Decentralized Multi-Agent Reinforcement Learning for Continuous-Space Stochastic Games[*]

Awni Altabaa[†]   Bora Yongacoglu[‡]   and Serdar Yüksel[‡]

Abstract

Stochastic games are a popular framework for studying multi-agent reinforcement learning (MARL). Recent advances in MARL have focused primarily on games with finitely many states. In this work, we study multi-agent learning in stochastic games with general state spaces and an information structure in which agents do not observe each other's actions. In this context, we propose a decentralized MARL algorithm and we prove the near-optimality of its policy updates. Furthermore, we study the global policy-updating dynamics for a general class of best-reply based algorithms and derive a closed-form characterization of convergence probabilities over the joint policy space.

1   Introduction

Multi-agent reinforcement learning (MARL) is the study of the learning dynamics of strategic agents that coexist in a shared environment, and is one of the important frontiers of machine learning and control. In this paper, we study MARL in *stochastic games*, also known as Markov games, a multi-agent generalization of Markov decision problems (MDPs) in which the cost-relevant history of the system is summarized by a state variable Shapley [1953]. Due to its ability to model both dynamic inter-temporal choice as well as strategic interaction, the stochastic games model has long been a popular framework for studying multi-agent learning Littman [1994].

In comparison to single-agent reinforcement learning, analysis of MARL is difficult due to several challenges inherent to multi-agent systems, including non-stationarity, conflicting interests, and decentralized information. As a result, fundamental understanding of multi-agent reinforcement learning theory has lagged behind its single-agent counterpart Zhang et al. [2021]. Owing partly to a variety of impressive empirical successes, such as those of Silver et al. [2016] and Vinyals et al. [2019], the subject of learning in stochastic games has

[†]Department of Statistics and Data Science at Yale University, New Haven, CT, United States:`awni.altabaa@yale.edu`

[‡]Department of Mathematics and Statistics at Queen's University, Kingston ON, Canada: `1bmy@queensu.ca, resp. yuksel@queensu.ca`

recently attracted considerable attention, and theoretical advances have been made along several research axes. In Section 2, we survey some of these advances in greater detail, and we observe that the bulk of theoretical progress in MARL has focused on games with finitely many states. By contrast, considerably less is known about learning in games with more general state spaces, which model various real-world settings that do not admit finite state space models.

In this paper, we study the problem of decentralized multi-agent reinforcement learning in stochastic games with standard Borel state spaces, which includes finite sets, countably infinite sets, and finite dimensional vector spaces as special cases. We are interested in the setting in which a generative model is not available for sampling state and (joint) action feedback and in which there is no offline data set available to the players. Instead, we focus on a setting in which players must learn using in-game feedback, updating their estimates and adjusting their behaviour only as new information arises from sequential interaction with the environment. We approach the problem with an eye on decentralized settings, in which agents may not be able to observe the actions of other agents, but must nevertheless attempt to respond optimally to the decision rules of others. In this context, we wish to propose an algorithm that is justifiable from an individual rationality point of view, but also comes with some long term guarantee of system-wide stability—with policies converging to some joint policy under self-play.

With these objectives in mind, we present a decentralized MARL algorithm for stochastic games with standard Borel state spaces, and we give a formal analysis of its properties. The algorithm proposed here extends the *Decentralized Q-learning* algorithm of Arslan and Yüksel [2017] to settings with general state spaces by quantizing the state variable. Although this approach is natural, one must implement it with care: by quantizing the state space, the algorithm designer artificially restricts the class of policies and may unintentionally limit the efficacy of its learning agents. We rigorously analyze such a state space quantization, and we provide sufficient conditions on the game that ensure such an approach entails an arbitrarily small loss in performance. To accomplish this analysis, we use recent theoretical advances in single-agent MDPs with general state spaces, building most prominently upon the quantized Q-learning results of Kara et al. [2021].

The algorithm presented here relies on the exploration phase technique, which requires that agents change their policies only at prescribed times, $t_0 < t_1 < \ldots$, while holding their policies fixed during the intervals $[t_k, t_{k+1} - 1]$, for $k \geq 0$. (That is, players do not change policies between update times.) The update times are predetermined and common to all agents, but outside of this no special coordination structure is assumed. In particular, all agents may change their policies simultaneously at an update time t_k or may choose to not change their policy. Besides this assumption on synchrony, agents need not communicate, observe one another's actions or cost realizations, or even be aware of the other agents or the fact that they are playing a game. Despite this decentralization, we show that policy updates of our algorithm are asymptotically optimal for each agent with respect to the most recently observed environment.

In addition to the near optimality of policy updates resulting from our algorithm, we

also analyze the joint policy dynamics for a class of best-reply-based updates. For such algorithms, we study convergence to equilibrium and give a characterization of the convergence behaviour when the game admits weakly acyclic structure. Finally, we describe additional settings in which self-play of our algorithm drives the joint policy process to ϵ-equilibrium policies.

Contributions:

- In Algorithm 1, we propose a decentralized MARL algorithm for environments with general state spaces. In Theorem 4.1, we prove that this algorithm yields policy updates which are near-optimal with respect to each agent's observed environment, under the assumptions of weak continuity of the transition kernel and continuity of the cost function.

- In Proposition 5.1, we theoretically analyze the global joint policy-updating dynamics of a general class of best-reply-based algorithms and study convergence to equilibria.

- In Section 7, we present a simulation study which corroborates our theoretical study.

Organization

The remainder of the paper is structured as follows. In Section 2 we discuss some related work on multi-agent reinforcement learning algorithms with theoretical guarantees. In Section 3 we give the relevant background on modeling multi-agent environments. In Section 4, we propose a continuous-space extension of the decentralized multi-agent Q-learning algorithm and present our accompanying theoretical results. In Section 5 we derive a complete characterization of the global policy-updating dynamics of our algorithms. In Section 6, we discuss the topic of convergence to equilibrium and suggest lines of future research. A simulation study of our algorithm is presented in Section 7.

Notation. We use $\mathbb{P}$ and $\mathbb{E}$ to denote the probability measure or expectation with subscripts and superscripts to denote the underlying probability space. $\mathcal{P}(S)$ denotes the set of probability measures on S and $\mathcal{P}(S'|S)$ denotes the set of transition kernels from S to S'. We will tend to use $\hat{\Box}$ to denote quantities associated with approximation due to quantization.

2 Related Work

Multi-agent learning is an interdisciplinary subject with contributions from various fields. In recent years, there has been a massive amount of new contributions to the subject, including contributions that are theoretical, empirical, and applied in nature. As such, a comprehensive overview of related literature is beyond the scope of this paper. Therefore, the discussion below takes a narrower view and attempts to focus on works that are closer to

the subject matter of this study. For recent surveys of multi-agent reinforcement learning, see Zhang et al. [2021] or Ozdaglar et al. [2021]. For references to older research on multi-agent learning from other perspectives, Cesa-Bianchi and Lugosi [2006], Fudenberg et al. [1998] and Young [2004].

In this paper, we adopt stochastic games as our formal model for multi-agent reinforcement learning. Stochastic games have long been a popular framework for studying MARL due to their ability to model both inter-temporal consequences of actions as well as multi-player strategic considerations Littman [1994]. The N-player stochastic game model generalizes both single-agent Markov decision problems (MDPs) as well as single-state repeated games.

In this paper, we study stochastic games with full state observability at each agent and no action-sharing between agents. This paradigm, in which agents have access only to the state variable and their own history of actions and cost realizations, is sometimes called the *independent learning* paradigm, though the terminology is not uniform Claus and Boutilier [1998], Laurent et al. [2011], Ozdaglar et al. [2021]. Under various names, independent learners have been the subject of considerable research recently, Sayin et al. [2021], Daskalakis et al. [2020], Yongacoglu et al. [2021, 2022], and important theoretical advances have been made.

We contrast the so-called independent learning paradigm studied here with the *joint action learner* paradigm, in which each agent observes the actions of all other agents. Under an assumption of perfect monitoring of one another's actions, many theoretical contributions were made for joint action learners. We selectively cite Littman [1994, 2001] and Hu and Wellman as a representative sampling of rigorous work on joint action learners in stochastic games.

Theoretical results for independent learners are comparatively more difficult to obtain than their counterparts for joint action learners, as salient information is hidden from agents. However, the perfect monitoring assumption can be too stringent for many (even mildly) decentralized settings. Moreover, the number of joint actions scales exponentially in the number of agents, meaning that joint action learning algorithms are often intractable.

While the preceding works study stochastic games with finite state spaces, this paper is concerned with a more general class of games that allows for standard Borel state spaces, which include finite state spaces and finite dimensional real vector spaces as special cases.

This work is in the tradition of *regret testing*, pioneered by Foster et al. [2006] for stateless repeated games and extended to multi-state stochastic games in Arslan and Yüksel [2017]. Other notable contributions in the regret testing tradition include Germano and Lugosi [2007] and Marden et al. [2009] for stateless repeated games and Yongacoglu et al. and Yongacoglu et al. [2022] for multi-state stochastic games.

Here, we discuss how one can modify the decentralized Q-learning algorithm of Arslan and Yüksel [2017] to accommodate for a general state space. To accomplish this generalization, we use tools from single-agent reinforcement learning theory. In particular, we build on the work of Saldi et al. [2017a], which studies quantization of state and action spaces in MDPs, and Kara et al. [2021] (building on Kara and Yüksel [2021]), which studies

Q-learning with quantization in MDPs with general state spaces.

3   Background on Stochastic Games

In this paper, we model our multi-agent environment as a *stochastic game*. A (discounted) stochastic game is a discrete-time controlled Markov process which models systems in which multiple agents interact in an environment, each agent pursuing its own objective.

A stochastic game is described by a tuple

$$(\mathcal{N}, \mathbb{X}, \{\mathbb{U}^i\}_{i \in \mathcal{N}}, \mathcal{T}, \{c^i\}_{i \in \mathcal{N}}, \{\beta^i\}_{i \in \mathcal{N}}),$$

where $\mathcal{N} = \{1, .., N\}$ is the set of N agents, $\mathbb{X}$ is a standard Borel state space, and for each $i \in \mathcal{N}$, $\mathbb{U}^i$ is agent i's (standard Borel) action space. We denote the joint action space by $\boldsymbol{U} := \mathbb{U}^1 \times \cdots \times \mathbb{U}^N$. State transitions are governed by $\mathcal{T}$, a stochastic kernel on $\mathbb{X}$ given $\mathbb{X} \times \boldsymbol{U}$. Player i's cost is determined by a stage-wise cost function $c^i : \mathbb{X} \times \boldsymbol{U} \to \mathbb{R}$, and player i uses a discount factor $\beta^i \in [0, 1)$ to aggregate its stream of stage costs.

At time $t = 0$, the system begins at state x_0. At each time $t \geq 0$, each agent i recalls the local history variable, to be described shortly, and selects an action $u_t^i \in \mathbb{U}^i$. After each agent has taken an action, agent i incurs a cost $c_t^i = c^i(x_t, \boldsymbol{u}_t)$, where $\boldsymbol{u}_t = (u_t^1, ..., u_t^N)$. The state randomly evolves according to the transition kernel as $x_{t+1} \sim \mathcal{T}(\cdot | x_t, \boldsymbol{u}_t)$.

In the independent learning paradigm, at each time step, agents may observe the global state variable, their own actions, and their own cost realization. The history of these observations may be used by each agent to choose their next action. We call this the the history/information variable, and denote player i's information variable at time t by $I_t^i = (x_0, u_0^i, c_0^i, \ldots, x_{t-1}, u_{t-1}^i, c_{t-1}^i, x_t)$. The information variable I_t^i takes values in the set $\mathbb{H}_t^i := \left(\mathbb{X} \times \mathbb{U}^i \times \mathbb{R}\right)^t \times \mathbb{X}$. Note that agents do not observe the actions of other agents.

An *policy* for agent i is a sequence of stochastic kernels $\pi^i = (\pi^i)_{t \geq 0}$ where $\pi^i \in \mathcal{P}(\mathbb{U}^i | \mathbb{H}_t^i)$. When following a policy π^i, player i selects its action $u_t^i \sim \pi_t^i(\cdot | I_t^i)$. A policy is said to be *stationary* if its action selection depends only on x_t.[1] A policy is said to be *deterministic* if the action is a deterministic function of the information variable. When the action sets $\{\mathbb{U}^i\}_i$ are taken to be compact, we define the ρ-perturbation of a policy π^i to be a randomized policy $\pi_\rho^i = (1 - \rho)\pi^i + \rho \mathrm{Uniform}(\mathbb{U}^i)$, where $\mathrm{Uniform}(\mathbb{U}^i)$ is the uniform distribution on the action set $\mathbb{U}^i$.

The objective of each agent is to minimize their own expected discounted cumulative cost, which depends on both their policy and the other agents' policies. The joint policy used by the agents $\boldsymbol{\pi}$ induces a probability measure on the trajectories of $(x_t, \boldsymbol{u}_t)_{t \geq 0}$, with associated expectation $\mathbb{E}^{\boldsymbol{\pi}}$. Then the expected discounted cost of agent i is given by

$$J_x^i(\boldsymbol{\pi}) := \mathbb{E}_x^{\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} (\beta^i)^t c^i(X_t, \boldsymbol{U_t}) \right]. \tag{1}$$

---

[1] Formally, π^i is stationary if for any $t, k \geq 0$ and information variables $I_t^i \in \mathbb{H}_t^i$, $h_k^i \in \mathbb{H}_k^i$, we have that $\pi_t^i(\cdot | I_t^i) = \pi_k^i(\cdot | h_k^i)$ whenever I_t^i and h_t^i agree in their final component.

Each agent's policy can be viewed as a response to the environment they see, which includes all other agents' policies. Let $\boldsymbol{\pi}$ be the joint policy of all agents, then denote agent i's policy by π^i and the other agents' policies by $\boldsymbol{\pi}^{-i}$.

Definition 3.1. *Let $\epsilon \geq 0$ and let Γ^i be a subset of player i's policies. A policy $\pi^{*i} \in \Gamma^i$ is an ϵ-best-reply to $\boldsymbol{\pi}^{-i}$ in Γ^i if*

$$J_x^i(\pi^{*i}, \boldsymbol{\pi}^{-i}) = \inf_{\pi^i \in \Gamma^i} J_x^i(\pi^i, \boldsymbol{\pi}^{-i}) + \epsilon, \ \forall x \in \mathbb{X}. \tag{2}$$

Furthermore, a 0-best-reply π^{*i} to $\boldsymbol{\pi}^{-i}$ is called a strict best-reply to $(\pi^i, \boldsymbol{\pi}^{-i})$ if $J_x^i(\pi^{*i}, \boldsymbol{\pi}^{-i}) < J_x^i(\pi^i, \boldsymbol{\pi}^{-i}), \quad$ for some $x \in \mathbb{X}$

Definition 3.2. *For $\epsilon \geq 0$, a policy $\boldsymbol{\pi}^* \in \boldsymbol{\Gamma}$ is an ϵ-equilibrium in $\boldsymbol{\Gamma}$ if π^{*i} is an ϵ-best-reply to $\boldsymbol{\pi}^{*-i}$ for all $i = 1, ..., N$*

When $\epsilon = 0$, a policy is simply called a best-reply instead of a 0-best-reply, and 0-equilibrium are simply called equilibrium.

4 Extending the Decentralized Q-Learning Algorithm to the Continuous-Spac Setting

In this section, we propose a decentralized multi-agent reinforcement learning algorithm for stochastic games with continuous spaces and provide our theoretical guarantees. We start by giving the relevant background on quantization of state and action spaces, and the quantized Q-learning algorithm for single-agent MDPs.

4.1 Quantization of State and Action Spaces

When the state space or the action space(s) in an MDP or a stochastic game are continuous, we may be interested in quantizing those spaces in order to run a Q-learning algorithm. At a high-level, we group similar states into a finite set of representative bins and learn a value function on these bins from which we can infer a nearly-optimal policy.

First, consider the quantization of the action space $\mathbb{U}$. Building on Saldi et al. [2017a], [Kara et al., 2021, Lemma 2.1] shows that any MDP with a weakly continuous transition kernel can be approximated by MDPs with finite action spaces. In view of this, we assume in this paper that the action spaces are finite, perhaps as a result of such a quantization.

We focus on the quantization of the state space. Building on Saldi et al. [2017a], we choose a partition of the state space $\mathbb{X}$ into M disjoint sets $\{B_i\}_{i=1}^M$, such that $\cup_i B_i = \mathbb{X}$, and $B_i \cap B_j = \emptyset$ for $i \neq j$. In each B_i, we choose (any) representative state $y_i \in B_i$, and denote the quantized finite state space by $\mathbb{Y} = \{y_1, ..., y_M\}$. We define the quantization mapping $q : \mathbb{X} \to \mathbb{Y}$ by $q(x) = y_i$ if $x \in B_i$.

We now define the *finite approximation MDP*. Let $\mu^* \in \mathcal{P}(\mathbb{X})$ be any weighting measure such that $\mu^*(B_i) > 0$ for all B_i. Now define a normalized measure on each B_i by $\hat{\mu}_{y_i}^*(A) =$

$\frac{\mu^*(A)}{\mu^*(B_i)}$ for each $A \subset B_i$. The stage-wise cost and transition kernel on $\mathbb{Y}$ are defined as an average of those quantities in the original MDP over B_i weighted by $\hat{\mu}^*_{y_i}$:

$$C^*(y_i, u) = \int_{B_i} c(x, u)\hat{\mu}^*_{y_i}(dx)$$

$$P^*(y_j|y_i, u) = \int_{B_i} \mathcal{T}(B_j|x, u)\hat{\mu}^*_{y_i}(dx)$$

(3)

We denote the optimal value function of this finite MDP by $\hat{J} : \mathbb{Y} \to \mathbb{R}$, and the optimal Q-function by $\hat{Q}^*$. This can be extended to the original state space $\mathbb{X}$ by making it constant over the quantization bins B_i. Slightly generalizing Saldi et al. [2017a], [Kara et al., 2021, Theorem 2.7] relates the finite approximation MDP to the original continuous-space MDP. It states that if $\mathcal{T}$ is weakly continuous in (x, u) and $c \colon \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ is continuous and bounded, then for any compact $K \subset \mathbb{X}$ we have

$$\sup_{x_0 \in K} |\hat{J}(x_0) - J^*(x_0)| \to 0, \quad \sup_{x_0 \in K} |J(x_0, \hat{\gamma}) - J^*(x_0)| \to 0$$

as $\max_{i=1,\dots,M-1} \sup_{x,x' \in B_i} \|x - x'\| =: \|L^-\|_\infty \to 0$, where $\hat{\gamma}$ denotes the optimal policy of the finite-state approximate model (extended to $\mathbb{X}$ via the quantization q). By assumption, $\mathbb{X}$ is a Borel subset of a Euclidean space, hence it is σ-compact. Thus, for each M, there exists a partition $\{B_i\}_{i=1}^M$ of $\mathbb{X}$ such that $\|L^-\|_\infty \to 0$ and $\cup_{i=1}^{M-1} B_i \nearrow \mathbb{X}$ as $M \to \infty$.

4.2 Single-Agent Quantized Q-Learning

In view of the above results on robustness to quantization, we consider the following quantized Q-learning algorithm for continuous-space MDPs:

$$Q_{t+1}(q(x), u) = (1 - \alpha_t(q(x), u)) Q_t(q(x), u)$$
$$+ \alpha_t(q(x), u) \left(c(x, u) + \beta \min_{v \in \mathbb{U}} Q_t(q(X_{t+1}), v) \right)$$

(4)

when $(X_t, U_t) = (x, u)$. Note that this is simply the standard Q-learning algorithm applied to the quantized states (i.e.: defined on the quantized state action pairs in $\mathbb{Y} \times \mathbb{U}$).

In Kara et al. [2021], the authors proved that under mild conditions on the transition kernel and cost function, this quantized version of the Q-learning algorithm converges to a limit which induces a near-optimal policy. In particular, the Q-learning iterations in Eq. (4) converge to the optimal Q-function of the finite MDP model in Eq. (3). This holds if the transition kernel is weakly continuous in the state and action, the cost function is continuous and bounded, the state process is positive Harris recurrent, and $\alpha_t(y, u) = (1 + \sum_{k=1}^t \mathbb{I}\{(y_k, u_k) = (y, u)\})^{-1}$. These results build on previous work in Kara and Yüksel [2021] on the convergence of finite-memory Q-learning for partially-observed Markov decision processes (POMDPs), by viewing quantization as a measurement kernel.

4.3 Decentralized Multi-Agent Q-Learning Algorithm for the Continuous-Space Setting

We now propose an extension of the decentralized multi-agent Q-learning algorithm to the continuous state space setting. Our proposed algorithm closely follows the algorithm in Arslan and Yüksel [2017] but incorporates quantization as in Kara et al. [2021]. The algorithmic description is given in Algorithm 1.

Like the finite space variant, the algorithm operates on two timescales. On the finer timescale, each agent learns an estimate of the optimal value function for their current environment, composed of the underlying stochastic game and the policies of the other agents. On the coarser timescale, agents switch policies to an estimated best-reply based on their learned value function.

In the extension to the continuous space setting, each agent has a state quantizer, q^i which maps the state space $\mathbb{X}$ to a finite set of representative states $\mathbb{Y}^i$. The agents' quantizers need not be the same. Hence, while all agents observe the same global state, the Q-learning algorithm they each run will in general see different quantizations of the state. Each agent's quantization induces a finite quantized *baseline* policy set, $\hat{\Pi}_q^i := \{\hat{\pi} : \mathbb{Y}^i \to \mathbb{U}^i\}$. When choosing actions, agents follow a perturbation of their baseline policies: with probability ρ^i, player i chooses its action uniformly at random. We denote the perturbed policy set by $\hat{\Pi}_{q,\rho}^i = \{(1 - \rho^i)\hat{\pi} + \rho^i\mathrm{Uniform}(\mathbb{U}^i) : \hat{\pi} \in \hat{\Pi}_q^i\}$.

Each environment that an agent can see corresponds to some *perturbed* quantized joint policy in the set $\hat{\mathbf{\Pi}}_{q,\rho} := \times_{i \in \mathcal{N}} \hat{\Pi}_{q,\rho}^i$. The *baseline* quantized policy set of agent i, $\hat{\Pi}_q^i$, is the set of policies that the agent considers when determining a response to the environment they see. It will be important to distinguish between these sets in the statement of our theoretical results.

During each exploration phase, the agents use some perturbed quantized policy in their policy space. This induces a stationary MDP environment from the perspective of each agent. Each agent runs the quantized Q-learning algorithm described in Section 4.2 on the MDP it sees. We will show that the quantized Q-learning algorithm converges to the optimal Q-function of the finite approximation MDP defined in Section 4.1, $\hat{Q}^{*i}_{\boldsymbol{\pi}_\rho^{-i}}$. Hence, for each 'environment' $\hat{\boldsymbol{\pi}}_\rho^{-i} \in \hat{\mathbf{\Pi}}_{q,\rho}^{-i}$, $\hat{Q}^{*i}_{\boldsymbol{\pi}_\rho^{-i}}$ induces a best-reply for agent i inside $\hat{\Pi}_q^i$, which is estimated using the learned Q_t^i. At the end of each exploration phase, each agent transitions to an estimated best-reply.

Assumption 4.1.

1. *$\mathcal{T}$ is weakly continuous in $(x, \boldsymbol{u})$*

2. *$c^i : \mathbb{X} \times \boldsymbol{U} \to \mathbb{R}_+$ is continuous and bounded for every i.*

3. *Under every stationary joint policy $\boldsymbol{\gamma} \in \hat{\mathbf{\Pi}}_{q,\rho}$, the state process $\{X_t\}_t$ is positive Harris recurrent, and thus admits a unique invariant measure, $\mu_{\boldsymbol{\gamma}}^*$*

Algorithm 1 Continuous-Space Decentralized Q-Learning Algorithm (for agent i)

Set Parameters

    q^i: agent i's quantizer of the state space $\mathbb{X}$; $\mathbb{Y}^i := q^i(\mathbb{X})$

    $\mathbb{Q}_q^i$: some compact subset of $\mathbb{R}^{|\mathbb{Y}^i| \times |\mathbb{U}^i|}$

    $\{T_k\}_{k \geq 0}$: sequence of exploration phase lengths in $\mathbb{Z}_{\geq 1}$

    $\rho^i \in (0,1)$: experimentation probability

    $\delta^i \in (0,\infty)$: tolerance for suboptimality

Initialize $\pi_0^i \in \hat{\Pi}_q^i$, $Q_0^i \in \mathbb{Q}_q^i$ (arbitrary)

Receive x_0

Iterate $k \geq 0$

    (k^{th} exploration phase)

    Iterate $t = 1, ..., T_k$

        Quantize state: $y_t^i = q^i(x_t)$

        Choose action: $u_t^i = \begin{cases} \pi_k^i(y_t) & \text{w.p. } 1 - \rho^i \\ \text{any } u^i \in \mathbb{U}^i & \text{w.p. } \rho^i \end{cases}$

        Receive $c^i(x_t, \boldsymbol{u}_t)$

        Receive $x_{t+1} \sim \mathcal{T}\left(\cdot \mid x_t, u_t^i, \boldsymbol{u}_t^{-i}\right)$,

        Quantize: $y_{t+1}^i = q^i(x_{t+1})$

        $\alpha_t^i(y_t^i, u_t^i) = \left(1 + \sum_{s=t_k}^{t} \mathbb{I}\left\{(y_s^i, u_s^i) = (y_t^i, u_t^i)\right\}\right)^{-1}$

        $Q_{t+1}^i(y_t^i, u_t^i) = \left(1 - \alpha_t^i(y_t^i, u_t^i)\right) Q_t^i(y_t^i, u_t^i)$

            $+ \alpha_t^i(y_t^i, u_t^i)\left(c^i(x_t, \boldsymbol{u}_t) + \beta \min_{v^i \in \mathbb{U}^i} Q_t^i(y_{t+1}^i, v^i)\right)$

        $Q_{t+1}^i(y^i, u^i) = Q_t^i(y^i, u^i), \quad \forall (y^i, u^i) \neq (y_t^i, u_t^i)$

    End

    $\widehat{\mathrm{BR}}_\delta^i(Q_t^i) = \left\{\hat{\gamma} \in \hat{\Pi}_q^i : Q_t^i(y, \hat{\gamma}(y)) \leq \min_{u \in \mathbb{U}^i} Q_t^i(y, u) + \delta^i, \forall y \in \mathbb{Y}_q^i\right\}$

    If $\pi_k^i \in \widehat{\mathrm{BR}}_\delta^i(Q_t^i)$

        $\pi_{k+1}^i = \pi_k^i$

    Else

        Choose $\pi_{k+1}^i \in \widehat{\mathrm{BR}}_\delta^i(Q_t^i)$

    End

    Reset Q_t^i to any $Q^i \in \mathbb{Q}_q^i$ (e.g.: $Q_t^i = 0$)

End

4. *For each stationary joint policy $\boldsymbol{\gamma} \in \hat{\boldsymbol{\Pi}}_{q,\rho}$, let $\mathbb{P}^{\boldsymbol{\gamma}}$ denote the probability measure associated with $\boldsymbol{\gamma}$. For every agent $i \in \mathcal{N}$, every observation-action pair $(y^i, u^i) \in \mathbb{Y}^i \times \mathbb{U}^i$, we suppose that (y^i, u^i) is visited infinitely often $\mathbb{P}^{\boldsymbol{\gamma}}$-almost surely.*

Theorem 4.1. *Suppose all players use Algorithm 1 to select their actions. For any $\epsilon > 0$, there exists $\tilde{T}$ such that $T_k \geq \tilde{T}$ implies*

$$\mathbb{P}\left[\left\|Q^i_{T_k} - \hat{Q}^{*i}_{\boldsymbol{\pi}^{-i}_{k,\rho}}\right\|_\infty < \epsilon\right] \geq 1 - \epsilon, \quad \forall k \geq 0, \tag{5}$$

where $\boldsymbol{\pi}_k$ is the baseline joint policy during the kth exploration phase and $\boldsymbol{\pi}_{k,\rho}$ is the perturbation of $\boldsymbol{\pi}_k$ that is used for action selection.

Furthermore, for any $\eta > 0$, there exists a fine-enough quantization q^i such that the greedy policy with respect to $\hat{Q}^{*i}_{\boldsymbol{\pi}^{-i}_\rho}$ is η-optimal in the MDP environment where the other agents act according to $\boldsymbol{\pi}^{-i}_\rho$.

Proof. Suppose each agent i uses the baseline policy $\pi^i \in \hat{\Pi}^i$. Denote the perturbed policy that agent i follows during the exploration phase by $\bar{\pi}^i \sim \rho^i \text{Unif}(\mathbb{U}^i) + (1 - \rho^i)\pi^i$, a mixture of the deterministic policy π^i and i.i.d. uniformly random actions.

Consider agent i's Q-learning updates. First, we show that agent i sees a stationary MDP. At each time step, agent i incurs the random cost $c^i(x, u^i, \boldsymbol{u}^{-i}), \boldsymbol{u}^{-i} \sim \bar{\boldsymbol{\pi}}^{-i}(x)$. Note that this is a random function of x, u and that the random component is independent and identically distributed. This is equivalent to the MDP with deterministic cost function $\tilde{c}^i(x, u^i) := \mathbb{E}_{\boldsymbol{u}^{-i} \sim \bar{\boldsymbol{\pi}}^{-i}(x)}\left[c(x, u^i, \boldsymbol{u}^{-i})\right]$. (As an aside, note that $\tilde{c}^i(x, u^i)$ converges to $c^i(x, u^i, \boldsymbol{\pi}^{-i}(x))$ as $\rho^j \to 0$ for $j \neq i$.)

Similarly, the distribution over next states seen by agent i when at state x and they take action u^i is given by the mixture $\mathbb{E}_{\boldsymbol{u}^{-i} \sim \bar{\boldsymbol{\pi}}^{-i}(x)}\left[\mathcal{T}(\cdot|x, u^i, \boldsymbol{u}^{-i})\right]$. (Note similarly that this converges weakly to the transition kernel $\mathcal{T}(\cdot|x, u^i, \boldsymbol{\pi}^{-i}(x))$ as $\rho^j \to 0$ for all $j \neq i$.)

Thus, agent i sees a stationary MDP (with a continuous state space), which we denote by $M^i_{\bar{\boldsymbol{\pi}}}$. Furthermore, since $\mathcal{T}(\cdot|x, \boldsymbol{u})$ is weakly continuous by assumption, the transition kernel of $M^i_{\bar{\boldsymbol{\pi}}}$ is weakly continuous in x, u^i. Also, since c^i is continuous and bounded in $\mathbb{X} \times \mathbb{U}$, $\tilde{c}$ is continuous and bounded in $\mathbb{X} \times \mathbb{U}^i$ (boundedness is immediate, and continuity follows from the dominated convergence theorem: $\lim_{k \to \infty} \tilde{c}(x_k, u^i_k) := \lim_{k \to \infty} \mathbb{E}\left[c(x_k, u^i_k, \boldsymbol{u}^{-i})\right] = \mathbb{E}\left[\lim_{k \to \infty} c(x_k, u^i_k, \boldsymbol{u}^{-i})\right] = \tilde{c}(x, u^i)$)

Invoking [Kara et al., 2021, Theorem 3.2, Corollary 3.3], we have that the quantized Q-learning algorithm for agent i converges almost surely to $\hat{Q}^{*i}_{\bar{\boldsymbol{\pi}}^{-i}}$. Furthermore, any greedy policy with respect to $\hat{Q}^{*i}_{\bar{\boldsymbol{\pi}}^{-i}}$ is nearly optimal. More precisely, if $\hat{\gamma}$ is such that $\hat{Q}^{*i}_{\bar{\boldsymbol{\pi}}^{-i}}(x, \hat{\gamma}(x)) = \min_{u^i \in \mathbb{U}^i} \hat{Q}^{*i}_{\bar{\boldsymbol{\pi}}^{-i}}(x, u^i)$, then for any compact $K \subset \mathbb{X}$, we have $\sup_{x_0 \in K} |J(x_0, \hat{\gamma}) - J^*(x_0)| \to 0$ as $\|L^-\| \to \infty$ where $\|L^-\|_\infty := \max_{j \in \{1,...,M-1\}} \sup_{x,x' \in B^i_j} \|x - x'\|$ and $\{B^i_j\}_j$ are the quantization bins of agent i's quantizer q^i. $\qquad\square$

The Q-factors that each agent learns, and hence their policy updates, correspond to the environment they see during that exploration phase, which includes the (perturbed)

policies that the other agents use. Thus, each agent is continually responding to their last-seen environment in a nearly-optimal way. This is a sensible goal for each agent to have, independently of all other agents. How close an agent's policy update is to optimality at the end of each exploration phase is determined by the 'resolution' of their quantizer. We can refer to the limiting state-action value function $\hat{Q}^{*i}_{\pi_{-i}}$ as agent i's 'subjective' value function since it is induced by their subjective observations via their quantizer.

In some applications, the goal might be to guarantee that the algorithm globally converges to an equilibrium. As the analysis in Section 5 shows, under certain further assumptions and adjustments to the algorithm, the algorithm converges to an equilibrium.

5 Policy-Updating Dynamics and Convergence to Equilibrium

In the previous section, our theorem was concerned with the algorithm's stage-wise behavior from the perspective of each agent. In particular, the convergence of each agent's Q-learning process and the near-optimality of the policy updates that those limiting Q-values induce. However, one may also be interested in understanding the algorithm's global dynamics over the joint policy space, and in particular, knowing whether the algorithm converges to some type of equilibrium.

In this section we derive a characterization of our algorithms' policy updating dynamics as a Markov chain over the joint policy space (when the exploration phases are long enough so that Q-learning converges). This characterization will apply to a general class of best-reply-based algorithms including the original finite-space decentralized Q-learning algorithm in Arslan and Yüksel [2017] and the continuous-space algorithm (Algorithm 1). Furthermore, under certain additional assumptions, we derive a closed-form expression for the probability of convergence to each equilibrium.

In what follows, we characterize the global policy-updating dynamics of the algorithms as a Markov chain, derive a closed-form expression for the probabilities of convergence to each equilibrium, and close by reflecting on best-reply-based MARL algorithms when the goal, in the case of the cooperative team setting, is achieving global optimality.

5.1 Global Policy-Updating Dynamics Modeled as a Markov Chain

The approach we take is to first analyze an idealized update process, then show that when the exploration phases are sufficiently long, the results derived for the idealized process hold with arbitrarily high probability and precision in the true stochastic algorithms.

Consider a general class of algorithms following the idealized policy updates in Algorithm 2. When an agent is best-responding, it does not switch its policy. When not best-responding, the agent randomly chooses a policy to switch to according to a distribution ψ which can depend on the best-reply set. In the finite-space algorithm from Arslan and Yüksel [2017], the agent stays at its current policy with probability λ^i ('inertia' parameter) and uniformly switches to a policy in its best-reply set otherwise. In Algorithm 1, the agent

always switches to a best-response (since their goal is to be individually best-responding rather than to converge to equilibrium necessarily). We consider ψ to be general in this analysis.

Note that in this section we will use BR to denote limiting best-reply sets and $\mathbf{\Pi}$ to denote the joint policy space (those quantities will depend on the agents' quantizers in the case of the quantized continuous-space algorithm).

Algorithm 2 Generalized Idealized Policy-Updating Process (for agent i)

Set Parameters
    $\psi^i(\cdot; \cdot)$ some prob. dist. over Π^i given a best-reply set
Initialize $\pi_0^i \in \Pi^i$ (arbitrarily)
Iterate $k \geq 0$
    If $\pi_k^i \in \mathrm{BR}^i(\boldsymbol{\pi}^{-i})$
        $\pi_{k+1}^i = \pi_k^i$
    Else
        $\pi_{k+1}^i = \gamma^i \in \Pi^i$ w.p. $\psi^i(\gamma^i; \mathrm{BR}^i(\boldsymbol{\pi}^{-i}))$
    End

We note that this is a stationary Markov chain in the space of joint policies, $\mathbf{\Pi}$. That is, when the agents are at joint policy $\boldsymbol{\pi_k}$, their transition to the next joint policy $\boldsymbol{\pi_{k+1}}$ is independent of all preceding policies, and is given by the dynamics in Algorithm 2. Furthermore, each agent updates its policy π^i in response to $\boldsymbol{\pi}^{-i}$ independently of the other agents' updates. We next derive the transition kernel of this Markov chain. To make the dependence explicit, we denote this transition matrix by $P(\boldsymbol{\psi}; \mathrm{BR}_G)$, where BR_G is the best-reply graph induced by the limiting Q functions.

For ease of computation, we first obtain an expression for the probability of transitioning from a given joint policy to a joint policy matching at agent i:

$$P_{\boldsymbol{\pi}, \pi'^i}^i := \mathbb{P}\left[\pi_{k+1}^i = \pi'^i | \boldsymbol{\pi}_k = \boldsymbol{\pi}\right]$$
$$= \begin{cases} 1 & \text{if } \pi^i \in \mathrm{BR}^i(\boldsymbol{\pi}^{-i}), \ \pi'^i = \pi^i, \\ \psi^i(\pi'^i; \mathrm{BR}^i(\boldsymbol{\pi}^{-i})), & \text{otherwise} \end{cases} \quad (6)$$

Now, for any pair of joint policies $(\boldsymbol{\pi}, \boldsymbol{\pi}')$, we get that the probability of transitioning from $\boldsymbol{\pi}$ to $\boldsymbol{\pi}'$ is:

$$P_{\boldsymbol{\pi}, \boldsymbol{\pi}'} := \mathbb{P}\left[\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}' | \boldsymbol{\pi}_k = \boldsymbol{\pi}\right] = \times_{i=1}^N P_{\boldsymbol{\pi}, \pi'^i}^i \quad (7)$$

5.2 Convergence to Equilibrium: Closed-form Probabilistic Characterization

Since our goal in this section is to characterize convergence to equilibrium, we require that $P(\boldsymbol{\psi}; \mathrm{BR}_G)$ is the transition matrix of an absorbing Markov chain. In Arslan and Yüksel [2017], the authors provide a sufficient condition on the class of stochastic games which

guarantees convergence to an equilibrium. This condition is termed 'weak acyclicity' and it assumes that there exists a sequence of policies from any joint policy to some equilibrium policy in which any two consecutive joint policies differ in only one agent and that agent's new policy is a best-response. In view of the characterization above, this condition clearly implies the constructed Markov chain is absorbing. However, this is not a necessary condition since, depending on ψ, there exists other finite-paths to equilibrium policies. For a fixed game with a particular best-reply graph, choosing ψ which gives positive probability to each policy in Π^i induces an absorbing Markov chain for the maximal set of stochastic games. For example, ψ could be such that the agent transitions to a best-reply w.p. $1 - \epsilon$ but explores their policy space uniformly w.p. ϵ. For the following results, we assume that ψ and BR_G are such that $P(\psi; \mathrm{BR}_G)$ is absorbing. This assumption is in the same spirit as the assumption of weak acyclicity in Arslan and Yüksel [2017], but is weaker and more general. In the next section, we will consider the special case of Algorithm 1. Using standard results on absorbing Markov chains, we can derive a closed-form expression for the probability of convergence to any equilibrium in the stochastic game. Under Algorithm 2, an equilibrium corresponds to an absorbing state in Markov chain terminology. From Eq. (6), we see that this occurs precisely when $\pi^i \in \mathrm{BR}^i(\boldsymbol{\pi}^{-i})$ for all $i \in \{1, ..., N\}$. All other joint policies are "transient states". Let t be the number of transient states and r the number of absorbing states. Thus, if we index the joint policy space appropriately, we obtain a transition matrix in absorbing canonical form:

$$P = \begin{bmatrix} Q & R \\ \mathbf{0} & I_r \end{bmatrix}$$

Where Q is the matrix of transition probabilities between transient states, R the matrix of transition probabilities from transient states to absorbing states, and I_r is the identity matrix corresponding to the fact that $P_{\boldsymbol{\pi}, \boldsymbol{\pi}} = 1$ for an absorbing state.

Remark. *In this section, what we mean by 'equilibrium' is an equilibrium according to BR_G. If the limiting Q-values correspond to the true state-action value function, this is a true equilibrium; if they differ, then it is a 'subjective' equilibrium.*

Proposition 5.1. *Let a stochastic game G be given, inducing a best-reply graph BR_G. Define $A(\psi, \mathrm{BR}_G)$ to be the vector giving the probability of converging to each joint policy in $\boldsymbol{\Pi}$. Suppose the transition matrix over $\boldsymbol{\Pi}$ is given in absorbing canonical form according to the derivation above. Then, the vector of convergence probabilities is given by*

$$A(\psi, \mathrm{BR}_G) = A_0 \begin{bmatrix} \mathbf{0} & (I_t - Q)^{-1} R \\ \mathbf{0} & I_r \end{bmatrix}, \tag{8}$$

where A_0 is a vector giving the initial distribution over the joint policy space (e.g.: uniform if each agent chooses their initially policy uniformly at random).

Proof. Let $(P_\infty)_{i,j}$ be the matrix giving the probability of converging to joint policy j if starting at joint policy i. By [Kemeny and Snell, 1983, Theorem 3.2.7], for all transient

13

initial states the probabilities are given by $(I_t - Q)^{-1}R$. For all absorbing initial states, we remain in the same state, which gives probabilities I_r. Finally, by [Kemeny and Snell, 1983, Theorem 3.1.1], the process cannot converge to a transient state, so those probabilities are 0. Thus,

$$P_\infty = \begin{bmatrix} \mathbf{0} & (I_t - Q)^{-1}R \\ \mathbf{0} & I_r \end{bmatrix}.$$

Finally, to get the absorption probabilities conditioned on the initial distribution over joint policies, we multiply by A_0. □

Thus, for the idealized update process of this general class of best-reply-based decentralized algorithms, we derived an explicit expression of the probability of converging to each equilibrium in $\mathbf{\Pi}$ (as a function of the algorithm parameters $\{\psi^i\}_{i \in \mathcal{N}}$ and the game's best-reply graph BR_G). Note that in addition to the probability of convergence to each equilibrium, absorbing Markov chain theory also tells us the expected number of policy updates required until convergence. In particular, the expected number of steps until convergence when starting at non-equilibrium policy i is $\left((I_t - Q)^{-1}\mathbf{1}\right)_i$.

As noted, the finite-space and continuous-space decentralized Q-learning algorithms are stochastic approximations of this idealized process, since the Q_t^i factors used to estimate their best-reply sets are noisy estimates of the limiting Q-values. Next, we show that for any algorithm approximating Algorithm 2 with Q-iterations which converge, our results hold approximately.

Consider a stochastic algorithm following Algorithm 2 with a particular form for $\{\psi^i\}_{i \in \mathcal{N}}$ but where the best-reply sets of each agent are estimated with learned Q-values. Let a stochastic dynamic game be given which induces a transition kernel $P(\psi; \mathrm{BR}_G)$ for the idealized policy-updating algorithm corresponding to an absorbing Markov chain in the joint policy space. Denote by P_k the stochastic realization of the transition kernel at exploration phase k for the algorithm in consideration. Note that P_k is stochastic and determined by the approximate best-reply sets estimated by the learned Q-factors Q_t^i at the end of each exploration phase. Denote by $P_{0:K}$ the matrix of terminating probabilities of the stochastic algorithm after K exploration phases. That is, $(P_{0:K})_{i,j}$ is the probability the algorithm is at joint policy j after K exploration phases given that the initial joint policy was i, $P_{0:K} = \times_{k=0}^{K} P_k$. $P_{0:K}$ is also stochastic and depends on the convergence of Q-learning for the first K exploration phases.

Proposition 5.2. *Consider a stochastic algorithm as described above approximating Algorithm 2 with corresponding $\psi = \{\psi^i\}_{i \in \mathcal{N}}$. Let the best-reply sets induced by the limiting Q-values of the algorithm be given by BR_G. Recall that the corresponding idealized best-reply process has a transition kernel $P(\psi; BR_G)$ and absorption matrix P_∞ (as defined above). Suppose that ψ and BR_G are such that $P(\psi; BR_G)$ is absorbing. Suppose that the stochastic algorithm's Q-iterations converge at every long-enough exploration phase (for this, the assumptions in [Arslan and Yüksel, 2017, Theorem 1] or Theorem 4.1 are sufficient for*

14

their respective algorithms). Then, for any $\epsilon > 0$, there exists $\bar{K}$, such that for any given $K > \bar{K}$, $K < \infty$, there exists $\bar{T}_K$ such that if $T_k > \bar{T}_K$ for $k = 0, ..., K$, then

$$\mathbb{P}\Big[\,|P_{0:K} - P_\infty|_\infty < \epsilon\,\Big] \geq 1 - \epsilon$$

Proof. Since the idealized best-reply process is an absorbing Markov chain, $\lim_{k\to\infty} P^k$ converges. So, for any $\epsilon > 0$, there exists $\bar{K}$ such that

$$\left|P^K - P_\infty\right|_\infty < \epsilon, \quad \text{for all } K > \bar{K}. \tag{9}$$

Note that the event

$$E_k := \left\{\omega \colon \left|Q^i_{\pi_k^{-i}} - Q^i_{\bar{\pi}_k^{-i}}\right|_\infty \leq \frac{1}{2}\min\{\delta^i, \bar{\delta} - \delta^i\}, \forall i\right\}$$

implies that $\Pi_k^i = \Pi_{\pi_k^{-i}}^i$ for all i. That is, the estimated best-reply set of each agent is the same as the true best-reply set. If the best-reply sets are equal at exploration phase k, then the transition kernel for the Q-learning algorithm is the same as that of the idealized best-reply process for that exploration phase ($P_k = P$).

[Arslan and Yüksel, 2017, Lemma 4] states that if T_k is large enough, then E_k can have probability arbitrarily close to one. Since K is finite, by [Arslan and Yüksel, 2017, Lemma 4], for $\epsilon > 0$, there exists $\bar{T}_K$ such that if $T_k > \bar{T}_K$ for all $k = 0, ..., K$ and $\delta^i \in (0, \bar{\delta}), \rho^i \in (0, \bar{\rho})$, for all i, then

$$\mathbb{P}\left[E_0, ..., E_K\right] \geq 1 - \epsilon. \tag{10}$$

Note that $\bar{T}_K$ depends on the specific choice of the number of exploration phases K, and not just the required lower bound $\bar{K}$.

Then, with $T_k > \bar{T}_K$ for all $k = 1, ..., K$, we have that

$$\begin{aligned}
Pr[P_{0:K} = P^K] &\geq Pr[P_i = P \text{ for all } i = 0, ..., K] \\
&\geq Pr[E_0, ..., E_k] \\
&\geq 1 - \epsilon,
\end{aligned}$$

since $E_0, ..., E_k \implies P_i = P$ for all $i = 0, ..., K \implies P_{0:K} = P^K$.

By Eq. (9), we have that for $K > \bar{K}$, $\left|P^K - P_\infty\right|_\infty < \epsilon$. This directly implies that

$$\mathbb{P}\Big[\,|P_{0:K} - P_\infty|_\infty < \epsilon\,\Big] \geq 1 - \epsilon.$$

$\square$

Remark. *Note that the finiteness of the number of exploration phases K is crucial. Thus, the result is not uniform over $K > \bar{K}$. Without finiteness, the algorithm still converges to equilibrium, but the absorption probabilities may deviate from this closed-form expression.*

Corollary 5.1. *In the above setup, suppose that the initial policies of each agent are chosen randomly according to A_0. Let A_k^s be the vector giving the probability that the decentralized Q-learning algorithm is at each policy after k exploration phases, $(A_k^s)_i := \mathbb{P}[\boldsymbol{\pi}_k = i \,|\, \boldsymbol{\pi}_0 \sim A_0]$.*

Suppose the assumptions in Proposition 5.1 hold. Then for any $\epsilon > 0$, there exists $\bar{T}_K$, $\bar{K}$ such that if $T_k \geq \bar{T}_K$, $K \geq \bar{K}$, then:

$$\mathbb{P}\Big[\,|A_K^s - A(\boldsymbol{\psi}, BR_G)|_\infty < \epsilon\Big] \geq 1 - \epsilon$$

Where $A(\boldsymbol{\psi}, BR_G)$ is the vector of absorption probabilities for the idealized process (given random initialization according to A_0) as defined in Proposition 5.1.

Proof. Recall that $A_K^s = U_{|\boldsymbol{\Pi}|} P_{0:K}$ and $A(\boldsymbol{\lambda}, \mathrm{BR}_G) = U_{|\boldsymbol{\Pi}|} P_\infty$, as per Proposition 5.1.

Then, by Proposition 5.2 for any $\epsilon > 0$, there exists $\bar{T}_K, \bar{K}$ such that $T_k \geq \bar{T}_K, K \geq \bar{K}$ implies $\mathbb{P}\Big[\,|A_K^s - A(G, \boldsymbol{\lambda})|_\infty < \epsilon\Big] \geq 1 - \epsilon$. $\qquad\square$

Remark. *Note that there are two sources of stochasticity in the above analysis. The first is the random policy initialization, and the second is the convergence of Q-learning within the required tolerance at each exploration phase.*

The above results make two main assumptions: that the noisy estimates of the limiting Q-values converge almost surely, and that $P(\boldsymbol{\psi}; \mathrm{BR}_G)$ is the transition kernel of an absorbing Markov chain. In the finite-space algorithm Arslan and Yüksel [2017] the first assumption is satisfied via the convergence of Q-learning for finite-MDPs, and the second assumption clearly holds for the class of weakly acyclic stochastic games which the authors consider. In the case of Algorithm 1 the first assumption follows from results on convergence of quantized Q-learning in continuous-space MDPs and is stated in Theorem 4.1. What remains, as will be studied in the following subsection, is the question of whether the second assumption holds.

Shortcomings of best-reply-based algorithms. In view of the analysis above, one shortcoming of the best-reply-based policy-updating dynamics is that the policy updates and the exploration of the policy space is determined only by instantaneous information in the best-reply graph. That is, agents following these algorithms switch to policies which are best-responses to their current environment, with no explicit record of prior history. Algorithms using such policy updates are able to guarantee convergence to player-by-player equilibria (local optimality), but they are unable to distinguish between global and local optimality in the cooperative team setting. In fact, the closed-form probabilities of convergence that we derive above show that, in the worst case, convergence to global optimality can occur with arbitrarily low probability. That is, there exists team problems and choices of parameters where convergence to a globally optimal solution only occurs if the random initialization of policies happens to be a global optimum. This behavior is particularly undesirable in team settings where the goal is finding a team-optimal joint policy. It can be shown that, in general, a player-by-player optimal solution can be arbitrarily worse than a team-optimal Yongacoglu et al. [2022].

16

6 Convergence to Equilibrium

As noted earlier, whether such MARL algorithms converge to a notion of equilibrium depends on the subjective state-value functions seen by each agent at each possible environment. We first answer the following question: if Algorithm 1 (perhaps with $\{\psi^i\}$ adjusted), what does it converge to? That is, what is a 'subjective equilibrium' according to this algorithm?

Proposition 6.1. *If the algorithm converges to a subjective equilibrium $\boldsymbol{\pi}$ for some game (i.e.: agents stop updating their policies), then the joint policy it converges to is an ϵ-equilibrium in $\boldsymbol{\Pi}$, where $\epsilon \to 0$ as the quantization gets finer.*

Proof. Suppose that for some stochastic game, the algorithm converges to a subjective equilibrium. Then each agent is best-responding with respect to their subjective value function, and hence by Theorem 4.1, agent i is within $\eta(q^i)$ of an optimal response to the environment, where $\eta(q^i) \to 0$ as q^i gets finer (i.e.: $\|L^-\| \to 0$). Hence, $\boldsymbol{\pi}$ is a $\max_{i \in \mathcal{N}} \eta(q^i)$-equilibrium in $\boldsymbol{\Pi}$. $\square$

Thus, if the algorithm converges, then the joint policy it has converged to is a near-equilibrium. The other side of the question of convergence is such: when will the algorithm converge? By our analysis in the previous section, this is equivalent to asking if there exists some path with positive probability in $P(\boldsymbol{\psi}; \mathrm{BR}_G)$ which terminates at a subjective equilibrium for each starting joint policy. If we choose ψ^i as some distribution which puts positive probability at every policy in $\hat{\Pi}^i$, then the set of paths with positive probability allowable by BR_G is maximized. The condition is then met if, for example, the underlying true best-reply graph has such a path, every best-reply is unique, and the quantization is fine-enough. More generally, this is closely related to the 'satisficing paths' property as defined in Yongacoglu et al. [2021]; we refer the reader to that reference for a treatment of how to establish such a property in certain classes of stochastic games.

6.1 Games with Convergence Guarantees

In the following, we discuss examples of games in which self-play of our algorithm is convergent. We begin by describing static stochastic teams with independent measurements, where the algorithm can be used without modification, and then we describe symmetric N-player stochastic games, for which a modified version of the algorithm drives policies to ϵ-equilibrium.

Decentralized stochastic team problems with independent measurements.

An important class of continuous-space stochastic games for which our algorithm is guaranteed to converge is static stochastic teams with independent measurements, viewed as repeated games where the state is independent and identically distributed (i.e.: the transition kernel is not a function of the current state or joint action). The independent measurements

condition is a mild one, as one can, under absolute continuity conditions of measurement kernels, via a change of measures argument building on Witsenhausen Witsenhausen [1988], reduce a dynamic team problem to one which is static and with independent measurements. In this setting, the agents affect each other only via their cost functions since there are no state dynamics. Since this is a static team problem, the existence of 'satisficing paths' to an equilibrium holds whenever an equilibrium exists (via a sequential update argument leading to a monotonically non-increasing expected cost sequence converging to a limit). [Yüksel, 2020, Theorem 5.2] guarantees the existence of an optimal solution for such static problems when the (reduced) cost is continuous in the action variables and the action spaces are compact, as well as the existence of an ϵ-optimal solution among quantized policies building on Saldi et al. [2017b]. Convergence of Q-learning updates to near local equilibria follows from the independence of measurements under static reduction.

Symmetric N-Player Games

The algorithm presented in this paper prescribed switching one's policy to a (subjectively) greedy policy whenever the agent did not estimate itself to be best-replying. Under this prescription, guarantees of convergence to equilibrium depend on the existence of suitable paths, as described earlier.

Two natural modifications of the algorithm as follows: (1) replace the 0-best-reply condition in the policy update with an ϵ-best-reply condition, so that the agent opts not to change policies whenever it is already (subjectively) ϵ-best-replying to its environment; (2) when not ϵ-best-replying to one's environment, replace the best-reply policy update with random search over the entire stationary randomized policy set.

The modifications described above were proposed in Yongacoglu et al. [2021] for stochastic games with finitely many states, and it was shown that the algorithm presented there drives policies to ϵ-equilibrium in *symmetric* N-player stochastic games. The analysis there depended on the existence of so-called ϵ-satisficing paths to ϵ-equilibrium from any initial policy, and leveraged symmetry to explicitly establish the existence of such paths in symmetric games with finitely many states.

When generalizing the main algorithm of Yongacoglu et al. [2021] to symmetric games with general state spaces, the translation of convergence guarantees is not automatic: indeed, ϵ-equilibrium need not exist when the state space is general Levy [2013], Levy and McLennan [2015]. However, when an ϵ-equilibrium, or even a subjective ϵ-equilibrium, can be guaranteed to exist in a general symmetric game, the results of Yongacoglu et al. [2021] can be invoked, and that algorithm will drive play to ϵ-equilibrium, provided all agents employ the same quantizer. We leave this extension as an interesting line for future research.
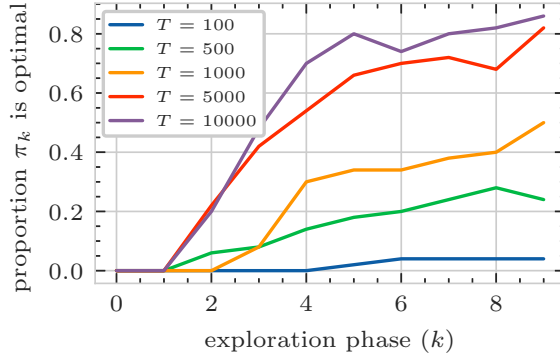
Figure 1: Simulation results: proportion of 50 trials where the policy at the kth exploration phase was optimal

7 Simulation Study

We now briefly present simulation results of Algorithm 1 on a simple 2-player stochastic team. Consider the stochastic team where $\mathbb{X} = [0,1]$, $\mathbb{U}^1 = \mathbb{U}^2 = \{-1,1\}$, $c^1(x, \boldsymbol{u}) = c^2(x, \boldsymbol{u}) = -x$, and the state dynamics are given by $x_{t+1} = [x_t + 0.1u_1u_2 + 0.1\xi\nu]_0^1$, where $\nu \sim \text{Uniform}(-1,1)$, $\xi \sim \text{Bernoulli}(0.1)$, and $[\cdot]_0^1$ denotes clipping to $[0,1]$. Thus, states with lower values have higher costs, and the team would want to head towards 1. The transition kernel is such that the state jumps forward by 0.1 if the agents cooperate (play the same action) and backwards by 0.1 otherwise, with some uniform random noise being added with probability 0.1. Each agent uses a uniform random quantizer onto the set $\mathbb{Y}^1 = \mathbb{Y}^2 = \{0, 0.25, 0.5, 0.75, 1.0\} = \mathbb{Y}$, via $q^1(x) = q^2(x) = \arg\min_{y \in \mathbb{Y}} |y - x|$.

We ran a version of Algorithm 1 with the modification that the agents choose not to update their policy ('inertia') with probability 0.25 and 0.75, respectively. We ran 50 trials for each exploration phase length T in $\{10^2, 10^3, 10^4, 10^5\}$. For each trial we kept track of the proportion of time that a team-optimal policy was used across the first 10 exploration phases (i.e. a joint policy in which the agents cooperate at every state value). For each trial, the initial joint policy is chosen such that the agents are anti-cooperating at every state, and the initial state is chosen uniformly at random in $[0,1]$, The values are averaged across trials and shown in Figure 1. Code for implementing Algorithm 1 and reproducing the simulation results can be found at `https://github.com/Awni00/decentralized-MARL-general-cts-spaces`.

8 Conclusion

In this paper, we proposed a decentralized multi-agent reinforcement learning algorithm for stochastic games with continuous spaces. We showed that as the resolution of each agent's quantizer increases, their policy-updates are asymptotically optimal responses to their most-

recently observed environment. We then analyzed the global policy-updating dynamics of the general class of best-reply-based algorithms and derived a closed-form expression for the probabilities of convergence to each equilibrium. We used this to analyze the convergence of the continuous-space algorithm to equilibrium.

References

G. Arslan and S. Yüksel. Decentralized Q-learning for stochastic teams and games. *IEEE Trans. Automat. Contr.*, 62(4):1545–1558, Apr. 2017. ISSN 0018-9286. doi: 10.1109/tac. 2016.2598476.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752):2, 1998.

C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

D. P. Foster, H. P. Young, et al. Regret testing: learning to play Nash equilibrium without knowing you have an opponent. pages 341–367, 2006.

D. Fudenberg, F. Drew, D. K. Levine, and D. K. Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

F. Germano and G. Lugosi. Global nash convergence of foster and young's regret testing. *Games and Economic Behavior*, 60(1):135–154, 2007.

J. Hu and M. P. Wellman. Nash Q-Learning for General-Sum Stochastic Games. page 31.

A. D. Kara and S. Yüksel. Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability. Mar. 2021.

A. D. Kara, N. Saldi, and S. Yüksel. Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity. Nov. 2021.

J. G. Kemeny and J. L. Snell. *Finite Markov Chains: With a New Appendix "Generalization of a Fundamental Matrix"*. Undergraduate Texts in Mathematics. Springer, New York, NY Heidelberg Berlin, reprint edition, 1983. ISBN 978-3-540-90192-1 978-0-387-90192-3.

G. J. Laurent, L. Matignon, L. Fort-Piat, et al. The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64, 2011.

Y. Levy. Discounted stochastic games with no stationary nash equilibrium: two examples. *Econometrica*, 81(5):1973–2007, 2013.

Y. J. Levy and A. McLennan. Corrigendum to "discounted stochastic games with no stationary nash equilibrium: two examples". *Econometrica*, 83(3):1237–1252, 2015.

M. L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufmann, 1994.

M. L. Littman. Friend or foe Q-Learning in general-sum games. In *In Proceedings of the 18th Int. Conf. on Machine Learning*, 2001.

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*, 48(1): 373–396, 2009.

A. Ozdaglar, M. O. Sayin, and K. Zhang. Independent learning in stochastic games. *arXiv preprint arXiv:2111.11743*, 2021.

N. Saldi, S. Yüksel, and T. Linder. On the Asymptotic Optimality of Finite Approximations to Markov Decision Processes with Borel Spaces. *Mathematics of Operations Research*, 42(4):945–978, Nov. 2017a. ISSN 0364-765X. doi: 10.1287/moor.2016.0832.

N. Saldi, S. Yüksel, and T. Linder. Finite model approximations and asymptotic optimality of quantized policies in decentralized stochastic control. *IEEE Transactions on Automatic Control*, 62:2360 – 2373, 2017b.

M. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar. Decentralized q-learning in zero-sum markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334, 2021.

L. S. Shapley. Stochastic Games*. *Proceedings of the National Academy of Sciences*, 39 (10):1095–1100, Oct. 1953. doi: 10.1073/pnas.39.10.1095.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, and R. Powell. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2, 2019.

H. S. Witsenhausen. Equivalent stochastic control problems. *Math. Control, Signals and Systems*, 1:3–11, 1988.

B. Yongacoglu, G. Arslan, and S. Yüksel. Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information.

B. Yongacoglu, G. Arslan, and S. Yüksel. Satisficing paths and independent multi-agent reinforcement learning in stochastic games. Oct. 2021.

B. Yongacoglu, G. Arslan, and S. Yüksel. Decentralized Learning for Optimality in Stochastic Dynamic Teams and Games With Local Control and Global State Information. *IEEE Transactions on Automatic Control*, 67(10):5230–5245, Oct. 2022. ISSN 1558-2523. doi: 10.1109/TAC.2021.3121228.

H. P. Young. *Strategic learning and its limits*. OUP Oxford, 2004.

S. Yüksel. A universal dynamic program and refined existence results for decentralized stochastic control. *SIAM Journal on Control and Optimization*, 58(5):2711–2739, 2020.

K. Zhang, Z. Yang, and T. Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, editors, *Handbook of Reinforcement Learning and Control*, volume 325, pages 321–384. Springer International Publishing, Cham, 2021. ISBN 978-3-030-60989-4 978-3-030-60990-0. doi: 10.1007/978-3-030-60990-0_12.