

SATISFICING PATHS AND INDEPENDENT MULTI-AGENT REINFORCEMENT LEARNING IN STOCHASTIC GAMES*

BORA YONGACOGLU[†], GÜRDAL ARSLAN[‡], AND SERDAR YÜKSEL[†]

Abstract. In multi-agent reinforcement learning (MARL), independent learners are those that do not observe the actions of other agents in the system. Due to the decentralization of information, it is challenging to design independent learners that drive play to equilibrium. This paper investigates the feasibility of using *satisficing* dynamics to guide independent learners to approximate equilibrium in stochastic games. For $\epsilon \geq 0$, an ϵ -satisficing policy update rule is any rule that instructs the agent to not change its policy when it is ϵ -best-responding to the policies of the remaining players; ϵ -satisficing paths are defined to be sequences of joint policies obtained when each agent uses some ϵ -satisficing policy update rule to select its next policy. We establish structural results on the existence of ϵ -satisficing paths into ϵ -equilibrium in both symmetric N -player games and general stochastic games with two players. We then present an independent learning algorithm for N -player symmetric games and give high probability guarantees of convergence to ϵ -equilibrium under self-play. This guarantee is made using symmetry alone, leveraging the previously unexploited structure of ϵ -satisficing paths.

Key words. Multi-agent reinforcement learning, independent learners, learning in games, stochastic games, decentralized systems

MSC codes. 91A15, 91A26, 60J20, 93A14

1. Introduction. Reinforcement learning (RL) algorithms use experience and feedback information to improve one’s performance in a control task [66]. In recent years, the field of RL has advanced tremendously both in terms of fundamental theoretical contributions (e.g. [1], [60, 61]) and successful applications (e.g. [63, 64], [49],[7]). These advances have led to the deployment of RL algorithms in large-scale systems in which many agents act, observe, and learn in a shared environment. Multi-agent reinforcement learning (MARL) is the study of emergent behaviour in complex, strategic environments, and is one of the important frontiers in modern artificial intelligence research.

The literature on MARL is relatively small when compared to that of single-agent RL, and this owes largely to the inherent challenges of learning in multi-agent settings. The first such challenge is of decentralized information: some relevant information will be unavailable to some of the players. This may occur due to strategic considerations, as competing agents may wish to hide their actions or knowledge from their rivals (as studied in [50]), or it may occur simply because of obstacles in communicating, observing, or storing large quantities of information in decentralized systems.

The second challenge inherent to MARL comes from the non-stationarity of the environment from the point of view of any individual agent (see, for instance, the survey by [29]). As an agent learns how to improve its performance, it will alter its behaviour, and this can have a destabilizing effect on the learning processes of the remaining agents, who may change their policies in response to outdated strategies. Notably, this issue arises when one tries to apply single-agent RL algorithms—which typically rely on state-action value estimates or gradient estimates that are made using historical data—in multi-agent settings. A number of studies, including [69] and [15], have reported non-convergent play when single-agent algorithms using local

*This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

[†]Department of Mathematics and Statistics, Queen’s University

[‡]Department of Electrical Engineering, University of Hawaii at Manoa

information are employed, without modification, in multi-agent settings.

Designing decentralized learning algorithms with desirable convergence properties is a task of great practical importance that lies at the intersection of the two challenges above. The notion of decentralization considered in this paper involves agents that observe a global state variable but do not observe the actions of other agents. Learning algorithms suitable for this information structure are called *independent learners* in the machine learning literature [77, 47, 46, 74]; they have also been called *payoff-based* and *radically uncoupled* in the control and game theory literatures, respectively, [43, 42, 24].

For our theoretical framework, we consider stochastic games with discounted costs. In this setting, our overarching goal is to provide MARL algorithms that are suitable for independent learners in a complex system, require little coordination among agents, and come with provable guarantees for long-run performance. To inform the development of such algorithms, this paper identifies structural properties of games that can be leveraged in algorithm design. We then illustrate the usefulness of the identified structure by providing an independent learning algorithm and proving that, under mild conditions, this algorithm leads to approximate equilibrium policies in self-play.

The structure we consider relates to *satisficing*, a natural approach to optimization that, as we discuss in §1.1 and §3, is used in several existing independent MARL algorithms. An agent that uses satisficing searches its policy space until it finds a policy that is deemed satisfactory and sufficient, at which point it settles on this policy. The agent continues to use this policy as long as the policy remains satisfactory. At a high level, the satisficing paths property formalized in §3 holds for a game and a subset of joint policies if there exist policy update rules of the satisficing variety that can drive play to equilibrium from any initial policy in the given policy subset. We show that two important classes of games—namely symmetric N -player games and general two-player games—admit this property within the set of stationary policies, which suggests that independent MARL algorithms that employ satisficing to update policies can be used to drive play to equilibrium in such games.

For N -player symmetric stochastic games, we build on this finding to present an algorithm that drives play to approximate equilibrium. This algorithm uses the exploration phase technique of [2] for policy evaluation, but differs in how players update their policies. Here, players discretize their policy space with a quantizer and use a satisficing rule to explore this quantized set, occasionally using random search when unsatisfied. Of note, here we do not restrict players to using deterministic stationary policies (pure strategies), as was done in [2], and allow for use of randomized stationary policies (mixed strategies), enabling convergence to near equilibrium in games that do not admit near equilibria in the set of deterministic stationary policies.

By relying on the satisficing paths property formalized in §3, our proof of convergence does not assume any further structure in the game beyond symmetry. To our knowledge, this is the first algorithm with formal convergence guarantees in this class of games: as we will discuss below, previous rigorous work on independent learners has focused on different—highly structured—classes of games, such as teams, potential games, weakly acyclic games, and two-player zero-sum games.

Contributions:

- (i) For any stochastic game and $\epsilon \geq 0$, we define ϵ -*satisficing paths* (Definition 3.4) and a related ϵ -*satisficing paths property* (Definition 3.5);
- (ii) In Theorem 3.6, we prove that symmetric games have the ϵ -satisficing paths property, for all $\epsilon \geq 0$. Moreover, our proof technique shows that, in symmet-

ric games, the ϵ -satisficing paths property is compatible with quantization, provided the quantization is sufficiently fine and symmetric;

- (iii) In Theorem 3.8, we prove that any two-player game has the ϵ -satisficing paths property, for all $\epsilon \geq 0$;
- (iv) We present Algorithm 2 for symmetric stochastic games and, in Theorem 5.1, we prove that self-play drives the policy process to ϵ -equilibrium.

1.1. Related Work. Beginning with Brown’s fictitious play algorithm [6, 53], the study of learning in games is nearly as old as game theory itself. There is a large and ongoing literature on fictitious play and its variants, with most works in this line considering a different information structure than the decentralized one studied here. The bulk of work on fictitious play focuses on settings with perfect monitoring of the actions of other players. This tradition includes the recent works of [37] and [59], which study fictitious play-type algorithms with perfect monitoring in stochastic games. Additionally, multiple recent studies have considered fictitious play-type algorithms for various decentralized information structures, such as [67, 20], and [58].

A number of early empirical works studied the behaviour resulting from independent RL agents coexisting in various shared environments, e.g. [69, 62, 15]. Contemporaneously, stochastic games were proposed as a theoretical framework for MARL [39]. Several *joint action learners* (learners that require access to the past actions of all other agents) were then proposed for playing stochastic games and proven to converge to equilibrium under various assumptions. A representative sampling of this stream of algorithms includes the Minimax Q-learning algorithm of [39], the Nash Q-learning algorithm of [31], and the Friend-or-Foe Q-learning algorithm of [40].

Early work on independent learners includes the following: [15] popularized the terminology of joint action learners and independent learners and stated conjectures; [35] presented an independent learner for fully cooperative games with deterministic state transitions and cost realizations and proved its convergence to optimality in that setting; and [5] proposed the WoLF-Policy Hill Climbing algorithm for general-sum stochastic games and conducted simulation studies.

Due in part to the challenges posed by non-stationarity and decentralized information, most contributions to the literature on independent learners focused either on the stateless case of repeated games and produced formal results, such as the works of [36, 24, 26, 12, 43, 42, 44], or otherwise studied the multi-state setting and presented only empirical results, such as the works [45, 46, 74].

More recently, a number of papers have studied independent learners for games with non-trivial state dynamics while still presenting rigorous guarantees. In [17], the authors studied the convergence of single-agent policy gradient algorithms employed in episodic two-player zero-sum games. It was shown that if the players’ policy updates satisfy a particular two-timescale rule, with one player updating quickly and the other updating slowly, then policies approach an approximate equilibrium. A complementary study was conducted in [58], where a different learning rule was proposed for non-episodic two-player zero-sum games. In this setting, a convergence result for the value function estimates was provided.

The preceding works produce rigorous results by taking advantage of the considerable structure of two-player zero-sum games, which are inherently adversarial strategic environments. Another class of games possessing very different exploitable structure is that of stochastic teams and their generalizations of weakly acyclic games and common interest games. An independent learning algorithm for weakly acyclic games was presented in [2]. By synchronized policy updating, this algorithm is able to

drive play to equilibrium via inertial best-response dynamics. In a recent paper [75], we modify this algorithm for use in common interest games and give high probability guarantees of convergence to team optimal policies in that setting.

Like the preceding works, this paper presents an independent learning algorithm for many state stochastic games and comes with convergence guarantees. However, this paper differs from those works in several ways. First, the class of games for which our learning algorithm has formal guarantees is distinct from those classes previously mentioned; at present, no algorithm comes with proven guarantees for general N -player symmetric games. Second, this paper also studies policy dynamics in games at large, beyond the learning setting. The structural results on ϵ -satisficing paths are of independent interest, and may be of use to other algorithm designers or to those studying equilibrium computation in stochastic games.

Organization. The remainder of the paper is organized as follows: Section 2 describes the stochastic games model and presents background results. Section 3 introduces ϵ -satisficing paths and proves structural results for symmetric N -player games and general two-player games. Building on these structural results, in Sections 4 and 5, we develop an independent learning algorithm for N -player symmetric games and give convergence guarantees. The results of a simulation study are summarized in Section 6. Additional discussion on related and future work is given in Section 7. The final section concludes. Proofs omitted from the body of the text are given in the appendices.

Notation. $\mathbb{Z}_{\geq 0}$ and \mathbb{N} denote the nonnegative and positive integers, respectively. For a finite set S , $\mathcal{P}(S)$ denotes the set of probability distributions over S . Given two 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 probability distributions on S' , with one distribution for each $s \in S$, and we write $\mathcal{T}(\cdot|s)$ for $s \in S$ to make the 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 $Y \sim \sum_{i=1}^n p_i f_i$. We use $\mathbf{1}\{\cdot\}$ to denote the indicator function of a given event. For a finite set S , $\text{Unif}(S)$ denotes the uniform distribution over S and 2^S denotes the set of subsets of S .

2. Background and Technical Preliminaries.

2.1. Stochastic Games. A finite stochastic game with discounted costs is described by the list

$$(2.1) \quad \mathcal{G} = (\mathcal{N}, \mathbb{X}, \{\mathbb{U}^i, c^i, \beta^i\}_{i \in \mathcal{N}}, P, \nu_0).$$

The components of \mathcal{G} are the following: \mathcal{N} is a finite set of $N \in \mathbb{N}$ players/agents. \mathbb{X} is a finite set of states. For agent $i \in \mathcal{N}$, \mathbb{U}^i is a finite set of actions, and we write $\mathbf{U} := \times_{i \in \mathcal{N}} \mathbb{U}^i$. An element $\mathbf{u} \in \mathbf{U}$ is called a *joint action*. For agent i , $c^i : \mathbb{X} \times \mathbf{U} \rightarrow \mathbb{R}$ is a stage cost function, and $\beta^i \in [0, 1)$ is a discount factor. A random initial state $x_0 \in \mathbb{X}$ is given by $x_0 \sim \nu_0$ where $\nu_0 \in \mathcal{P}(\mathbb{X})$. State transitions are governed by the transition kernel $P \in \mathcal{P}(\mathbb{X}|\mathbb{X} \times \mathbf{U})$.

At time $t \in \mathbb{Z}_{\geq 0}$, the state variable is denoted by $x_t \in \mathbb{X}$. Each player $i \in \mathcal{N}$ observes its local observation variable y_t^i , to be described shortly, and selects its action $u_t^i \in \mathbb{U}^i$. The joint action is denoted by $\mathbf{u}_t = (u_t^i)_{i \in \mathcal{N}} \in \mathbf{U}$. Upon selection of the joint action \mathbf{u}_t , each player $i \in \mathcal{N}$ observes its realized cost $c^i(x_t, \mathbf{u}_t) \in \mathbb{R}$, and the system transitions to state x_{t+1} , where $x_{t+1} \sim P(\cdot|x_t, \mathbf{u}_t)$.

To complete the probabilistic description of the game, we now discuss how the sequence of joint actions is generated. Each player $i \in \mathcal{N}$ uses a *policy* to select its sequence of actions $\{u_t^i\}_{t \geq 0}$, using only information that is locally available at the time of each decision. We use y_t^i to denote the observation variable for player i at time $t \geq 0$, and we let h_t^i denote player i 's information variable at time t , according to which player i selects u_t^i . We make the following assumption throughout this paper.

ASSUMPTION 1 (Independent Learners). *For each $i \in \mathcal{N}$, player i 's observation variables $\{y_t^i\}_{t \geq 0}$ and information variables $\{h_t^i\}$ are given by*

- $y_0^i = x_0$ and $y_{t+1}^i = (u_t^i, c^i(x_t, \mathbf{u}_t), x_{t+1})$ for $t \geq 0$;
- $h_0^i = y_0^i$ and $h_{t+1}^i = (h_t^i, y_{t+1}^i)$ for $t \geq 0$;

We note that this is the standard informational assumption in the literature on *independent learners* [77, 47, 46, 74, 17]. The independent learner (IL) paradigm is one of the principal alternatives to the joint action learner (JAL) paradigm, studied previously in [39, 40, 31] among many others. The key difference is that the JAL paradigm assumes that each agent i gets to observe the complete joint action profile \mathbf{u}_t after it is played; i.e. $y_{t+1}^i = (\mathbf{u}_t, c^i(x_t, \mathbf{u}_t), x_{t+1})$. In contrast, in the IL paradigm, player i does not view \mathbf{u}_t directly at any time.

DEFINITION 2.1 (Policies). *For player $i \in \mathcal{N}$, define $\mathbb{Y}^i := \mathbb{U}^i \times \mathbb{R} \times \mathbb{X}$ and $H_t^i := \mathbb{X} \times (\mathbb{Y}^i)^t$ for $t \geq 0$. A sequence $\pi^i = \{\pi_t^i\}_{t \geq 0}$ of stochastic kernels is called a policy for i if $\pi_t^i \in \mathcal{P}(\mathbb{U}^i | H_t^i)$ for every $t \geq 0$. The set of policies for i is denoted by Γ^i .*

When following the policy π^i , agent i selects its action u_t^i by sampling $u_t^i \sim \pi_t^i(\cdot | h_t^i)$. Although agents can, in principle, use arbitrarily complicated, history-dependent policies to select their actions, we will restrict our analysis to the subset of stationary policies, defined below. Such a restriction entails no loss in optimality for a particular player, provided the remaining players use stationary policies. Focusing on stationary policies is quite natural: we refer the reader to [38] for an excellent elaboration.

DEFINITION 2.2. *A policy $\pi^i \in \Gamma^i$ is called stationary if the following holds for any $t, k \geq 0$: if $h_t^i = (x_0, u_0^i, c^i(x_0, \mathbf{u}_0), \dots, x_{t-1}, u_{t-1}^i, c^i(x_{t-1}, \mathbf{u}_{t-1}), x_t) \in H_t^i$ and $\tilde{h}_k^i = (\tilde{x}_0, \tilde{u}_0^i, c^i(\tilde{x}_0, \tilde{\mathbf{u}}_0), \dots, \tilde{x}_{k-1}, \tilde{u}_{k-1}^i, c^i(\tilde{x}_{k-1}, \tilde{\mathbf{u}}_{k-1}), \tilde{x}_k) \in H_k^i$ are such that $x_t = \tilde{x}_k$, then $\pi_t^i(\cdot | h_t^i) = \pi_k^i(\cdot | \tilde{h}_k^i)$.*

In words, a stationary policy selects each action according to a probability distribution that depends only on the present state and not on the history or the time index. For player $i \in \mathcal{N}$, we denote the set of stationary policies by Γ_S^i and identify Γ_S^i with $\mathcal{P}(\mathbb{U}^i | \mathbb{X})$. To ease the notational burden, in the sequel, we treat stationary policies for player i as stochastic kernels on \mathbb{U}^i given \mathbb{X} , without reference to the complete information or history variables. Henceforth, unqualified reference to a policy shall be understood to mean a stationary policy.

We use boldface characters to denote joint objects, such as $\mathbf{u}_t = (u_t^i)_{i \in \mathcal{N}}$ above. To isolate the role of a particular player $i \in \mathcal{N}$, a joint object with i 's component removed is written using $-i$ in the agent index, e.g. $\mathbf{u}^{-i} = (u^j)_{j \in \mathcal{N}, j \neq i}$.

Given a joint policy $\boldsymbol{\pi}$, we use $\Pr^{\boldsymbol{\pi}}$ to denote the resulting probability measure on trajectories $\{(x_t, u_t)\}_{t \geq 0}$ and we use $E^{\boldsymbol{\pi}}$ to denote the associated expectation.¹ The

¹In principle, we should also introduce notation for the initial distribution in the probability measure, such as $\Pr_{\nu_0}^{\boldsymbol{\pi}}$. We omit such notation throughout this paper, because we typically condition on an initial state, making the dependence on the initial distribution redundant. In instances where we do not explicitly condition on an initial state, it should be understood that the stated property holds for any initial distribution.

objective of agent $i \in \mathcal{N}$ is to find a policy that minimizes the expectation of its series of discounted costs, given by

$$(2.2) \quad J^i(\boldsymbol{\pi}, x) := E^{\boldsymbol{\pi}} \left[\sum_{t \geq 0} (\beta^t)^t c^i(x_t, u_t^i, \mathbf{u}_t^{-i}) \mid x_0 = x \right]$$

for all $x \in \mathbb{X}$. Note that agent i controls only its own policy, π^i , but its objective function is affected by the policies of the remaining agents. This motivates the following definitions.

DEFINITION 2.3. *Let $i \in \mathcal{N}$, $\epsilon \geq 0$, and let $\Pi^i \subseteq \Gamma^i$. For $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}^{-i}$, a policy $\pi^{*i} \in \Pi^i$ is called an ϵ -best-response to $\boldsymbol{\pi}^{-i}$ over Π^i if*

$$(2.3) \quad J^i(\pi^{*i}, \boldsymbol{\pi}^{-i}, x) \leq \inf_{\pi^i \in \Pi^i} J^i(\pi^i, \boldsymbol{\pi}^{-i}, x) + \epsilon, \quad \forall x \in \mathbb{X}.$$

DEFINITION 2.4. *For fixed $i \in \mathcal{N}$, $\epsilon \geq 0$, $\Pi^i \subseteq \Gamma^i$, and $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}^{-i}$, we let $\text{BR}_\epsilon^i(\boldsymbol{\pi}^{-i}, \Pi^i)$ denote player i 's (possibly empty) set of ϵ -best-responses to $\boldsymbol{\pi}^{-i}$ over Π^i .*

DEFINITION 2.5. *Let $\epsilon \geq 0$. A joint policy $\boldsymbol{\pi}^* \in \boldsymbol{\Gamma}$ constitutes an ϵ -equilibrium if $\pi^{*i} \in \text{BR}_\epsilon^i(\boldsymbol{\pi}^{*-i}, \Gamma^i)$ for every player $i \in \mathcal{N}$.*

When $\epsilon = 0$, a 0-best-response is simply called a best-response and a 0-equilibrium is called an equilibrium. When the set Π^i over which i is optimizing is clear from context (typically, $\Pi^i = \Gamma_S^i$), we may omit ‘‘over Π^i ’’ and simply write $\text{BR}_\epsilon^i(\boldsymbol{\pi}^{-i})$.

For $\epsilon \geq 0$, we let $\boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}$ denote the set of ϵ -equilibrium policies. It is well-known that, for any finite stochastic game with discounted costs, the set of 0-equilibrium policies is non-empty [22].

The following definition will be useful in the coming sections.

DEFINITION 2.6. *Let $\xi > 0$ and $i \in \mathcal{N}$. A stationary policy $\pi^i \in \Gamma_S^i$ is called ξ -soft if $\pi^i(a^i | x) \geq \xi$ for every $x \in \mathbb{X}$ and $a^i \in \mathcal{U}^i$. The policy $\pi^i \in \Gamma_S^i$ is called soft if it is ξ -soft for some $\xi > 0$.*

2.2. Symmetric Games. In some applications, the strategic environment being modelled exhibits symmetry in the agents. To model such settings, we define a class of symmetric games with the following properties: (1) each agent has the same set of actions; (2) the state dynamics depend only on the profile of actions taken by all players, without special dependence on the identities of the agents. That is, permuting the agents' actions in a joint action leaves the conditional probabilities for the next state unchanged; (3) such a permutation results in a corresponding permutation of costs incurred. We formalize and clarify these points in the definition below.

First, we introduce additional notation: if $\mathcal{U}^i = \mathcal{U}^j$ for all $i, j \in \mathcal{N}$, given a permutation $\sigma : \mathcal{N} \rightarrow \mathcal{N}$ and joint action $\mathbf{a} = (a^i)_{i \in \mathcal{N}}$, we define $\sigma(\mathbf{a}) \in \mathcal{U}$ to be the joint action in which i 's component is given by $a^{\sigma(i)}$. That is, player i 's action in $\sigma(\mathbf{a})$ is given by player $\sigma(i)$'s action in \mathbf{a} : $\sigma(\mathbf{a})^i = a^{\sigma(i)}$.

DEFINITION 2.7 (Symmetric Game). *A stochastic game \mathcal{G} given by (2.1) is called symmetric if the following holds:*

- *There exists a set \mathcal{U} and a constant $\beta \in (0, 1)$ such that $\mathcal{U}^i = \mathcal{U}$ and $\beta^i = \beta$ for all $i \in \mathcal{N}$;*
- *For any $i \in \mathcal{N}$, permutation σ on \mathcal{N} , and $(x, \mathbf{a}) \in \mathbb{X} \times \mathcal{U}$, we have*

$$c^i(x, \sigma(\mathbf{a})) = c^{\sigma(i)}(x, \mathbf{a}), \quad \text{and} \quad P(\cdot | x, \mathbf{a}) = P(\cdot | x, \sigma(\mathbf{a})).$$

Observe the following useful facts about symmetric games.

LEMMA 2.8. *Let \mathcal{G} be a symmetric game and let $\pi \in \Gamma_S$ be a stationary joint policy. For $i, j \in \mathcal{N}$, if $\pi^i = \pi^j$, then $J^i(\pi^i, \pi^{-i}, x) = J^j(\pi^j, \pi^{-j}, x)$, for any $x \in \mathbb{X}$.*

Proof. Let σ be the permutation on \mathcal{N} such that $\sigma(i) = j$, $\sigma(j) = i$, and $\sigma(p) = p$ for all $p \in \mathcal{N} \setminus \{i, j\}$. For any $t \geq 0$ and joint action $\mathbf{a} \in \mathbf{U}$, it follows from $\pi^i = \pi^j$ that $\Pr^\pi(\mathbf{u}_t = \mathbf{a}) = \Pr^\pi(\mathbf{u}_t = \sigma(\mathbf{a}))$. Then, since $c^i(x, \sigma(\mathbf{a})) = c^j(x, \mathbf{a})$ for any state $x \in \mathbb{X}$, we have the following:

$$\begin{aligned} E^\pi [\beta^t c^i(x_t, \mathbf{u}_t) | x_t = x] &= \sum_{\mathbf{a} \in \mathbf{U}} \Pr^\pi(\mathbf{u}_t = \sigma(\mathbf{a}) | x_t = x) \beta^t c^i(x, \sigma(\mathbf{a})) \\ &= \sum_{\mathbf{a} \in \mathbf{U}} \Pr^\pi(\mathbf{u}_t = \mathbf{a} | x_t = x) \beta^t c^j(x, \mathbf{a}) = E^\pi [\beta^t c^j(x_t, \mathbf{u}_t) | x_t = x]. \end{aligned}$$

It follows that $E^\pi[\beta^t c^i(x_t, \mathbf{u}_t)] = E^\pi[\beta^t c^j(x_t, \mathbf{u}_t)]$, and the result follows by summing over times $t \geq 0$ and taking limits. \square

COROLLARY 2.9. *Let \mathcal{G} be a symmetric game and let $\pi \in \Gamma_S$ be a stationary joint policy. For $i, j \in \mathcal{N}$, if $\pi^i = \pi^j$, then*

$$\pi^i \in \text{BR}_\epsilon^i(\pi^{-i}, \Gamma^i) \iff \pi^j \in \text{BR}_\epsilon^j(\pi^{-j}, \Gamma^j)$$

2.3. Background on Learning Algorithms.

2.3.1. Learning in MDPs. In independent learning settings, pertinent information for policy selection is not available to the players. Player i does not know the policy used by players $-i$, the value of its current policy against those of the other players, or whether its current policy is an ϵ -best-response. We now review how Q-learning can be used to address these uncertainties.

Markov decision processes (MDPs) can be viewed as a stochastic game with one player, i.e. $|\mathcal{N}| = 1$. In standard Q-learning [71], a single agent interacts with its MDP environment using some policy and maintains a vector of Q-factors, the t^{th} iterate denoted $Q_t \in \mathbb{R}^{\mathbb{X} \times \mathbf{U}}$. Upon selecting action u_t at state x_t and observing the subsequent state $x_{t+1} \sim P(\cdot | x_t, u_t)$ and cost $c(x_t, u_t)$, the Q-learning agent updates its Q-factors as follows:

(2.4)

$$Q_{t+1}(x_t, u_t) = Q_t(x_t, u_t) + \theta_t(x_t, u_t) \left[c(x_t, u_t) + \beta \min_{a \in \mathbf{U}} Q_t(x_{t+1}, a) - Q_t(x_t, u_t) \right]$$

where $\theta_t(x_t, u_t) \in [0, 1]$ is a random step-size parameter and $Q_{t+1}(s, a) = Q_t(s, a)$ for all $(s, a) \neq (x_t, u_t)$.²

Under mild conditions, $Q_t \rightarrow Q^*$ almost surely as $t \rightarrow \infty$, where $Q^* \in \mathbb{R}^{\mathbb{X} \times \mathbf{U}}$ is variously called the (state-)action value function or the Q-function [72, 70]. The value $Q^*(s, a)$ represents the expected discounted cost-to-go from the initial state s , assuming that the agent initially chooses action a and follows an optimal policy thereafter. The function Q^* is given by

$$Q^*(s, a) = E^{\pi^*} \left[\sum_{t=0}^{\infty} \beta^t c(x_t, u_t) \middle| x_0 = s, u_0 = a \right] \quad \forall (s, a) \in \mathbb{X} \times \mathbf{U},$$

²We are interested in the tabular, online variant, where access to the state, action, and cost feedback arrive piece-by-piece as the agent interacts with its environment.

where π^* is any optimal policy. The vector Q^* can then be used to construct any optimal policy $\tilde{\pi}^*$ in a state-by-state manner by setting

$$\tilde{\pi}^*(a^*|x) = 1, \text{ where } a^* \in \left\{ u \in \mathbb{U} : Q^*(x, u) = \min_{a \in \mathbb{U}} Q^*(x, a) \right\} \quad \forall x \in \mathbb{X}.$$

2.3.2. Learning in Stochastic Games. In the single-agent literature, the MDP is fixed and the Q^* notation is used, but one could also introduce notation to identify the MDP. Returning to the game setting, if all agents except i follow a stationary policy $\pi^{-i} \in \Gamma_S^{-i}$, agent i faces an environment that is equivalent to an MDP that depends on π^{-i} . We denote agent i 's t^{th} Q-factor iterate by Q_t^i and i 's optimal Q-factors when playing against π^{-i} by $Q_{\pi^{-i}}^{*i} \in \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i}$. With this notation, $Q_{\pi^{-i}}^{*i}(x, u^i)$ represents agent i 's expected discounted cost-to-go from the initial state x assuming that agent i initially chooses u^i and uses an optimal policy thereafter while the other agents use π^{-i} , a fixed stationary policy. We note that an optimal policy for i is guaranteed to exist since i faces a finite, discounted MDP, and that any optimal policy for i in this MDP is a 0-best-response to π^{-i} in the underlying game \mathcal{G} . More generally, we have the following fact: for any $i \in \mathcal{N}$, $\pi^{-i} \in \Gamma_S^{-i}$,

$$\pi^i \in \text{BR}_\epsilon^i(\pi^{-i}, \Gamma_S^i) \iff J^i(\pi^i, \pi^{-i}, x) \leq \min_{a^i \in \mathbb{U}^i} Q_{\pi^{-i}}^{*i}(x, a^i) + \epsilon, \quad \forall x \in \mathbb{X}.$$

2.4. Continuity of Value Functions and Quantized Policies. We now present some useful results on the continuity of the various value functions introduced above. We begin by metrizing the policy sets Γ_S and Γ_S^i for each player $i \in \mathcal{N}$. For $i \in \mathcal{N}$, we define a metric d^i on Γ_S^i by setting

$$d^i(\pi^i, \tilde{\pi}^i) := \max \{ |\pi^i(a^i|s) - \tilde{\pi}^i(a^i|s)| : s \in \mathbb{X}, a^i \in \mathbb{U}^i \}, \quad \forall \pi^i, \tilde{\pi}^i \in \Gamma_S^i.$$

We then define the metric \mathbf{d} on Γ_S by $\mathbf{d}(\pi, \tilde{\pi}) := \max_{i \in \mathcal{N}} d^i(\pi^i, \tilde{\pi}^i)$. The sets of policies $\{\Gamma_S^i\}_{i \in \mathcal{N}}$ are then compact in the topology induced by the metrics $\{d^i\}_{i \in \mathcal{N}}$, and similarly Γ_S is compact in the topology induced by \mathbf{d} .

LEMMA 2.10. *For any player $i \in \mathcal{N}$ and state $s \in \mathbb{X}$, the function $\varphi_s^i : \Gamma_S \rightarrow \mathbb{R}$ given by*

$$\varphi_s^i(\pi) = J^i(\pi, s) \quad \forall \pi \in \Gamma_S$$

is continuous.

LEMMA 2.11. *For any player $i \in \mathcal{N}$ and state-action $(s, a^i) \in \mathbb{X} \times \mathbb{U}^i$, the mapping $\phi_{(s, a^i)}^i : \Gamma_S^{-i} \rightarrow \mathbb{R}$ given by*

$$\phi_{(s, a^i)}^i(\pi^{-i}) = Q_{\pi^{-i}}^{*i}(s, a^i), \quad \forall \pi^{-i} \in \Gamma_S^{-i}$$

is continuous.

LEMMA 2.12. *For any player $i \in \mathcal{N}$ and state $s \in \mathbb{X}$, we have that the mapping $f_s^i : \Gamma_S^{-i} \rightarrow \mathbb{R}$ given by*

$$f_s^i(\pi^{-i}) = \min_{a^i \in \mathbb{U}^i} Q_{\pi^{-i}}^{*i}(s, a^i), \quad \forall \pi^{-i} \in \Gamma_S^{-i}$$

is continuous.

LEMMA 2.13. For any player $i \in \mathcal{N}$ and fixed policy $\pi^{-i} \in \Gamma_S^{-i}$, we have that the mapping $g_{\pi^{-i}}^i : \Gamma_S^i \rightarrow \mathbb{R}$ given by

$$g_{\pi^{-i}}^i(\pi^i) = \max_{s \in \mathbb{X}} \left(J^i(\pi^i, \pi^{-i}, s) - \min_{a^i \in \mathbb{U}^i} Q_{\pi^{-i}}^{*i}(s, a^i) \right), \quad \forall \pi^i \in \Gamma_S^i$$

is continuous.

The proofs of Lemmas 2.10–2.13 can be found in Appendix B.

2.4.1. Quantizing Policy Sets. For algorithm design purposes, we now consider the effects of restricting each player to select its policy from a finite subset of its stationary policies. The finite subset of policies will be obtained by a fine quantization of the set of all stationary policies. Due to the preceding results on continuity of the many value functions, if this quantization is sufficiently fine, then restriction to this finite subset of policies entails only a small loss in performance.

DEFINITION 2.14. Let $\xi > 0$, and $i \in \mathcal{N}$. A mapping $q^i : \Gamma_S^i \rightarrow \Gamma_S^i$ is called a ξ -quantizer if

- (i) the set $q^i(\Gamma_S^i) = \{q^i(\pi^i) : \pi^i \in \Gamma_S^i\}$ is finite, and
- (ii) For all $\pi^i \in \Gamma_S^i$, we have that $d^i(\pi^i, q^i(\pi^i)) < \xi$.

We remark that if $0 < \xi_1 < \xi_2$, then any ξ_1 -quantizer is automatically a ξ_2 -quantizer, by part (ii) in the preceding definition.

DEFINITION 2.15. Let $i \in \mathcal{N}$ and $\xi > 0$. A subset $\Pi^i \subseteq \Gamma_S^i$ is called a ξ -quantization of Γ_S^i if $\Pi^i = q^i(\Gamma_S^i)$ for some ξ -quantizer q^i .

We extend this terminology to also refer to subsets $\mathbf{\Pi} \subseteq \mathbf{\Gamma}_S$ as ξ -quantizations of $\mathbf{\Gamma}_S$ when, for each component $i \in \mathcal{N}$, Π^i is a ξ -quantization of Γ_S^i . We note that for any $\xi > 0$ and $i \in \mathcal{N}$, by the compactness of Γ_S^i , there always exists some ξ -quantization Π^i of Γ_S^i such that all policies $\pi^i \in \Pi^i$ are soft.

Since the sets $\{\Gamma_S^i\}_{i \in \mathcal{N}}$ and $\mathbf{\Gamma}_S$ are compact, it follows from Lemma 2.10 that the cost functionals are uniformly continuous on $\mathbf{\Gamma}_S$. That is, for any $\delta > 0$, there exists $\xi = \xi(\delta)$ such that for any $i \in \mathcal{N}$, $x \in \mathbb{X}$, and joint policies $\pi, \tilde{\pi} \in \mathbf{\Gamma}_S$, if $\mathbf{d}(\pi, \tilde{\pi}) < \xi$, then we have $|J^i(\pi, x) - J^i(\tilde{\pi}, x)| < \delta$.

For $\epsilon > 0$, a quantization $\mathbf{\Pi}$ of $\mathbf{\Gamma}_S$ into bins of radius less than $\xi(\frac{\epsilon}{3})$ has the desirable property that the quantization Π^i always contains an $\frac{\epsilon}{3}$ -best-response to any policy $\pi^{-i} \in \Gamma_S^{-i}$ of the remaining players. That is, $\Pi^i \cap \text{BR}_{\epsilon/3}^i(\pi^{-i}, \Gamma_S^i) \neq \emptyset$. Moreover, we are guaranteed at least one ϵ -equilibrium in the quantization $\mathbf{\Pi}$.

COROLLARY 2.16. For any $\epsilon > 0$, there exists $\xi = \xi(\epsilon, \mathcal{G}) > 0$ such that if $\mathbf{\Pi}$ is a ξ -quantization of $\mathbf{\Gamma}_S$, then we have $\mathbf{\Pi} \cap \mathbf{\Gamma}_S^{\epsilon\text{-eq}} \neq \emptyset$.

We note that, by a previous remark, Corollary 2.16 holds for any ξ' -quantization of $\mathbf{\Gamma}_S$, where $\xi' < \xi(\epsilon, \mathcal{G})$. Furthermore, such a quantization can always be selected so as to only contain soft policies.

3. Policy Dynamics and ϵ -Satisficing. This section studies discrete-time dynamical systems on the set of stationary joint policies $\mathbf{\Gamma}_S$. In particular, we focus on those dynamical systems whose trajectories are obtained by a *policy revision process*, in which each agent changes its policy according some update rule. While we do not restrict ourselves to studying systems in which all agents use the same update rule, we do focus on rules of the so-called ϵ -satisficing variety, to be defined shortly. Broadly speaking, we wish to address the following question: when can it be guaranteed that some ϵ -satisficing policy revision process drives play to ϵ -equilibrium irrespective of

the initial joint policy? We present positive results for N player symmetric games and general two player games.

For the following definitions, we let \mathcal{G} be a stochastic game given by (2.1).

DEFINITION 3.1. *For player $i \in \mathcal{N}$, a function $T^i : \Gamma_S \rightarrow \Gamma_S^i$ is called a policy update rule for player i .*

Given a collection $\mathbf{T} = \{T^i : i \in \mathcal{N}\}$ of policy update rules for each player, we will study the discrete-time orbits of \mathbf{T} . For $\boldsymbol{\pi} \in \Gamma_S$, we define $\mathbf{T}(\boldsymbol{\pi})$ to be the stationary joint policy where i 's component is given by $T^i(\boldsymbol{\pi})$. That is, $\mathbf{T}(\boldsymbol{\pi}) := (T^i(\boldsymbol{\pi}))_{i \in \mathcal{N}}$. Furthermore, for all $\boldsymbol{\pi} \in \Gamma_S$, we put $\mathbf{T}^0(\boldsymbol{\pi}) = \boldsymbol{\pi}$ and

$$\mathbf{T}^{k+1}(\boldsymbol{\pi}) = \mathbf{T}(\mathbf{T}^k(\boldsymbol{\pi})), \quad \forall k \geq 0.$$

DEFINITION 3.2. *Let $\mathbf{T} = \{T^i : i \in \mathcal{N}\}$ be a collection of policy update rules for each player, and let $\mathbb{T} : \Gamma_S \rightarrow \{(\boldsymbol{\pi}_k)_{k \geq 0} : \boldsymbol{\pi}_k \in \Gamma_S \text{ for all } k \geq 0\}$ be a mapping from joint policies to sequences of joint policies.*

If $\mathbb{T}(\boldsymbol{\pi}) = (\mathbf{T}^k(\boldsymbol{\pi}))_{k \geq 0}$ for every $\boldsymbol{\pi} \in \Gamma_S$, we say that \mathbb{T} is the policy revision process associated to \mathbf{T} .

Policy revision processes arising from specific update rules have received considerable attention in the past, both in discrete time and continuous time settings. Special interest has been given to best-response dynamics and its variants (see, for instance, [37, 48, 54]), replicator dynamics [25, 30], and the relationship between these dynamics and game theoretic learning.

Relatively fewer works focus on entire classes of policy revision processes. Two pioneering works in this tradition are [27] and [28]. In [27], the authors consider continuous time policy revision processes in normal form games and give a non-existence result: if the policy update rules satisfy certain regularity properties as well as a condition called uncoupledness, then the associated policy revision process may not converge to Nash equilibrium. In [28], the authors consider discrete time stochastic policy revision processes arising from uncoupled dynamics and provide a positive result, contingent on the policy update rules also allowing for use of memory.

Like [27] and [28], we will study a class of policy revision processes, rather than focusing on a revision process associated to a particular update rule. Instead of focusing on uncoupled dynamics, however, we will study policy update rules that instruct an agent to keep using its current policy whenever that policy is an ϵ -best-response to the prevailing joint policy. Such update rules, which we formalize as ϵ -satisficing update rules below, have the desirable property that ϵ -equilibrium policies are stable points for the associated policy revision processes.

DEFINITION 3.3. *Let $\epsilon \geq 0$ and let T^i be a policy update rule for player $i \in \mathcal{N}$. T^i is said to be ϵ -satisficing if, for any $(\pi^i, \boldsymbol{\pi}^{-i}) \in \Gamma_S$, we have that $\pi^i \in \text{BR}_\epsilon^i(\pi^i, \boldsymbol{\pi}^{-i})$ implies $T^i(\boldsymbol{\pi}) = \pi^i$.*

Our chosen terminology is inspired by [65], where ‘‘satisficing’’ refers to becoming satisfied and halting search when a sufficiently good input has been found in an optimization problem. Satisficing has a long history in both single-agent decision theory (e.g. [52], [10]) and also multi-agent game theory (e.g. [11]). Recently, there has been renewed interest in studying dynamics arising from particular ϵ -satisficing policy update rules; for example, see [8, Section 5] and [14].

A policy revision process \mathbb{T} is called ϵ -satisficing if it is associated to a collection of policy update rules $\mathbf{T} = \{T^i\}_{i \in \mathcal{N}}$ such that T^i is ϵ -satisficing for each player $i \in \mathcal{N}$. It is natural to ask the following: what assumptions must be made on a game in order

to guarantee that *some* ϵ -satisficing revision process can drive the joint policy process to $\Gamma_S^{\epsilon\text{-eq}}$ irrespective of the initial policy? With this question in mind, we state the following definitions.

DEFINITION 3.4. *Let $\epsilon \geq 0$. A (possibly finite) sequence $(\boldsymbol{\pi})_{k \geq 0}$ of stationary joint policies is called an ϵ -satisficing path if, for every $k \geq 0$ and $i \in \mathcal{N}$, $\pi_k^i \in \text{BR}_\epsilon^i(\boldsymbol{\pi}_k^{-i})$ implies $\pi_{k+1}^i = \pi_k^i$.*

DEFINITION 3.5. *Let $\epsilon \geq 0$ and let $\mathbf{\Pi} \subseteq \Gamma_S$ be a subset of stationary joint policies. A game \mathcal{G} is said to have the ϵ -satisficing paths property within $\mathbf{\Pi}$ if for every $\boldsymbol{\pi} \in \mathbf{\Pi}$, there exists an ϵ -satisficing path $(\boldsymbol{\pi}_t)_{t \geq 0}$ and an integer $K = K(\boldsymbol{\pi})$, such that (i) $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$, (ii) $\boldsymbol{\pi}_t \in \mathbf{\Pi}$ for every $t \geq 0$, and (iii) $\boldsymbol{\pi}_K \in \Gamma_S^{\epsilon\text{-eq}}$.*

We note that Definitions 3.4 and 3.5 are not attached to any particular policy revision process or collection of policy update rules. In particular, we note that Definition 3.4 does not require that a player must switch to a best-response when not already ϵ -best-responding. Consequently, one may interpret the ϵ -satisficing paths property as a necessary condition for convergence to $\Gamma_S^{\epsilon\text{-eq}}$ when players use arbitrary ϵ -satisficing update rules. It is therefore useful to establish whether this property holds (or fails to hold): in games where the ϵ -satisficing paths property does not hold within Γ_S , the use of ϵ -satisficing policy update rules may be inappropriate, as there exist initial joint policies from which $\Gamma_S^{\epsilon\text{-eq}}$ cannot be reached in finite time by following an ϵ -satisficing path.

3.1. Satisficing Paths in N -Player Symmetric Games.

THEOREM 3.6. *Let \mathcal{G} be a symmetric stochastic game given by (2.1). Then \mathcal{G} has the ϵ -satisficing paths property in Γ_S for all $\epsilon \geq 0$.*

In the proof below, we construct an ϵ -satisficing path of finite length from any initial policy into the set $\Gamma_S^{\epsilon\text{-eq}}$. Intuitively, beginning from an arbitrary policy, unsatisfied players (i.e. players not ϵ -best-responding) can change policies to match the policy of some other (not necessarily satisfied) player. We create a cohort of players using the same policy and progressively grow the cohort—either by adding an unsatisfied player to the cohort by switching its policy, or by switching the policy of every member of the cohort to match that of some other player—until a stopping condition is met. We stop either because we have found an ϵ -equilibrium or because no player is satisfied, which allows us to move in one step to an arbitrary ϵ -equilibrium.

Proof. Let $\boldsymbol{\pi}_0 \in \Gamma_S$ be an initial policy. We claim that there exists some ϵ -satisficing path of finite length from $\boldsymbol{\pi}_0$ to $\Gamma_S^{\epsilon\text{-eq}}$. Put $C_{-1} = \emptyset$ and select a player $i(0) \in \mathcal{N}$ arbitrarily. We define our first cohort, C_0 to be the subset of players whose policy matches that of player $i(0)$: $C_0 := \{j \in \mathcal{N} : \pi_0^j = \pi_0^{i(0)}\}$.

For some $n \geq 0$, suppose that we have a sequence of joint policies $\{\boldsymbol{\pi}_k\}_{k=0}^n$ and player subsets $\{C_k\}_{k=0}^n$ such that items (1)–(4) below hold for each $k \in \{0, \dots, n\}$:

- (1) All players in C_k use the same policy, i.e. $\pi_k^i = \pi_k^j$ for all $i, j \in C_k$;
- (2) $C_{k-1} \subset C_k$ and $|C_k| \geq |C_{k-1}| + 1$;
- (3) If player $i \in C_k$, $j \notin C_k$, then $\pi_k^j \neq \pi_k^i$;
- (4) $\boldsymbol{\pi}_0, \dots, \boldsymbol{\pi}_k$ is an ϵ -satisficing path.

If $\boldsymbol{\pi}_n \in \Gamma_S^{\epsilon\text{-eq}}$, then $(\boldsymbol{\pi}_0, \dots, \boldsymbol{\pi}_n)$ is an ϵ -satisficing path of finite length from $\boldsymbol{\pi}_0$ to $\Gamma_S^{\epsilon\text{-eq}}$. If $\boldsymbol{\pi}_n \notin \Gamma_S^{\epsilon\text{-eq}}$ and $C_n = \mathcal{N}$, then by Corollary 2.9, $\pi_n^i \notin \text{BR}_\epsilon^i(\boldsymbol{\pi}_n^{-i})$ for all $i \in \mathcal{N}$, and so all players may change their policies. It follows that for any $\boldsymbol{\pi}^* \in \Gamma_S^{\epsilon\text{-eq}}$, the sequence $(\boldsymbol{\pi}_0, \dots, \boldsymbol{\pi}_n, \boldsymbol{\pi}^*)$ is an ϵ -satisficing path of finite length from $\boldsymbol{\pi}_0$ to $\Gamma_S^{\epsilon\text{-eq}}$.

We now focus on the final case: $\pi_n \notin \Gamma_S^{\epsilon\text{-eq}}$ and $C_n \neq \mathcal{N}$. In this case, the sequence $(\pi_k, C_k)_{k=0}^n$ is submaximal, in that there exists a policy π_{n+1} and a player subset C_{n+1} such that the extended sequence $(\pi_k, C_k)_{k=0}^{n+1}$ satisfies (1)–(4) for each $k \in \{0, \dots, n+1\}$. We produce such a policy π_{n+1} and player set C_{n+1} now, proceeding in cases.

Since $\pi_n \notin \Gamma_S^{\epsilon\text{-eq}}$, there exists some player who is not ϵ -best-responding at π_n , i.e. there exists $i \in \mathcal{N}$ such that $\pi_n^i \notin \text{BR}_\epsilon^i(\pi_n^{-i})$. We have two sub-cases to consider: by Corollary 2.9, either (a) $\pi_n^j \in \text{BR}_\epsilon^j(\pi_n^{-j})$ for every $j \in C_n$, or (b) $\pi_n^j \notin \text{BR}_\epsilon^j(\pi_n^{-j})$ for every $j \in C_n$.

In sub-case (a), all players in C_n are ϵ -best-responding. Select an unsatisfied player $j(n+1) \in \mathcal{N} \setminus C_n$, and construct a successor policy π_{n+1} by putting $\pi_{n+1}^{j(n+1)} = \pi_n^{i^*}$, where $i^* \in C_n$ is any player in C_n , and put $\pi_{n+1}^i = \pi_n^i$ for all players $i \neq j(n+1)$. We then put $C_{n+1} = C_n \cup \{j(n+1)\}$, and the sequence $\{\pi_k, C_k\}_{k=0}^n$ is extended to $\{\pi_k, C_k\}_{k=0}^{n+1}$ while preserving (1)–(4) for all $k \in \{0, \dots, n+1\}$.

In sub-case (b), all players in C_n are allowed to switch their policy while preserving the ϵ -satisficing property of the path. Select a player $h(n+1) \in \mathcal{N} \setminus C_n$ and define the successor policy π_{n+1} as follows:

$$\pi_{n+1}^i = \begin{cases} \pi_n^i & \text{if } i \notin C_n \\ \pi_n^{h(n+1)} & \text{if } i \in C_n. \end{cases}$$

We note that the player $h(n+1)$ may be selected arbitrarily from $\mathcal{N} \setminus C_n$ and need not be ϵ -best-responding to $\pi_n^{-h(n+1)}$. Next, we define $C_{n+1} = C_n \cup \{j \in \mathcal{N} : \pi_n^j = \pi_n^{h(n+1)}\}$. Thus, the sequence $\{\pi_k, C_k\}_{k=0}^n$ has been extended to $\{\pi_k, C_k\}_{k=0}^{n+1}$ while preserving (1)–(4) for all $k \in \{1, \dots, n+1\}$.

The preceding extension process—of obtaining policy π_{n+1} and cohort C_{n+1} from π_n and C_n —can be repeated at most finitely many times before one of the two aforementioned stopping conditions (namely, $C_{n+1} = \mathcal{N}$ or $\pi_{n+1} \in \Gamma_S^{\epsilon\text{-eq}}$) is met. If the stopping condition $\pi_{n+1} \in \Gamma_S^{\epsilon\text{-eq}}$ is satisfied, we have produced an ϵ -satisficing path of finite length from π_0 into $\Gamma_S^{\epsilon\text{-eq}}$. Otherwise, the stopping condition $C_{n+1} = \mathcal{N}$ is satisfied while $\pi_{n+1} \notin \Gamma_S^{\epsilon\text{-eq}}$, in which case all players may switch policies and $(\pi_0, \dots, \pi_{n+1}, \pi^*)$ is an ϵ -satisficing into $\Gamma_S^{\epsilon\text{-eq}}$ for any $\pi^* \in \Gamma^{\epsilon\text{-eq}}$. \square

In fact, the argument in the proof of Theorem 3.6 can be used, without modification, to prove the following result, in which the policy set is restricted.

THEOREM 3.7. *Let \mathcal{G} be a symmetric game given by (2.1) and let $\epsilon \geq 0$. Let $\Pi \subseteq \Gamma_S$ be a subset of stationary joint policies satisfying $\Pi^i = \Pi^j$ for all $i, j \in \mathcal{N}$ and suppose $\Pi \cap \Gamma_S^{\epsilon\text{-eq}} \neq \emptyset$. Then, \mathcal{G} has the ϵ -satisficing paths property within Π .*

Although this result may initially appear to be only a modest generalization of Theorem 3.6, we will see in the next section that it has important consequences for algorithm design. In particular, we will see that the existence of ϵ -satisficing paths within a finite subset of policies is a sufficient condition for the convergence of some learning processes in symmetric games.

3.2. Satisficing paths in General Two-Player Games. In this subsection, we state and prove our second structural result, which is that general two-player stochastic games have the ϵ -satisficing paths property for any $\epsilon > 0$. This result assumes no symmetry in the game, and therefore requires a rather different proof technique than the one used for Theorem 3.6. The proof used here is non-constructive and relies on the continuity properties of various value functions.

THEOREM 3.8. *Let \mathcal{G} be a stochastic game given by (2.1) with $|\mathcal{N}| = 2$. Then, \mathcal{G} has the ϵ -satisficing paths property within Γ_S for any $\epsilon > 0$.*

Proof. Fix $\pi_0 \in \Gamma_S$, and let $\text{Sat}_\epsilon(\pi_0) := \{i \in \mathcal{N} : \pi_0^i \in \text{BR}_\epsilon^i(\pi_0^{-i})\}$. We will argue that there exists an ϵ -satisficing path (π_0, π_1, π_2) such that $\pi_2 \in \Gamma_S^{\epsilon\text{-eq}}$.

We proceed in three cases: either (1) $|\text{Sat}_\epsilon(\pi_0)| = 0$, (2) $|\text{Sat}_\epsilon(\pi_0)| = 1$, or (3) $|\text{Sat}_\epsilon(\pi_0)| = 2$. Cases (1) and (3) are straightforward—in Case (1) select any $\pi^* \in \Gamma_S^{\epsilon\text{-eq}}$ and put $\pi_1 = \pi_2 = \pi^*$; in Case (3), put $\pi_0 = \pi_1 = \pi_2$. Then, in either case we have that (π_0, π_1, π_2) is an ϵ -satisficing path that $\pi_2 \in \Gamma_S^{\epsilon\text{-eq}}$.

We focus now on Case (2), where exactly one player is ϵ -best-responding initially. Let $i \in \text{Sat}_\epsilon(\pi_0)$ denote the player who is ϵ -best-responding, and let $j \in \mathcal{N} \setminus \text{Sat}_\epsilon(\pi_0)$ denote the player who is not ϵ -best-responding at π_0 . There are two sub-cases to consider:

- (a) There exists $\pi^j \in \Gamma_S^j \setminus \text{BR}_\epsilon^j(\pi_0^i)$ such that $\pi_0^i \notin \text{BR}_\epsilon^i(\pi^j)$;
- (b) For all $\pi^j \in \Gamma_S^j \setminus \text{BR}_\epsilon^j(\pi_0^i)$, we have that $\pi_0^i \in \text{BR}_\epsilon^i(\pi^j)$.

In case (a), suppose $\pi_1^j \in \Gamma_S^j \setminus \text{BR}_\epsilon^j(\pi_0^i)$ is such that $\pi_0^i \notin \text{BR}_\epsilon^i(\pi_1^j)$, and put $\pi_1 = (\pi_0^i, \pi_1^j)$. Then, (π_0, π_1) is an ϵ -satisficing path, since only j changed its policy from π_0 to π_1 . By the condition defining case (a), neither player is ϵ -best-responding at π_1 , and so both may change policies. Thus, for any $\pi^* \in \Gamma_S^{\epsilon\text{-eq}}$, we have that (π_0, π_1, π^*) is an ϵ -satisficing path into $\Gamma_S^{\epsilon\text{-eq}}$.

For case (b), we will argue that the premise implies that there exists some $\pi_1^{*j} \in \text{BR}_\epsilon^j(\pi_0^i)$ such that $\pi_1 = (\pi_0^i, \pi_1^{*j}) \in \Gamma_S^{\epsilon\text{-eq}}$.

Note that for any $\pi^j \in \Gamma_S^j$, we have that $\pi^j \in \text{BR}_\epsilon^j(\pi_0^i)$ if and only if

$$\max_{s \in \mathbb{X}} \left(J^j(\pi^j, \pi_0^i, s) - \min_{a^j \in \mathbb{U}^j} Q_{\pi_0^i}^{*j}(s, a^j) \right) \leq \epsilon,$$

and an analogous condition characterizes ϵ -best-responding for player i .

We will use this formulation of ϵ -best-responding to construct a continuous function $\Phi : [0, 1] \rightarrow \mathbb{R}$. Fix some 0-best-response $\tilde{\pi}^j \in \text{BR}_0^j(\pi_0^i)$. For each $\lambda \in [0, 1]$, we define a policy $\pi_{(\lambda)}^j \in \Gamma_S^j$ as

$$\pi_{(\lambda)}^j(\cdot|x) = (1 - \lambda)\pi_0^j(\cdot|x) + \lambda\tilde{\pi}^j(\cdot|x), \quad \forall x \in \mathbb{X}.$$

We define $\Phi : [0, 1] \rightarrow \mathbb{R}$ by

$$\Phi(\lambda) := \max_{s \in \mathbb{X}} \left(J^j \left(\pi_{(\lambda)}^j, \pi_0^i, s \right) - \min_{a^j \in \mathbb{U}^j} Q_{\pi_0^i}^{*j}(s, a^j) \right), \quad \forall \lambda \in [0, 1].$$

Note that $0 = \Phi(1) < \epsilon$, since $\pi_{(1)}^j = \tilde{\pi}^j \in \text{BR}_0^j(\pi_0^i)$, and that $\epsilon < \Phi(0)$, since $\pi_{(0)}^j = \pi_0^j \notin \text{BR}_\epsilon^j(\pi_0^i)$. As it is the composition of continuous functions (Lemmas 2.10–2.13), we have that Φ is continuous. By the intermediate value theorem, there exists $\lambda \in (0, 1)$ such that $\Phi(\lambda) = \epsilon$. We put $\lambda^* = \inf\{\lambda \in (0, 1) : \Phi(\lambda) = \epsilon\} > 0$. It follows that for any $\lambda < \lambda^*$, we have $\pi_{(\lambda)}^j \notin \text{BR}_\epsilon^j(\pi_0^i)$, otherwise the minimality of λ^* is contradicted.

We take an increasing sequence $\{\lambda_n\}_{n \geq 0}$ such that $\lambda_n \uparrow \lambda^*$, and we define the policies $\gamma_n^j := \pi_{(\lambda_n)}^j$ for each $n \geq 0$. We have that $\gamma_n^j \rightarrow \pi_{(\lambda^*)}^j$, and furthermore $\gamma_n^j \in \Gamma_S^j \setminus \text{BR}_\epsilon^j(\pi_0^i)$ for each $n \geq 0$, while $\pi_{(\lambda^*)}^j \in \text{BR}_\epsilon^j(\pi_0^i)$.

By the condition defining case (b), we have that $\pi_0^i \in \text{BR}_\epsilon^i(\gamma_n^j)$ for all $n \geq 0$. Equivalently,

$$\max_{s \in \mathbb{X}} \left(J^i(\pi_0^i, \gamma_n^j, s) - \min_{a^i \in \mathbb{U}^i} Q_{\gamma_n^j}^{*i}(s, a^i) \right) \leq \epsilon \quad \forall n \geq 0.$$

By continuity, this holds taking the limit as $n \rightarrow \infty$, and so $\pi_0^i \in \text{BR}_\epsilon^i(\pi_{(\lambda^*)}^j)$. Thus, in case (b), we may put $\pi_1 = \pi_2 = (\pi_0^i, \pi_{(\lambda^*)}^j) \in \Gamma_S^{\epsilon\text{-eq}}$, which completes the proof. \square

Remarks. We now offer some intuition about the argument used to prove Theorem 3.8, and discuss difficulties that arise when one attempts to generalize this proof method to N -player games, with $N \geq 3$, or to restricted policy subsets $\Pi \subset \Gamma_S$.

Cases (1) and (3) in the preceding proof are intuitively simple. In case (1), neither agent is presently ϵ -best-responding and therefore both agents may switch their policies to any successor. In case (3), both agents are presently ϵ -best-responding and the existence of a path to ϵ -equilibrium is trivial. This leaves case (2), where exactly one agent is ϵ -best-responding, as the only remaining case.

In intuitive terms, we analyze case (2) by asking whether the unsatisfied player (player j in the proof above) can destabilize the other player *without making itself satisfied*. The ability to induce mutual dissatisfaction is the defining property of case (2a), and leads to a very simple analysis: if the ϵ -unsatisfied player can make both players unsatisfied, then both are free to switch policies in the next period. The remaining sub-case, case (2b), is simply the logical negation of case (2a). This final case involves a satisfied player, i , whose satisfaction cannot be destabilized by the unsatisfied player j as long as j remains unsatisfied. This satisfied disposition allows the unsatisfied player to approach a best-response without unsettling the already satisfied player.

Unfortunately, generalizing this proof technique to games with more than two players is rather challenging. When considering an N -player game with $N > 2$, in addition to the trivial cases—where no players are ϵ -satisfied and where all players are ϵ -satisfied—there are $N - 1$ middling cases, where there are exactly k satisfied players, with $1 \leq k \leq N - 1$. When $k > 1$, the condition analogous to case (2a) remains easy to analyze, but its logical negation fails to be useful. In the sub-case analogous to case (2b), the set of k initially satisfied players is not a monolith: changing policies may leave some of the formerly satisfied players satisfied while making others unsatisfied. As a result, the technique of taking limits and relying on continuity properties of the value functions may not yield the desired result.

As another matter, the proof technique used to prove Theorem 3.8 does not readily apply to (finite) policy subsets. That is, this proof technique does not immediately lead to a generalization of Theorem 3.8 in the way that Theorem 3.7 generalized Theorem 3.6. As we will see in the coming sections, a consequence is that algorithm design for general two-player stochastic games is rather more involved than in symmetric games with N -players.

4. Exploiting the ϵ -Satisficing Paths Property. In this section and the next, we demonstrate how one can design independent learners that exploit the ϵ -satisficing paths property of symmetric games. We begin, in this section, by developing intuition in the simplified setting where learning is black-boxed and players update their policies using an ϵ -satisficing rule that incorporates random search when not ϵ -best-responding. A complete independent learning algorithm suitable for online learning in symmetric

games is then presented in the next section and analyzed as a two-timescale, noisy implementation of the black-boxed process.

4.1. Policy revision with oracle. In Algorithm 1, we present an procedure in which players randomly revise their policies in discrete time, resulting in a time homogenous Markov chain on Γ_S . There is no learning in this idealized process: at each time step, player $i \in \mathcal{N}$ receives the relevant state value and action value information from an oracle, and uses this information to select its successor policy. We assume that the policy updates are jointly independent across agents, conditional on the information given by the oracle. That is, we do not assume shared randomness.

The relevant parameters and objects used in Algorithm 1 are the following:

- $\Pi^i \subset \Gamma_S^i$: a finite subset of policies from which i selects its policy. Π^i will be taken to be a fine quantization of Γ_S^i ;
- $\text{UpdateRule}^i \in \mathcal{P}(\Pi^i | \Pi^i \times \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i} \times \mathbb{R}^{\mathbb{X}})$ is a stochastic kernel that is used to select a (candidate) successor policy. The distribution over Π^i will depend on the current policy of player i , its Q-factors (learned or received from an oracle), and its value function estimates (learned or received from an oracle);
- $d^i \geq 0$ is a tolerance for sub-optimality. This is included to account for learning error in the next section; in this section, since there is no learning error, we take $d^i = 0$.
- An experimentation probability $e^i \in (0, 1)$: when a player is not ϵ -best-responding, it selects its successor policy according to a mixture distribution, with mixture parts UpdateRule^i and the uniform distribution on Π^i ;

Algorithm 1: Randomized Policy Revision for agent $i \in \mathcal{N}$ (with oracle)

1 **Set Parameters**

- 2 $\Pi^i \subset \Gamma_S^i$: a fine quantization of Γ_S^i
3 $\text{UpdateRule}^i \in \mathcal{P}(\Pi^i | \Pi^i \times \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i} \times \mathbb{R}^{\mathbb{X}})$: (described above)
4 $e^i \in (0, 1)$: experimentation probability when not ϵ -best-responding
5 $d^i = 0$: tolerance for sub-optimality

6 **Initialize** $\pi_0^i \in \Pi^i$: initial policy

7 **for** $k \geq 0$ (k^{th} policy update)

8 **Receive** $Q_{\pi_k}^{*i}$ and $J_{\pi_k}^i$ given by $J_{\pi_k}^i(x) = J^i(\pi_k, x)$ for all $x \in \mathbb{X}$.

9 **if** $J_{\pi_k}^i(x) \leq \min_{a^i \in \mathbb{U}^i} Q_{\pi_k}^{*i}(x, a^i) + \epsilon + d^i \quad \forall x \in \mathbb{X}$ **then**

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

11 **else** $\pi_k^i \notin \text{BR}_\epsilon^i(\pi_k^{-i})$

12 | $\pi_{k+1}^i \sim (1 - e^i)\text{UpdateRule}^i(\cdot | \pi_k^i, Q_{\pi_k}^{*i}, J_{\pi_k}^i) + e^i\text{Unif}(\Pi^i)$

13 Go to $k + 1$

LEMMA 4.1. *Let \mathcal{G} be an N -player stochastic game and let $\epsilon \geq 0$. Suppose that $\Pi \subset \Gamma_S$ is a finite subset of policies such that \mathcal{G} has the ϵ -satisficing paths property within Π . If all players update their policies according to Algorithm 1, then*

$$\lim_{k \rightarrow \infty} \Pr(\pi_k \in \Gamma_S^{\epsilon\text{-eq}}) = 1.$$

Proof. We have that the stochastic process $\{\boldsymbol{\pi}_k\}_{k \geq 0}$ is a time homogenous Markov chain on $\boldsymbol{\Pi}$, and that any policy $\boldsymbol{\pi}^* \in \boldsymbol{\Pi} \cap \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}$ is an absorbing state for this Markov chain.

By hypothesis, we have that for each $\tilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}$, there exists an ϵ -satisficing path of finite length from $\tilde{\boldsymbol{\pi}}$ into $\boldsymbol{\Pi} \cap \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}$. For each $\tilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}$, we let $L_{\tilde{\boldsymbol{\pi}}} < \infty$ denote the shortest path of positive probability from $\tilde{\boldsymbol{\pi}}$ to some ϵ -equilibrium policy in $\boldsymbol{\Pi}$, and we let $p_{\tilde{\boldsymbol{\pi}}} > 0$ be the probability that the Markov chain $\{\boldsymbol{\pi}_k\}_{k \geq 0}$ follows this path in $L_{\tilde{\boldsymbol{\pi}}}$ steps conditional on $\boldsymbol{\pi}_0 = \tilde{\boldsymbol{\pi}}$.

We then define $L := \max\{L_{\tilde{\boldsymbol{\pi}}} : \tilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}\}$ and $p := \min\{p_{\tilde{\boldsymbol{\pi}}} : \tilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}\}$. We have $L < \infty$ and $p > 0$ by the finiteness of the set $\boldsymbol{\Pi}$. Then, for any $k \geq 0$ we have

$$\Pr\left(\bigcap_{j=1}^M \{\boldsymbol{\pi}_{k+jL} \notin \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}\} \mid \boldsymbol{\pi}_k\right) \leq (1-p)^M \rightarrow 0, \text{ as } M \rightarrow \infty. \quad \square$$

COROLLARY 4.2. *Let \mathcal{G} be an N -player symmetric game and let $\epsilon > 0$. Suppose $\boldsymbol{\Pi} \subset \boldsymbol{\Gamma}_S$ is a sufficiently fine quantization of $\boldsymbol{\Gamma}_S$ such that $\boldsymbol{\Pi} \cap \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}} \neq \emptyset$ and $\Pi^i = \Pi^j$ for all $i, j \in \mathcal{N}$. If all players update their policies according to Algorithm 1, then*

$$\lim_{k \rightarrow \infty} \Pr(\boldsymbol{\pi}_k \in \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}) = 1.$$

We note that, by Corollary 2.16, a policy subset $\boldsymbol{\Pi}$ satisfying the conditions of Corollary 4.2 can always be found by taking a sufficiently fine quantization of $\boldsymbol{\Gamma}_S$.

Remarks. From Lemma 4.1, one can see that the existence of ϵ -satisficing paths within a finite subset of policies is, in fact, a sufficient condition for the convergence to ϵ -equilibrium of any ϵ -satisficing process that incorporates random search with positive probability, provided that agents restrict their policies to the finite set in question.

The significance of symmetric games in Corollary 4.2 is that—because of Corollary 2.9 and the proof of Theorem 3.6—it can be guaranteed that satisficing and quantization are compatible: whenever the quantization $\boldsymbol{\Pi}$ is sufficiently fine and symmetric ($\Pi^i = \Pi^j$ for all $i, j \in \mathcal{N}$), \mathcal{G} has the ϵ -satisficing paths property within $\boldsymbol{\Pi}$, by Theorem 3.7.

When symmetry is not assumed, it is not immediately clear that a given fine quantization will admit ϵ -satisficing paths to ϵ -equilibrium. For this reason, we do not provide an analog of Corollary 4.2 for two-player general sum games, and the learning algorithm and results in the next section are only presented for symmetric N -player games.

Although Algorithm 1 cannot be used *as is* by online independent learners, it does offer insights for the design of independent learners. In particular, both the ϵ -best-responding condition of Line 9 in Algorithm 1 as well as the policy update rule in Line 12 depend on the unobserved joint policy $\boldsymbol{\pi}_k$ only through the quantities $Q_{\boldsymbol{\pi}_k}^{*i-i}$ and $J_{\boldsymbol{\pi}_k}^i$, which we define by $J_{\boldsymbol{\pi}_k}^i(x) := J^i(\boldsymbol{\pi}_k, x)$ for each $x \in \mathbb{X}$. Crucially, both of these quantities can be estimated during play via learning *without* observing the joint action sequence $\{\mathbf{u}_t^{-i}\}_{t \geq 0}$. In the next section, we combine the adaptation mechanism of Algorithm 1 with learning to replace the oracle and achieve guarantees on finding ϵ -equilibrium even with independent learners in an online setting.

4.2. Choice of Parameters and Update Rule. We conclude with a brief discussion on the choices of $e^i \in (0, 1)$, Π^i , $\text{UpdateRule}^i \in \mathcal{P}(\Pi^i | \Pi^i \times \mathbb{R}^{\mathbb{X} \times \cup^i} \times \mathbb{R}^{\mathbb{X}})$ in Algorithm 1 and their effects on convergence to $\boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}$.

From the proof of Lemma 4.1, one sees that—from the ϵ -satisficing paths property within Π alone—random search over Π is sufficient to find and stay at an ϵ -equilibrium when using Algorithm 1. This leads to the requirement that Π is selected in such a manner that the game has the ϵ -satisficing paths property within Π .

In the context of symmetric stochastic games, this can be done by selecting Π to be a symmetric ξ -quantization of \mathbf{F}_S with ξ -sufficiently small so that Π contains an ϵ -equilibrium. From a practical point of view, taking the coarsest such quantization is sensible, because it narrows the search space. The required fineness of quantization, as measured by ξ , can be determined using the system data (cost functions, discount factors, and the transition kernel) by posing the question as an optimal quantization problem. This is an interesting question that we leave for future research.

Once an appropriate choice of Π has been made, we require that $e^i \in (0, 1)$ for each player $i \in \mathcal{N}$, so that we are guaranteed that random search can drive play to ϵ -equilibrium within Π . This is done to avoid *requiring* sophisticated policy update rules that leverage intimate knowledge of the game at hand. However, selecting e^i too large may be prohibitively slow, and the choice of UpdateRule^i may result in significant speed-up in driving play to ϵ -equilibrium.

Thus, the choice of UpdateRule^i represents one area in which algorithm designers can incorporate knowledge of the system being controlled when selecting the particular update rules for the various agents. For instance, inertial best-response dynamics may be appropriate in cooperative settings (as in [2]), while update rules employing gradient ascent/descent may be more appropriate in adversarial settings (as in [16]).

5. Synchronized two timescale learning algorithm. In this section, we present Algorithm 2, an independent learning algorithm suitable for online play of symmetric stochastic games under the decentralized information structure of Assumption 1. This algorithm can be interpreted as a noise-perturbed, two-timescale variant of Algorithm 1, where now action values and state values are estimated rather than obtained from an oracle.

The algorithm design approach used here builds on a technique presented in [2], which decouples learning and adaptation: players fix their policies for long intervals of time called exploration phases, during which they update their learning iterates. At the end of an exploration phase, players update their policies using the learned information and then reset their learning iterates ahead of the next exploration phase. This decoupled design is used to mitigate the challenges related to learning in a non-stationary environment, which is among the fundamental difficulties in MARL [29, 77]. At its core, this approach consists of four parts:

1. Time is partitioned into intervals called “exploration phases,” the k^{th} lasting $T_k \in \mathbb{N}$ stage games, beginning with the stage game at time $t_k := \sum_{l=0}^{k-1} T_l$ and ending after the stage game at $t_k + T_k - 1$;
2. *Within* an exploration phase, agent $i \in \mathcal{N}$ follows a fixed policy and obtains feedback data on state-action-cost trajectories;
3. *Within* an exploration phase, agent i processes feedback data for policy evaluation, estimation of best-response sets, and estimation of state-action values. This is done without access to the joint action information or knowledge of the joint policy;
4. *Between* the k^{th} and $(k + 1)^{\text{th}}$ exploration phases, agent i uses the learned information to update its policy from π_k^i to π_{k+1}^i . We focus here on ϵ -satisficing update rules.

Since the policy of each player is held constant within an exploration phase,

this algorithm is not a two-timescale algorithm in the traditional sense (wherein two separate sequences of iterates are updated at every step, with one iterate sequence being updated using an asymptotically larger learning rate than the other sequence), but it is two-timescale in spirit: policy adjustment is done on the slow timescale (indexed by exploration phases) while learning iterates are updated on the fast timescale (indexed by stage games).

Algorithm 2: Independent Learning with ϵ -satisficing (for agent i)

```

1 Set Parameters
2    $\Pi^i \subset \Gamma_S^i$ : a fine quantization of  $\Gamma_S^i$  satisfying Assumption 2
3   UpdateRule $^i \in \mathcal{P}(\Pi^i | \Pi^i \times \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i} \times \mathbb{R}^{\mathbb{X}})$ : described in §4
4    $\mathbb{Q}^i \subset \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i}$  and  $\mathbb{J}^i \subset \mathbb{R}^{\mathbb{X}}$ : compact sets
5    $\{T_k\}_{k \geq 0}$ : a sequence in  $\mathbb{N}$  of exploration phase lengths
6   Put  $t_0 = 0$  and  $t_{k+1} = t_k + T_k$  for all  $k \geq 0$ .
7    $e^i \in (0, 1)$ : random policy updating probability
8    $d^i \in (0, \infty)$ : tolerance level for sub-optimality

9 Initialize  $\pi_0^i \in \Pi^i$ ,  $\widehat{Q}_0^i \in \mathbb{Q}^i$ ,  $\widehat{J}_0^i \in \mathbb{J}^i$  (all arbitrary)
10 Receive Initial state  $x_0 \in \mathbb{X}$ 
11 for  $k \geq 0$  ( $k^{\text{th}}$  exploration phase)
12   for  $t = t_k, t_k + 1, \dots, t_{k+1} - 1$  // Policy evaluation loop
13     Select  $u_t^i \sim \pi_k^i$ 
14     Receive cost  $c_t^i := c^i(x_t, u_t^i, \mathbf{u}_t^{-i})$ , state  $x_{t+1}$ 
15     Set  $n_t^i := \sum_{\tau=t_k}^t \mathbf{1}\{(x_\tau, u_\tau^i) = (x_t, u_t^i)\}$ 
16     Set  $m_t^i := \sum_{\tau=t_k}^t \mathbf{1}\{x_\tau = x_t\}$ 
17      $\widehat{Q}_{t+1}^i(x_t, u_t^i) = \left(1 - \frac{1}{n_t^i}\right) \widehat{Q}_t^i(x_t, u_t^i) + \frac{1}{n_t^i} \left[c_t^i + \beta^i \min_{v^i} \widehat{Q}_t^i(x_{t+1}, v^i)\right]$ 
18      $\widehat{Q}_{t+1}^i(x, a^i) = \widehat{Q}_t^i(x, a^i)$ , for all  $(x, a^i) \neq (x_t, u_t^i)$ 
19      $\widehat{J}_{t+1}^i(x_t) = \left(1 - \frac{1}{m_t^i}\right) \widehat{J}_t^i(x_t) + \frac{1}{m_t^i} \left[c_t^i + \beta^i \widehat{J}_t^i(x_{t+1})\right]$ 
20      $\widehat{J}_{t+1}^i(x) = \widehat{J}_t^i(x)$ , for all  $x \neq x_t$ 
    // Policy Update
21   if  $\widehat{J}_{t_{k+1}}^i(x) \leq \min_{a^i} \widehat{Q}_{t_{k+1}}^i(x, a^i) + \epsilon + d^i \quad \forall x \in \mathbb{X}$  then
22     |  $\pi_{k+1}^i \leftarrow \pi_k^i$ 
23   else
24     |  $\pi_{k+1}^i \sim (1 - e^i)\text{UpdateRule}^i(\cdot | \pi_k^i, \widehat{Q}_{t_{k+1}}^i, \widehat{J}_{t_{k+1}}^i) + e^i \text{Unif}(\Pi^i)$ 
25   Reset  $\widehat{Q}_{t_{k+1}}^i$  to any  $Q^i \in \mathbb{Q}^i$  (e.g., project  $\widehat{Q}_{t_{k+1}}^i$  on  $\mathbb{Q}^i$ )
26   Reset  $\widehat{J}_{t_{k+1}}^i$  to any  $J^i \in \mathbb{J}^i$ 

```

5.1. Convergence Result for Algorithm 2. We now offer a formal guarantee for the performance of Algorithm 2 under self-play. Throughout the remainder of this section, we fix the symmetric game \mathcal{G} and the constant $\epsilon > 0$. We make the following assumptions.

ASSUMPTION 2. For a fixed symmetric game \mathcal{G} and $\epsilon > 0$, the set of policies

$\mathbf{\Pi} \subset \mathbf{\Gamma}_S$ is a quantization of $\mathbf{\Gamma}_S$ with the following properties:

- $\Pi^i = \Pi^j$ for each player $i, j \in \mathcal{N}$;
- For any $i \in \mathcal{N}$ and $\pi^i \in \Pi^i$, the policy π^i is soft;
- The set $\mathbf{\Pi} \cap \mathbf{\Gamma}_S^{\epsilon\text{-eq}}$ is non-empty

Such a policy subset $\mathbf{\Pi}$ always exists by Corollary 2.16 and the discussion of §??.

Next, we introduce a constant $\bar{d} = \bar{d}(\mathbf{\Pi}, \epsilon)$ that depends on the game \mathcal{G} , $\mathbf{\Pi}$, and ϵ . We require that the tolerance for sub-optimality d^i is positive to account for noise in the action value and state value estimates; however, d^i cannot be taken too large, otherwise player i may mistakenly suppose it is ϵ -best-responding when using policy that is truly ϵ -suboptimal policy. We let $\bar{d} := \min(S \setminus \{0\})$, where S is a finite set defined as

$$S := \left\{ \left| \epsilon - \left(J^i(\boldsymbol{\pi}, x) - \min_{a^i \in \mathbb{U}^i} Q_{\boldsymbol{\pi}^{-i}}^{*i}(x, a^i) \right) \right| : i \in \mathcal{N}, (\pi^i, \boldsymbol{\pi}^{-i}) \in \mathbf{\Pi}, x \in \mathbb{X} \right\}.$$

Prior to taking absolute values, elements of the set S can be interpreted as shifted sub-optimality gaps: for each player $i \in \mathcal{N}$, policy $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i}) \in \mathbf{\Pi}$, and state $x \in \mathbb{X}$, the quantity $J^i(\boldsymbol{\pi}, x) - \min_{a^i \in \mathbb{U}^i} Q_{\boldsymbol{\pi}^{-i}}^{*i}(x, a^i) \geq 0$ measure the degree to which player i 's policy π^i is sub-optimal against $\boldsymbol{\pi}^{-i}$ in state x . Thus, measuring

$$\left| \epsilon - \left(J^i(\boldsymbol{\pi}, x) - \min_{a^i \in \mathbb{U}^i} Q_{\boldsymbol{\pi}^{-i}}^{*i}(x, a^i) \right) \right|$$

captures how close player i is to *exactly* ϵ -best-responding in state x , and \bar{d} is the minimum non-zero discrepancy from exact ϵ -best-responding.

ASSUMPTION 3. For all $i \in \mathcal{N}$, $d^i \in (0, \bar{d})$.

We also make the following standard assumption on the state transition kernel P , which is necessary to ensure that no proper subset of states is absorbing.

ASSUMPTION 4. For any two state $x, x' \in \mathbb{X}$, there exists $H = H(x, x') \in \mathbb{N}$ and sequences of states $\{s_k\}_{k=0}^{H+1}$ and joint actions $\{\mathbf{a}_k\}_{k=0}^H$ such that $s_0 = x$, $s_{H+1} = x'$, and

$$\prod_{j=0}^H P(s_{j+1} | s_j, \mathbf{a}_j) > 0,$$

Assumption 4 requires that there is a non-zero probability of transitioning from any initial state to any other state in a finite number of steps, provided the players select an appropriate sequence of joint actions. Both the number of steps and the sequence of joint actions in question are allowed to depend on the pair of states. As such, this assumption is quite weak, and is commonly made elsewhere in the literature. (See, for instance, [58, Assumption 2-i], for an equivalent assumption on the transition kernel.) This assumption is necessary to ensure that no proper subset of states is absorbing.

THEOREM 5.1. Let \mathcal{G} be a symmetric game, $\epsilon > 0$. Suppose all players use Algorithm 2 to play the game \mathcal{G} , and that Assumptions 2, 3, and 4 hold. Then, for any $\psi > 0$, there exists $\tilde{T} = \tilde{T}(\psi, \epsilon, \mathbf{\Pi}, \{d^i\}_{i \in \mathcal{N}})$ such that if $T_k \geq \tilde{T}$ for all exploration phase indices $k \geq 0$, then

$$\Pr(\boldsymbol{\pi}_k \in \mathbf{\Pi} \cap \mathbf{\Gamma}_S^{\epsilon\text{-eq}}) \geq 1 - \psi, \quad \text{for all sufficiently large } k.$$

The proof of Theorem 5.1 is given in Appendix A.

Remarks. We have chosen to present Algorithm 2 and Theorem 5.1 in terms of symmetric games and the quantized set of policies $\mathbf{\Pi}$ to take advantage of the structural properties established in earlier sections. It is worth noting that the algorithm and convergence guarantees above can be applied more generally: in analogy to Lemma 4.1 and the remarks following Corollary 4.2, Algorithm 2 can be used to drive play to ϵ -equilibrium in non-symmetric games, provided the game has the ϵ -satisficing paths property within the finite set $\mathbf{\Pi}$ and the policies in $\mathbf{\Pi}$ are soft. These conditions can also be shown to hold for (non-symmetric) stochastic teams and exact potential games, among other classes of games.

6. Simulations. We now present the results of a simulation study of Algorithm 2 applied to a symmetric stochastic, described below.

| | | |
|-------|-------|-------|
| | a_0 | a_1 |
| a_0 | 5,5 | 0,0 |
| a_1 | 0,0 | 5,5 |

| | | |
|-------|------------|----------|
| | a_0 | a_1 |
| a_0 | 0.75, 0.75 | 1.5, 0.5 |
| a_1 | 0.5, 1.5 | 1, 1 |

(a) State s_0 : a coordination game

(b) State s_1 : prisoner's dilemma

Fig. 1: The stage games for a two-state stochastic game. Player 1 (2) picks a row (column), and its reward, to be maximized, is the 1st (2nd) entry of the chosen cell.

Here, the player set is $\mathcal{N} = \{1, 2\}$, the state space $\mathbb{X} = \{s_0, s_1\}$, and the action sets $\mathbb{U}^1 = \mathbb{U}^2 = \{a_0, a_1\}$. The cost functions are taken to be the negatives of the stage reward functions described in Figure 1. The discount factor is $\beta^i = 0.8$ for each $i \in \mathcal{N}$, and the transition kernel P is fully described by the following equations.

$$\Pr(x_{t+1} = s_0 | x_t = s_0, u_t^1 = b^1, u_t^2 = b^2) = \begin{cases} 0.8, & \text{if } b^1 = b^2 \\ 0.2, & \text{otherwise.} \end{cases}$$

$$\Pr(x_{t+1} = s_0 | x_t = s_1, u_t^1 = b^1, u_t^2 = b^2) = \begin{cases} 0.9, & \text{if } b^1 = b^2 = a_1 \\ 0.25, & \text{otherwise.} \end{cases}$$

The game described above involves two states with rather different strategic qualities: in state s_0 , the players are incentivized to coordinate actions, so as to receive a high reward and remain in the high-value state s_0 . By contrast, in state s_1 , players face the prisoner's dilemma stage game, with the added consideration that successfully cooperating (playing action a_1) drives play back to the high-value state s_0 with high probability, but cooperating while the other player defects (plays a_0) results in a lower immediate reward and does not increase the likelihood of transitioning to state s_0 .

Parameter Choices. For each $i \in \mathcal{N}$, we chose the quantized policy sets $\mathbf{\Pi}^i$ as follows. First, we define $\tilde{\mathbf{\Pi}}^i \subset \Gamma_S^i$ as $\tilde{\mathbf{\Pi}}^i := \{\pi^i \in \Gamma_S^i : 10^2 \pi^i(a^i | s) \in \mathbb{Z}, \forall s \in \mathbb{X}, a^i \in \mathbb{U}^i\}$. That is, policies in $\tilde{\mathbf{\Pi}}^i$ can be described using two digit precision after the decimal point in each component probability distribution. We then define a function $\text{Soft} : \Gamma_S^i \rightarrow \Gamma_S^i$ by

$$\text{Soft}(\pi^i) := \begin{cases} \pi^i, & \text{if } \pi^i \text{ is 0.025-soft,} \\ 0.9\pi^i + 0.1 \cdot \pi_{\text{unif}}^i & \text{otherwise,} \end{cases}$$

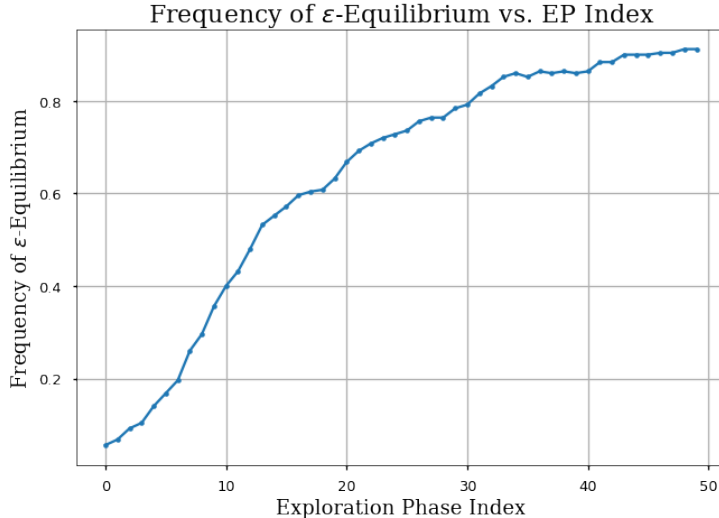


Fig. 2: Frequency of $\{\pi_k \in \Gamma_S^{\epsilon\text{-eq}} \cap \Pi\}$, averaged over 250 trials.

| EP Index | 1 | 10 | 25 | 40 | 45 | 50 |
|--|-------|-------|-------|------|-----|-------|
| $\frac{1}{250} \sum_{k=1}^{250} \mathbf{1}\{\pi_k \in \Gamma^{\epsilon\text{-eq}}\}$ | 0.056 | 0.356 | 0.728 | 0.86 | 0.9 | 0.912 |

Table 1: Selected frequencies of ϵ -equilibrium at various exploration phase indices

where π_{unif}^i is the policy that selects actions uniformly at random in each state. Finally, we put $\Pi^i = \text{Soft}(\tilde{\Pi}^i)$.

We put $\epsilon = 2$. Although this may appear to be a rather large choice of ϵ , the relatively large discount factor of $\beta^i = 0.8$ leads to aggregate long-run rewards that are an order of magnitude larger. Empirically, we found that performance within $\epsilon = 2$ of a 0-best-response entailed achieving over 80% of one’s optimal return in state s_1 and over 85% of one’s optimal return in state s_0 .

We ran 250 trials of Algorithm 2. Each trial consisted of 50 exploration phases of length 50,000, for a total of 2.5 million stage games per trial. For our choice of policy update rule, we set UpdateRule^i to be a gradient ascent-type policy update, in which the player increments its policy toward a 0-best-response using a small step size. Our empirical results, in which we observe the frequency of ϵ -equilibrium rising to over 90% of trials, are summarized in Figure 2 and Table 1.

7. Discussion. Algorithm 2 is in the tradition of Foster and Young’s regret testing algorithm [24]. Their seminal algorithm was developed for stateless repeated games and came with convergence guarantees for general (stateless, finite) two-player games. A variant of the regret testing algorithm was studied in [26], where it was shown that the variant algorithm converges to equilibrium in any *generic* N -player

stateless finite game.³ Other papers in the regret testing tradition include [43] and [2], which studied algorithms in weakly acyclic games.

Unlike earlier contributions in the regret testing tradition, the line of proof used here relies on the novel structure of ϵ -satisficing paths. Similarly, while other rigorous contributions to the study of independent learners in stochastic games have relied heavily on various structural assumptions (most notably focusing on two-player zero-sum games, N -player teams, and potential games), it appears that the rich structure of ϵ -satisficing paths has been under-exploited. To our knowledge, Algorithm 2 is the first independent learner for general symmetric stochastic games that comes with formal guarantees of convergence to ϵ -equilibrium.

Like the previous algorithms in the regret testing tradition, one shortcoming of Algorithm 2 is that players update their policies synchronously at the end of each synchronized exploration phase. This may be justifiable in certain settings where time is naturally partitioned (e.g. competitions with a natural off-season; games where certain stage game decision must be made quickly while others can be made slowly) or in cooperative applications; however, in other settings it may be inappropriate. Empirically, it appears that perfect synchrony is not needed, and [43] offers a possible approach to formalizing this, but as of yet, no proof has been given. This issue can potentially be addressed by using a conventional two-timescale algorithm, in which agents update their policies after every stage game with a slow learning rate while they update their learning iterates using a fast learning rate.

7.1. Future Work. It is natural to ask whether all N -player stochastic games have the ϵ -satisficing paths property for any $\epsilon \geq 0$. It is clear that the proof of Theorem 3.6 is not amenable to generalization, since in general games players will not have matching policy sets and therefore cannot imitate one another’s policies to grow the size of the cohort. Generalizing the proof of Theorem 3.8 is perhaps more promising, but the case-by-case breakdown used there becomes significantly complicated when $N > 2$. We leave this question open for future research.

The study of learning in symmetric games is applicable to several other technical fields. One such field is the emerging area of mean field game theory, which has been used to model various large-scale decentralized decision-making systems, such as traffic networks [13]. Mean-field games (see e.g., [33, 32, 34, 9]) can be viewed as limit models of symmetric games with finitely many agents and weakly coupled interactions. A number of works have investigated the deep connection between finite symmetric games and mean-field games with a continuum of players, e.g. [23, 56, 57, 55]. Therefore, results on equilibrium, learning, and dynamics in symmetric games are consequential for developing a theory of learning in large-scale, decentralized systems. Some of the ideas presented in this paper have been applied to N -player mean field games in [76].

Pertaining specifically to the algorithms presented here, an important question for future work is to determine how one should select the quantizer determining the finite policy sets $\{\Pi^i\}_{i \in \mathcal{N}}$ so as to achieve optimal performance with the algorithm. When selecting a ξ -quantizer, one must balance selecting sufficiently small quantizer bin radius ξ , which is needed to guarantee that some ϵ -equilibrium exists in the resulting quantization, with ensuring that the cardinality of the quantized set is relatively small, to avoid searching a needlessly large search space and slowing down convergence to ϵ -equilibrium.

In the same vein, another important question for future work involves generalizing

³We note that the class of generic games is a proper subclass of all games. As such, this guarantee may not hold for finite stateless games in general.

the learning procedure. In this paper, we have focused on (standard) tabular Q-learning and state value estimation using linear learning rates to produce each player’s estimate of whether it is ϵ -best-responding. This choice was made for simplicity of exposition and in order to make rigorous claims about convergence behaviour. Since Q-learning can be quite slow to converge (see, for instance, [68, 21]), it would be desirable to study variants of our algorithm that employ other learning algorithms, such as Speedy Q-learning [3], Zap Q-learning [19], or some form of function approximation in an attempt to shorten exploration phase lengths.

7.2. On Terminology and Other Related Work. We wish to conclude our discussion by situating our work in the broader literature on multi-agent reinforcement learning.

In using the terminology of “independent learners” to describe learners in stochastic games with full state observability at each agent but no action-sharing between agents, we follow the terminological conventions of [15] and [46]. Given that the field of MARL is relatively young, this convention, like many others, is not uniform. Some authors use the term “independent learner” to describe self-interested learners who employ best-response or policy gradient-type algorithms in stochastic games, and this terminology does not convey any assumptions about action-sharing. For examples of work that uses this language, see the recent survey by [51] and the references therein. Using the language of [51], the object of our study is model-free MARL in the minimal information setting.

Many studies in MARL, including those presented here, are concerned with asymptotic convergence guarantees. A separate line of results is concerned with giving non-asymptotic guarantees in stochastic games, such as regret bounds for the performance of a particular agent in a multi-agent system. We selectively cite [73, 4] and [41] as results in this second line, and we refer the reader to [51, Section 5] for an excellent review of recent work in this area.

A third line of research in MARL is concerned with the complexity of computing various equilibrium concepts in stochastic games, including the recent work of [18], which establishes the PPAD-hardness of computing (stationary Markov) ϵ -coarse correlated equilibria in stochastic games. We wish to point out that there is no inherent conflict in the negative results of [18] and the positive results presented here: firstly, we study randomized algorithms and give high probability guarantees, and, secondly, we do not offer complexity analysis for our algorithms. Furthermore, although we prove the existence of short ϵ -satisficing paths to ϵ -equilibrium in two-player general sum games and in N -player symmetric games, we do not study the complexity of computing such a path.

8. Conclusions. In this paper, we introduced the ϵ -satisficing paths property, a useful structural property pertaining to policy revision dynamics in stochastic games. We proved that two important classes of games, namely N -player symmetric games and two-player general games, both have this property. In the case of N -player symmetric games, we have exploited this structure to design an independent learner and showed that this algorithm drives play to near equilibrium with arbitrarily high probability under self-play. This is the first result of its kind for this class of games, and we believe that a similar design approach can be used to establish analogous results in other classes of games admitting the ϵ -satisficing paths property for $\epsilon > 0$.

Appendix A. Proof of Theorem 5.1. In this section, we prove Theorem 5.1. To study the evolution of the policy process $\{\pi_k\}_{k \geq 0}$ obtained by Algorithm 2, we first study the convergence behaviour of the learning iterates $\{\widehat{Q}_t^i\}_{t \geq 0}$ and $\{\widehat{J}_t^i\}_{t \geq 0}$ and then argue that properly selected parameters result in policy updates similar to those obtained by the oracle process studied in Lemma 4.1.

We note that when agent i uses Algorithm 2 to select its actions, it is using a particular randomized, non-stationary policy. When all agents employ Algorithm 2, we use \Pr (with no policy index in the superscript) to denote the probability measure on trajectories of states and joint actions. In contrast, for all other joint policies $\pi \in \Gamma$, we use \Pr^π to denote the associated probability measure on trajectories of states and joint actions. This distinction is made to facilitate comparing and analyzing various stochastic learning iterates.

A.1. Convergence Behaviour of Learning Iterates. We begin by studying the convergence of player learning iterates $\{\widehat{Q}_t^i\}_{t \geq 0}$ and $\{\widehat{J}_t^i\}_{t \geq 0}$. Since the policy updates are informed by the iterate sequences only at the end of each exploration phase, we give special attention to the iterate values at the sample times $\{t_{k+1}\}_{k \geq 0}$. That is, we are interested in the sequences $\{\widehat{Q}_{t_{k+1}}^i\}_{k \geq 0}$ and $\{\widehat{J}_{t_{k+1}}^i\}_{k \geq 0}$.

In the interest of comparing the iterate sequences $\{\widehat{Q}_t^i\}_t$ and $\{\widehat{J}_t^i\}_t$ with easily analyzed iterate sequences, we now introduce two related stochastic processes, $\{\bar{Q}_t^i\}_{t \geq 0}$ and $\{\bar{J}_t^i\}_{t \geq 0}$. The latter sequences are obtained using the state-action-cost trajectories using Algorithm 3, below.

Algorithm 3: Q- and J-factor Updating Without Resetting

1 Inputs

2 $\bar{Q}_0^i \in \mathbb{Q}^i \subset \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i}$, where \mathbb{Q}^i is compact

3 $\bar{J}_0^i \in \mathbb{J}^i \subset \mathbb{R}^{\mathbb{X}}$, where \mathbb{J}^i is compact

4 Trajectory $\{(x_t, u_t^i, c_t^i)\}_{t \geq 0}$, where $c_t^i = c^i(x_t, u_t^i, \mathbf{u}_t^{-i})$ for all $t \geq 0$

5 for $t \geq 0$

6 $n_t^i := \sum_{\tau=0}^t \mathbf{1}\{(x_\tau, u_\tau^i) = (x_t, u_t^i)\}$

7 $m_t^i := \sum_{\tau=0}^t \mathbf{1}\{x_\tau = x_t\}$

8 $\bar{Q}_{t+1}^i(x_t, u_t^i) = (1 - 1/n_t^i)\bar{Q}_t^i(x_t, u_t^i) + (1/n_t^i) [c_t^i + \beta \min_{\tilde{a}^i \in \mathbb{U}^i} \bar{Q}_t^i(x_{t+1}, \tilde{a}^i)]$

9 $\bar{Q}_{t+1}^i(s, a^i) = \bar{Q}_t^i(s, a^i)$ for all $(s, a^i) \neq (x_t, u_t^i)$.

10 $\bar{J}_{t+1}^i(x_t) = (1 - 1/m_t^i)\bar{J}_t^i(x_t) + (1/m_t^i) [c_t^i + \beta \bar{J}_t^i(x_{t+1})]$.

11 $\bar{J}_{t+1}^i(s) = \bar{J}_t^i(s)$ for all $s \neq x_t$.

The sequences $\{\widehat{Q}_t^i\}_{t \geq 0}$ and $\{\bar{Q}_t^i\}_{t \geq 0}$ are related through the Q-factor update. There are, however, two major differences. First, Algorithm 2 instructs player i to reset its counters at the end of the k^{th} exploration phase (i.e. after the update at time t_{k+1} , before the update at time $t_{k+1} + 1$), meaning the step sizes differ for the two iterate sequences $\{\widehat{Q}_t^i\}_{t \geq 0}$ and $\{\bar{Q}_t^i\}_{t \geq 0}$ even when the state-action-cost trajectories are identical. Second, Algorithm 2 instructs player i to reset its Q-factors at the end of the k^{th} exploration phase, while Algorithm 3 does not involve any resetting.

Consequently, one sees that the process $\{\widehat{Q}_t^i\}_{t \geq 0}$ depends on the initial condition \widehat{Q}_0^i , the state-action trajectory, and the choice of reset values after each exploration phase. In contrast, the process $\{\bar{Q}_t^i\}_{t \geq 0}$ depends only on the initial value \bar{Q}_0^i and the

state-action trajectory. Analogous remarks hold relating $\{\widehat{J}_t^i\}_{t \geq 0}$ and $\{\bar{J}_t^i\}_{t \geq 0}$.

Recall that the k^{th} exploration phase begins with the stage game at time t_k and ends before the stage game at time $t_{k+1} = t_k + T_k$. During the k^{th} exploration phase, the sequences $\{\widehat{Q}_t^i\}_{t=t_k}^{t_k+T_k}$ and $\{\widehat{J}_t^i\}_{t=t_k}^{t_k+T_k}$ depend only on the initial conditions $\widehat{Q}_{t_k}^i, \widehat{J}_{t_k}^i$, and the state-action trajectory $x_{t_k}, \mathbf{u}_{t_k}, \dots, \mathbf{u}_{t_k+T_k-1}, x_{t_k+T_k}$. This leads to the following useful observation: for any $(s_0, s_1, \dots, s_{T_k}) \in \mathbb{X}^{T_k+1}$ and $(\mathbf{a}_0, \dots, \mathbf{a}_{T_k-1}) \in \mathbf{U}^{T_k}$, we have that

$$\begin{aligned} & \Pr \left(\left\{ x_{t_k+T_k} = s_{T_k} \right\} \bigcap_{j=0}^{T_k-1} \left\{ x_{t_k+j} = s_j, \mathbf{u}_{t_k+j} = \mathbf{a}_j \right\} \middle| x_{t_k} = x, \boldsymbol{\pi}_k = \boldsymbol{\pi} \right) \\ &= \Pr^{\boldsymbol{\pi}} \left(\left\{ x_{T_k} = s_{T_k} \right\} \bigcap_{j=0}^{T_k-1} \left\{ x_j = s_j, \mathbf{u}_j = \mathbf{a}_j \right\} \middle| x_0 = x \right). \end{aligned}$$

In words, once players following Algorithm 2 select a policy $\boldsymbol{\pi}$ for the k^{th} exploration phase, then the conditional probabilities of the trajectories restricted to time indices in that exploration phase can be described by $\Pr^{\boldsymbol{\pi}}$, with the indices of the events suitably shifted to start at time 0. This leads to a series of useful lemmas, which we include below for completeness.

In the lemmas below, for notational convenience, we define $J_{\boldsymbol{\pi}}^i : \mathbb{X} \rightarrow \mathbb{R}$ as $J_{\boldsymbol{\pi}}^i(s) = J^i(\boldsymbol{\pi}, s)$ for all $s \in \mathbb{X}$ and any $i \in \mathcal{N}$, $\boldsymbol{\pi} \in \boldsymbol{\Gamma}_S$.

LEMMA A.1. *Let $\boldsymbol{\pi} \in \boldsymbol{\Pi} \subset \boldsymbol{\Gamma}_S$ be some joint policy and let $i \in \mathcal{N}$. For any initial conditions $Q^i \in \mathbb{Q}^i$, $J^i \in \mathbb{J}^i$, and state $x \in \mathbb{X}$, we have the following:*

$$\Pr \left(\widehat{Q}_{t_k+T_k}^i \in \cdot \middle| \widehat{Q}_{t_k}^i = Q^i, \boldsymbol{\pi}_k = \boldsymbol{\pi}, x_{t_k} = x \right) = \Pr^{\boldsymbol{\pi}} \left(\bar{Q}_{T_k}^i \in \cdot \middle| \bar{Q}_0^i = Q^i, x_0 = x \right),$$

and

$$\Pr \left(\widehat{J}_{t_k+T_k}^i \in \cdot \middle| \widehat{J}_{t_k}^i = J^i, \boldsymbol{\pi}_k = \boldsymbol{\pi}, x_{t_k} = x \right) = \Pr^{\boldsymbol{\pi}} \left(\bar{J}_{T_k}^i \in \cdot \middle| \bar{J}_0^i = J^i, x_0 = x \right).$$

LEMMA A.2. *For any joint policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, player $i \in \mathcal{N}$, we have the following:*

- (i) $\Pr^{\boldsymbol{\pi}} \left(\lim_{t \rightarrow \infty} \bar{Q}_t^i = Q_{\boldsymbol{\pi}^{-i}}^{*i} \middle| \bar{Q}_0^i = Q^i, x_0 = x \right) = 1$, for any $x \in \mathbb{X}$, $Q^i \in \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i}$;
- (ii) $\Pr^{\boldsymbol{\pi}} \left(\lim_{t \rightarrow \infty} \bar{J}_t^i = J_{\boldsymbol{\pi}}^i \middle| \bar{J}_0^i = J^i, x_0 = x \right) = 1$, for any $x \in \mathbb{X}$, $J^i \in \mathbb{R}^{\mathbb{X}}$;
- (iii) *If $\mathbb{Q}^i \subset \mathbb{R}^{\mathbb{X} \times \mathbb{U}^i}$ and $\mathbb{J}^i \subset \mathbb{R}^{\mathbb{X}}$ are compact sets, then items (i) and (ii) hold uniformly in their initial conditions. That is, for any $\xi > 0$, there exists $T = T(i, \xi, \boldsymbol{\pi}) \in \mathbb{N}$ such that*

$$\Pr^{\boldsymbol{\pi}} \left(\sup_{t \geq T} \left\| \bar{Q}_t^i - Q_{\boldsymbol{\pi}^{-i}}^{*i} \right\|_{\infty} < \xi \middle| \bar{Q}_0^i = Q^i, x_0 = x \right) \geq 1 - \xi,$$

and

$$\Pr^{\boldsymbol{\pi}} \left(\sup_{t \geq T} \left\| \bar{J}_t^i - J_{\boldsymbol{\pi}}^i \right\|_{\infty} < \xi \middle| \bar{J}_0^i = J^i, x_0 = x \right) \geq 1 - \xi,$$

for any $Q^i \in \mathbb{Q}^i, J^i \in \mathbb{J}^i, x \in \mathbb{X}$.

Proof. Items (i) and (ii) are proved using standard stochastic approximation arguments, e.g. those in [70]. We note that each state-joint action pair is visited infinitely often $\Pr^{\boldsymbol{\pi}}$ -almost surely. This is obtained by Assumptions 2 and 4: the former which guarantees that $\boldsymbol{\pi}^j$ is soft for every player $j \in \mathcal{N}$, and the latter guarantees that all states are mutually accessible.

The uniformity claim, in item (iii), was proven for Q-factors alone in [2, Lemma A1]. The same line of argument can be used for the iterates $\{\widehat{J}_t^i\}_{t \geq 0}$. \square

Combining Lemmas A.1 and A.2, we get the following result on conditional probabilities

LEMMA A.3. *Let $k, \ell \in \mathbb{Z}_{\geq 0}$ such that $\ell \geq k$. Let \mathcal{F}_k denote the σ -algebra generated by the random variables $\boldsymbol{\pi}_k, \{\widehat{Q}_{t_k}^i, \widehat{J}_{t_k}^i\}_{i \in \mathcal{N}}$, and x_{t_k} . For any $\xi > 0$, there exists $T = T(\xi) < \infty$ such that if $T_\ell \geq T$, then Pr-almost surely, we have*

$$\Pr \left(\bigcap_{i \in \mathcal{N}} \left\{ \left\| \widehat{Q}_{t_{\ell+1}}^i - Q_{\boldsymbol{\pi}_\ell}^{*i} \right\|_\infty < \xi \right\} \cap \left\{ \left\| \widehat{J}_{t_{\ell+1}}^i - J_{\boldsymbol{\pi}_\ell}^i \right\|_\infty < \xi \right\} \middle| \mathcal{F}_k \right) \geq 1 - \xi.$$

A.2. Proof Details. Let $\Xi := \frac{1}{2} \min_{j \in \mathcal{N}} \{d^j, \bar{d} - d^j\}$, and for $k \geq 0$, let E_k denote the event that each player i learned its Q- and J-factors to within Ξ of their fixed points during the k^{th} exploration phase. That is, for any $k \geq 0$,

$$E_k := \bigcap_{i \in \mathcal{N}} \left\{ \left\| \widehat{Q}_{t_{k+1}}^i - Q_{\boldsymbol{\pi}_k}^{*i} \right\|_\infty < \Xi \right\} \cap \left\{ \left\| \widehat{J}_{t_{k+1}}^i - J_{\boldsymbol{\pi}_k}^i \right\|_\infty < \Xi \right\}.$$

For any $\ell \geq 0$, let $E_{k:k+\ell} := E_k \cap E_{k+1} \cap \dots \cap E_{k+\ell}$. For any $k \geq 0$, we also define $G_k := \{\boldsymbol{\pi}_k \in \boldsymbol{\Pi} \cap \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}\}$ to be the event that the policy of the k^{th} exploration phase is an ϵ -equilibrium.

By our definition of \bar{d} , assumption that $d^i \in (0, \bar{d})$ for each $i \in \mathcal{N}$ (Assumption 3), and our choice of Ξ , we have that, given E_k , each player $i \in \mathcal{N}$ correctly ascertains whether $\pi_k^i \in \text{BR}_\epsilon^i(\boldsymbol{\pi}_k^{-i})$ by verifying whether $\widehat{J}_{t_{k+1}}^i(x) \leq \min_{a^i} \widehat{Q}_{t_{k+1}}^i + \epsilon + d^i$ for each $x \in \mathbb{X}$. From this, it follows that

$$(A.1) \quad \Pr(G_{k+\ell} | G_k \cap E_{k:k+\ell}) = 1, \forall \ell \geq 0.$$

Recall the quantity $L := \max\{L_{\boldsymbol{\pi}_0} : \boldsymbol{\pi}_0 \in \boldsymbol{\Pi}\}$, where for each $\boldsymbol{\pi}_0 \in \boldsymbol{\Pi}$, $L_{\boldsymbol{\pi}_0}$ is defined as the shortest ϵ -satisficing path within $\boldsymbol{\Pi}$ beginning at $\boldsymbol{\pi}_0$ and ending in $\boldsymbol{\Pi} \cap \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}$. For any $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, there exists an ϵ -satisficing path from $\boldsymbol{\pi}$ into $\boldsymbol{\Pi} \cap \boldsymbol{\Gamma}_S^{\epsilon\text{-eq}}$, and if this path has length less than L , it may be extended to have length L by repeating its final term. Thus, for any $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, we have the following inequality:

$$(A.2) \quad \Pr(G_{k+L} | \{\boldsymbol{\pi}_k = \boldsymbol{\pi}\} \cap E_{k:k+L}) \geq p_{\min} > 0,$$

where $p_{\min} := \prod_{i \in \mathcal{N}} \left(\frac{e^i}{|\boldsymbol{\Pi}^i|} \right)^L$. The quantity p_{\min} is a loose lower bound obtained as follows: starting at $\boldsymbol{\pi}_k = \boldsymbol{\pi}$, at each step, suppose that every unsatisfied player chooses to experiment (with probability e^i at each step) and selects the policy that follows the specified ϵ -satisficing path by (with probability $1/|\boldsymbol{\Pi}^i|$ at each step). There are at most L such steps, therefore the probability of following the specified path by random experimentation alone is no less than p_{\min} .

Fix $u^* \in (0, 1)$ such that $\frac{u^* p_{\min}}{1 - u^* + u^* p_{\min}} > 1 - \psi/2$. As a consequence of Lemma A.3, there exists $\tilde{T} < \infty$ such that if $T_l \geq \tilde{T}$ for every $l \geq 0$, then we have $\Pr(E_{k:k+L} | \boldsymbol{\pi}_k = \boldsymbol{\pi}) \geq u^*$ for all $k \geq 0$ and any $\boldsymbol{\pi} \in \boldsymbol{\Pi}$. Thus, for any $k \geq 0$, we have that

$$\Pr(E_{k:k+L} | G_k) \geq u^* \text{ and } \Pr(E_{k:k+L} | G_k^c) \geq u^*.$$

For any $k \geq 0$, we can lower bound $\Pr(G_{k+L})$ by conditioning on G_k and its complement:

$$\Pr(G_{k+L}) = \Pr(G_{k+L} | G_k) \Pr(G_k) + \Pr(G_{k+L} | G_k^c) (1 - \Pr(G_k)).$$

We lower bound the constituent terms above by conditioning again with $E_{k:k+L}$ and invoking (A.1) and (A.2) to get

$$\Pr(G_{k+L}) \geq 1 \cdot \Pr(E_{k:k+L}|G_k) \cdot \Pr(G_k) + p_{\min} \cdot \Pr(E_{k:k+L}|G_k^c) \cdot (1 - \Pr(G_k)).$$

Assuming $T_l \geq \tilde{T}$ for all $l \geq 0$, this gives

$$\Pr(G_{k+L}) \geq u^* \cdot \Pr(G_k) + p_{\min} \cdot u^* \cdot (1 - \Pr(G_k)), \quad \forall k \geq 0.$$

For each $k \in \{0, 1, \dots, L-1\}$, define $y_0^{(k)} := \Pr(G_k)$, and for $m \geq 0$, define $y_{m+1}^{(k)} = u^* y_m^{(k)} + u^* p_{\min} (1 - y_m^{(k)})$. One can use an inductive argument to show that

$$(A.3) \quad \Pr(G_{k+mL}) \geq y_m^{(k)} \quad \forall m \geq 0.$$

Observe that $y_{m+1}^{(k)}$ can be written as

$$y_{m+1}^{(k)} = (u^* - u^* p_{\min})^{m+1} y_0^{(k)} + u^* p_{\min} \sum_{j=0}^m (u^* - u^* p_{\min})^j.$$

Since $0 < u^* - u^* p_{\min} < 1$, we have that $\lim_{m \rightarrow \infty} y_m^{(k)} = \frac{u^* p_{\min}}{1 - u^* + u^* p_{\min}} > 1 - \frac{\psi}{2}$. Thus, by (A.3), we have that $\Pr(G_{k+mL}) > 1 - \psi/2$ holds for all sufficiently large m , which proves the result.

Appendix B. Proof of Lemmas 2.10–2.13.

LEMMA B.1. Fix $i \in \mathcal{N}$ and $T \in \mathbb{N}$. For any $(s, a^i) \in \mathbb{X} \times \mathbb{U}^i$ and $(\tilde{s}_k, \tilde{\mathbf{a}}_k)_{k=0}^T \in (\mathbb{X} \times \mathbb{U})^{T+1}$, the mapping $F : \Gamma_S \rightarrow \mathbb{R}$ given by

$$F(\boldsymbol{\pi}) := \Pr^{\boldsymbol{\pi}} \left[\bigcap_{k=0}^T \{x_k = \tilde{s}_k, \mathbf{u}_k = \tilde{\mathbf{a}}_k\} \middle| x_0 = s, u_0^i = a^i \right], \quad \forall \boldsymbol{\pi} \in \Gamma_S$$

is continuous.

Proof. We write the specified probability as

$$\begin{aligned} & \Pr^{\boldsymbol{\pi}} \left[\bigcap_{j=0}^T \{x_j = \tilde{s}_j, \mathbf{u}_j = \tilde{\mathbf{a}}_j\} \middle| x_0 = s, u_0^i = a^i \right] \\ &= \mathbf{1}\{\tilde{s}_0 = s, \tilde{a}_0^i = a^i\} \cdot \prod_{j \neq i} \pi^j(\tilde{a}_0^j | \tilde{s}_0) \times \prod_{k=0}^{T-1} P(\tilde{s}_{k+1} | \tilde{s}_k, \tilde{\mathbf{a}}_k) \cdot \prod_{k=1}^T \boldsymbol{\pi}(\tilde{\mathbf{a}}_k | \tilde{s}_k), \end{aligned}$$

where $\boldsymbol{\pi}(\tilde{\mathbf{a}}_k | \tilde{s}_k) = \prod_{j \in \mathcal{N}} \pi^j(\tilde{a}_k^j | \tilde{s}_k)$. Since the conditional probability under $\Pr^{\boldsymbol{\pi}}$ of a given history of finite length is a finite product involving the components of $\boldsymbol{\pi}$, continuity follows. \square

Proof of Lemma 2.10. Fix player $i \in \mathcal{N}$, $s \in \mathbb{X}$, $\epsilon > 0$, and choose $T \in \mathbb{N}$ large enough that

$$\frac{(\beta^i)^T}{1 - \beta^i} \cdot \max \{ |c^i(\tilde{s}, \mathbf{a})| : (\tilde{s}, \mathbf{a}) \in \mathbb{X} \times \mathbb{U} \} < \frac{\epsilon}{4}.$$

Then, since

$$J^i(\boldsymbol{\pi}, s) = E^\pi \left[\sum_{t=0}^T (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right] + E^\pi \left[\sum_{t>T} (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right]$$

for any $\boldsymbol{\pi} \in \boldsymbol{\Gamma}_S$, we have that, for any $\boldsymbol{\pi}, \tilde{\boldsymbol{\pi}} \in \boldsymbol{\Gamma}_S$,

$$\begin{aligned} & |J^i(\boldsymbol{\pi}, s) - J^i(\tilde{\boldsymbol{\pi}}, s)| \leq \\ & \left| E^\pi \left[\sum_{t=0}^T (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right] - E^{\tilde{\boldsymbol{\pi}}} \left[\sum_{t=0}^T (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right] \right| + \frac{\epsilon}{2}. \end{aligned}$$

Define a function $C^i : (\mathbb{X} \times \mathbf{U})^{T+1} \rightarrow \mathbb{R}$ by

$$C^i(\tilde{s}_0, \tilde{\mathbf{a}}_0, \dots, \tilde{s}_T, \tilde{\mathbf{a}}_T) := \sum_{t=0}^T (\beta^i)^t c^i(\tilde{s}_t, \tilde{\mathbf{a}}_t), \quad \forall (\tilde{s}_k, \tilde{\mathbf{a}}_k)_{k=0}^T \in (\mathbb{X} \times \mathbf{U})^{T+1}.$$

We thus write

$$E^\pi \left[\sum_{t=0}^T (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right] = \sum_{\omega \in (\mathbb{X} \times \mathbf{U})^{T+1}} C^i(\omega) \Pr^\pi \left((x_k, \mathbf{u}_k)_{k=0}^T = \omega \middle| x_0 = s \right),$$

for any $\boldsymbol{\pi} \in \boldsymbol{\Gamma}_S$. From the latter expression and Lemma B.1, we see that the mapping $\boldsymbol{\pi} \mapsto E^\pi \left(\sum_{t=0}^T (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right)$ is continuous on $\boldsymbol{\Gamma}_S$.

Thus, taking $\boldsymbol{\pi}, \tilde{\boldsymbol{\pi}} \in \boldsymbol{\Gamma}_S$ sufficiently close, under the metric \mathbf{d} , we have that

$$\left| E^\pi \left[\sum_{t=0}^T (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right] - E^{\tilde{\boldsymbol{\pi}}} \left[\sum_{t=0}^T (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s \right] \right| < \frac{\epsilon}{2}.$$

From our choice of T , we then also get that $|J^i(\boldsymbol{\pi}, s) - J^i(\tilde{\boldsymbol{\pi}}, s)| < \epsilon$, proving the lemma.

Proof of Lemma 2.11. Fix player $i \in \mathcal{N}$ and $(s, a^i) \in \mathbb{X} \times \mathbb{U}^i$. We will show that the mapping $\boldsymbol{\pi}^{-i} \mapsto Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a^i)$ is continuous on $\boldsymbol{\Gamma}_S^{-i}$. Recall that for any $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$, the Q-factors are given by

$$Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a^i) := E^{(\boldsymbol{\pi}^{*i}, \boldsymbol{\pi}^{-i})} \left[\sum_{t=0}^{\infty} (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s, u_0^i = a^i \right],$$

where $\boldsymbol{\pi}^{*i} \in \text{BR}_0^i(\boldsymbol{\pi}^{-i})$ is any best-response to $\boldsymbol{\pi}^{-i}$. Since $\boldsymbol{\pi}^{-i} \in \boldsymbol{\Gamma}_S^{-i}$ is stationary, player i faces an MDP with controlled state process $\{x_t\}_{t \geq 0}$. Therefore, some stationary and deterministic best-response policy exists. We denote the set of stationary and deterministic policies for player i by $\text{DS}^i \subset \boldsymbol{\Gamma}_S^i$:

$$\text{DS}^i := \{ \boldsymbol{\pi}^i \in \boldsymbol{\Gamma}_S^i \mid \forall \tilde{s} \in \mathbb{X}, \exists \tilde{a}^i \in \mathbb{U}^i : \boldsymbol{\pi}^i(\tilde{a}^i \mid \tilde{s}) = 1 \}.$$

Note that the set DS^i is finite and can be identified with the finite set $\{f^i : \mathbb{X} \rightarrow \mathbb{U}^i\}$. Let $\boldsymbol{\pi}^{*i} \in \text{BR}_0^i(\boldsymbol{\pi}^{-i}) \cap \text{DS}^i$. We then have that, for any $\tilde{s} \in \mathbb{X}$,

$$(B.1) \quad J^i(\boldsymbol{\pi}^{*i}, \boldsymbol{\pi}^{-i}, \tilde{s}) = \inf_{\tilde{\boldsymbol{\pi}}^i \in \boldsymbol{\Gamma}^i} J^i(\tilde{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}, \tilde{s}) = \min_{\tilde{\boldsymbol{\pi}}^i \in \text{DS}^i} J^i(\tilde{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}, \tilde{s}).$$

For each $\tilde{s} \in \mathbb{X}$, the mapping $\boldsymbol{\pi}^{-i} \mapsto \min_{\tilde{\boldsymbol{\pi}}^i \in \text{DS}^i} J^i(\tilde{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}, \tilde{s})$ is continuous on $\boldsymbol{\Gamma}_S^{-i}$ by Lemma 2.10, as it is the pointwise minimum of finitely many continuous functions.

Letting $\boldsymbol{\pi}^* = (\boldsymbol{\pi}^{*i}, \boldsymbol{\pi}^{-i})$, we write $Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a^i)$ as

$$\begin{aligned} Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a^i) &= E^{\boldsymbol{\pi}^*} \left[\sum_{t=0}^{\infty} (\beta^i)^t c^i(x_t, \mathbf{u}_t) \middle| x_0 = s, u_0^i = a^i \right] \\ &= E^{\boldsymbol{\pi}^*} [c^i(x_0, \mathbf{u}_0) | x_0 = s, u_0^i = a^i] + \beta^i \cdot E^{\boldsymbol{\pi}^*} [J^i(\boldsymbol{\pi}^*, x_1) | x_0 = s, u_0^i = a^i], \end{aligned}$$

where the latter term is obtained by the tower property (conditioning on x_0, u_0^i , and x_1) and using the fact that $\boldsymbol{\pi}^*$ is stationary to simplify the resulting conditional expectation. Using (B.1), we then have

$$\begin{aligned} &Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a^i) \\ &= E^{\boldsymbol{\pi}^*} [c^i(x_0, \mathbf{u}_0) | x_0 = s, u_0^i = a^i] + \beta^i \cdot E^{\boldsymbol{\pi}^*} \left[\min_{\tilde{\boldsymbol{\pi}}^i \in \text{DS}^i} J^i(\tilde{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}, x_1) \middle| x_0 = s, u_0^i = a^i \right]. \end{aligned}$$

We observe that the first term does not depend on $\boldsymbol{\pi}^{*i}$ and can be re-written as

$$E^{\boldsymbol{\pi}^*} [c^i(x_0, \mathbf{u}_0) | x_0 = s, u_0^i = a^i] = \sum_{\mathbf{a}^{-i} \in \mathbf{U}^{-i}} \boldsymbol{\pi}^{-i}(\mathbf{a}^{-i} | s) c^i(s, a^i, \mathbf{a}^{-i}),$$

where $\boldsymbol{\pi}^{-i}(\mathbf{a}^{-i} | s) = \prod_{j \neq i} \boldsymbol{\pi}^j(a^j | s)$. Note, further, that the mapping

$$\boldsymbol{\pi}^{-i} \mapsto \sum_{\mathbf{a}^{-i} \in \mathbf{U}^{-i}} \boldsymbol{\pi}^{-i}(\mathbf{a}^{-i} | s) c^i(s, a^i, \mathbf{a}^{-i})$$

is continuous on $\boldsymbol{\Gamma}_S^{-i}$. Indeed, the second term does not depend on $\boldsymbol{\pi}^{*i}$ either: using iterated expectations and conditioning additionally on \mathbf{u}_0^{-i} , we write

$$\begin{aligned} &E^{\boldsymbol{\pi}^*} \left[\min_{\tilde{\boldsymbol{\pi}}^i \in \text{DS}^i} J^i(\tilde{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}, x_1) \middle| x_0 = s, u_0^i = a^i \right] \\ &= \sum_{\mathbf{a}^{-i} \in \mathbf{U}^{-i}} \boldsymbol{\pi}^{-i}(\mathbf{a}^{-i} | s) \left(\sum_{s' \in \mathbb{X}} \min_{\tilde{\boldsymbol{\pi}}^i \in \text{DS}^i} J^i(\tilde{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}, s') \cdot P(s' | s, a^i, \mathbf{a}^{-i}) \right). \end{aligned}$$

From this, one sees that the mapping

$$\boldsymbol{\pi}^{-i} \mapsto \beta^i E^{\boldsymbol{\pi}^*} \left[\min_{\tilde{\boldsymbol{\pi}}^i \in \text{DS}^i} J^i(\tilde{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}, x_1) \middle| x_0 = s, u_0^i = a^i \right]$$

is also continuous on $\boldsymbol{\Gamma}_S^{-i}$. Thus, $Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a^i)$ is the sum of two functions that are continuous on $\boldsymbol{\Gamma}_S^{-i}$ and so is itself continuous on $\boldsymbol{\Gamma}_S^{-i}$, completing the proof.

Proofs of Lemmas 2.12 and 2.13. Lemma 2.12 follows from Lemma 2.11, as the function under consideration is the pointwise minimum of finitely many continuous functions.

Lemma 2.13 follows from Lemmas 2.10 and 2.12, as the function under consideration is the pointwise maximum of finitely many continuous functions.

REFERENCES

- [1] A. AGARWAL, S. M. KAKADE, J. D. LEE, AND G. MAHAJAN, *On the theory of policy gradient methods: Optimality, approximation, and distribution shift*, Journal of Machine Learning Research, 22 (2021), pp. 1–76.
- [2] G. ARSLAN AND S. YÜKSEL, *Decentralized Q-learning for stochastic teams and games*, IEEE Transactions on Automatic Control, 62 (2017), pp. 1545–1558.
- [3] M. G. AZAR, R. MUNOS, M. GHAVAMZADEH, AND H. KAPPEN, *Speedy Q-learning*, in Advances in Neural Information Processing Systems, 2011.
- [4] Y. BAI, C. JIN, AND T. YU, *Near-optimal reinforcement learning with self-play*, Advances in Neural Information Processing Systems, 33 (2020), pp. 2159–2170.
- [5] M. BOWLING AND M. VELOSO, *Multiagent learning using a variable learning rate*, Artificial Intelligence, 136 (2002), pp. 215–250.
- [6] G. W. BROWN, *Iterative solution of games by fictitious play*, Activity Analysis of Production and Allocation, 13 (1951), pp. 374–376.
- [7] N. BROWN AND T. SANDHOLM, *Superhuman AI for heads-up no-limit poker: Libratus beats top professionals*, Science, 359 (2018), pp. 418–424.
- [8] O. CANDOGAN, A. ÖZDAGLAR, AND P. A. PARRILO, *Near-potential games: Geometry and dynamics*, ACM Transactions on Economics and Computation (TEAC), 1 (2013), pp. 1–32.
- [9] R. CARMONA, F. DELARUE, AND D. LACKER, *Mean field games with common noise*, The Annals of Probability, 44 (2016), pp. 3740–3803.
- [10] R. CASSIDY, C. FIELD, AND M. KIRBY, *Solution of a satisficing model for random payoff games*, Management Science, 19 (1972), pp. 266–271.
- [11] A. CHARNES AND W. W. COOPER, *Deterministic equivalents for optimizing and satisficing under chance constraints*, Operations Research, 11 (1963), pp. 18–39.
- [12] G. C. CHASPARIS, A. ARAPOSTATHIS, AND J. S. SHAMMA, *Aspiration learning in coordination games*, SIAM Journal on Control and Optimization, 51 (2013), pp. 465–490.
- [13] G. CHEVALIER, J. LE NY, AND R. MALHAMÉ, *A micro-macro traffic model based on mean-field games*, in 2015 American Control Conference (ACC), IEEE, 2015, pp. 1983–1988.
- [14] S. CHIEN AND A. SINCLAIR, *Convergence to approximate Nash equilibria in congestion games*, Games and Economic Behavior, 71 (2011), pp. 315–327.
- [15] 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, 1998, pp. 746–752.
- [16] C. DASKALAKIS, D. J. FOSTER, AND N. GOLOWICH, *Independent policy gradient methods for competitive reinforcement learning*, Advances in Neural Information Processing Systems, 33 (2020), pp. 5527–5540.
- [17] C. DASKALAKIS, D. J. FOSTER, AND N. GOLOWICH, *Independent policy gradient methods for competitive reinforcement learning*, arXiv preprint arXiv:2101.04233, (2021).
- [18] C. DASKALAKIS, N. GOLOWICH, AND K. ZHANG, *The complexity of Markov equilibrium in stochastic games*, arXiv preprint arXiv:2204.03991, (2022).
- [19] A. M. DEVRAJ AND S. MEYN, *Zap Q-learning*, Advances in Neural Information Processing Systems, 30 (2017).
- [20] C. EKSIN AND A. RIBEIRO, *Distributed fictitious play for multiagent systems in uncertain environments*, IEEE Transactions on Automatic Control, 63 (2017), pp. 1177–1184.
- [21] E. EVEN-DAR AND Y. MANSOUR, *Learning rates for Q-learning*, Journal of Machine Learning Research, 5 (2003), pp. 1–25.
- [22] A. M. FINK, *Equilibrium in a stochastic n-person game*, Journal of Science of the Hiroshima University, Series AI (Mathematics), 28 (1964), pp. 89–93.
- [23] M. FISCHER, *On the connection between symmetric n-player games and mean field games*, The Annals of Applied Probability, 27 (2017), pp. 757–810.
- [24] D. FOSTER AND H. P. YOUNG, *Regret testing: Learning to play Nash equilibrium without knowing you have an opponent*, Theoretical Economics, 1 (2006), pp. 341–367.
- [25] A. GAUNERSDORFER AND J. HOFBAUER, *Fictitious play, Shapley polygons, and the replicator equation*, Games and Economic Behavior, 11 (1995), pp. 279–303.
- [26] F. GERMANO AND G. LUGOSI, *Global Nash convergence of Foster and Young’s regret testing*, Games and Economic Behavior, 60 (2007), pp. 135–154.
- [27] S. HART AND A. MAS-COLELL, *Uncoupled dynamics do not lead to Nash equilibrium*, American Economic Review, 93 (2003), pp. 1830–1836.
- [28] S. HART AND A. MAS-COLELL, *Stochastic uncoupled dynamics and Nash equilibrium*, Games and Economic Behavior, 57 (2006), pp. 286–303.
- [29] 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).
- [30] J. HOFBAUER AND K. SIGMUND, *Evolutionary game dynamics*, Bulletin of the American Mathematical Society, 40 (2003), pp. 479–519.
- [31] J. HU AND M. P. WELLMAN, *Nash Q-learning for general-sum stochastic games*, Journal of Machine Learning Research, 4 (2003), pp. 1039–1069.
- [32] M. HUANG, P. E. CAINES, AND R. P. MALHAMÉ, *Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ -Nash equilibria*, IEEE Transactions on Automatic Control, 52 (2007), pp. 1560–1571.
- [33] M. HUANG, R. P. MALHAMÉ, AND P. E. CAINES, *Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle*, Communications in Information & Systems, 6 (2006), pp. 221–252.
- [34] J.-M. LASRY AND P.-L. LIONS, *Mean field games*, Japanese Journal of Mathematics, 2 (2007), pp. 229–260.
- [35] 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.
- [36] D. LESLIE AND E. COLLINS, *Individual Q-learning in normal form games*, SIAM Journal on Control and Optimization, 44 (2005), pp. 495–514.
- [37] D. S. LESLIE, S. PERKINS, AND Z. XU, *Best-response dynamics in zero-sum stochastic games*, Journal of Economic Theory, 189 (2020), p. 105095.
- [38] Y. LEVY, *Discounted stochastic games with no stationary Nash equilibrium: two examples*, Econometrica, 81 (2013), pp. 1973–2007.
- [39] M. L. LITTMAN, *Markov games as a framework for multi-agent reinforcement learning*, in Machine Learning Proceedings 1994, Elsevier, 1994, pp. 157–163.
- [40] M. L. LITTMAN, *Friend-or-foe Q-learning in general-sum games*, in ICML, vol. 1, 2001, pp. 322–328.
- [41] Q. LIU, T. YU, Y. BAI, AND C. JIN, *A sharp analysis of model-based reinforcement learning with self-play*, in International Conference on Machine Learning, PMLR, 2021, pp. 7001–7010.
- [42] J. R. MARDEN AND J. S. SHAMMA, *Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation*, Games and Economic Behavior, 75 (2012), pp. 788–808.
- [43] 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 (2009), pp. 373–396.
- [44] J. R. MARDEN, H. P. YOUNG, AND L. Y. PAO, *Achieving Pareto optimality through distributed learning*, SIAM Journal on Control and Optimization, 52 (2014), pp. 2753–2770.
- [45] 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, IEEE, 2007, pp. 64–69.
- [46] L. MATIGNON, G. J. LAURENT, AND N. LE FORT-PIAT, *Coordination of independent learners in cooperative Markov games.*, HAL preprint hal-00370889, (2009).
- [47] L. MATIGNON, G. J. LAURENT, AND N. LE FORT-PIAT, *Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems.*, Knowledge Engineering Review, 27 (2012), pp. 1–31.
- [48] A. MATSUI, *Best response dynamics and socially stable strategies*, Journal of Economic Theory, 57 (1992), pp. 343–362.
- [49] V. MNIH, K. KAVUKCUOGLU, D. SILVER, A. A. RUSU, J. VENESS, M. G. BELLEMARE, A. GRAVES, M. RIEDMILLER, A. K. FIDJELAND, G. OSTROVSKI, ET AL., *Human-level control through deep reinforcement learning*, Nature, 518 (2015), pp. 529–533.
- [50] M. ORNIK AND U. TOPCU, *Deception in optimal control*, in 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2018, pp. 821–828.
- [51] A. OZDAGLAR, M. O. SAYIN, AND K. ZHANG, *Independent learning in stochastic games*, arXiv preprint arXiv:2111.11743, (2021).
- [52] R. RADNER, *Satisficing*, in Optimization Techniques IFIP Technical Conference, Springer, 1975, pp. 252–263.
- [53] J. ROBINSON, *An iterative method of solving a game*, Annals of Mathematics, 54 (1951), pp. 296–301.
- [54] T. ROUGHGARDEN, *Algorithmic game theory*, Communications of the ACM, 53 (2010), pp. 78–86.
- [55] S. SANJARI, N. SALDI, AND S. YÜKSEL, *Optimality of independently randomized symmetric policies for exchangeable stochastic teams with infinitely many decision makers*, arXiv

- preprint arXiv:2008.11570, (2020).
- [56] S. SANJARI AND S. YÜKSEL, *Optimal solutions to infinite-player stochastic teams and mean-field teams*, IEEE Transactions on Automatic Control, 66 (2020), pp. 1071–1086.
 - [57] S. SANJARI AND S. YÜKSEL, *Optimal policies for convex symmetric stochastic dynamic teams and their mean-field limit*, SIAM Journal on Control and Optimization, 59 (2021), pp. 777–804.
 - [58] 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 (2021), pp. 18320–18334.
 - [59] M. O. SAYIN, F. PARISE, AND A. OZDAGLAR, *Fictitious play in zero-sum stochastic games*, SIAM Journal on Control and Optimization, 60 (2022), pp. 2095–2114.
 - [60] J. SCHULMAN, S. LEVINE, P. ABBEEL, M. JORDAN, AND P. MORITZ, *Trust region policy optimization*, in International Conference on Machine Learning, PMLR, 2015, pp. 1889–1897.
 - [61] J. SCHULMAN, F. WOLSKI, P. DHARIWAL, A. RADFORD, AND O. KLIMOV, *Proximal policy optimization algorithms*, arXiv preprint arXiv:1707.06347, (2017).
 - [62] S. SEN, M. SEKARAN, AND J. HALE, *Learning to coordinate without sharing information*, in Proceedings of the 12th National Conference on Artificial Intelligence, 1994, pp. 426–431.
 - [63] D. SILVER, A. HUANG, C. J. MADDISON, A. GUEZ, L. SIFRE, G. VAN DEN DRIESSCHE, J. SCHRITTWIESER, I. ANTONOGLU, V. PANNEERSHELVAM, M. LANCTOT, ET AL., *Mastering the game of Go with deep neural networks and tree search*, Nature, 529 (2016), pp. 484–489.
 - [64] D. SILVER, J. SCHRITTWIESER, K. SIMONYAN, I. ANTONOGLU, A. HUANG, A. GUEZ, T. HUBERT, L. BAKER, M. LAI, A. BOLTON, ET AL., *Mastering the game of Go without human knowledge*, Nature, 550 (2017), pp. 354–359.
 - [65] H. A. SIMON, *Rational choice and the structure of the environment.*, Psychological review, 63 (1956), p. 129.
 - [66] R. S. SUTTON AND A. G. BARTO, *Reinforcement learning: An introduction*, MIT press, 2018.
 - [67] B. SWENSON, C. EKSIN, S. KAR, AND A. RIBEIRO, *Distributed inertial best-response dynamics*, IEEE Transactions on Automatic Control, 63 (2018), pp. 4294–4300.
 - [68] C. SZEPESVÁRI, *The asymptotic convergence-rate of Q-learning*, Advances in neural information processing systems, 10 (1997).
 - [69] M. TAN, *Multi-agent reinforcement learning: Independent vs. cooperative agents*, in Proceedings of the Tenth International Conference on Machine Learning, 1993, pp. 330–337.
 - [70] J. N. TSITSIKLIS, *Asynchronous stochastic approximation and Q-learning*, Machine Learning, 16 (1994), pp. 185–202.
 - [71] C. WATKINS, *Learning from Delayed Rewards*, PhD thesis, Cambridge University, 1989.
 - [72] C. WATKINS AND P. DAYAN, *Q-Learning*, Machine Learning, 8 (1992), pp. 279–292.
 - [73] C.-Y. WEI, Y.-T. HONG, AND C.-J. LU, *Online reinforcement learning in stochastic games*, Advances in Neural Information Processing Systems, 30 (2017).
 - [74] E. WEI AND S. LUKE, *Lenient learning in independent-learner stochastic cooperative games*, The Journal of Machine Learning Research, 17 (2016), pp. 2914–2955.
 - [75] 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 (2021), pp. 5230–5245.
 - [76] B. YONGACOGLU, G. ARSLAN, AND S. YÜKSEL, *Independent learning and subjectivity in mean-field games*, in 2022 IEEE 61st Conference on Decision and Control (CDC), IEEE, 2022, pp. 2845–2850.
 - [77] K. ZHANG, Z. YANG, AND T. BAŞAR, *Multi-agent reinforcement learning: A selective overview of theories and algorithms*, Handbook of Reinforcement Learning and Control, (2021), pp. 321–384.