is a strict best reply with respect to  $\pi_k$ .

**Definition 7.** A stochastic game is called weakly acyclic under multi-DM strict best replies (or simply weakly acyclic) if there is a multi-DM strict best reply path starting from



each deterministic joint policy and ending at a deterministic equilibrium policy.

The notion of weak acyclicity used here is with respect to stationary deterministic policies for stochastic games, and generalizes the notion of weak acyclicity introduced in [55] for single-stage games. All teams are weakly acyclic; however, a common interest game need not be. See [56] for other examples of single-stage weakly acyclic games.

In weakly acyclic games, inertial best reply dynamics [16] lead to equilibrium policies. If the policy update functions satisfy  $h^i = R^{i,\lambda^i}$  for each  $DM^i$ , the IUP introduced in the previous section can be regarded as a perturbed inertial best reply dynamics, where  $\{\pi_k \in \Pi_{\text{opt}}\}$  can be replaced with any arbitrary event if the game is not a common interest game, provided the induced Markov chain is time-homogenous.

**Assumption 3.** For every  $DM^i$ ,  $h^i = R^{i,\lambda^i}$ , where  $\lambda^i \in (0, 1)$ .

Under Assumption 3, each  $DM^i$  always best replies with inertia when not experimenting.

**Lemma 3.** Consider a weakly acyclic game. Suppose that each  $DM^i$  updates its policy according to the IUP of Algorithm 1, and let Assumption 3 hold. Let  $A_{\gamma,\kappa}$  denote the matrix of the transition probabilities for the induced time homogenous Markov chain on  $\Pi$ . Denote the unique stationary distribution associated to this Markov chain by  $\mu_{\gamma,\kappa}^*$ . For any  $\epsilon > 0$ , there exists  $\bar{\kappa}_\epsilon \in (0, 1)$  such that  $\max\{\gamma^i, \kappa^i\} \in (0, \bar{\kappa}_\epsilon)$ , for all  $i$ , implies

$$\mu_{\gamma,\kappa}^*(\Pi_{\text{eq}}) \geq 1 - \epsilon/4.$$

Moreover, uniformly over all such  $\gamma, \kappa$ , there exists  $\bar{m} \in \mathbb{N}$  such that

$$\inf_{m \geq \bar{m}, \mu_0 \in \mathcal{P}(\Pi)} (\mu_0 A_{\gamma,\kappa}^m)(\Pi_{\text{eq}}) \geq 1 - \epsilon/2.$$

*Proof.* For all  $\pi^* \in \Pi_{\text{eq}}$ ,

$$A_{\gamma,\kappa}(\pi^*, \pi^*) \geq \prod_i (1 - \max\{\gamma^i, \kappa^i\}) \quad (4)$$

Let  $L_\pi < |\Pi|$  be the length of a multi-DM strict best reply path of minimal length from  $\pi \in \Pi \setminus \Pi_{\text{eq}}$  to some  $\tilde{\pi} \in \Pi_{\text{eq}}$ , and  $L := \max_{\pi \in \Pi \setminus \Pi_{\text{eq}}} L_\pi$ . For any  $\pi \notin \Pi_{\text{eq}}$ , consider a path  $\pi = \pi_0, \pi_1, \dots, \pi_L$  where  $\pi_0, \pi_1, \dots, \pi_{L_\pi}$  is a multi-DM strict best reply path and  $\pi_{L_\pi} = \dots = \pi_L = \tilde{\pi} \in \Pi_{\text{eq}}$ . In each transition  $\pi_k \rightarrow \pi_{k+1}$ , some DMs switch to one of their strict best replies and the others stay put. Therefore, from any  $\pi \notin \Pi_{\text{eq}}$ , the IUP with  $\gamma = \kappa \equiv 0$  generates such a path  $\pi_0, \pi_1, \dots, \pi_L$  with probability at least  $p_{\min} := \prod_{i=1}^N \min\{\lambda^i, (1 - \lambda^i)/|\Pi^i|\}^L \in (0, 1)$ . By taking  $\gamma^i > 0$ ,  $\kappa^i > 0$  into account, this leads to

$$\sum_{\tilde{\pi} \in \Pi_{\text{eq}}} (A_{\gamma,\kappa})^L(\pi, \tilde{\pi}) \geq p_{\min} \prod_i (1 - \max\{\gamma^i, \kappa^i\})^L \quad (5)$$

for all  $\pi \in \Pi \setminus \Pi_{\text{eq}}$ . Writing  $A = A_{\gamma,\kappa}$ , from (4)-(5), we have, for all  $k \in \mathbb{N}$ ,

$$\begin{aligned} (\mu_0 A^{k+L})(\Pi \setminus \Pi_{\text{eq}}) &\leq L \sum_i \max\{\gamma^i, \kappa^i\} \\ &\quad + (\mu_0 A^k)(\Pi \setminus \Pi_{\text{eq}})(1 - p_{\min}). \end{aligned}$$

This leads to, for all  $j, k \in \mathbb{N}$ ,

$$(\mu_0 A^{k+jL})(\Pi \setminus \Pi_{\text{eq}}) \leq L \sum_i \max\{\gamma^i, \kappa^i\} / p_{\min} + (1 - p_{\min})^j.$$

Since  $|1 - p_{\min}| < 1$ , the desired result follows.  $\square$

For small experimentation probabilities, the IUP under Assumption 3 leads to equilibrium policies in the long run. We will use this to show that Algorithm 2 under Assumptions 1-3 has the same long run behavior.

For weakly acyclic games, decentralized learning algorithms which assign arbitrarily high probabilities to equilibrium policies in the long run are presented in [16]. However, these algorithms do not provide any guarantee on achieving team-optimality when implemented in teams or common interest games. We now strengthen a result of [16] with respect to team-optimality.

**Theorem 2.** Consider a weakly acyclic game in which each  $DM^i$  uses Algorithm 2, and let Assumptions 1-3 hold. For any  $\epsilon > 0$ , there exist

$$\begin{aligned} \bar{\kappa}_\epsilon \in (0, 1), \quad \tilde{\gamma}_\epsilon(\kappa) \in (0, 1), \\ \tilde{W}_\epsilon(\gamma, \kappa) \in \mathbb{N}_+, \quad \tilde{T}_\epsilon(\gamma, \kappa, W_{\max}) \in \mathbb{N}_+, \end{aligned}$$

where  $W_{\max} = \max_i W^i$ , such that if, for all  $i, k \in \mathbb{N}$ ,

$$\begin{aligned} \kappa^i \in (0, \bar{\kappa}_\epsilon), \quad \gamma^i \in (0, \tilde{\gamma}_\epsilon(\kappa)), \\ W^i \geq \tilde{W}_\epsilon(\gamma, \kappa), \quad T_k \geq \tilde{T}_\epsilon(\gamma, \kappa, W_{\max}) \end{aligned}$$

then

$$\liminf_{k \in \mathbb{N}} Pr(\pi_k \in \Pi_{\text{eq}}) \geq 1 - \epsilon. \quad (6)$$

Moreover, if the game is a common interest game, then  $\Pi_{\text{eq}}$  can be replaced by  $\Pi_{\text{opt}}$  in (6).

*Proof.* See Appendix B.  $\square$

## VI. LEARNING WITH CONSTANT ASPIRATIONS

In this section, we introduce Algorithm 3, a variant of Algorithm 2 in which every  $DM^i$  employs a constant aspiration level  $\Lambda^i \in \mathbb{R}$  throughout, i.e.,  $\Lambda_k^i = \Lambda^i$  for every exploration phase  $k \in \mathbb{N}$ . Pre-setting the aspiration levels is motivated by applications where each DM has the prior knowledge of a conservative estimate of its achievable cost. Such prior knowledge may be available to DMs, for example, from previous experience or through an initial phase of experimentation, and can be used to heuristically discern “good” from “bad” performance. One implication of this assumption is that if there is indeed a set of joint policies each simultaneously outperforming all pre-set aspiration levels (i.e., the cost estimates) and the other joint policies fail to satisfy any DM, we show that DMs using Algorithm 3 will almost surely outperform their aspiration levels in the long run (part (1)-(2) of Theorem 3). This is the case, for example, in a common interest game when the aspiration levels are between the dominant costs and the other costs. In contrast, DMs using Algorithm 2 adaptively adjust their aspiration levels and achieve optimal performance but only in common interest games and in the weaker sense of eventually assigning arbitrarily high probability to the set of optimal policies (Theorem 1). In addition, unlike in Algorithm 2, we characterize the long term behavior of Algorithm 3 in all games regardless of whether or not the pre-set aspiration levels are achievable. Loosely speaking, DMs using Algorithm 3 in any game are likely to use a certain minimal set of policies in the long run, which are closed under multi-agent strict best-replies (part (3)-(4) of

Theorem 3). This minimal set of policies reduces to the set  $\Pi_{\text{eq}}$  of equilibrium policies in any weakly-acyclic game. Thus, in part (3)-(4) of Theorem 3, we characterize the long-term behavior of Algorithm 3 in a manner analogous to and, in fact, more general than Theorem 2, which characterizes the long-term behavior of Algorithm 2 as  $\Pi_{\text{eq}}$  in weakly acyclic games.

**Algorithm 3:** for DM<sup>i</sup>

```

1 Set Parameters
2  $\mathbb{J}^i \subset \mathbb{R}^X, \mathbb{Q}^i \subset \mathbb{R}^{X \times U^i}$ : compact sets
3  $\{T_k\}_{k \geq 0}$ : a sequence in  $\mathbb{N}_+$  of exploration phase lengths (common to all DMs)
4 Set  $t_0 = 0$  and  $t_{k+1} = t_k + T_k$  for all  $k \geq 0$ .
5  $\rho^i \in (0, 1)$ : action experimentation probability
6  $\delta^i > 0$ : tolerance for sub-optimality when constructing best-reply sets
7  $\{\gamma_n^i\}_{n \in \mathbb{N}}, \kappa^i$ : policy experimentation probabilities
8  $\lambda^i \in (0, 1)$ : inertia parameter for policy update
9  $g^i, h^i \in \mathcal{P}(\Pi^i | \Pi^i \times 2^{\Pi^i})$ : policy update kernels
10  $\Lambda^i \in \mathbb{R}$ : an aspiration level
11  $\{\alpha_n^i\}_{n \geq 0}$ : step sizes such that  $\alpha_n^i \in [0, 1]$ ,  $\sum_n \alpha_n^i = \infty, \sum_n (\alpha_n^i)^2 < \infty$ 
12 Initialize (arbitrary)  $\pi_0^i \in \Pi^i, J_0^i \in \mathbb{J}^i, Q_0^i \in \mathbb{Q}^i$ 
13 Receive  $x_0$ 
14 for  $k \geq 0$  ( $k^{\text{th}}$  exploration phase)
15   for  $t = t_k, t_k + 1, \dots, t_{k+1} - 1$ 
16     Select  $u_t^i \sim (1 - \rho^i)\mathbb{I}_{\pi_k^i(x_t)} + \rho^i \text{Unif}(U^i)$ 
17     Receive cost  $c^i(x_t, u_t^i, \mathbf{u}_t^{-i})$ 
18     Receive state  $x_{t+1} \sim P(\cdot | x_t, \mathbf{u}_t)$ 
19     Set  $m_t^i =$  number of visits to  $x_t$  in  $[t_k, t]$ 
20      $J_{t+1}^i(x_t) = (1 - \alpha_{m_t^i}^i)J_t^i(x_t) + \alpha_{m_t^i}^i(c^i(x_t, u_t^i) + \beta^i J_t^i(x_{t+1}))$ 
21      $J_{t+1}^i(x) = J_t^i(x), \forall x \neq x_t$ 
22     Set  $n_t^i =$  number of visits to  $(x_t, u_t^i)$  in  $[t_k, t]$ 
23      $Q_{t+1}^i(x_t, u_t^i) = (1 - \alpha_{n_t^i}^i)Q_t^i(x_t, u_t^i) + \alpha_{n_t^i}^i[c^i(x_t, u_t^i, \mathbf{u}_t^{-i}) + \beta^i \min_{v^i} Q_t^i(x_{t+1}, v^i)]$ 
24      $Q_{t+1}^i(x, u^i) = Q_t^i(x, u^i), \forall (x, u^i) \neq (x_t, u_t^i)$ 
25   end
26    $\text{BR}_k^i = \{\pi^i \in \Pi^i : Q_{t_{k+1}}^i(x, \pi^i(x)) \leq \min_{v^i} Q_{t_{k+1}}^i(x, v^i) + \delta^i, \forall x \in \mathbb{X}\}$ 
27    $\tilde{S}_k^i = \sum_{x \in \mathbb{X}} J_{t_{k+1}}^i(x)$ 
28   if  $\tilde{S}_k^i \leq \Lambda^i$  then
29      $\pi_{k+1}^i \sim (1 - \gamma_k^i)g^i(\cdot | \pi_k^i, \text{BR}_k^i) + \gamma_k^i \text{Unif}(\Pi^i)$ 
30   else
31      $\pi_{k+1}^i \sim (1 - \kappa^i)h^i(\cdot | \pi_k^i, \text{BR}_k^i) + \kappa^i \text{Unif}(\Pi^i)$ 
32   end
33   end
34   end
35   end
36   end
37   end
38   end
39   end
40   end
41   end
42   end
43   end
44   end
45   end
46   end
47   end
48   end
49   end
50   end
51   end
52   end
53   end
54   end
55   end
56   end
57   end
58   end
59   end
60   end
61   end
62   end
63   end
64   end
65   end
66   end
67   end
68   end
69   end
70   end
71   end
72   end
73   end
74   end
75   end
76   end
77   end
78   end
79   end
80   end
81   end
82   end
83   end
84   end
85   end
86   end
87   end
88   end
89   end
90   end
91   end
92   end
93   end
94   end
95   end
96   end
97   end
98   end
99   end
100  end

```

The following definitions are introduced to describe the long-term behavior of Algorithm 3.

**Definition 8.** For any  $i, \eta \in \Delta, \pi \in \Pi, \Lambda \in \mathbb{R}^N$ , let

$$\tilde{S}^i(\eta) := \sum_x J_x^i(\eta).$$

(i) Let  $\widetilde{\text{BR}}(\pi) :=$

$$\{\tilde{\pi} \in \Pi : \tilde{\pi}^i \neq \pi^i \Rightarrow \tilde{\pi}^i \text{ is a strict best reply to } \pi, \forall i\}.$$

A nonempty set of policies  $\tilde{\Pi} \subset \Pi$  is closed under multi-DM strict best replies, or a *cumber set*, if

$$\pi \in \tilde{\Pi} \Rightarrow \widetilde{\text{BR}}(\pi) \subset \tilde{\Pi}.$$

A *cumber set* is minimal if it does not properly contain another *cumber set*.

(iii) Let

$$\widetilde{\text{BR}}^\Lambda(\pi) := \{\tilde{\pi} \in \Pi : \tilde{\pi}^i \neq \pi^i \Rightarrow \tilde{S}^i(\tilde{\pi}) > \Lambda^i \text{ and } \tilde{\pi}^i \text{ is a strict best reply to } \pi, \forall i\}.$$

A nonempty set of policies  $\tilde{\Pi} \subset \Pi$  is closed under multi-DM strict best replies with aspiration levels  $\Lambda = \{\Lambda^i\}_{i=1}^N$ , or a  $\Lambda$ -*cumber set*, if

$$\pi \in \tilde{\Pi} \Rightarrow \widetilde{\text{BR}}^\Lambda(\pi) \subset \tilde{\Pi}.$$

A  $\Lambda$ -*cumber set* is minimal if it does not properly contain another  $\Lambda$ -*cumber set*.

Let  $\Pi_{\text{cumber}}$  and  $\Pi_{\text{cumber}}^\Lambda$  denote the union of minimal *cumber sets* and the union of  $\Lambda$ -minimal *cumber sets*, respectively.

The repeated game ( $|\mathbb{X}| = 1$ ) with the stage cost functions shown in Figure 2 is a common interest game for  $\beta^1 = \beta^2$ . The minimal *cumber sets* are  $\{(1,1), (2,1), (2,2), (1,2)\}$  (which is also a strict best-reply path) and  $\{(3,3)\}$ , which are also the minimal  $\Lambda$ -*cumber sets* for  $\Lambda^1 = \Lambda^2 < 7$ . For  $\Lambda^1 = \Lambda^2 \in [7, 10)$ , there are three minimal  $\Lambda$ -*cumber sets*:  $\{(2,1)\}, \{(1,2)\}$ , and  $\{(3,3)\}$ . For  $\Lambda^1 = \Lambda^2 \in [10, 20)$ , there are five minimal  $\Lambda$ -*cumber sets*:  $\{(1,1)\}, \{(2,1)\}, \{(2,2)\}, \{(1,2)\}$ , and  $\{(3,3)\}$ . For  $\Lambda^1 = \Lambda^2 \geq 20$ , any singleton  $\{\pi\}$ , where  $\pi \in \Pi$ , is a minimal  $\Lambda$ -*cumber set*. On the other hand, for  $\Lambda^1 \geq 10, \Lambda^2 < 7$ , the minimal  $\Lambda$ -*cumber sets* are  $\{(1,1)\}, \{(2,2)\}$ , and  $\{(3,3)\}$ .

		$u_t^2$ :		
		1	2	3
$u_t^1$ :	1	10, 3	5, 7	20, 20
	2	5, 7	10, 3	20, 20
	3	20, 20	20, 20	0, 0

Fig. 2. Stage cost for a two-DM game where DM<sup>1</sup> (DM<sup>2</sup>) chooses a row (a column) and its cost is the first (the second) entry in the chosen cell.

Allowing only single-DM best replies in the definition of a *cumber set* results in the notion of a *cusber set* introduced in [45]. The following are true, for any  $\Lambda \in \mathbb{R}^N$ .

- $\Pi$  is both a *cumber set* and a  $\Lambda$ -*cumber set*.
- $\pi \in \Pi_{\text{eq}} \Leftrightarrow \{\pi\}$  is a (minimal) *cumber set*.
- $\pi \in \Pi_{\text{eq}} \Rightarrow \{\pi\}$  is a (minimal)  $\Lambda$ -*cumber set*.
- $(\pi \in \Pi, \tilde{S}^i(\pi) \leq \Lambda^i, \forall i) \Rightarrow \{\pi\}$  is a (minimal)  $\Lambda$ -*cumber set*.
- There is a multi-DM strict best reply path from any  $\pi \in \Pi \setminus \Pi_{\text{cumber}}$  to  $\Pi_{\text{cumber}}$ .
- There is a multi-DM strict best reply path from any  $\pi \in \Pi \setminus \Pi_{\text{cumber}}^\Lambda$  to  $\Pi_{\text{cumber}}^\Lambda$ .
- $\Pi_{\text{cumber}} = \Pi_{\text{eq}} \Leftrightarrow$  the game is weakly acyclic under multi-DM strict best replies.

Let  $\bar{L}_\pi < |\Pi|$  be the length of a multi-DM strict best reply path of minimal length from  $\pi \in \Pi \setminus \Pi_{\text{cumber}}$  to some  $\tilde{\pi} \in \Pi_{\text{cumber}}$ , and  $\bar{L} := \max_{\pi \in \Pi \setminus \Pi_{\text{cumber}}} \bar{L}_\pi$ .

**Assumption 4.** Assume, for all  $i$ ,  $\delta^i \in (0, \bar{\delta})$ ,  $\rho^i \in (0, \rho^\Lambda)$ , where  $\bar{\delta}$ ,  $\rho^\Lambda$  are constants defined in Appendix A, C, respectively ( $\bar{\delta}$  depends only on the game, whereas,  $\rho^\Lambda$  depends on the game and  $\Lambda$ ). Assume further that, for all  $i$ ,  $n \in \mathbb{N}$ ,  $\gamma_n^i \in [0, 1]$ ,  $\sum_{n \in \mathbb{N}} \gamma_n^i < \infty$ , and  $\kappa^i \in (0, 1)$ .

**Theorem 3.** Consider a discounted stochastic game where each  $DM^i$  updates its policies by Algorithm 3, and let Assumptions 1 and 4 hold.

1) Suppose that  $g^i = R^{i,1}$ ,  $\forall i$ , and that there exists a nonempty set  $\Pi^\Lambda \subset \Pi$  satisfying

$$\tilde{S}^i(\pi^*) < \Lambda^i < \tilde{S}^i(\tilde{\pi}), \forall i, \pi^* \in \Pi^\Lambda, \tilde{\pi} \in \Pi \setminus \Pi^\Lambda. \quad (7)$$

Then, there exist  $\tilde{T}_k \in \mathbb{N}_+$ ,  $k \in \mathbb{N}$ , such that if  $T_k \geq \tilde{T}_k$ ,  $\forall k$ , then

$$Pr(\pi_k \rightarrow \pi^*, \text{ for some } \pi^* \in \Pi^\Lambda) = 1.$$

2) Suppose that  $g^i = R^{i,\lambda^i}$ ,  $\forall i$ , and that there exists a cumber set  $\Pi^\Lambda$  satisfying (7). Then there exists  $\tilde{T}_k \in \mathbb{N}_+$ ,  $k \in \mathbb{N}$ , such that if  $T_k \geq \tilde{T}_k$ ,  $\forall k$ , then

$$Pr(\pi_k \rightarrow \Pi^*, \text{ for a minimal cumber set } \Pi^* \subset \Pi^\Lambda) = 1.$$

3) Suppose that  $g^i = R^{i,1}$ ,  $h^i = R^{i,\lambda^i}$ ,  $\forall i$ . Then,

$$\liminf_{k \in \mathbb{N}} Pr(\pi_k \in \Pi_{\text{cumber}}^\Lambda) \geq 1 - (\bar{L}/\bar{p}_{\min}) \sum_i \kappa^i$$

for some  $\bar{p}_{\min} \in (0, 1)$  which is independent of  $\sum_i \kappa^i$ .

4) Suppose that  $g^i = h^i = R^{i,\lambda^i}$ ,  $\forall i$ . Then,

$$\liminf_{k \in \mathbb{N}} Pr(\pi_k \in \Pi_{\text{cumber}}) \geq 1 - (\bar{L}/\bar{p}_{\min}) \sum_i \kappa^i$$

for some  $\bar{p}_{\min} \in (0, 1)$  which is independent of  $\sum_i \kappa^i$ .

*Proof.* See Appendix C.  $\square$

Algorithm 3 prescribes each  $DM^i$  to update its policy differently (using the policy update kernels  $g^i$  or  $h^i$  coupled with different experimentation probabilities  $\gamma_k^i$  or  $\kappa^i$ ) depending on  $DM^i$ 's assessment of whether its aspiration is achieved or not. The experimentation probability needs to vanish asymptotically for the former case but be positive throughout<sup>3</sup> for the latter case. In practice, the experimentation probabilities for either case are envisioned to be (asymptotically) small so that the policy updates are primarily governed by  $g^i$  and  $h^i$ . With this in mind, Theorem 3 can be interpreted as follows.

The first part of Theorem 3 assumes (i) each  $DM^i$  stays with its policy when it assesses that its aspiration is achieved, (ii) each policy  $\pi \in \Pi$  either simultaneously achieves every  $DM^i$ 's aspiration (i.e.,  $\pi \in \Pi^\Lambda$ ) or not a single  $DM^i$ 's aspiration (i.e.,  $\pi \notin \Pi^\Lambda$ ). With this (and regardless of  $h^i$ ),  $DM^i$ s converge almost surely to an aspiration achieving joint policy. Note that this does not rule out convergence to a strictly dominated policy.

The second part assumes that the aspiration achieving policies are closed under multi-DM strict best replies. That is, it assumes  $\Pi^\Lambda$  is a cumber set. Under this condition (and regardless of  $h^i$ ),  $DM^i$ s converge almost surely to a subset of the aspiration achieving joint policies, which is a minimal cumber set. Note that this rules out neither persistent oscillations within a minimal cumber set (inside the aspiration

achieving policies) nor convergence to a set of strictly dominated policies. However, in a weakly-acyclic game (under multi-DM strict best replies), convergence to an aspiration achieving equilibrium is guaranteed; in particular, the equilibrium policies not achieving  $DM^i$ 's aspirations are ruled out. This implies convergence to an optimal policy in teams if the aspiration levels are between the cost of suboptimal and optimal equilibria. If  $\Pi^\Lambda$  is not a cumber set,  $DM^i$ s can leave  $\Pi^\Lambda$  through multi-DM strict best replies and the result may not hold.

Theorem 3 also predicts the long-term behavior of Algorithm 3 when the joint policies  $\Pi$  cannot be partitioned as aspiration achieving policies ( $\Pi^\Lambda$ ) and the other policies in the sense of (7). The third part of Theorem 3 assumes that each  $DM^i$  stays with its policy when its aspiration is achieved, otherwise best replies with inertia, i.e.,  $g^i = R^{i,1}$ ,  $h^i = R^{i,\lambda^i}$ . With this (and regardless of the game),  $DM^i$ 's long-term probability of choosing a policy in a minimal  $\Lambda$ -cumber set (a minimal set that  $DM^i$ s cannot exit through the strict best replies of those whose aspirations are not achieved) can be arbitrarily close to one if the experimentation probabilities are sufficiently small. The fourth part assumes that each  $DM^i$  always best replies with inertia when it is not experimenting, i.e.,  $g^i = h^i = R^{i,\lambda^i}$ . With this, and regardless of the game,  $DM^i$ s tend to choose policies in a minimal cumber set (the equilibria and the minimal multi DM strict best reply cycles) for small experimentation probabilities. Under the conditions of the third or the fourth part,  $DM^i$ s may not consistently achieve their aspirations.

## VII. A SIMULATION STUDY

We consider the following two DM stochastic team with  $\mathbb{U}^1 = \mathbb{U}^2 = \mathbb{X} = \{1, 2\}$  and common discount factor  $\beta = 0.8$ . The stage cost for each state is presented in Figure 3.

		$u_2^2:$				$u_2^1:$	
		1	2			1	2
$u_1^1:$	1	1, 1	3, 3	2	$u_1^2:$	10, 10	10, 10
	2	3, 3	1, 1			10, 10	13, 13
		$x_t = 1$				$x_t = 2$	

Fig. 3. Stage cost for a two-DM game where  $DM^1$  ( $DM^2$ ) chooses a row (a column) and its cost is the first (the second) entry in the chosen cell.

$x_t = 1$  is the low cost state and  $x_t = 2$  is the high cost state. The transition probabilities, given below, are constructed so that when  $DM^i$ s successfully coordinate their decisions (in a state-dependent manner) the state transitions with high probability to the low cost state. Otherwise, the state transitions with high probability to the high cost state. The transition kernel is fully described by

$$P(1|x, a^1, a^2) = 0.95, \text{ if } x = a^1 = a^2$$

$$P(2|x, a^1, a^2) = 0.95, \text{ if } x \neq a^1 \text{ or } a^1 \neq a^2$$

In particular, when  $x_t = 2$ ,  $DM^i$ s are faced with the choice between on the one hand incurring a lower short term cost 10 and likely remaining in the high cost state and on the other hand paying a higher short term cost 13 with the hopes of transitioning to the low cost state and avoiding sustained high costs.

For sufficiently large discount factors, including  $\beta = 0.8$  as selected, the unique team optimal policy is for both  $DM^i$ s

<sup>3</sup> $\kappa^i$  can be time-varying as long as it stay uniformly above zero.

to coordinate as  $u_t^1 = u_t^2 = x_t$ , for all  $t \in \mathbb{N}$ . However, there are three suboptimal equilibrium policies, namely (i)  $u_t^1 = u_t^2 = 1$ , for all  $t \in \mathbb{N}$ , (ii)  $u_t^1 = u_t^2 = 2$ , for all  $t \in \mathbb{N}$ , (iii)  $u_t^1 = u_t^2 \neq x_t$ , for all  $t \in \mathbb{N}$ .

We simulated Algorithm 2 and 3 with the following parameter choices.

Case A: Algorithm 2,  $h^i = R^{i,\lambda^i}$ ,  $\lambda^i \in (0, 1)$

$$\kappa^i = \gamma^i + 0.1, W^i = 30, T_k = 10000$$

Case B: Algorithm 2,  $h^i = R^{i,\lambda^i}$ ,  $\lambda^i = 1$

$$\kappa^i = 1, W^i = 50, T_k = 5000$$

Case C: Algorithm 3,  $g^i = h^i = R^{i,\lambda^i}$ ,  $\lambda^i \in (0, 1), \Lambda^i = 30$

$$\kappa^i = \gamma^i + 0.2, T_k = 7500$$

Case D: Algorithm 3,  $g^i = h^i = R^{i,\lambda^i}$ ,  $\lambda^i = 1, \Lambda^i = 30$

$$\kappa^i = \gamma^i + 0.2, T_k = 7500$$

where the aspiration level  $\Lambda^i$  used in case C and D was chosen without extensive tuning.

The algorithms performed generally as expected. The disparity across different cases owes largely to the parameter selections. In each case, the percentage of time where the joint policies are team optimal, i.e.,  $\pi_k \in \Pi_{\text{opt}}$ , are shown below.

Case	$\gamma = 0.05$	$\gamma = 0.01$	$\gamma = 0.005$	$\gamma = 0.001$
A	0.638	0.902	0.922	0.972
B	0.432	0.776	0.864	0.952
C	0.648	0.908	0.960	0.984
D	0.242	0.564	0.720	0.914

As the experimentation probability  $\gamma$  is reduced, the empirical frequency of the event  $\pi_k \in \Pi_{\text{opt}}$  increases and, for  $\gamma = 0.001$ , the joint policies are team optimal for more than 90% of the time. These numerical results confirm the theoretical results.

### VIII. CONCLUDING REMARKS

In this paper, we presented learning algorithms for stochastic teams and common interest games under a decentralized information structure in which players do not share actions with one another. While previous studies have focused on repeated games, or otherwise used a large degree of control sharing among decision makers to obtain convergence results, we have provided a method for achieving team optimality in teams and stochastic common interest without any control sharing during play and with limited prior information about the game.

The proof methods used in this paper center on approximating the true joint policy-valued stochastic process using time homogenous Markov chains through a novel Dobrushin's coefficient based analysis. The algorithms presented are amenable to further variations and can be modified as needed, and the Markov chain analysis used for the convergence guarantees can likewise be easily modified for more general applications.

We chose to focus on games with full state observations available to each agent since there are few formal results on multi-agent learning even under this simplifying assumption. The partially observed information structure, in which each player has access to only local state information, is an important and challenging direction for future research.

### APPENDIX A: PROOF OF THEOREM 1

Let  $\sigma(A) \in [0, 1]$  denote the Dobrushin coefficient of an  $n \times n$  right stochastic matrix  $A$ , defined in [57] as

$$\sigma(A) := \min_{i,k \in \{1, \dots, n\}} \sum_{j=1}^n \min\{A(i, j), A(k, j)\}. \quad (8)$$

**Lemma 4.** Consider an  $n \times n$  right stochastic matrix  $A$  with  $\sigma(A) > 0$ , and a sequence of  $n \times n$  right stochastic matrices  $\{A_k\}_{k \in \mathbb{N}}$ . For any  $\epsilon \in (0, 1)$ , if

$$\sup_{k \in \mathbb{N}} \|A_k - A\|_{\infty} \leq \tau := \frac{\sigma(A)\epsilon}{2n} \quad (9)$$

then, for any probability vector  $\mu_0$  of dimension  $n$ ,

$$\limsup_{k \in \mathbb{N}} \|\mu_0 A_0 \cdots A_k - \mu^*\|_1 \leq \epsilon$$

where  $\mu^*$  is the unique probability vector satisfying  $\mu^* = \mu^* A$ .

*Proof.* Recall that  $\|\mu A - \nu A\|_1 \leq (1 - \sigma(A))\|\mu - \nu\|_1$ , for all probability vectors  $\mu, \nu$ ; see [57]. Since  $\sigma(A) > 0$ , by Banach's fixed point theorem, there exists a unique probability vector  $\mu^*$  satisfying  $\mu^* = \mu^* A$ , and  $\lim_k \mu_0 A^k = \mu^*$ , for any probability vector  $\mu_0$ .

From (8)-(9), we have  $\sup_{k \in \mathbb{N}} |\sigma(A_k) - \sigma(A)| \leq n\tau$ , which implies  $\sup_{k \in \mathbb{N}} (1 - \sigma(A_k)) \leq \xi := 1 - \sigma(A)/2$ . Note  $\xi \in (0, 1)$ . We write

$$\begin{aligned} \|\mu_0 A_0 - \mu^*\|_1 &= \|\mu_0 A_0 - \mu^* A\|_1 \\ &\leq \|\mu_0 A_0 - \mu^* A_0\|_1 + \|\mu^* A_0 - \mu^* A\|_1 \\ &\leq (1 - \sigma(A_0))\|\mu_0 - \mu^*\|_1 + n\tau \\ &\leq \xi \|\mu_0 - \mu^*\|_1 + n\tau \end{aligned}$$

Repeated application of these inequalities result in

$$\|\mu_0 A_0 \cdots A_{k-1} - \mu^*\|_1 \leq \xi^k \|\mu_0 - \mu^*\|_1 + \epsilon, \quad \forall k,$$

where  $\epsilon = n\tau \frac{1}{1-\xi}$ , which is consistent with (9). As  $\lim_k \xi^k \|\mu_0 - \mu^*\|_1 = 0$ , the lemma follows.  $\square$

*Proof of Theorem 1*

Let  $\epsilon \in (0, 1)$  and  $\kappa \in (0, 1)^N$ . By Lemma 2, there exists  $\bar{\gamma}_\epsilon(\kappa)$  such that  $\max_i \gamma^i \in (0, \bar{\gamma}_\epsilon(\kappa))$  implies  $\mu_{\gamma, \kappa, h}^*(\Pi_{\text{opt}}) \geq 1 - \epsilon/2$ , where  $\mu_{\gamma, \kappa, h}^*$  is the unique invariant measure of the Markov chain induced by the IUP. Assume  $\max_i \gamma^i \in (0, \bar{\gamma}_\epsilon(\kappa))$ .

For all  $k \in \mathbb{N}$ ,  $\pi, \pi' \in \Pi$ , we define

$$\mu_k(\pi) := \Pr(\pi_k = \pi) \quad (10)$$

$$A_k(\pi, \pi') := \Pr(\pi_{k+1} = \pi' | \pi_k = \pi) \quad (11)$$

where  $\pi_k$  is the joint baseline policy during the  $k^{\text{th}}$  exploration phase of Algorithm 2. Note that  $\mu_{k+1} = \mu_0 A_0 \cdots A_k$ . To prove the theorem, we will show

$$\limsup_{k \in \mathbb{N}} \|\mu_k - \mu_{\gamma, \kappa, h}^*\|_1 \leq \epsilon/2.$$

Due to Lemma 4 and  $\sigma(A_{\gamma, \kappa, h}) > 0$ <sup>4</sup>, it is sufficient to show

$$\|A_k - A_{\gamma, \kappa, h}\|_{\infty} \leq \tau := \frac{\sigma(A_{\gamma, \kappa, h})\epsilon}{4|\Pi|} \quad (12)$$

<sup>4</sup> $A_{\gamma, \kappa, h}(\pi, \pi') \geq \prod_{i=1}^N \min\{\gamma^i, \kappa^i\} / |\Pi|^i > 0, \forall \pi, \pi' \in \Pi$ , due to uniform experimentation by each DM<sup>i</sup> with probability  $\gamma^i > 0$  or  $\kappa^i > 0$ . By (8), this implies  $\sigma(A_{\gamma, \kappa, h}) > 0$ .

for all but finitely many  $k \in \mathbb{N}$ . We note that  $\sigma(A_{\gamma, \kappa, h}) > 0$  since all entries of  $A_{\gamma, \kappa, h}$  are strictly positive, as the IUP updates policies using uniform randomization with strictly positive probability owing to  $\gamma^i, \kappa^i > 0$  for every  $\text{DM}^i$ .

To ensure (12), we will introduce an event  $R_k$  such that, for all  $\pi, \pi' \in \Pi$ , and all but finitely many  $k \in \mathbb{N}$ ,

$$\Pr(\pi_{k+1} = \pi' | \pi_k = \pi, R_k) = A_{\gamma, \kappa, h}(\pi, \pi') \quad (13)$$

and we will show that

$$\Pr(R_k | \pi_k = \pi) \geq 1 - \tau \quad (14)$$

by choosing the parameters of Algorithm 2 appropriately. Note that (13)-(14) imply (12) as follows:

$$\begin{aligned} & A_{\gamma, \kappa, h}(\pi, \pi') - \tau \\ & \leq A_{\gamma, \kappa, h}(\pi, \pi')(1 - \tau) \\ & \leq A_{\gamma, \kappa, h}(\pi, \pi') \Pr(R_k | \pi_k = \pi) \\ & \quad + \Pr(\pi_{k+1} = \pi' | \pi_k = \pi, R_k^c) \Pr(R_k^c | \pi_k = \pi) \\ & = A_k(\pi, \pi') \\ & \leq A_{\gamma, \kappa, h}(\pi, \pi') \cdot 1 + 1 \cdot P(R_k^c | \pi_k = \pi) \\ & \leq A_{\gamma, \kappa, h}(\pi, \pi') + \tau \end{aligned}$$

where  $R_k^c$  denotes the complement of  $R_k$ .

Define

$$\begin{aligned} \bar{\delta} & := \min\{|Q_{\pi^i}^{*i}(x, u) - Q_{\pi^i}^{*i}(x, v)| > 0 : \\ & \quad i, \pi^i \in \Pi^i, x \in \mathbb{X}, u, v \in \mathbb{U}^i\} \quad (15) \\ S^i(\pi) & := \sum_{x \in \mathbb{X}} Q_{\pi^i}^{*i}(x, \pi^i(x)), \quad \forall \pi \in \Delta \\ \bar{d} & := \frac{1}{2} \min\{|S^i(\pi) - S^i(\tilde{\pi})| > 0 : i, \pi, \tilde{\pi} \in \Pi\}. \end{aligned}$$

Let  $\bar{\pi}_k^i \in \Delta^i$  denote the policy used by  $\text{DM}^i$  in the  $k^{\text{th}}$  exploration phase, i.e.,

$$\bar{\pi}_k^i(\cdot | x) := (1 - \rho^i) \mathbb{I}_{\bar{\pi}_k^i(x)} + \rho^i \text{Unif}(\mathbb{U}^i), \quad \forall x \in \mathbb{X}.$$

Let  $\bar{\rho} > 0$  be such that  $\max_i \rho^i \in (0, \bar{\rho})$  implies

$$\begin{aligned} \|Q_{\bar{\pi}_k^i}^{*i} - Q_{\bar{\pi}_k^i}^{*i}\|_{\infty} & < \frac{1}{2} \min\{\delta^i, \bar{\delta} - \delta^i\}, \quad \forall i, k \in \mathbb{N} \\ |S^i(\bar{\pi}_k) - S^i(\bar{\pi}_k)| & < \frac{1}{2} \min\{d^i, \bar{d} - d^i\}, \quad \forall i, k \in \mathbb{N}. \end{aligned}$$

Such  $\bar{\rho}$  exists due to [16, Lemma 3]. Assume that, for all  $i$ ,  $d^i \in (0, \bar{d})$ ,  $\delta^i \in (0, \bar{\delta})$ ,  $\rho^i \in (0, \bar{\rho})$ . Assume further that

$$W^i \geq \bar{W}_{\epsilon}(\gamma, \kappa) := \min\{W \in \mathbb{N} : (1 - \phi)^W < \phi\tau/3\}, \quad \forall i$$

where  $\phi := \prod_i \min\{\gamma^i/|\Pi^i|, \kappa^i/|\Pi^i|\} \in (0, 1)$ .

For any time  $k \geq W_{\max}$ , we define the event

$$R_k := F_k \cap \bigcap_{\ell=0}^{W_{\max}} G_{k-\ell} \cap \bigcup_{\ell=1}^{\bar{W}_{\epsilon}(\gamma, \kappa)} H_{k-\ell}$$

where, for any  $\ell \in \mathbb{N}$ , we define

$$\begin{aligned} F_{\ell} & := \{\|Q_{\bar{\pi}_{\ell+1}^i}^i - Q_{\bar{\pi}_{\ell+1}^i}^{*i}\|_{\infty} < \min\{\delta^i, \bar{\delta} - \delta^i\}/2, \forall i\} \\ G_{\ell} & := \{|S_{\ell}^i - S^i(\bar{\pi}_{\ell})| < \min\{d^i, \bar{d} - d^i\}/2, \forall i\} \\ H_{\ell} & := \{\pi_{\ell} \in \Pi_{\text{opt}}\}. \end{aligned}$$

Conditioned on  $R_k$ ,  $k \geq W_{\max}$ , we have, for all  $i$ ,

$$\text{BR}_k^i = \text{BR}^i(\bar{\pi}_k^i)$$

and

$$S_k^i \leq \Lambda_k^i \iff \pi_k \in \Pi_{\text{opt}}.$$

This implies (13), for all  $k \geq W_{\max}$ . Intuitively, the event  $R_k$  guarantees that (i) all players have sufficiently reliable Q-factors during the  $k^{\text{th}}$  exploration phase, due to  $F_k$ ; (ii) for every  $\text{DM}^i$ , the estimated cost scores are sufficiently close to the true cost scores during every exploration phase in  $\text{DM}^i$ 's most recent memory window, by  $G_k$ ; (iii) an optimal baseline policy was visited recently enough that all players remember its cost score, by  $H_k$ .

We will now show (14) for sufficiently large exploration lengths  $\{T_{\ell}\}_{\ell}$ . (Since the events  $G_{\ell}$  and  $F_{\ell}$  are defined in terms of Q-factors, this is a statement about the long term behavior of the Q-factor iterates *within an exploration phase*.)

Note that within the  $k^{\text{th}}$  exploration phase of Algorithm 2 (and Algorithm 3), the environment faced by each  $\text{DM}^i$ , that is determined by  $\bar{\pi}_k^i$ , is a stationary MDP (with finite state and control spaces) and satisfies the usual conditions of stochastic approximation theory. In such a setting, it is well-known that the sequence of Q-factors produced by the standard Q-learning algorithm from any initial condition is bounded and convergent with probability one [2]. Since each exploration phase in Algorithm 2 (and in Algorithm 3) starts with re-initialized Q-factors (and what we may call J-factors in the case of Algorithm 3) within the compact sets  $\{Q^i\}_{i=1}^N$  (and  $\{J^i\}_{i=1}^N$ ), the Q-factors (and J-factors) produced by Algorithm 2 (and Algorithm 3) during any exploration phase remain bounded with probability one; c.f. Lemma 1 in [16].

Furthermore, Lemma 1 in [16] shows that, uniformly in the initial conditions within  $\{Q^i\}_{i=1}^N$  (and  $\{J^i\}_{i=1}^N$ ), the Q-factors (and J-factors) produced by Algorithm 2 (and Algorithm 3) enter any arbitrarily small neighborhood of their limits with arbitrarily high probability at the end of any sufficiently long exploration phase.

By [16, Lemma 4], there exists  $T_{\epsilon}(\gamma, \kappa, W_{\max}) \in \mathbb{N}_{+}$  such that if  $\min_{k \in \mathbb{N}} T_k \geq T_{\epsilon}(\gamma, \kappa, W_{\max})$ , we have

$$\Pr(F_k), \Pr(G_k) \geq 1 - \phi\tau/(3W_{\max}).$$

This implies, for  $k \geq W_{\max}$ ,

$$\Pr(\bigcap_{\ell=0}^{W_{\max}} G_{k-\ell}) \geq 1 - \phi\tau/3.$$

In addition, we have, for  $k \geq W_{\max}$ ,

$$\Pr(H_k) \geq 1 - (1 - \phi)^{\bar{W}_{\epsilon}(\gamma, \kappa)} \geq 1 - \phi\tau/3.$$

All together, the preceding imply, for  $k \geq W_{\max}$ ,

$$\Pr(R_k) \geq 1 - \phi\tau. \quad (16)$$

Since  $\inf_{k \in \mathbb{N}, \pi \in \Pi} \Pr(\pi_k = \pi) \geq \phi > 0$ , (16) implies (14), because (16) implies  $\Pr(R_k \cap \{\pi_k = \pi\}) \geq (1 - \tau)\Pr(\pi_k = \pi)$  for any  $k, \pi$ .

We have shown that (13) and (14) hold. In turn, this implies (12) holds, and invoking Lemma 4 completes the proof.  $\square$

## APPENDIX B: PROOF OF THEOREM 2

**Lemma 5.** Consider an  $n \times n$  right stochastic matrix  $A$ , and a sequence of  $n \times n$  right stochastic matrices  $\{A_k\}_{k \in \mathbb{N}}$ . For any  $\epsilon \in (0, 1)$ ,  $m \in \mathbb{N}$ , if

$$\sup_{k \in \mathbb{N}} \|A_k - A\|_{\infty} \leq \epsilon/(2nm) \quad (17)$$

then

$$\sup_{k \in \mathbb{N}, \mu_0} \|\mu_0 A_k \cdots A_{k+m-1} - \mu_0 A^m\|_1 \leq \epsilon/2$$

where  $\mu_0$  is any probability vector of dimension  $n$ .

*Proof of Theorem 2*

Let  $\epsilon \in (0, 1)$ . Assume

$$0 < \kappa^i < \tilde{\kappa}_\epsilon := \min\{\bar{\kappa}_\epsilon, \epsilon/(4|\mathbf{\Pi}|\bar{m}N)\}, \quad \forall i$$

where  $\bar{\kappa}_\epsilon$  and  $\bar{m}$  are as in Lemma 3. Then, assume

$$0 < \gamma^i < \tilde{\gamma}_\epsilon(\boldsymbol{\kappa}) := \min\{\bar{\gamma}_\epsilon(\boldsymbol{\kappa}), \tilde{\kappa}_\epsilon\}, \quad \forall i$$

where  $\bar{\gamma}_\epsilon(\boldsymbol{\kappa})$  is as in Lemma 2. With these choices of  $\gamma, \boldsymbol{\kappa}$ , Lemma 3 holds, i.e.,

$$\inf_{\mu_0 \in \mathcal{P}(\mathbf{\Pi})} (\mu_0 A_{\gamma, \boldsymbol{\kappa}}^{\bar{m}})(\mathbf{\Pi}_{\text{eq}}) \geq 1 - \epsilon/2. \quad (18)$$

For any  $k \in \mathbb{N}$ , defining  $A_k$  as in (11), we have

$$\begin{aligned} \|A_k - A_{\gamma, \boldsymbol{\kappa}, h}\|_\infty &\leq 1 - \prod_i (1 - \max\{\gamma^i, \kappa^i\}) \\ &\quad \times \min_{\boldsymbol{\pi} \in \mathbf{\Pi}} \Pr(\text{BR}_k^i = \text{BR}^i(\boldsymbol{\pi}^{-i}), \forall i | \boldsymbol{\pi}_k = \boldsymbol{\pi}) \end{aligned} \quad (19)$$

By [16, Lemma 4], there exists  $T_\epsilon \in \mathbb{N}_+$  such that if  $T_k \geq T_\epsilon$ ,

$$\min_{\boldsymbol{\pi} \in \mathbf{\Pi}} \Pr(\text{BR}_k^i = \text{BR}^i(\boldsymbol{\pi}^{-i}), \forall i | \boldsymbol{\pi}_k = \boldsymbol{\pi}) \geq 1 - \frac{\epsilon}{4|\mathbf{\Pi}|\bar{m}N}. \quad (20)$$

Assume that, for all  $i, k \in \mathbb{N}$ ,

$$\begin{aligned} W^i &\geq \tilde{W}_\epsilon(\gamma, \boldsymbol{\kappa}) := \bar{W}_\epsilon(\gamma, \boldsymbol{\kappa}) \\ T_k &\geq \tilde{T}_\epsilon(\gamma, \boldsymbol{\kappa}, W_{\max}) := \max\{T_\epsilon, \bar{T}_\epsilon(\gamma, \boldsymbol{\kappa}, W_{\max})\} \end{aligned}$$

where  $\bar{W}_\epsilon(\gamma, \boldsymbol{\kappa}), \bar{T}_\epsilon(\gamma, \boldsymbol{\kappa}, W_{\max})$  are as in Theorem 1. By (19)-(20) and the assumptions on  $\gamma, \boldsymbol{\kappa}, \{T_k\}_{k \in \mathbb{N}}$ , we have

$$\sup_{k \in \mathbb{N}} \|A_k - A_{\gamma, \boldsymbol{\kappa}}\|_\infty \leq \epsilon/(2|\mathbf{\Pi}|\bar{m}N).$$

Lemma 5 implies

$$\sup_{k \in \mathbb{N}, \mu_0 \in \mathcal{P}(\mathbf{\Pi})} \|\mu_0 A_k \cdots A_{k+\bar{m}-1} - \mu_0 A_{\gamma, \boldsymbol{\kappa}}^{\bar{m}}\|_1 \leq \epsilon/2. \quad (21)$$

The desired result for weakly acyclic games follows from (18)-(21). Note that the parameter choices satisfy the hypothesis of Theorem 1; hence, the results of Theorem 1 also hold.

### APPENDIX C: PROOF OF THEOREM 3

Let  $\bar{\delta}$  be as in (15), and let  $\rho^\Lambda \in (0, 1)$  be such that  $\rho^i \in (0, \rho^\Lambda)$ , for all  $i$ , implies

$$\|Q_{\boldsymbol{\pi}_k}^{*i} - Q_{\bar{\boldsymbol{\pi}}_k}^{*i}\|_\infty < \frac{1}{2} \min\{\delta^i, \bar{\delta} - \delta^i\}, \quad \forall i, k \in \mathbb{N}$$

and

$$|\tilde{S}^i(\boldsymbol{\pi}_k) - \tilde{S}^i(\bar{\boldsymbol{\pi}}_k)| < \min_{\boldsymbol{\pi} \in \mathbf{\Pi}} |\Lambda^i - S^i(\boldsymbol{\pi})|, \quad \forall i, k \in \mathbb{N}$$

where  $\bar{\boldsymbol{\pi}}_k^i(\cdot | x_k) = (1 - \rho^i) \mathbb{I}_{\boldsymbol{\pi}_k^i(x_k)} + \rho^i \text{Unif}(\mathbb{U}^i)$ . Such  $\rho^\Lambda \in (0, 1)$  exists due to [16, Lemma 3].

Let  $\epsilon_k \in (0, 1)$ ,  $k \in \mathbb{N}$ , be such that  $\sum_{k \in \mathbb{N}} \epsilon_k < \infty$ . Due to [16, Lemma 1], there exists finite integers  $\bar{T}_k \in \mathbb{N}_+$ ,  $k \in \mathbb{N}$  such that if  $T_k \geq \bar{T}_k$ , for all  $k \in \mathbb{N}$ ,

$$\Pr(|\tilde{S}_k^i - \tilde{S}^i(\bar{\boldsymbol{\pi}}_k)| < \epsilon_k, \|Q_{t_{k+1}}^i - Q_{\bar{\boldsymbol{\pi}}_k}^{*i}\|_\infty < \epsilon_k, \forall i) \geq 1 - \epsilon_k.$$

Assume now  $T_k \geq \bar{T}_k$ , for all  $k \in \mathbb{N}$ . Hence, there exists  $\tilde{k} \in \mathbb{N}_+$  such that, for all  $i, k \geq \tilde{k}$ ,

$$\Pr(E_k) \geq 1 - \epsilon_k$$

where

$$\begin{aligned} E_k &:= \{((\boldsymbol{\pi}_k \in \mathbf{\Pi}^\Lambda, \tilde{S}_k^i < \Lambda^i) \text{ or } (\boldsymbol{\pi}_k \notin \mathbf{\Pi}^\Lambda, \tilde{S}_k^i > \Lambda^i)), \\ &\quad \text{BR}_k^i = \text{BR}^i(\boldsymbol{\pi}_k^{-i}), \forall i\}. \end{aligned}$$

1) We have, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\boldsymbol{\pi}_{k+1} \in \mathbf{\Pi}^\Lambda | \boldsymbol{\pi}_k \in \mathbf{\Pi}^\Lambda) &\geq (1 - \epsilon_k) \prod_i (1 - \gamma_k^i) \\ \Pr(\boldsymbol{\pi}_{k+1} \in \mathbf{\Pi}^\Lambda | \boldsymbol{\pi}_k \notin \mathbf{\Pi}^\Lambda) &\geq (1 - \epsilon_k) \prod_i (\kappa^i / |\mathbf{\Pi}^i|). \end{aligned}$$

This leads to, with some algebra, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\boldsymbol{\pi}_{k+1} \notin \mathbf{\Pi}^\Lambda) &\leq \left(1 - \prod_i (\kappa^i / |\mathbf{\Pi}^i|)\right) \Pr(\boldsymbol{\pi}_k \notin \mathbf{\Pi}^\Lambda) \\ &\quad + \epsilon_k + \sum_i \gamma_k^i. \end{aligned}$$

Since  $\left|1 - \prod_i (\kappa^i / |\mathbf{\Pi}^i|)\right| < 1$ ,  $\sum_{k \in \mathbb{N}} \epsilon_k < \infty$ , and  $\sum_{i, k \in \mathbb{N}} \gamma_k^i < \infty$ , we have  $\sum_{k \in \mathbb{N}} \Pr(\boldsymbol{\pi}_k \notin \mathbf{\Pi}^\Lambda) < \infty$ . Borel-Cantelli Lemma implies

$$\Pr(\boldsymbol{\pi}_k \notin \mathbf{\Pi}^\Lambda, \text{ for infinitely many } k \in \mathbb{N}) = 0.$$

Also,  $\sum_{k \in \mathbb{N}} \Pr((\boldsymbol{\pi}_k \in \mathbf{\Pi}^\Lambda, \tilde{S}_k^i \geq \Lambda^i) < \infty$ , hence

$$\Pr(\boldsymbol{\pi}_k \in \mathbf{\Pi}^\Lambda, S_k^i \geq \Lambda^i, \text{ for infinitely many } k \in \mathbb{N}) = 0.$$

This proves the first part.

2) We have, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\boldsymbol{\pi}_{k+1} \in \mathbf{\Pi}^\Lambda \cap \mathbf{\Pi}_{\text{cumber}} | \boldsymbol{\pi}_k \in \mathbf{\Pi}^\Lambda \cap \mathbf{\Pi}_{\text{cumber}}) \\ \geq (1 - \epsilon_k) \prod_i (1 - \gamma_k^i). \end{aligned}$$

There exists  $\bar{p}_{\min} \in (0, 1)$  (which depends only on  $\lambda^1, \dots, \lambda^N, |\mathbf{\Pi}^1|, \dots, |\mathbf{\Pi}^N|, \bar{L}$ ) such that, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\boldsymbol{\pi}_{k+\bar{L}} \in \mathbf{\Pi}^\Lambda \cap \mathbf{\Pi}_{\text{cumber}} | \boldsymbol{\pi}_k \in \mathbf{\Pi}^\Lambda \setminus \mathbf{\Pi}_{\text{cumber}}) \\ \geq \bar{p}_{\min} \prod_{n=k}^{k+\bar{L}-1} (1 - \epsilon_n) \prod_i (1 - \gamma_n^i) \end{aligned} \quad (22)$$

and

$$\begin{aligned} \Pr(\boldsymbol{\pi}_{k+\bar{L}} \in \mathbf{\Pi}^\Lambda \cap \mathbf{\Pi}_{\text{cumber}} | \boldsymbol{\pi}_k \notin \mathbf{\Pi}^\Lambda) \\ \geq \prod_i (\kappa^i / |\mathbf{\Pi}^i|) \prod_{n=k}^{k+\bar{L}-1} (1 - \epsilon_n) \prod_i (1 - \gamma_n^i). \end{aligned}$$

This leads to, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\boldsymbol{\pi}_{k+\bar{L}} \notin \mathbf{\Pi}^\Lambda \cap \mathbf{\Pi}_{\text{cumber}}) \\ \leq \left(1 - \min\left\{\bar{p}_{\min}, \prod_i (\kappa^i / |\mathbf{\Pi}^i|)\right\}\right) \\ \times \Pr(\boldsymbol{\pi}_k \notin \mathbf{\Pi}^\Lambda \cap \mathbf{\Pi}_{\text{cumber}}) \\ + \sum_{n=k}^{k+\bar{L}-1} \epsilon_n + \sum_{n=k}^{k+\bar{L}-1} \sum_i \gamma_n^i. \end{aligned}$$

Since  $\left|1 - \min\left\{\bar{p}_{\min}, \prod_i \frac{\kappa^i}{\prod_i^i}\right\}\right| < 1$ ,  $\sum_{k \in \mathbb{N}} \epsilon_k < \infty$ , and  $\sum_{i, k \in \mathbb{N}} \gamma_k^i < \infty$ , we have  $\sum_{j \in \mathbb{N}} \Pr(\pi_{k+j\bar{L}} \notin \Pi^\Lambda \cap \Pi_{\text{cumber}}) < \infty$ , for all  $k \in \mathbb{N}$ . This results in  $\sum_{k \in \mathbb{N}} \Pr(\pi_k \notin \Pi^\Lambda \cap \Pi_{\text{cumber}}) < \infty$ . Borel-Cantelli Lemma implies

$\Pr(\pi_k \notin \Pi^\Lambda \cap \Pi_{\text{cumber}}, \text{ for infinitely many } k \in \mathbb{N}) = 0$ .

Also  $\sum_{k \in \mathbb{N}} \Pr(\text{BR}_k^i \neq \text{BR}^i(\pi_k^{-i})) < \infty$ , hence,

$\Pr(\text{BR}_k^i \neq \text{BR}^i(\pi_k^{-i}), \text{ for infinitely many } k \in \mathbb{N}) = 0$ .

This proves the second part.

3) We have, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\pi_{k+\bar{L}} \in \Pi_{\text{cumber}}^\Lambda | \pi_k \in \Pi_{\text{cumber}}^\Lambda) \\ \geq \prod_{n=k}^{k+\bar{L}-1} (1 - \epsilon_n) \prod_i (1 - \max\{\gamma_n^i, \kappa^i\}). \end{aligned}$$

We also have, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\pi_{k+\bar{L}} \in \Pi_{\text{cumber}}^\Lambda | \pi_k \notin \Pi_{\text{cumber}}^\Lambda) \\ \geq \bar{p}_{\min} \prod_{n=k}^{k+\bar{L}-1} (1 - \epsilon_n) \prod_i (1 - \max\{\gamma_n^i, \kappa^i\}) \end{aligned}$$

where  $\bar{p}_{\min} \in (0, 1)$  is as in (22). This leads to, for all  $k \geq \tilde{k}$ ,

$$\begin{aligned} \Pr(\pi_{k+\bar{L}} \notin \Pi_{\text{cumber}}^\Lambda) \leq \sum_{n=k}^{k+\bar{L}-1} \left( \epsilon_n + \sum_i \max\{\gamma_n^i, \kappa^i\} \right) \\ + (1 - \bar{p}_{\min}) \Pr(\pi_k \notin \Pi_{\text{cumber}}^\Lambda). \end{aligned}$$

Since  $|1 - \bar{p}_{\min}| < 1$  and  $\lim_{k \in \mathbb{N}} \sum_{n=k}^{k+\bar{L}-1} \epsilon_n = 0$ , we have, for all  $k \in \mathbb{N}$ ,

$$\begin{aligned} \limsup_{j \in \mathbb{N}} \Pr(\pi_{k+j\bar{L}} \notin \Pi_{\text{cumber}}^\Lambda) \\ \leq \limsup_{j \in \mathbb{N}} \sum_{n=k+j\bar{L}}^{k+(j+1)\bar{L}-1} \sum_i \max\{\gamma_n^i, \kappa^i\} / \bar{p}_{\min}. \end{aligned}$$

This proves the third part.

4) It follows exactly the same as the third part by replacing  $\Pi_{\text{cumber}}^\Lambda$  with  $\Pi_{\text{cumber}}$ .

## REFERENCES

- [1] B. Yongacoglu, G. Arslan, and S. Yüksel, "Reinforcement learning for decentralized stochastic control," in *58th Conference on Decision and Control (CDC)*, pp. 5556–5561, IEEE, 2019.
- [2] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, no. 3, pp. 185–202, 1994.
- [3] C. Watkins, *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.
- [4] C. Watkins and P. Dayan, "Q-Learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.
- [5] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, "A survey of learning in multiagent environments: Dealing with non-stationarity," *arXiv preprint arXiv:1707.09183*, 2017.
- [6] E. Altman, "Non zero-sum stochastic games in admission, service and routing control in queueing systems," *Queueing Systems*, vol. 23, no. 1–4, pp. 259–279, 1996.
- [7] J. Ding, M. Kamgarpour, S. Summers, A. Abate, J. Lygeros, and C. Tomlin, "A stochastic games framework for verification and control of discrete time stochastic hybrid systems," *Automatica*, vol. 49, no. 9, pp. 2665–2674, 2013.
- [8] A. M. Fink, "Equilibrium in a stochastic  $n$ -person game," *Journal of Science of the Hiroshima University, Series A1 (Mathematics)*, vol. 28, no. 1, pp. 89–93, 1964.
- [9] F. L. Lewis and D. Liu, *Hybrid Learning in Stochastic Games and Its Application in Network Security*. IEEE, 2013.
- [10] D. Fudenberg and J. Tirole, "Game theory, 1991," *Cambridge, Massachusetts*, vol. 393, no. 12, 1991.
- [11] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine Learning Proceedings 1994*, pp. 157–163, Elsevier, 1994.
- [12] Y.-C. Ho, "Team decision theory and information structures," *Proceedings of the IEEE*, vol. 68, no. 6, pp. 644–654, 1980.
- [13] S. Yüksel and T. Başar, *Stochastic Networked Control Systems*. Birkhäuser, 2013.
- [14] R. J. Aumann and S. Sorin, "Cooperation and bounded recall," *Games and Economic Behavior*, vol. 1, no. 1, pp. 5–39, 1989.
- [15] S. Takahashi, "Infinite horizon common interest games with perfect information," *Games and Economic Behavior*, vol. 53, no. 2, pp. 231–247, 2005.
- [16] G. Arslan and S. Yüksel, "Decentralized Q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2017.
- [17] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337, 1993.
- [18] S. Sen, M. Sekaran, and J. Hale, "Learning to coordinate without sharing information," in *Proceedings of the 12th National Conference on Artificial Intelligence*, pp. 426–431, 1994.
- [19] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the Tenth Innovative Applications of Artificial Intelligence Conference, Madison, Wisconsin*, pp. 746–752, July 1998.
- [20] M. L. Littman and C. Szepesvári, "A generalized reinforcement-learning model: Convergence and applications," in *ICML*, vol. 96, pp. 310–318, 1996.
- [21] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *Journal of Machine Learning Research*, vol. 4, no. Nov, pp. 1039–1069, 2003.
- [22] J. Hu, M. P. Wellman, et al., "Multiagent reinforcement learning: theoretical framework and an algorithm.," in *ICML*, vol. 98, pp. 242–250, Citeseer, 1998.
- [23] M. L. Littman, "Friend-or-foe Q-learning in general-sum games," in *ICML*, vol. 1, pp. 322–328, 2001.
- [24] M. L. Littman, "Value-function reinforcement learning in Markov games," *Cognitive Systems Research*, vol. 2, no. 1, pp. 55–66, 2001.
- [25] A. Greenwald, K. Hall, and R. Serrano, "Correlated Q-learning," in *ICML*, vol. 20, p. 242, 2003.
- [26] G. Tesauro, "Extending Q-learning to general adaptive multi-agent systems," in *Advances in Neural Information Processing Systems*, pp. 871–878, 2004.
- [27] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal Nash equilibrium in team Markov games," in *Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS'02*, pp. 1603–1610, 2002.
- [28] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, 1993.
- [29] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems.," *The Knowledge Engineering Review*, vol. 27, no. 1, pp. 1–31, 2012.
- [30] P. Stone and M. Veloso, "Multiagent systems: A survey from a machine learning perspective," *Autonomous Robots*, vol. 8, no. 3, pp. 345–383, 2000.
- [31] D. L. Leotta, J. Ruiz-del Solar, and R. Babuška, "Decentralized reinforcement learning of robot behaviors," *Artificial Intelligence*, vol. 256, pp. 130–159, 2018.
- [32] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.
- [33] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *In Proceedings of the Seventeenth International Conference on Machine Learning*, Citeseer, 2000.
- [34] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, no. 2, pp. 215–250, 2002.
- [35] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Hysteretic Q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 64–69, IEEE, 2007.

- [36] S. Kapetanakis and D. Kudenko, "Reinforcement learning of coordination in cooperative multi-agent systems," *AAAI/IAAI*, vol. 2002, pp. 326–331, 2002.
- [37] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Coordination of independent learners in cooperative Markov games," *HAL preprint hal-00370889*, 2009.
- [38] L. Panait, K. Sullivan, and S. Luke, "Lenient learners in cooperative multiagent systems," in *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 801–803, 2006.
- [39] E. Wei and S. Luke, "Lenient learning in independent-learner stochastic cooperative games," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2914–2955, 2016.
- [40] G. C. Chasparis, A. Arapostathis, and J. S. Shamma, "Aspiration learning in coordination games," *SIAM J. Control and Optimization*, vol. 51, no. 1, pp. 465–490, 2013.
- [41] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 373–396, 2009.
- [42] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," *SIAM J. Control Optim.*, vol. 52, no. 5, p. 27532770, 2014.
- [43] J. R. Marden and J. S. Shamma, "Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation," *Games and Economic Behavior*, vol. 75, no. 2, pp. 788–808, 2012.
- [44] B. S. Pradelski and H. P. Young, "Learning efficient Nash equilibria in distributed systems," *Games and Economic Behavior*, vol. 75, no. 2, pp. 882 – 897, 2012.
- [45] J. Josephson and A. Matros, "Stochastic imitation in finite games," *Games and Economic Behavior*, vol. 49, no. 2, pp. 244 – 259, 2004.
- [46] L. Matignon, G. J. Laurent, N. Le Fort-Piat, and Y.-A. Chapuis, "Designing decentralized controllers for distributed-air-jet mems-based micromanipulators by reinforcement learning," *Journal of Intelligent & Robotic Systems*, vol. 59, no. 2, pp. 145–166, 2010.
- [47] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for interference control in OFDMA-based femtocell networks," in *2010 IEEE 71st Vehicular Technology Conference*, pp. 1–5, IEEE, 2010.
- [48] S. Nie, Z. Fan, M. Zhao, X. Gu, and L. Zhang, "Q-learning based power control algorithm for D2D communication," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, IEEE, 2016.
- [49] X. Lin, Y. Tang, X. Lei, J. Xia, Q. Zhou, H. Wu, and L. Fan, "MARL-based distributed cache placement for wireless networks," *IEEE Access*, vol. 7, pp. 62606–62615, 2019.
- [50] W. Wang, A. Kwasinski, D. Niyato, and Z. Han, "A survey on applications of model-free strategy learning in cognitive wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1717–1757, 2016.
- [51] O. Hernandez-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.
- [52] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 287–308, 2000.
- [53] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*, pp. 5872–5881, PMLR, 2018.
- [54] V. S. Borkar, "Stochastic approximation with two time scales," *Systems & Control Letters*, vol. 29, no. 5, pp. 291–294, 1997.
- [55] H. P. Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. New Jersey: Princeton University Press, 1998.
- [56] A. Fabrikant, A. D. Jaggard, and M. Schapira, "On the structure of weakly acyclic games," in *Algorithmic Game Theory*, pp. 126–137, Springer, 2010.
- [57] R. L. Dobrushin, "Central limit theorem for nonstationary Markov chains. 1," *Theory Probab. Appl.*, vol. 1, no. 1, pp. 65–80, 1956.



**Bora Yongacoglu** received his B.A. degree from McGill University in 2016, with majors in Mathematics and Economics. He received his M.Sc. in Applied Mathematics in 2018 from Queen's University, where he is currently a Ph.D. student in Applied Mathematics. His research interests include stochastic control, decentralized control, and learning in multi-agent systems.



**Gürdal Arslan** received Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign in 2001. From 2001 to 2004, he was an Assistant Researcher in the Department of Mechanical and Aerospace Engineering, University of California, Los Angeles. In August 2004, he joined the Electrical Engineering Department at the University of Hawaii, Manoa, where he is currently an Associate Professor. His current research interests lie in the design of cooperative multi-agent systems using game theoretic methods. He received the National Science Foundation CAREER Award in May 2006. He currently serves as an Associate Editor for *Automatica*.



**Serdar Yüksel** (S'02, M'11) received his B.Sc. degree in Electrical and Electronics Engineering from Bilkent University in 2001; M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign in 2003 and 2006, respectively. He was a post-doctoral researcher at Yale University before joining Queen's University as an Assistant Professor in the Department of Mathematics and Statistics, where he is now an Associate Professor. He has been awarded the 2013 CAIMS/PIMS Early Career Award in Applied Mathematics. His research interests are on stochastic control, decentralized control, information theory and probability. He has been an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *Automatica*, and *Systems and Control Letters*.