# Asynchronous Decentralized Q-Learning in Stochastic Games

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel

***Abstract*— Non-stationarity is a fundamental challenge in multi-agent reinforcement learning (MARL), where agents update their behaviour as they learn and individual agents may have an incomplete view of the actions of others. Many theoretical advances in MARL avoid the challenge of non-stationarity by coordinating the policy updates of agents in various ways, including synchronizing times at which agents are allowed to revise their policies. In this paper, we study MARL in stochastic games, and show that a recent decentralized Q-learning algorithm can be modified to accommodate asynchronous policy updates while continuing to give high probability guarantees of equilibrium. In this generalization, players need not agree on the schedule of policy update times, and may change their policies at their own, separately selected times. This allows for greater decentralization and tames non-stationarity without imposing the coordination assumptions of prior work.**

## I. Introduction

Multi-agent systems are characterized by the coexistence of several autonomous agents acting in a shared environment. In multi-agent reinforcement learning (MARL), agents in the system change their behaviour in response to the feedback information received in previous interactions. When agents do not directly observe the actions of their counterparts, the system is non-stationary from any one agent's perspective, and agents attempt to optimize their performance against a moving target [1]. The non-stationarity of MARL environments has been identified as one of the fundamental problems in MARL [2]. In contrast to the rich literature on single-agent learning theory, the theory of MARL is relatively underdeveloped, due in large part to its inherent challenges of non-stationarity and decentralized information.

This paper studies learning algorithms for stochastic games, a common framework for MARL in which the cost-relevant history of the system is summarized by a state variable. In this paper, we focus on stochastic games in which each agent fully observes the system's state variable but does not observe the actions of other agents, exacerbating the challenge of non-stationarity.

Early rigorous work in MARL in stochastic games avoided the problem of non-stationarity by studying applications in which joint actions were observed by all agents [3], [4], [5]. More recently, there has been interest in the *individual action learner* setting, where actions are not shared between agents. In the individual action learner setting, several rigorous contributions have been made recently, including [6], [7], [8], [9], [10], to be discussed shortly.

Of the recent advances in MARL in the individual action learner setting, many of the algorithms with strong guarantees have circumvented the challenge of non-stationarity by relying, implicitly or explicitly, on coordination between the agents. In particular, several algorithms rely on some form of synchrony, whereby agents agree on the times at which they may revise their behaviour and are constrained to fix their policies at other times. While this is justifiable in some settings, it can be restrictive in others, including applications where parameters are selected independently. As such, it would be desirable to provide MARL algorithms that do not require synchrony but still come with rigorous performance guarantees in the individual action learner setting.

**Contributions:** In this paper, we study a modification of the *decentralized Q-learning* algorithm of [6], a recent algorithm proposed for weakly acyclic $N$-player stochastic games. By employing a constant learning rate in the Q-learning algorithm, we show that inertial best-response dynamics provide a mechanism for taming non-stationarity without coordinating players' parameter choices ahead of play. Under appropriate parameter selection, we show that this algorithm drives policies to equilibrium with arbitrarily high probability.

**Notation:** For a standard Borel space $A$, we let $\mathcal{P}(A)$ denote the set of probability measures on $A$ with its Borel $\sigma$-algebra. For standard Borel spaces $A$ and $B$, we let $\mathcal{P}(A|B)$ denote the set of transition kernels on $A$ given $B$.

### A. Related Work

This paper studies stochastic games in which each agent fully observes the system state but does not observe the actions of other players.[1] As such, we are interested in MARL algorithms that make use only of one's history of state observations, *individual* actions, and cost feedback. This information paradigm is common in the literature in MARL, with examples such as [7], [8], [9], [10], [6], [11], [12], and is sometimes called the independent learning paradigm. This terminology is, however, not uniform, as independent learning has also recently been used to refer to learners that update their policies in a (perhaps myopic) self-interested manner, such as via best-response updates or gradient-like updates [13].[2]

G. Arslan is with the Department of Electrical Engineering, University of Hawaii at Manoa, email: gurdal@hawaii.edu. B. Yongacoglu and S. Yüksel are with the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada, email: {1bmy,yuksel@queensu.ca}.

[1] We use the terms players and agents interchangeably.

[2] In response to the changing conventions, we recommend that learners who condition only on their *individual* action history should be called *individual action learners*, in analogy to their counterpart of *joint action learners*.

At least two challenges emerge under imperfect monitoring, when the joint action is not observed. First, agents cannot form estimates about the policies of other agents. In extreme examples, players may be unaware of the very existence of their counterparts. Second, players cannot form estimates about joint action values, yielding several promising joint action learning algorithms unusable.

Rather than estimating a global joint action Q-function, several works have studied the prospect of greedily changing one's policy, either in a best-response sense using a local action Q-function, or in a gradient sense. To handle the challenge of non-stationarity, some authors, e.g. [7], have proposed the use of a *multi-timescale approach*, whereby some agents change their policies faster than others, possibly in an alternating manner. In these works, agreement on a particular schedule for policy updating may be interpreted as an implicit form of parameter coordination.

An alternative approach involves responding to one's environment without accounting for the existence of other players *at all*. Works in this tradition follow the regret testing paradigm of Foster and Young [14], which presented an algorithm for stateless games. This approach was later studied by [15] and [16] among others in the context of stateless repeated games, where impressive convergence guarantees can be made due to the absence of state dynamics that complicate value estimation.

In [6], the regret testing paradigm of Foster and Young was modified for multi-state stochastic games, where one must account for both the immediate cost of an action and also the cost-to-go, which depends on the evolution of the state variable. The *decentralized Q-learning* algorithm of [6] instructs agents to agree on an increasing sequence of policy update times, $(t_k)_{k\geq 0}$, and to fix one's policy within the intervals $[t_k, t_{k+1})$, called the $k^{th}$ *exploration phase*. In so doing, the joint policy process is fixed over each exploration phase, and within each exploration phase, each agent faces an MDP. Provided the exploration phase lengths $T_k := t_{k+1} - t_k$ are sufficiently long, this allows for analysis of learning iterates using single-agent learning theory.

In effect, the exploration phase technique of [6] decouples learning from adaptation, and allows for separate analysis of learning iterates and policy dynamics. This allows for approximation arguments to be used, whereby the dynamics of the policy process resemble those of an idealized process in which players obtain noise-free learning iterates for use in their policy updates. This has lead to a series of theoretical contributions in MARL that all make use of the exploration phase technique, including [9] and [10].

One natural criticism of the exploration phase technique described above is the synchronization of policy updates. In the description above, agents agree on the policy update times $\{t_k\}_{k\geq 0}$ *exactly*, and no agent ever updates its policy in the interval $(t_k, t_{k+1} - 1]$. This can be justified in some settings, but is demanding in decentralized settings where parameters are selected independently across players. Indeed, the assumption of synchrony is made in various works in the regret testing tradition, including [14], [15] and [16].

Intuitively, asynchrony may be problematic for regret testers because the action-value estimates of players depend on historical data from each player's most recent exploration phase. As such, if other players change their policies during an individual's exploration phase, the individual receives feedback from different sources, and its learning iterates may not approximate any quantity relevant to the prevailing environment at the time of the agent's policy update. These changes of policy during an exploration phase constitute disruptions of a player's learning, and analysis of the overall joint policy process is difficult when players do not reliably learn accurate action-values.

In [16], a heuristic argument suggested that the use of inertia in policy updating may allow one to relax the assumption of perfect synchrony in regret testing algorithms for stateless repeated games. The premise of this argument is such: if players occasionally abstain from changing their policies due to random inertia, then they will abstain from disrupting the learning of other agents. If the exploration phase of a given individual is allowed to proceed for a sufficiently long time without disruptions, then any errors may be corrected, and the learned estimates should be approximately correct. In this way, it is argued that inertia acts as a random coordination mechanism, and that perfect synchrony may not be necessary. In this paper, we formalize this heuristic argument and show that it is essentially correct, with one caveat: our analysis reveals that the value estimation protocol must be modified to account for the non-stationarity in the environment. In particular, the algorithm of this paper uses a constant learning rate to ensure that learning iterates rapidly overcome outdated feedback data.

## II. Model

### A. Stochastic Games

We study stochastic games with finitely many players. Formally, a stochastic game $\mathcal{G}$ is described by a list:

$$\mathcal{G} = \left(\mathcal{N}, \mathbb{X}, \{\mathbb{U}^i, c^i, \beta^i : i \in \mathcal{N}\}, P, \nu_0\right). \tag{1}$$

The components of $\mathcal{G}$ are as follows: $\mathcal{N} = \{1, 2, \ldots, N\}$ is a finite set of $N$ agents. The set $\mathbb{X}$ is a finite set of system states. For each player $i \in \mathcal{N}$, $\mathbb{U}^i$ is $i$'s finite set of actions. We write $\mathbf{U} = \times_{i\in\mathcal{N}}\mathbb{U}^i$, and refer to elements of $\mathbf{U}$ as *joint actions*. For each player $i$, a function $c^i : \mathbb{X} \times \mathbf{U} \to \mathbb{R}$ determines player $i$'s stage costs, which are aggregated using a discount factor $\beta^i \in [0, 1)$. The initial system state has distribution $\nu_0 \in \mathcal{P}(\mathbb{X})$, and state transitions are governed by a 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 $x_t$, and each player $i$ selects an action $u_t^i \in \mathbb{U}^i$ according to its policy, to be described shortly. The joint action at time $t$ is denoted $\mathbf{u}_t$. Each player $i$ then incurs a cost $c_t^i := c^i(x_t, \mathbf{u}_t)$, and state variable evolves according to $x_{t+1} \sim P(\cdot|x_t, \mathbf{u}_t)$. This process is then repeated at time $t + 1$, and so on.

A policy is a rule for selecting actions according to the observed history of the system. Here, we assume that at time $t \geq 0$, player $i$ observes the following information:

$$I_t^i = (x_0, u_0^i, c_0^i, x_1, \ldots, c_{t-1}^i, x_t).$$

Player $i$ fully observes the system state, its own actions, and its own cost *realizations,* but does not observe the actions of other players directly. We do not assume that player $i$ knows the function $c^i$.

In general, action selection can incorporate randomness, and players may use arbitrarily complicated, history-dependent policies. However, our analysis will focus on stationary (Markov) policies, a subset of policies that randomly select actions in a time invariant manner that conditions only on the currently observed state variable. The set of stationary policies for player $i \in \mathcal{N}$ is denoted $\Gamma_S^i$ and we identify $\Gamma_S^i$ with $\mathcal{P}(\mathbb{U}^i|\mathbb{X})$, the set of transition kernels on $\mathbb{U}^i$ given $\mathbb{X}$. Henceforth, unqualified reference to a policy shall be understood to mean a stationary policy.

*Definition 1:* For $i \in \mathcal{N}$, $\xi > 0$, a policy $\pi^i \in \Gamma_S^i$ is called *$\xi$-soft* if $\pi^i(a^i|x) \geq \xi$ for all $(x, a^i) \in \mathbb{X} \times \mathbb{U}^i$. A policy $\pi^i \in \Gamma_S^i$ is called *soft* if it is $\xi$-soft for some $\xi > 0$.

*Definition 2:* A policy $\pi^i \in \Gamma_S^i$ is called *deterministic* if for each $x \in \mathbb{X}$, there exists $a^i \in \mathbb{U}^i$ such that $\pi^i(a^i|x) = 1$.

The set of deterministic stationary policies for player $i$ is denoted by $\Gamma_{SD}^i$ and is identified with the set of functions from $\mathbb{X}$ to $\mathbb{U}^i$.

**Notation:** We let $\mathbf{\Gamma}_S := \times_{i\in\mathcal{N}}\Gamma_S^i$ denote the set of *joint policies*. To isolate player $i$'s component in a particular joint policy $\boldsymbol{\pi} \in \mathbf{\Gamma}_S$, we write $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i})$, where $-i$ is used in the agent index to represent all agents other than $i$. Similarly, we write the joint policy set as $\mathbf{\Gamma}_S = \Gamma_S^i \times \mathbf{\Gamma}_S^{-i}$, and so on.

For any joint policy $\boldsymbol{\pi}$ and initial distribution $\nu \in \mathcal{P}(\mathbf{X})$, there is a unique probability measure on the set $(\mathbf{X} \times \mathbf{U})^\infty$. We denote this measure by $\Pr_\nu^{\boldsymbol{\pi}}$, and let $E_\nu^{\boldsymbol{\pi}}$ denote its expectation. We use this to define player $i$'s value function:

$$J^i(\boldsymbol{\pi}, \nu) := E_\nu^{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\beta^t c_t^i\right] = E_\nu^{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\beta^t c^i(x_t, \mathbf{u}_t)\right].$$

When $\nu = \delta_s$ places full probability on some state $s \in \mathbb{X}$, we write $J^i(\boldsymbol{\pi}, s)$ instead of $J^i(\boldsymbol{\pi}, \delta_s)$. For $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i})$, we will also write $J^i(\pi^i, \boldsymbol{\pi}^{-i}, \nu)$ to isolate the role of $\pi^i$.

*Definition 3:* Let $\epsilon \geq 0$, $i \in \mathcal{N}$. A policy $\pi^{*i} \in \Gamma_S^i$ is called an *$\epsilon$-best-response to* $\boldsymbol{\pi}^{-i} \in \mathbf{\Gamma}_S^{-i}$ if, for every $s \in \mathbb{X}$,

$$J^i(\pi^{*i}, \boldsymbol{\pi}^{-i}, s) \leq \inf_{\tilde{\pi}^i \in \Gamma_S^i} J^i(\tilde{\pi}^i, \boldsymbol{\pi}^{-i}, s) + \epsilon.$$

The set of $\epsilon$-best-responses to $\boldsymbol{\pi}^{-i}$ is denoted $\mathrm{BR}_\epsilon^i(\boldsymbol{\pi}^{-i})$. It is well-known that for any $\boldsymbol{\pi}^{-i} \in \mathbf{\Gamma}_S^{-i}$, player $i$'s set of 0-best-responses $\mathrm{BR}_0^i(\boldsymbol{\pi}^{-i})$ is non-empty, and the infimum above is in fact attained.

*Definition 4:* Let $\epsilon \geq 0$. A joint policy $\boldsymbol{\pi}^* \in \mathbf{\Gamma}_S$ is called an *$\epsilon$-equilibrium* if $\pi^{*i} \in \mathrm{BR}_\epsilon^i(\boldsymbol{\pi}^{*-i})$ for all $i \in \mathcal{N}$.

For $\epsilon \geq 0$, we let $\mathbf{\Gamma}_S^{\epsilon\text{-eq}} \subseteq \mathbf{\Gamma}_S$ denote the set of $\epsilon$-equilibrium policies. It is known that the set $\mathbf{\Gamma}_S^{0\text{-eq}}$ is non-empty [17]. We also let $\mathbf{\Gamma}_{SD}^{\epsilon\text{-eq}} \subset \mathbf{\Gamma}_{SD}$ denote the subset of stationary deterministic $\epsilon$-equilibrium policies, which may be empty in general.

### B. Weakly Acyclic Stochastic Games

We now introduce weakly acyclic games, an important subclass of games that will be the main focus of this paper.

*Definition 5:* A sequence $\{\boldsymbol{\pi}_k\}_{k\geq 0}$ in $\mathbf{\Gamma}_{SD}$ is called a *strict best-response path* if for any $k \geq 0$ there is a unique player $i \in \mathcal{N}$ such that $\pi_{k+1}^i \neq \pi_k^i$ and $\pi_{k+1}^i \in \mathrm{BR}_0^i(\boldsymbol{\pi}_k^{-i})$.

*Definition 6:* The stochastic game $\mathcal{G}$ is weakly acyclic if (i) $\mathbf{\Gamma}_{SD}^{0\text{-eq}} \neq \varnothing$, and (ii) for any $\boldsymbol{\pi}_0 \in \mathbf{\Gamma}_{SD}$, there is a strict best-response path from $\boldsymbol{\pi}_0$ to some $\boldsymbol{\pi}^* \in \mathbf{\Gamma}_{SD}^{0\text{-eq}}$.

The multi-state formulation above was stated in [6], though weakly acyclic games had previously been studied in stateless games [18]. An important special case is that of stochastic teams, where $c^i = c^j$ for each $i, j$, and the interests of all agents are perfectly aligned. Markov potential games, [19], [20], [21] constitute another special case of weakly acyclic games.

### C. Q-Functions in Stochastic Games

In the stochastic game $\mathcal{G}$, when the other players use a stationary policy $\boldsymbol{\pi}^{-i} \in \mathbf{\Gamma}_S^{-i}$, player $i$ faces an environment that is equivalent to a single-agent MDP. The MDP in question depends on the policy $\boldsymbol{\pi}^{-i}$ as well as the game $\mathcal{G}$, and (stationary Markov) optimal policies for this MDP are equivalent to 0-best-responses to $\boldsymbol{\pi}^{-i}$ in the game $\mathcal{G}$.

Player $i$'s best-responses to a policy $\boldsymbol{\pi}^{-i} \in \mathbf{\Gamma}_S^{-i}$ can be characterized using an appropriately defined *Q-function*, $Q_{\boldsymbol{\pi}^{-i}}^{*i} : \mathbb{X} \times \mathbb{U}^i \to \mathbb{R}$.[3] The function $Q_{\boldsymbol{\pi}^{-i}}^{*i}$ can be defined by a fixed point equation of a Bellman operator, but here we give an equivalent definition in terms of the optimal policy of the corresponding MDP:

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

for all $(x, a^i) \in \mathbb{X} \times \mathbb{U}^i$, where $\boldsymbol{\pi}^* = (\pi^{*i}, \boldsymbol{\pi}^{-i})$ and $\pi^{*i} \in \mathrm{BR}_0^i(\boldsymbol{\pi}^{-i}) \cap \Gamma_{SD}^i$.

*Definition 7:* For $Q^i : \mathbb{X} \times \mathbb{U}^i \to \mathbb{R}$ and $\epsilon \geq 0$, we define

$$\widehat{\mathrm{BR}}_\epsilon^i(Q^i) := \{\pi^{*i} \in \Gamma_{SD}^i : Q^i(x, \pi^{*i}(x)) \leq \min_{a^i \in \mathbb{U}^i} Q^i(x, a^i) + \epsilon, \forall x \in \mathbb{X}\}.$$

The set $\widehat{\mathrm{BR}}_\epsilon^i(Q^i)$ is the set of stationary deterministic policies that are $\epsilon$-greedy with respect to $Q^i$. The function $Q^i$ plays the role of an action-value function, and for $Q^i = Q_{\boldsymbol{\pi}^{-i}}^{*i}$, we have $\widehat{\mathrm{BR}}_\epsilon^i(Q_{\boldsymbol{\pi}^{-i}}^{*i}) = \mathrm{BR}_\epsilon^i(\boldsymbol{\pi}^{-i}) \cap \Gamma_{SD}^i$.

When the remaining players follow a stationary policy, player $i$ can use Q-learning to estimate its action-values, which can then be used to estimate a 0-best-response policy. The situation is more complicated when the remaining players revise their policies over time. Under this non-stationarity, Q-learning may not be guaranteed to converge, and this procedure for estimating a best-response may be ineffective. These issues were considered by [6], who proposed the *Decentralized Q-learning algorithm* as a means of estimating best-response policies in the presence of policy updating, but required synchronized policy updating. In the next section,

[3] We use the terms Q-function, action-value function, and state-action value function interchangeably.

we present Algorithm 1, a modification of Decentralized Q-learning that allows for decentralized parameter selection and can tolerate non-stationarity of the learning environment.

## III. Asynchronous Decentralized Q-Learning

An asynchronous variant of Decentralized Q-learning is presented in Algorithm 1. Unlike in the original decentralized Q-learning algorithm, Algorithm 1 allows for the sequence of exploration phase lengths $\{T_k^i\}_{k\geq 0}$ to vary by agent, employs constant learning rate, and does not reset Q-factors at the end of an exploration phase. These discrepancies will be addressed in Section V.

**Algorithm 1:** Asynchronous Decentralized Q-Learning

**Set Parameters**
- $\mathbb{Q}^i \subset \mathbb{R}^{\mathbb{X}\times\mathbb{U}^i}$: a compact set
- $\{T_k^i\}_{k\geq 0}$: a sequence in $\mathbb{N}$ of learning phase lengths
  - Put $t_0^i = 0$ and $t_{k+1}^i = t_k^i + T_k^i$ for all $k \geq 0$.
- $\rho^i \in (0,1)$: experimentation probability
- $\lambda^i \in (0,1)$: inertia during policy update
- $\delta^i \in (0,\infty)$: tolerance level for sub-optimality
- $\alpha^i \in (0,1)$: step-size parameter

**Initialize** $\pi_0^i \in \Gamma_{SD}^i$, $\widehat{Q}_0^i \in \mathbb{Q}^i$ (arbitrary)

**for** $k \geq 0$ ($k^{th}$ exploration phase for agent $i$ )

  **for** $t = t_k^i, t_k^i + 1, \ldots, t_{k+1}^i - 1$

    Observe $x_t$

    Select $u_t^i = \begin{cases} \pi_k^i(x_t), & \text{w.p. } 1-\rho^i \\ u^i \sim \text{Unif}(\mathbb{U}^i), & \text{w.p. } \rho^i \end{cases}$

    Observe cost $c_t^i := c(x_t, \mathbf{u}_t)$, state $x_{t+1}$

    Put $\Delta_t^i = c_t^i + \beta^i \min_{a^i} \widehat{Q}_t^i(x_{t+1}, a^i)$

    $\widehat{Q}_{t+1}^i(x_t, u_t^i) = (1-\alpha^i)\widehat{Q}_t^i(x_t, u_t^i) + \alpha^i \Delta_t^i$

    $\widehat{Q}_{t+1}^i(x, u^i) = \widehat{Q}_t^i(x, u^i)$, for all $(x, u^i) \neq (x_t, u_t^i)$

  **if** $\pi_k^i \in \widehat{\text{BR}}_{\delta^i}^i(\widehat{Q}_{t_{k+1}^i}^i)$, **then**

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

  **else**

    $\pi_{k+1}^i \leftarrow \begin{cases} \pi_k^i, & \text{w.p. } \lambda^i \\ \pi^i \in \widehat{\text{BR}}_{\delta^i}^i(\widehat{Q}_{t_{k+1}^i}^i), & \text{w.p. } 1-\lambda^i \end{cases}$

### A. Primitive Random Variables

We now introduce several collections of *primitive* random variables that will be used in describing the assumptions and implementation of Algorithm 1. For any player $i \in \mathcal{N}$ and $t \geq 0$, we define the following random variables:

- $\{W_t\}_{t\geq 0}$ is an identically distributed, $[0,1]$-valued stochastic process. For some $f : \mathbb{X} \times \mathbf{U} \times [0,1] \to \mathbb{X}$, state transitions are driven by $\{W_t\}_{t\geq 0}$ via $f$:

$$\begin{aligned}&\Pr(x_{t+1} = s' | x_t = s, \mathbf{u}_t = \mathbf{a}) = P(s'|s,\mathbf{a}) \\ =&\Pr\left(W_t \in \{w : f(s,\mathbf{a},w) = s'\}\right),\end{aligned}$$

  for any $(s, \mathbf{a}, s') \in \mathbb{X} \times \mathbf{U} \times \mathbb{X}$ and $t \geq 0$;
- $\tilde{u}_t^i \sim \text{Unif}(\mathbb{U}^i)$;
- $\tilde{\rho}_t^i \sim \text{Unif}([0,1])$;
- $\tilde{\lambda}_t^i \sim \text{Unif}([0,1])$;
- For non-empty $B^i \subseteq \Gamma_{SD}^i$, $\tilde{\pi}_t^i(B^i) \sim \text{Unif}(B^i)$;
- $T_t^i$ is an $\mathbb{N}$-valued random variable, elaborated below.

*Assumption 1:* The collection of primitive random variables $\mathcal{V}_1 \cup \mathcal{V}_2$ is mutually independent, where

$$\begin{aligned}\mathcal{V}_1 &:= \bigcup_{i\in\mathcal{N}, t\geq 0} \left\{W_t, \tilde{\rho}_t^i, \tilde{u}_t^i, \tilde{\lambda}_t^i, T_t^i\right\}, \\ \mathcal{V}_2 &:= \bigcup_{i\in\mathcal{N}, t\geq 0} \left\{\tilde{\pi}_t^i(B^i) : B^i \subseteq \Gamma_{SD}^i, B^i \neq \varnothing\right\}.\end{aligned}$$

**Remark:** The random variables in $\mathcal{V}_1 \cup \mathcal{V}_2$ are taken to be primitive random variables that, with the exception of exploration phase lengths $\{T_k^i\}_{i\in\mathcal{N}, k\geq 0}$, do not depend on any player's choice of hyperparameters. We also note that the primitive random variables $\{\tilde{u}_t^i : i \in \mathcal{N}, t \geq 0\}$ should not be conflated with the action process $\{u_t^i : i \in \mathcal{N}, t \geq 0\}$, which depends on the sample path and is not a collection of primitive random variables.

### B. Assumptions

In order to state our main result, Theorem 1, we now impose some assumptions on the underlying game $\mathcal{G}$ and on the choices of hyperparameters at each player.

*Assumption 2:* For any pair of states $(s, s') \in \mathbb{X}\times\mathbb{X}$, there exists $H = H(s,s') \in \mathbb{N}$ and a sequence of joint actions $\mathbf{a}_0, \ldots, \mathbf{a}_H \in \mathbf{U}$ such that

$$\Pr(x_{H+1} = s' | x_0 = s, \mathbf{u}_0 = \mathbf{a}_0, \ldots, \mathbf{u}_H = \mathbf{a}_H) > 0.$$

Assumption 2 requires that the state process can be driven from any initial state to any other state in finitely many steps, provided a suitable selection of joint actions is made. This is a rather weak assumption on the underlying transition kernel $P$, and is quite standard in the theory of MARL (c.f. [13, Assumption 4.1, Case iv]).

Our next assumption restricts the hyperparameter selections in Algorithm 1. Let $\bar{\delta} := \min(\mathfrak{A} \setminus \{0\})$, where

$$\begin{aligned}\mathfrak{A} := \big\{\big|&Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a_1^i) - Q_{\boldsymbol{\pi}^{-i}}^{*i}(s, a_2^i)\big| : \\ &i \in \mathcal{N}, \boldsymbol{\pi}^{-i} \in \mathbf{\Gamma}_{SD}^{-i}, s \in \mathbb{X}, a_1^i, a_2^i \in \mathbb{U}^i\big\}.\end{aligned}$$

The quantity $\bar{\delta}$, defined originally by [6] and recalled above, is the minimum *non-zero* separation between two optimal Q-factors with matching states, minimized over all agents $i \in \mathcal{N}$ and over all policies $\pi^{-i} \in \Pi^{-i}$.

For any baseline policy $\boldsymbol{\pi} \in \mathbf{\Gamma}_{SD}$ and fixed exploration parameters $\{\rho^i\}_{i\in\mathcal{N}}$, we use the notation $\hat{\boldsymbol{\pi}} \in \mathbf{\Gamma}_S$ to denote a corresponding behaviour policy, which is stationary but not deterministic. When using $\hat{\pi}^i$, agent $i \in \mathcal{N}$ follows $\pi^i$ with probability $1-\rho^i$ and mixes uniformly over $\mathbb{U}^i$ with probability $\rho^i$. [6, Lemma B3] shows that the optimal Q-factors for these two environments will be close provided $\rho^i$ is sufficiently small for all $i$. In particular, there exists $\bar{\rho} > 0$ such that if $\rho^i \in (0, \bar{\rho})\ \forall i \in \mathcal{N}$, then

$$\|Q_{\boldsymbol{\pi}^{-j}}^{*j} - Q_{\hat{\boldsymbol{\pi}}^{-j}}^{*j}\|_\infty < \frac{\min\{\delta^i, \bar{\delta} - \delta^i\}}{4}, \forall j \in \mathcal{N}, \boldsymbol{\pi}^{-j} \in \mathbf{\Gamma}_{SD}^{-j}.$$

*Assumption 3:* For all $i \in \mathcal{N}$, $\delta^i \in (0, \bar{\delta})$ and $\rho^i \in (0, \bar{\rho})$.

*Assumption 4:* There exists integers $R, T \in \mathbb{N}$ such that

$$\Pr\left(\cap_{i\in\mathcal{N},k\geq 0}\left\{T_k^i \in [T, RT]\right\}\right) = 1.$$

When all players use Algorithm 1, we view the resulting sequence of policies $\{\pi_k^i\}_{k\geq 0}$ as player $i$'s *baseline* policy process, where $\pi_k^i$ is $i$'s baseline policy during $[t_k^i, t_{k+1}^i)$, player $i$'s $k^{th}$ exploration phase. This policy process $\{\pi_k^i\}_{k\geq 0}$ is indexed on the coarser timescale of exploration phases. For ease of reference, we also introduce a sequence of baseline policies indexed by the finer timescale of stage games. For $t \geq 0$, we let $\phi_t^i = \pi_k^i$ whenever $t \in [t_k^i, t_{k+1}^i)$ denote player $i$'s baseline policy during the stage game at time $t$. The baseline joint policy at stage game $t$ is then denoted $\boldsymbol{\phi}_t = (\phi_t^i)_{i\in\mathcal{N}}$. Furthermore, we refer to the collection of Q-factor step-size parameters $\{\alpha^i\}_{i\in\mathcal{N}}$ as $\boldsymbol{\alpha} \in (0,1)^N$.

*Theorem 1:* Let $\mathcal{G}$ be a weakly acyclic game and suppose each player $i \in \mathcal{N}$ uses Algorithm 1 to play $\mathcal{G}$. Suppose Assumptions 1–4 hold, and let $\epsilon > 0$. There exists $\bar{\alpha}_\epsilon > 0$ and a function $\bar{T}_\epsilon(0,1)^N \times \mathbb{N} \to \mathbb{N}$ such that if

$$\max_{i\in\mathcal{N}} \alpha^i < \bar{\alpha}_\epsilon, \text{ and } T \geq \bar{T}_\epsilon(\boldsymbol{\alpha}, R),$$

then $\Pr(\boldsymbol{\phi}_t \in \boldsymbol{\Gamma}_{SD}^{0\text{-eq}}) \geq 1-\epsilon$, for all sufficiently large $t \in \mathbb{N}$.

The remainder of this paper is devoted to proving Theorem 1 and discussing insights from the analysis of the proof.

## IV. Proof of theorem 1

In the analysis to follow, we study the performance of Algorithm 1 under a fixed realization of the primitive exploration phase random variables $\mathcal{T} = \{T_k^i : i \in \mathcal{N}, k \geq 0\}$. We now introduce several objects to facilitate the analysis, and we suppress any dependence on the realization of $\mathcal{T}$.

### A. Various Notations and Constructions for the Proof

For each $i \in \mathcal{N}$, we arbitrarily order non-empty subsets of $\Gamma_{SD}^i$ as $B^{i,1}, \dots, B^{i,m_i}$, where $m_i = |2^{\Gamma_{SD}^i} \setminus \varnothing|$. We introduce the following new quantities for each $t \geq 0$:

- $\omega_t^i := (\tilde{\rho}_t^i, \tilde{u}_t^i, \tilde{\lambda}_t^i, \tilde{\pi}_t^i(B^{i,1}), \dots, \tilde{\pi}_t^i(B^{i,m_i})), \quad \forall i \in \mathcal{N}$;
- $\boldsymbol{\omega}_t := (W_t, \omega_t^1, \dots, \omega_t^N)$;
- $\boldsymbol{\varpi}_t := (\boldsymbol{\omega}_t, \boldsymbol{\omega}_{t+1}, \dots)$;
- $\widehat{\mathbf{Q}}_t := (\widehat{Q}_t^1, \dots, \widehat{Q}_t^N)$;
- $h_t := (x_0, \boldsymbol{\phi}_0, \widehat{\mathbf{Q}}_0, \dots, x_t, \boldsymbol{\phi}_t, \widehat{\mathbf{Q}}_t)$;
- $H_t := \left(\mathbb{X} \times \boldsymbol{\Gamma}_{SD} \times \mathbb{R}^{\mathbb{X}\times\mathbb{U}^1} \times \cdots \times \mathbb{R}^{\mathbb{X}\times\mathbb{U}^N}\right)^{t+1}$;
- $H_{t,\text{eq}} := \{h_t \in H_t : \boldsymbol{\phi}_t \in \boldsymbol{\Gamma}_{SD}^{0\text{-eq}}\}$;

For $i \in \mathcal{N}, t \geq 0$, let the functions $\mathcal{Q}_t^i$ and $\Phi_t$ be constructed such that $Q_t^i = \mathcal{Q}_t^i(h_s, \boldsymbol{\varpi}_s)$, and $\boldsymbol{\phi}_t = \Phi_t(h_s, \boldsymbol{\varpi}_s)$, for all $0 \leq s \leq t$.

Next, for any $s, t \geq 0$ and any $i \in \mathcal{N}$, we introduce a function $\bar{Q}_{t+s}^i(h_t, \boldsymbol{\varpi}_t)$ that reports the *hypothetical Q-factors* player $i$ *would have* obtained if the baseline policies were frozen at time $t$. That is, the history up to time $t$ is given by $h_t$, the primitive random variables from $t$ on are given by $\boldsymbol{\varpi}_t$, and we obtain the hypothetical ($\boldsymbol{\varpi}_t$-measurable) continuation trajectory $(\bar{x}_t, \bar{\mathbf{u}}_t, \dots, \bar{x}_{t+s}, \bar{\mathbf{u}}_{t+s})$, as $\bar{x}_t := x_t$,

$$\bar{x}_{t+m+1} = f(\bar{x}_{t+m}, \bar{\mathbf{u}}_{t+m}, W_{t+m}), \quad \forall\, 0 \leq m \leq s,$$

where for each player $j$ and time $t+m \geq t$,

$$\bar{u}_{t+m}^j := \begin{cases} \phi_t^i(\bar{x}_{t+m}), & \text{if } \tilde{\rho}_{t+m}^i > \rho^i \\ \tilde{u}_{t+m}^i & \text{otherwise.} \end{cases}$$

Note that the index of $\phi_t^i$ is not $t+m$, which reflects that, in this hypothetical continuation, the baseline policies were frozen at time $t$. Each hypothetical Q-factor estimate $\bar{Q}_{t+s}^i(s, a^i)$ is then (analytically) built out of this hypothetical trajectory using the regular Q-learning update with the initial condition prescribed by $\widehat{\mathbf{Q}}_t$.

### B. Supporting Results on Q-Learning Iterates

Recall that for each player $i$, the set $\mathbb{Q}^i \subset \mathbb{R}^{\mathbb{X}\times\mathbb{U}^i}$ is compact, and so $M_i := \sup\{\|Q^i\|_\infty : Q^i \in \mathbb{Q}^i\} < \infty$.

*Lemma 1:* We have that

$$\max_{i\in\mathcal{N}} \sup_{t\geq 0} \left\|\widehat{Q}_t^i\right\|_\infty \leq \max_{i\in\mathcal{N}} \left\{\frac{\|c^i\|_\infty}{(1-\beta^i)}, M_i\right\} < \infty.$$

**Proof:** Let $M := \max_{i\in\mathcal{N}} \left\{\frac{\|c^i\|_\infty}{(1-\beta^i)}, M_i\right\}$. For all $i \in \mathcal{N}$, we have

$$\|\widehat{Q}_{t+1}^i\|_\infty \leq$$
$$\max\left\{(1-\alpha^i)\|\widehat{Q}_t^i\|_\infty + \alpha^i\big(\|c^i\|_\infty + \beta^i\|\widehat{Q}_t^i\|_\infty\big), \|\widehat{Q}_t^i\|_\infty\right\}.$$

If $\|\widehat{Q}_t^i\|_\infty \leq M$, then

$$\begin{aligned}\|\widehat{Q}_{t+1}^i\|_\infty &\leq \max\left\{(1-\alpha^i)M + \alpha^i\big(\|c^i\|_\infty + \beta M\big), M\right\} \\ &= \max\left\{M + \alpha^i\big(\underbrace{\|c^i\|_\infty - (1-\beta^i)M}_{\leq 0}\big), M\right\} \\ &= M.\end{aligned}$$

This proves the lemma since $\|\widehat{Q}_0^i\|_\infty \leq M$.

The following lemma says that players can learn their optimal Q-factors accurately when they use sufficiently small step sizes and when their learning is not disrupted by policy updates for a sufficiently long number of steps. It is worded in terms of primitive random variables and hypothetical continuation Q-factors for reasons described in Section V.

*Lemma 2:* For any $\xi > 0$, there exists $\hat{\alpha}_\xi > 0$ and function $\hat{T}_\xi : (0,1)^N \to \mathbb{N}$ such that if (i) $\alpha^i \in (0, \hat{\alpha}_\xi)$ for all $i \in \mathcal{N}$, and (ii) $\tilde{T} \geq \hat{T}_\xi(\boldsymbol{\alpha})$, then

$$\Pr(\Omega_{t:t+\tilde{T}}) \geq 1-\xi, \quad \forall\, t \geq 0, h_t \in H_t,$$

where

$$\Omega_{t:t+\tilde{T}} = \left\{\boldsymbol{\varpi}_t : \max_{i\in\mathcal{N}} \|\bar{Q}_{t+\tilde{T}}^i(h_t, \boldsymbol{\varpi}_t) - \widehat{Q}_{\boldsymbol{\phi}_t^{-i}}^{*i}\|_\infty < \xi\right\}.$$

We fix $\hat{\alpha}_\xi$ and $\hat{T}_\xi(\cdot)$ with the properties outlined in Lemma 2.

**Proof:** Since each player $i$'s policy is $\rho^i$-soft, our Assumption 2 implies the persistent excitation assumption of

[23, Assumption 1]. The result then follows from Lemma 1 and [23, Theorem 3.4], using the Markov inequality. □

### C. A Supporting Result Controlling Update Frequencies

A core challenge of analyzing non-synchronized multi-agent learning is that when one player updates its policy, it changes the environment for others. That is, policy updates for one player constitute potential destabilizations of learning for others. We now introduce a sequence of time intervals $\{[\tau_k^{\min}, \tau_k^{\max}]\}_{k\geq 0}$ that will be useful in quantifying and analyzing the effects of such disruptions.

*Definition 8:* Let $\tau_0^{\min} = \tau_0^{\max} := 0$. For $k \geq 0$, define

$$\tau_{k+1}^{\min} := \inf\{t_n^i : t_n^i > \tau_k^{\max}, i \in \mathcal{N}, n \geq 0\}$$
$$\tau_{k+1}^{\max} := \inf\big\{t \geq \tau_{k+1}^{\min} : \forall\, i \in \mathcal{N}, \exists\, n \in \mathbb{N} \text{ s.t. } t_n^i \in [\tau_{k+1}^{\min}, t], \text{ and } \inf\{t_{\bar{n}}^i > t : i \in \mathcal{N}, \bar{n} \in \mathbb{N}\} \geq t + T/N\big\},$$

where $T$ is the constant appearing in Assumption 4.

The intervals $[\tau_k^{\min}, \tau_k^{\max}]$ represent active phases, during which players may change their policies. The $(k+1)^{th}$ active phase starts at $\tau_{k+1}^{\min}$, which is defined as the first time after $\tau_k^{\max}$ at which some agent has an opportunity to revise its policy. As a consequence, no policy updates occur in $(\tau_k^{\max}, \tau_{k+1}^{\min})$.

The definition of $\tau_{k+1}^{\max}$ is slightly more involved: it requires that (a) each agent has an opportunity to switch policies in $[\tau_{k+1}^{\min}, \tau_{k+1}^{\max}]$ and (b) that no agent finishes its current exploration phase in the next $T/N$ stage games.

These active phases are characterized by the end times of each player's exploration phases, $\{t_n^i : i \in \mathcal{N}, n \geq 1\}$, and therefore depend on the exploration phase length parameters $\mathcal{T} = \{T_n^i : i \in \mathcal{N}, n \geq 0\}$. We recall that the collection $\mathcal{T}$ is generated at the outset of play, independently of the other primitive random variables. In the analysis to follow, we fix a particular realization of $\mathcal{T}$ satisfying Assumption 4 and our notation suppresses the dependence of $\{\tau_k^{\min}, \tau_k^{\max}\}_{k\geq 0}$ on this realization. All statements are understood to hold almost surely.

*Lemma 3:* The sequences $\{\tau_k^{\min}\}_{k\geq 0}$ and $\{\tau_k^{\max}\}_{k\geq 0}$ are well-defined (i.e. the infimum defining each term is achieved by a finite integer), and, for any $k \geq 0$, we have that

(a) $\tau_{k+1}^{\min} \geq \tau_k^{\max} + T/N$;

(b) For each $i \in \mathcal{N}$, we have

$$\sum_{n\geq 0} \mathbf{1}\{t_n^i \in [\tau_k^{\min}, \tau_k^{\max}]\} \leq R + 1.$$

**Proof:**

For some $k \geq 0$, suppose that (a) and (b) hold for all $l \leq k$. That is,

(a) $\tau_{l+1}^{\min} \geq \tau_l^{\max} + T/N$, for all $0 \leq l \leq k$;

(b) for each $i \in \mathcal{N}, l \leq k$, we have

$$\sum_{n\geq 0} \mathbf{1}\{t_n^i \in [\tau_l^{\min}, \tau_l^{\max}]\} \leq R + 1.$$

We observe that this holds for $k = 0$. We will show the following: (1) $\tau_{k+1}^{\max} < \infty$; (2) $\tau_{k+2}^{\min} \geq \tau_{k+1}^{\max} + T/N$; (3) For all $i \in \mathcal{N}$

$$\sum_{n\geq 0} \mathbf{1}\{t_n^i \in [\tau_{k+1}^{\min}, \tau_{k+1}^{\max}]\} \leq R + 1.$$

For each $i \in \mathcal{N}$, let $n_i \in \mathbb{Z}_{\geq 0}$ denote the index such that

$$t_{n_i-1}^i < \tau_{k+1}^{\min} \leq t_{n_i}^i.$$

By minimality of $\tau_{k+1}^{\min}$, we have that $t_{n_i-1}^i \leq \tau_k^{\max}$. By Assumption 4, we have $T_{n_i-1}^i \leq RT$ and thus

$$t_{n_i}^i = t_{n_i-1}^i + T_{n_i-1}^i \leq \tau_k^{\max} + RT.$$

This, in turn, implies $t_{n_i}^i - \tau_{k+1}^{\min} \leq RT - T/N$, since $\tau_{k+1}^{\min} \geq \tau_k^{\max} + T/N$ by hypothesis.

We put $\hat{t}_0 := \max_i t_{n_i}^i$, and let $j(0) \in \mathcal{N}$ denote an agent with $t_{n_{j(0)}}^{j(0)} = \hat{t}_0$.

For each $l \in \{0, 1, \ldots, N-1\}$, we define $\hat{t}_{l+1}$ as

$$\hat{t}_{l+1} = \min\{t_{\bar{k}}^j > \hat{t}_l : j \in \mathcal{N}, \bar{k} \geq 0\}.$$

Observe that $\hat{t}_0 \leq \tau_{k+1}^{\max}$. Moreover, if there exists some $l \leq N-1$ such that $\hat{t}_{l+1} \geq \hat{t}_l + T/N$, then $\tau_{k+1}^{\max} \leq \hat{t}_l$, and thus $\tau_{k+1}^{\max} < \infty$. We now argue that such $l$ exists.

For the sake of a contradiction, suppose

$$\max\{\hat{t}_1 - \hat{t}_0, \cdots, \hat{t}_N - \hat{t}_{N-1}\} < T/N \qquad (\dagger)$$

This implies

$$\hat{t}_N - \hat{t}_0 = \sum_{l=0}^{N-1} \left(\hat{t}_{l+1} - \hat{t}_l\right) < N \cdot \frac{T}{N} = T \leq \min_{i,n} T_n^i.$$

One concludes that the minima defining each $\{\hat{t}_l : 1 \leq l \leq N\}$ are attained by $N$ distinct minimizing agents, one of whom is $j(0)$. But then, for some $l$, we have

$$\hat{t}_0 = t_{n_{j(0)}}^{j(0)}, \text{ and } t_{n_{j(0)}}^{j(0)} + T_{n_{j(0)}}^{j(0)} = \hat{t}_l \leq \hat{t}_N,$$

which implies $T_{n_{j(0)}}^{j(0)} < T$, contradicting Assumption 4.

We have thus shown that the set $\mathfrak{T} \neq \varnothing$, where $\mathfrak{T}$ is given by

$$\mathfrak{T} := \{l \in \{0, 1, \ldots, N-1\} : \hat{t}_{l+1} \geq \hat{t} + T/N\}.$$

Let $l^* = \min \mathfrak{T}$, and note that, if $l^* \neq 0$, then $\hat{t}_{k+1} < \hat{t}_k + T/N$ for all $k < l^*$. It follows that (1) $\tau_{k+1}^{\max} = \hat{t}_{l^*} < \infty$ and (2) $\tau_{k+2}^{\min} = \hat{t}_{l^*+1} \geq \tau_{k+1}^{\max} + T/N$.

We conclude by showing that (3) holds. That is,

$$\sum_{n\geq 0} \mathbf{1}\{t_n^i \in [\tau_{k+1}^{\min}, \tau_{k+1}^{\max}]\} \leq R + 1, \quad \forall i \in \mathcal{N}.$$

By Assumption 4, it suffices to show that

$$\tau_{k+1}^{\max} - \tau_{k+1}^{\min} < (R+1)T.$$

We have already shown $\hat{t}_0 - \tau_{k+1}^{\min} \leq RT - T/N$, which handles the case where $l^* = 0$, and we focus on $l^* > 0$. Since $\hat{t}_{k+1} < \hat{t}_k + T/N$ for all $k < l^*$ and $l^* \leq N-1$, we have $\tau_{k+1}^{\max} - \tau_{k+1}^{\min} = \hat{t}_{l^*} - \tau_{k+1}^{\min}$ and

$$\hat{t}_{l^*} - \tau_{k+1}^{\min} = \sum_{l=0}^{l^*-1} \left[\hat{t}_{l+1} - \hat{t}_l\right] + \hat{t}_0 - \tau_{k+1}^{\min}$$

$$\leq l^*(T/N) + (RT - T/N) < (R+1)T,$$

which concludes the proof. □

For $k \geq 0$, we define $B_k$ to be the event that the baseline policy is a fixed equilibrium policy throughout the $k^{th}$ active phase, $[\tau_k^{\min}, \tau_k^{\max}]$. That is,

$$B_k := \left\{ \boldsymbol{\phi}_{\tau_k^{\min}} = \cdots = \boldsymbol{\phi}_{\tau_k^{\max}} \in \mathbf{\Gamma}_{SD}^{\text{0-eq}} \right\}.$$

We define $L := \ell^* + 1$, where $\ell^* = \max\{\ell(\boldsymbol{\pi}) : \boldsymbol{\pi} \in \mathbf{\Gamma}_{SD}\}$ and $\ell(\boldsymbol{\pi})$ denotes the length of a shortest strict best-response path from $\boldsymbol{\pi} \in \mathbf{\Gamma}_{SD}$ to an equilibrium policy in $\mathbf{\Gamma}_{SD}^{\text{0-eq}}$.

*Lemma 4:* Let $\theta > 0$, and define

$$\xi := \frac{1}{(R+1)NL} \min\left\{ \theta, \frac{1}{2} \min_{i \in \mathcal{N}} \{\delta^i, \bar{\delta} - \delta^i\} \right\}.$$

Suppose $\max_{i \in \mathcal{N}} \alpha^i < \hat{\alpha}_\xi$ and $T \geq N\hat{T}_\xi(\boldsymbol{\alpha})$, where $\hat{\alpha}_\xi$ and $\hat{T}_\xi(\cdot)$ are the objects specified in Lemma 2. For any $k \geq 0$ and history $h_{\tau_k^{\max}} \in H_{\tau_k^{\max},\text{eq}}$, we have

$$\Pr\left(B_{k+1} | h_{\tau_k^{\max}}\right) \geq 1 - \theta/L.$$

**Remark:** Since we are studying a particular realization of the exploration phase lengths $\mathcal{T}$, we have that $\tau_k^{\max} \in \mathbb{Z}_{\geq 0}$ is some constant, and $h_{\tau_k^{\max}}$ is a $\tau_k^{\max}$-history,

$$h_{\tau_k^{\max}} = (x_0, \boldsymbol{\phi}_0, \widehat{\mathbf{Q}}_0, \ldots, x_{\tau_k^{\max}}, \boldsymbol{\phi}_{\tau_k^{\max}}, \widehat{\mathbf{Q}}_{\tau_k^{\max}}),$$

with $\boldsymbol{\phi}_{\tau_k^{\max}} \in \mathbf{\Gamma}_{SD} \cap \mathbf{\Gamma}^{\text{0-eq}}$.

**Proof:** We put $\tilde{t}_0 := \tau_{k+1}^{\min}$ and recursively define

$$\tilde{t}_{l+1} := \min\{t_n^i > \tilde{t}_l : i \in \mathcal{N}, n \geq 0\}, \quad \forall l \geq 0.$$

We define $m \in \mathbb{Z}_{\geq 0}$ to be the index achieving $\tilde{t}_m = \tau_{k+1}^{\max}$, and note that $m = 0$ is possible when $\tau_{k+1}^{\max} = \tau_{k+1}^{\min}$.

Note that $m$ counts the number of stage games after $\tau_{k+1}^{\min}$ and on/before $\tau_{k+1}^{\max}$ at which *any* player ends an exploration phase. From the proof of Lemma 3 we have that

$$m \leq \sum_{i \in \mathcal{N}} \sum_{n \geq 0} \mathbf{1}\{t_n^i \in [\tau_{k+1}^{\min}, \tau_{k+1}^{\max}]\} < (R+1)N.$$

For each $l \in \{0, \cdots, m\}$, let

$$A_l := \{i \in \mathcal{N} : \exists\, n(l) \in \mathbb{N} \text{ s.t. } t_{n(l)}^i = \tilde{t}_l\}.$$

From our choices of $\xi, \boldsymbol{\alpha}$, and $T \geq N\hat{T}_\xi(\boldsymbol{\alpha})$ and the fact that $\tilde{t}_l \geq \tau_{k+1}^{\min} \geq \tau_k^{\max} + T/N$ for all $l \in \{0, 1, \ldots, m\}$, we have, by Lemma 2, that

$$\Pr\left(\Omega_{\tau_k^{\max}:\tilde{t}_l}\right) \geq 1 - \xi, \quad \forall l \in \{0, 1, \ldots, m\},$$

where, for any $s, t \geq 0$, we recall that $\Omega_{t:t+s}$ is given by

$$\Omega_{t:t+s} = \left\{ \boldsymbol{\varpi}_t : \max_{i \in \mathcal{N}} \left\| \bar{Q}_{t+s}^i(h_t, \boldsymbol{\varpi}_t) - Q_{\boldsymbol{\phi}_t^{-i}}^{*i} \right\|_\infty < \xi \right\}.$$

Note that $\tilde{t}_0 = \tau_{k+1}^{\min}$, and the minimality defining $\tau_{k+1}^{\min}$ implies that no player updates its policy during the interval $[\tau_k^{\max}, \tau_{k+1}^{\min})$. It follows that the hypothetical continuation trajectory defining each player $i$'s hypothetical Q-factors $\bar{Q}_{\tilde{t}_0}^i(h_{\tau_k^{\max}}, \boldsymbol{\varpi}_{\tau_k^{\max}})$ coincides with the sample trajectory defining $\widehat{\mathbf{Q}}_{\tilde{t}_0}$, since action selections and state transitions are decided by $\boldsymbol{\phi}_{\tau_k^{\max}}$ and $\boldsymbol{\varpi}_{\tau_k^{\max}}$. Thus, by our choice of $\tilde{t}_0 = \tau_{k+1}^{\min}$,

$$\bar{Q}_{\tilde{t}_0}^i(h_{\tau_k^{\max}}, \boldsymbol{\varpi}_{\tau_k^{\max}}) = \widehat{Q}_{\tilde{t}_0}^i, \quad \forall i \in \mathcal{N}.$$

For $\boldsymbol{\varpi}_{\tau_k^{\max}} \in \Omega_{\tau_k^{\max}:\tilde{t}_0}$, each player $i \in A_0$ recovers its estimated Q-factors $\widehat{Q}_{\tilde{t}_0}^i$ within $\xi$ of $Q_{\boldsymbol{\phi}_{\tau_k^{\max}}^{-i}}^{*i}$, since

$$\widehat{Q}_{\tilde{t}_0}^i = \mathcal{Q}_{\tilde{t}_0}^i(h_{\tau_k^{\max}}, \boldsymbol{\varpi}_{\tau_k^{\max}}).$$

Since we have hypothesized that $\boldsymbol{\phi}_{\tau_k^{\max}} \in \mathbf{\Gamma}_{SD}^{\text{0-eq}}$ and chosen $\xi < \frac{1}{2}\min\{\delta^i, \bar{\delta} - \delta^i\}$, it follows that agent $i$ does not change its policy at time $\tilde{t}_0 = \tau_{k+1}^{\min}$ when given the opportunity to revise its policy. In symbols, that is to say

$$\begin{aligned}
&\boldsymbol{\varpi}_{\tau_k^{\max}} \in \Omega_{\tau_k^{\max}:\tilde{t}_0} \\
\Rightarrow &\max_{i \in \mathcal{N}} \left\| \mathcal{Q}_{\tilde{t}_0}^i(h_{\tau_k^{\max}}, \boldsymbol{\varpi}_{\tau_k^{\max}}) - Q_{\boldsymbol{\phi}_{\tau_k^{\max}}^{-i}}^{*i} \right\|_\infty < \xi \\
\Rightarrow &\boldsymbol{\phi}_{\tilde{t}_0} = \Phi_{\tilde{t}_0}(h_{\tau_k^{\max}}, \boldsymbol{\varpi}_{\tau_k^{\max}}) = \boldsymbol{\phi}_{\tau_k^{\max}}.
\end{aligned}$$

Repeating this logic, one has that if

$$\boldsymbol{\varpi}_{\tau_k^{\max}} \in \bigcap_{0 \leq l \leq m} \Omega_{\tau_k^{\max}:\tilde{t}_l}$$

then $\boldsymbol{\phi}_{\tilde{t}_l} = \Phi(h_{\tau_k^{\max}}, \boldsymbol{\varpi}_{\tau_k^{\max}}) = \boldsymbol{\phi}_{\tau_k^{\max}}$ for each $l \leq m$.

The probability of this intersection is lower bounded using the union bound and Lemma 2:

$$\Pr\left( \bigcap_{0 \leq l \leq m} \Omega_{\tau_k^{\max}:\tilde{t}_l} \right) \geq 1 - (m+1)\xi \geq 1 - \theta/L,$$

as desired. □

*Lemma 5:* Let $\theta > 0$, and define

$$\xi := \frac{1}{(R+1)NL} \min\left\{ \theta, \frac{1}{2} \min_{i \in \mathcal{N}} \{\delta^i, \bar{\delta} - \delta^i\} \right\}.$$

Suppose $\max_{i \in \mathcal{N}} \alpha^i < \hat{\alpha}_\xi$ and $T \geq N\hat{T}_\xi(\boldsymbol{\alpha})$, where $\hat{\alpha}_\xi$ and $\hat{T}_\xi(\cdot)$ are the objects specified in Lemma 2. For any $k \geq 0$ and history $h_{\tau_k^{\max}} \in H_{\tau_k^{\max}} \setminus H_{\tau_k^{\max},\text{eq}}$, we have

$$\Pr\left(B_{k+L} | h_{\tau_k^{\max}}\right) \geq p_{\min}(1 - \theta),$$

where $p_{\min} := \prod_{j \in \mathcal{N}} \min\left\{ \frac{1-\lambda^j}{|\Gamma_{SD}^j|}, \lambda^j \right\}^{(R+1)L} > 0$.

**Proof** Let $\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_\ell$ be a (shortest) strict best-response path from $\boldsymbol{\pi}_0 := \boldsymbol{\phi}_{\tau_k^{\max}}$ to $\boldsymbol{\pi}_\ell \in \mathbf{\Gamma}_{SD}^{\text{0-eq}}$, and $\ell \in \{1, \ldots, \ell^*\}$. For each pair of neighbouring joint policies $\boldsymbol{\pi}_l$ and $\boldsymbol{\pi}_{l+1}$, there exists exactly one player $i(l)$ who switches its policy. That is, $\pi_{l+1}^j = \pi_l^j$ for all $j \neq i(l)$.

Consider the event that no agent updates its policy during $[\tau_k^{\max}, \tau_{k+1}^{\max}]$ due to inertia except player $i(0)$, who updates its policy exactly once to $\pi_1^{i(0)}$ and remains inert at all other update opportunities in $[\tau_k^{\max}, \tau_{k+1}^{\max}]$.

By Lemma 2 and Lemma 3, the conditional probability of this event given $h_{\tau_k^{\max}}$ is lower bounded by

$$(1-\theta/L)\prod_{j\in\mathcal{N}}\min\left\{\frac{1-\lambda^j}{|\Pi^j|},\lambda^j\right\}^{R+1}.$$

The same lower bound similarly applies to each transition along $\boldsymbol{\pi}_0,\ldots,\boldsymbol{\pi}_\ell$, which leads to

$$\begin{aligned}&\Pr\left(\boldsymbol{\phi}_{\tau_{k+\ell}^{\max}}=\boldsymbol{\pi}_\ell\middle|h_{\tau_k^{\max}}\right)\\ \geq&(1-\theta/L)^\ell\prod_{j\in\mathcal{N}}\min\left\{\frac{1-\lambda^j}{|\Pi^j|},\lambda^j\right\}^{(R+1)\ell}.\end{aligned}$$

Next, consider the event that no agent updates its policy during $[\tau_{k+\ell}^{\max},\tau_{k+L}^{\max}]$ due to inertia. The conditional probability of this event given $h_{\tau_{k+\ell}^{\max}}$ is lower bounded by

$$\prod_{j\in\mathcal{N}}(\lambda^j)^{(R+1)(L-\ell)}.$$

This results in

$$\Pr(B_{k+L}|h_{\tau_k^{\max}})\geq(1-\theta/L)^L p_{\min}.$$

This proves the lemma since $(1-\theta/L)^L\geq 1-\theta$. □

### D. Proof of Theorem 1

Given $\epsilon>0$, let $\theta>0$ be the unique solution to

$$\frac{(1-\theta)p_{\min}}{\theta+(1-\theta)p_{\min}}-\theta=1-\epsilon$$

and

$$\xi=\frac{1}{(R+1)NL}\min\left\{\theta,\frac{1}{2}\min_{i\in\mathcal{N}}\{\delta^i,\bar{\delta}-\delta^i\}\right\}.$$

Suppose that

$$\max_{i\in\mathcal{N}}\alpha^i\leq\bar{\alpha}_\epsilon:=\hat{\alpha}_\xi\ \text{ and }\ \ T\geq\bar{T}_\epsilon(\boldsymbol{\alpha},R):=N\hat{T}_\xi(\boldsymbol{\alpha}).$$

By Lemmas 4 and 5, for all $k\geq 0$,

$$\Pr(B_{k+L}|B_k)\geq 1-\theta, \tag{2}$$
$$\Pr(B_{k+L}|B_k^c)\geq(1-\theta)p_{\min}. \tag{3}$$

Let $p_k:=\Pr(B_k)$ for all $k\geq 0$. The subsequent details lower bounding $p_{k+mL}$ for large $m$ and every $k<L$ are omitted, as they resemble the proofs of [6] or [10].

□

## V. Discussion and Insights

At its core, the proof of Theorem 1 amounts to first showing that the probability of transiting to equilibrium in a finite number of steps can be uniformly lower bounded and then showing that the probability of remaining at equilibrium can be made arbitrarily large with respect to this lower bound. Once this can be done, the mechanics of the proof parallel those of [6] or [10]. However, the analysis of [6] or [10] must be considerably modified to lower bound the transition probabilities in the way described above.

The first discrepancy is such: in the fully synchronized version, where $T_n^i=T_n^j$ for all players $i,j$, we have that the active phase are trivial, i.e. $\tau_k^{\min}=\tau_k^{\max}$ for each $k\geq 0$. As such, there is no need to refer to baseline policies $\boldsymbol{\phi}_t$ at time $t$, as one can analyze policies on the coarser timescale of active phases, indexed by exploration phases rather than stage games. The desirable event, in this case, can be defined as $B_k=\{\boldsymbol{\pi}_k\in\boldsymbol{\Gamma}_{SD}^{0\text{-eq}}\}$ rather than the more convoluted definition of $B_k$ defined in Section IV.

In the un-synchronized variant, one intuitively expects that returning to equilibrium requires that agents learn their equilibrium Q-functions with sufficient accuracy. However, this can only be guaranteed when agents spend a sufficiently long time learning against their equilibrium environment. From this, one sees that defining $B_k=\{\boldsymbol{\phi}_{\tau_k^{\max}}\}$ is ill-suited for our purposes, as it does not account for the event where play arrives at equilibrium immediately before $\tau_k^{\max}$. In the latter event, a particular agent may have spent a long timespan learning against a now outdated environment for which its equilibrium policy may not be a best-response.

To preclude this obstacle, we defined $B_k$ to be the event where the baseline policy $\boldsymbol{\phi}_t$ is fixed at some equilibrium throughout the $k^{th}$ active phase $[\tau_k^{\min},\tau_k^{\max}]$, and we have defined $\tau_k^{\max}$ in way that guarantees each player will learn against $\boldsymbol{\phi}_{\tau_k^{\max}}$ for an appreciable length of time before evaluating and revising its policy. Additionally, we have replaced the decreasing learning rate of [6] with a constant learning $\alpha^i>0$ so that agent $i$ can rapidly correct errors in its Q-factor estimates that contain information about outdated environments.

With these two modifications—of using a constant learning rate and studying the policy process during active phases rather than during exploration phases—the first of two main ideas suggested by [16] goes through: with positive probability, the policy process follows a strict best-response path, along which most agents are inert at a given time while the active agent learns its Q-factors accurately against an unchanging environment. (C.f. Lemma 5.)

The second of the main ideas suggested by [16]—to argue that $\Pr(B_{k+L}|B_k)$ can be made large relative to the lower bound of transiting to equilibrium—requires a more nuanced analysis in the presence of asynchrony. The natural strategy is to argue that, with high probability conditional on $B_k$, every agent accurately learns its Q-factors during $[\tau_{k+l}^{\min},\tau_{k+l}^{\max}]$. Unfortunately, conditioning on the event $B_k$ carries information that agents opted not to switch policies, possibly for a reason other than inertia. In turn, this carries information about the state-action trajectory before $\tau_k^{\max}$ and indeed before $\tau_k^{\min}$, which will influence the ensuing Q-factor estimates and may have a confounding effect on the conditional probabilities.

This point reveals why we have approached the problem as we have, and why Lemma 2 is stated in terms of the hypothetical continuation Q-factors, which depend only on primitive random variables and the history up to the time of conditioning, rather than in terms of the Q-factor estimates, which depend on the sample path of realized states and actions.

## VI. Conclusions

In this paper, we considered an asynchronous variant of the Decentralized Q-learning algorithm of [6]. We have shown that Decentralized Q-learning can still drive policies to equilibrium in weakly acyclic stochastic games without making strong coordination assumptions such as synchronizing the schedules on which players update their policies. To accommodate asynchronous policy updating and non-stationarity in each agent's learning environment, we have introduced a constant learning rate that can rapidly overcome errors in learning estimates that are artifacts of outdated information.

## References


[1] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 750–797, 2019.

[2] 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.

[3] M. L. Littman and C. Szepesvári, "A generalized reinforcement-learning model: Convergence and applications," in *ICML*, vol. 96, pp. 310–318, Citeseer, 1996.

[4] M. L. Littman, "Friend-or-foe Q-learning in general-sum games," in *ICML*, vol. 1, pp. 322–328, 2001.

[5] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *Journal of Machine Learning Research*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[6] G. Arslan and S. Yüksel, "Decentralized Q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2017.

[7] C. Daskalakis, D. J. Foster, and N. Golowich, "Independent policy gradient methods for competitive reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5527–5540, 2020.

[8] M. Sayin, K. Zhang, D. Leslie, T. Başar, and A. Ozdaglar, "Decentralized Q-learning in zero-sum Markov games," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18320–18334, 2021.

[9] 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*, vol. 67, no. 10, pp. 5230–5245, 2022.

[10] B. Yongacoglu, G. Arslan, and S. Yüksel, "Satisficing paths and independent multi-agent reinforcement learning in stochastic games," *arXiv preprint arXiv:2110.04638*, 2022.

[11] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the Tenth Innovative Applications of Artificial Intelligence Conference, Madison, Wisconsin*, pp. 746–752, 1998.

[12] 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*, vol. 27, no. 1, pp. 1–31, 2012.

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

[14] D. Foster and H. P. Young, "Regret testing: Learning to play Nash equilibrium without knowing you have an opponent," *Theoretical Economics*, vol. 1, pp. 341–367, 2006.

[15] F. Germano and G. Lugosi, "Global Nash convergence of Foster and Young's regret testing," *Games and Economic Behavior*, vol. 60, no. 1, pp. 135–154, 2007.

[16] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 373–396, 2009.

[17] A. M. Fink, "Equilibrium in a stochastic $n$-person game," *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, vol. 28, no. 1, pp. 89–93, 1964.

[18] H. P. Young, *Strategic Learning and its Limits*. Oxford University Press., 2004.

[19] D. H. Mguni, Y. Wu, Y. Du, Y. Yang, Z. Wang, M. Li, Y. Wen, J. Jennings, and J. Wang, "Learning in nonzero-sum stochastic games with potentials," in *International Conference on Machine Learning*, pp. 7688–7699, PMLR, 2021.

[20] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras, "Global convergence of multi-agent policy gradient in Markov potential games," *arXiv preprint arXiv:2106.01969*, 2021.

[21] R. Zhang, Z. Ren, and N. Li, "Gradient play in multi-agent Markov stochastic games: Stationary points and convergence," *arXiv preprint arXiv:2106.00198*, 2021.

[22] B. Yongacoglu, G. Arslan, and S. Yüksel, "Asynchronous decentralized Q-learning in stochastic games," *Preprint online*, https://yongac.github.io/files/asynchronous.pdf, 2023.

[23] C. L. Beck and R. Srikant, "Error bounds for constant step-size Q-learning," *Systems & Control Letters*, vol. 61, no. 12, pp. 1203–1208, 2012.