

Supplementary Material for Sample Complexity of Partially Observable Decentralized Q-learning for Cooperative Games

Overview

This document provides supplementary material for the paper titled "Adaptive Policy Selection in Multi-Agent Reinforcement Learning." It includes detailed proofs, derivations, and additional results referenced in the main text.

1 Appendix A

In this section, we provide the proof of Proposition 1. We remind the reader of the proposition.

Proposition 1. *Given Assumption ??, if agent i follows the exploration strategy in Section ??, then $P(\lim_{k \rightarrow \infty} \pi_i^k \in \Pi_i^*) \stackrel{a.s.}{=} 1$.*

Proof. We first introduce the concept of the empirical average reward, which serves as a key component in analyzing the convergence properties of the policy selection by the agent.

Specifically, let $\hat{r}_i^k(\pi_i^k) = \frac{1}{T_i^k(\pi_i^k)} \sum_{n=1}^k \sum_{t=2nT+T}^{2nT+2T-1} r_i^t(s_i^t, \pi_i^n, \boldsymbol{\pi}_{-i}^n) \mathbb{I}(\pi_i^n = \pi_i^k)$ be the empirical average reward of policy π_i^k up to episode k , where $T_i^k(\pi_i^k)$ is the number of times π_i^k is selected. At each episode k , the probability that agent i selects a suboptimal policy satisfies

$$P(\pi_i^k \notin \Pi_i^*) \leq \frac{\beta_i^k |\Pi_i \setminus \{\Pi_i^*\}|}{|\Pi_i|} + (1 - \beta_i^k) P(\hat{r}_i^k(\pi_i^k) \geq \bar{r}_i^k(\pi_i^*)), \quad (1.1)$$

where β_i^k is the exploration rate in (??), $|\Pi_i \setminus \{\Pi_i^*\}|$ is the number of suboptimal policies, and $\bar{r}_i(\pi_i)$ is the long-term average reward of π_i^k under the limiting distribution of $\boldsymbol{\pi}_{-i}^k$, defined as $\bar{r}_i(\pi_i) = \mathbb{E}_{\boldsymbol{\pi}_{-i} \sim \boldsymbol{\pi}_{-i}^\infty} \left[\frac{1}{T} \sum_{t=1}^T r_i^t(\pi_i, \boldsymbol{\pi}_{-i}) \right]$. The inequality is due to $P(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^*))$ including all cases where π_i^k appears better than π_i^* , even if it is suboptimal. We now present the following lemma, which establishes the convergence of the empirical reward.

We recall the following fundamental result

Theorem 1 ([?]). *If X_n is a martingale with $\sup_n \mathbb{E}[|X_n|^p] < \infty$ where $p > 1$, then $X_n \xrightarrow{a.s.} X$ and in L^p , where $X = \limsup_n X_n$.*

Lemma 1. Under weak stabilization of π_{-i}^k , for all $\pi_i^k \in \Pi_i$, $\lim_{k \rightarrow \infty} \hat{r}_i^k(\pi_i^k) \stackrel{a.s.}{=} \bar{r}_i(\pi_i^k)$.

Proof. We aim to show that $\hat{r}_i^k(\pi_i^k)$, the empirical average reward, converges almost surely (a.s.) to $\bar{r}_i(\pi_i^k)$, the long-term average reward.

First, the weak stabilization of π_{-i}^k (i.e., $\pi_{-i}^k \xrightarrow{d} \pi_{-i}^\infty$ as $k \rightarrow \infty$) follows because the policy space of the other agents, Π_{-i} , is finite. Since $\{\pi_{-i}^k\}$ lies in a compact subset of \mathbb{R}^{N-1} , it has a convergent subsequence by the Bolzano-Weierstrass Theorem [?]. Under Assumption ??, the joint policy process (π_i^k, π_{-i}^k) induces an irreducible and aperiodic Markov chain with a unique stationary distribution. By the properties of finite-state Markov chains and weak stabilization, the marginal distribution of π_{-i}^k converges weakly to the stationary distribution, i.e., $\pi_{-i}^k \xrightarrow{d} \pi_{-i}^\infty$. Next, under weak stabilization, the time-average reward $\hat{r}_i^k(\pi_i^k)$ converges to the ensemble average $\bar{r}_i(\pi_i^k)$, which is defined with respect to the limiting distribution of π_{-i}^∞ , as $\bar{r}_i(\pi_i^k) = \mathbb{E}_{\pi_{-i} \sim \pi_{-i}^\infty} [r_i(s, \pi_i^k, \pi_{-i})]$. Finally, the sequence $M_k = \hat{r}_i^k(\pi_i^k) - \bar{r}_i(\pi_i^k)$ forms a martingale with respect to the history \mathcal{H}_k , as $\mathbb{E}[M_{k+1} \mid \mathcal{H}_k] = M_k$ and its increments are bounded. By Assumption 1, each reward is bounded as $r_i^t \leq r_{i,\max}$, which ensures that the cumulative reward within an episode satisfies $r_i^k \leq T r_{i,\max}$, where T is the duration of an episode. Consequently, the magnitude of the martingale satisfies $|M_k| \leq T(r_{i,\max} - r_{i,\min})$, which guarantees that M_k is square-integrable since $\mathbb{E}[M_k^2] < \infty$. By the Martingale Convergence Theorem, $M_k \xrightarrow{a.s.} 0$, ensuring that $\hat{r}_i^k(\pi_i^k) \xrightarrow{a.s.} \bar{r}_i(\pi_i^k)$, completing the proof. \square

We now examine the behavior of policy switches. The exploration rate $\beta_i^k \sim \frac{1}{j}$, where j is the number of consecutive plays of the current policy, implies that the expected number of exploration-driven switches up to episode k is $S_k^{\text{explore}} \sim \sum_{j=1}^k \beta_i^j \sim \sum_{j=1}^k \frac{1}{j} \sim \log k$, yielding $S_k^{\text{explore}} = O(\log k)$. During exploitation $(1 - \beta_i^k)$, policy switches occur when the empirical average reward suggests switching from the current policy to another one. By Lemma 1, the empirical rewards $\hat{r}_i^k(\pi_i^k)$ converge almost surely to the long-term average rewards $\bar{r}_i(\pi_i^k)$, and the convergence rate depends on the number of times a policy is played, $T_i^k(\pi_i^k)$. For a policy π_i^k played $T_i^k(\pi_i^k)$ times, the probability of incorrectly switching due to empirical reward errors decreases exponentially, as $P(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^*)) \leq e^{-\theta T_i^k(\pi_i^k)}$, where $\theta > 0$ is a constant. Since the total number of plays for a policy scales as $T_i^k(\pi_i^k) \sim \frac{k}{|\Pi_i|}$ for uniformly explored policies, the likelihood of exploitation-driven switches decays rapidly. This ensures that the number of exploitation-driven switches satisfies $S_k^{\text{exploit}} = O(\log k)$. Since $S_k^{\text{explore}} = O(\log k)$ and $S_k^{\text{exploit}} = O(\log k)$, the total number of policy switches satisfies $S_k = S_k^{\text{explore}} + S_k^{\text{exploit}} = O(\log k)$. This result shows that S_k is sublinear, as $\frac{S_k}{k} \rightarrow 0$ as $k \rightarrow \infty$.

To formalize the relationship between the sublinear growth of policy switches and the decay of the exploration rate, we present the following lemma.

Lemma 2. Suppose that the number of policy switches S_k grows sublinearly such that $S_k = O(k)$, then $\lim_{k \rightarrow \infty} \beta_i^k \stackrel{a.s.}{=} 0$

Proof. If $\max\{j \geq 1 : \pi_i^{1,k-j+1} = \pi_i^{1,k}\} \rightarrow \infty$, $\beta_i^k \rightarrow 0$. If $\max\{j \geq 1 : \pi_i^{1,k-j+1} = \pi_i^{1,k}\}$ is reset infinitely often, $S_k \rightarrow \infty$. Sublinear growth $S_k = o(k)$ implies $m_k \geq k/S_k \rightarrow \infty$, so $\beta_i^k \rightarrow 0$. \square

Now, we are ready to finalize the proof of Proposition 1. Using the starting inequality (1.1). The first term, $\frac{\beta_i^k |\Pi_i \setminus \{\Pi_i^*\}|}{|\Pi_i|}$, represents the probability of selecting a suboptimal policy during exploration. By Lemma 1, $\beta_i^k \rightarrow 0$, so:

$$\lim_{k \rightarrow \infty} \frac{\beta_i^k |\Pi_i \setminus \{\Pi_i^*\}|}{|\Pi_i|} = 0.$$

The second term, $(1 - \beta_i^k) P(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^*))$, represents the probability of selecting a suboptimal policy during exploitation. If $\pi_i^k \notin \Pi_i^*$, Lemma 2 ensures

$$P(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^*)) \leq e^{-\theta T_i^k(\pi_i^*)},$$

for some $\theta > 0$. This bound relies on the boundedness of rewards (Assumption ??) and the convergence of empirical rewards (Lemma 2). Since $T_i^k(\pi_i^*) \geq \frac{k}{|\Pi_i|}$ (uniform exploration lower bound), this term is $O(\frac{1}{k})$.

Combining the bounds, we get

$$P(\pi_i^k \notin \Pi_i^*) \leq \frac{\beta_i^k |\Pi_i \setminus \{\Pi_i^*\}|}{|\Pi_i|} + (1 - \beta_i^k) e^{-\theta T_i^k(\pi_i^*)}.$$

As $k \rightarrow \infty$, $\beta_i^k \rightarrow 0$ by Lemma 2, and $e^{-\theta T_i^k(\pi_i^*)} \rightarrow 0$ since $T_i^k(\pi_i^*) \rightarrow \infty$. Thus, $P(\lim_{k \rightarrow \infty} \pi_i^k \in \Pi_i^*) = 1$. □

2 Appendix B

We restate Theorem 1 for clarity,

Theorem 2. *Consider a partially observable discounted cooperative SG. Let Assumptions ?? and ?? hold. Suppose that each agent updates its policy using Algorithm ??. Then:*

(i) *For any $0 < \delta < 1$, $|Q_i^t - Q_i^*| < \epsilon$ holds with probability exceeding $1 - \delta$ for all*

$$t > t_{mix} + \frac{C}{\epsilon^2} \log \left(\frac{|\mathcal{N}| |\mathcal{S}_i| |\mathcal{A}_i|}{\delta - \exp \left(-\frac{(kT - t_{mix})(\Delta_{\rho,i})^2}{2\sigma_2^2} \right)} \right), \quad (2.1)$$

where $C = \max \left(\frac{443}{\mu_{\min}}, \frac{2\sigma_1^2}{\mu_{\min}}, \frac{(\gamma V_i^* \eta_{i,\max})^2}{2}, \frac{2\sigma_2^2}{\mu_{\min}(1 - (1 - \eta_i^t)^T)^2} \right)$, and $\Delta_{\rho,i} = \rho^* - \rho^{2,k}$. Here, σ_1^2 is the variance of the reward deviation caused by non-stationarity, satisfying $\sigma_1^2 \leq (C^t)^2 \eta_{i,\max}^2 (r_{\max} - r_{\min})^2$, and σ_2^2 is the variance of the feedback error, bounded by $\sigma_2^2 \leq \left(\frac{C^t \eta_{i,\max}}{2} \right)^2$.

(ii) *Additionally, $\Pr(\lim_{t \rightarrow \infty} Q_i^t = Q_i^*) \stackrel{a.s.}{=} 1$.*

Proof. In order to define the Multi-Agent decomposition error term we first define the error term for each agent i as the difference between the Q-function at time t and the optimal Q-function,

$$\begin{aligned}
\Delta_i^t &:= Q_i^t - Q_i^* \\
&= (1 - \eta_i^t)Q_i^{t-1} + \eta_i^t (r_i + \gamma P_i^t V_i^{t-1}) - Q_i^* \\
&= (1 - \eta_i^t)(Q_i^{t-1} - Q_i^*) + \eta_i^t (r_i + \gamma P_i^t V_i^{t-1} - Q_i^*) \\
&\stackrel{(*)}{=} (1 - \eta_i^t)\Delta_i^{t-1} + \eta_i^t \left(\xi_r^t + \beta^k (2\mathbb{I}(t) - 1) + \right. \\
&\quad \left. \gamma (P_i^t V_i^{t-1} - P V^*) \right) \\
&= (1 - \eta_i^t)\Delta_i^{t-1} + \eta_i^t \xi_r^t + \eta_i^t \beta (2\mathbb{I}(t) - 2) + \eta_i^t \gamma (P_i^t - P) V^* \\
&\quad + \eta_i^t \gamma P_i^t (V_i^{t-1} - V^*),
\end{aligned} \tag{2.2}$$

where $(*)$ is due to the fact that the optimal policy is consistent during the entire episode $\pi_i^{1,*} = \pi_i^{2,*}$ and $\xi_r^t = r_i(s_i^t, a_i^t, \mathbf{a}_{-i}^t) - r_i^*(s_i^t, a_i^t, \mathbf{a}_{-i}^t)$, where $r_i^*(s_i^t, \pi_i^*, \boldsymbol{\pi}_{-i}^*)$ is an optimal reward given a joint optimal policy $\boldsymbol{\pi}^*$.

Next, by applying the recursion iteratively by expressing Δ_i^{t-1} in terms of Δ_i^{t-2} , and continuing until we reach Δ_i^0 , we obtain:

$$\begin{aligned}
\Delta_i^t &= \underbrace{\prod_{j=1}^t (1 - \eta_i^j) \Delta_i^0}_{e_0^t} + \underbrace{\sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l \xi_r^l}_{e_1^t} + \underbrace{\sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l \gamma (P_i^l - P) V^*}_{e_2^t} + \\
&\quad \underbrace{\sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l \beta^k (2\mathbb{I}(l) - 2)}_{e_3^t} + \underbrace{\sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l \gamma P_i^l (V_i^{l-1} - V^*)}_{e_4^t}
\end{aligned} \tag{2.3}$$

By the triangle inequality, we have

$$|\Delta_i^t| \leq |e_0^t| + |e_1^t| + |e_2^t| + |e_3^t| + |e_4^t| \tag{2.4}$$

To bound $|\Delta_i^t|$, we analyze each error term $e_0^t, e_1^t, e_2^t, e_3^t$, and e_4^t individually. We start with e_0^t following a similar proof technique as in [?].

Lemma 3. *For any $\delta > 0$, suppose $t > \frac{443t_{mix}}{\mu_{min}} \log \frac{4|\mathcal{S}||\mathcal{A}||\mathcal{N}|}{\delta}$. Then with probability greater than $1 - \delta$ one has*

$$|e_0^t| \leq (1 - \eta_i^t)^{\frac{1}{2}t\mu_{min}} |\Delta_i^0| \tag{2.5}$$

Proof. From the definition of e_0^t and the Q-learning update rule in (??), one can easily see that,

$$\begin{aligned}
\left| \prod_{j=1}^t (1 - \eta_i^j) \Delta_i^0 \right| &= \prod_{j=1}^{C^t(s_i, a_i)} (1 - \eta_i^j) |\Delta_i^0| \\
&\leq (1 - \eta_{i,min}^t)^{C^t(s_i, a_i)} |\Delta_i^0|,
\end{aligned} \tag{2.6}$$

Now using Lemma 8 in [?], and applying union bound over the state space \mathcal{S}_i , the action space \mathcal{A}_i and the set of agents \mathcal{N} , one has, with probability greater than $1 - \delta$, that,

$$C^t(s_i, a_i) \geq t\mu_{min}/2 \tag{2.7}$$

Using the fact that, any aperiodic and irreducible Markov chain on a finite state space is uniformly ergodic [?]. Thus, (2.7) holds uniformly over all (s_i, a_i) and all agents $i \in \mathcal{N}$ and all $\frac{443t_{mix}}{\mu_{min}} \log \frac{4|\mathcal{S}||\mathcal{A}||\mathcal{N}|}{\delta} \leq t \leq T$, where μ_{min} is the stationary distribution of the Markov chain (s_i^0, s_i^1, \dots) . Then, we have, $|e_0^t| \leq (1 - \eta_{i,min})^{\frac{t\mu_{min}}{2}} |\Delta_i^0|$. For learning rates $\eta_i^j \in (0, 1)$ with $\eta_{i,max} = \max_j \eta_i^j$, and $\eta_{i,min} = \min_j \eta_i^j$, the following inequality holds, $\sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l \leq C^t \eta_{i,max}$. To derive this result, we first bound the product term as follows $\prod_{j=l+1}^t (1 - \eta_i^j) \leq (1 - \eta_{i,min})^{t-l}$. Substituting this bound into the summation and utilizing the definition of C^t , we obtain

$$\begin{aligned} \sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l &\leq \eta_i^l \frac{1 - (1 - \eta_{i,min})^{C^t}}{\eta_{i,min}} \\ &\stackrel{(*)}{\leq} \eta_{i,max} \cdot \frac{C^t \eta_{i,min}}{\eta_{i,min}} \\ &= C^t \eta_{i,max}. \end{aligned} \tag{2.8}$$

Here, $(*)$ follows from Bernoulli's inequality, $(1 - \eta_{i,min})^{C^t} \geq 1 - C^t \eta_{i,min}$, which yields $1 - (1 - \eta_{i,min})^{C^t} \leq C^t \eta_{i,min}$.

Next, we analyze the deviation between the observed reward and the optimal reward under the joint optimal policy:

$$\xi_r^k = r_i(s_i^t, a_i^t, \mathbf{a}_{-i}^t) - r_i^*(s_i^t, a_i^t, \mathbf{a}_{-i}^t), \tag{2.9}$$

where $r_i^*(s_i^t, a_i^t, \mathbf{a}_{-i}^t)$ represents the optimal reward under the joint optimal policy.

Lemma 4. *Let $\delta > 0$ and $\epsilon > 0$ be the confidence and error thresholds, respectively. Then, for any $\delta > 0$ and $\epsilon > 0$, one has $|e_1^t| \leq \epsilon$ with probability at least $1 - \delta$, $\forall i \in \mathcal{N}$, $s_i \in \mathcal{S}_i$, $a_i \in \mathcal{A}_i$, and $t \geq t_{mix} + \frac{2\sigma_1^2}{\mu_{min}\epsilon^2} \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)$, where $\sigma_1^2 = \text{Var}[e_1^t] \leq C^t \eta_{i,max} (r_{max} - r_{min})^2$.*

Proof. Observe that the error term ξ_r^t is bounded, $|\xi_r^t| \leq r_{max} - r_{min}$, therefore e_1^t has a well-defined variance $\text{Var}[e_1^t] = \sigma^2 \leq C^{t2} \eta_{i,max}^2 (r_{max} - r_{min})^2$. Furthermore, $\{e_1^t\}$ forms a martingale difference sequence with $\mathbb{E}[e_1^t | \mathcal{H}_i^{k-1}] = 0$, satisfying the conditions for Bernstein's inequality,

$$P(|e_1^t| > \epsilon) \leq 2 \exp \left(\frac{-C^t \epsilon^2}{2\sigma^2 + \frac{2}{3}(r_{max} - r_{min})\epsilon} \right). \tag{2.10}$$

We extend the probability bound over all agents and state-action pairs using the union bound:

$$P \left(\max_{i, s_i, a_i} |e_1^t| > \epsilon \right) \leq 2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i| \exp \left(\frac{-C^t \epsilon^2}{2\sigma^2 + \frac{2}{3}(r_{max} - r_{min})\epsilon} \right). \tag{2.11}$$

Solving for t , $C^t \approx \mu_{min}(t - t_{mix})$, where, $\mu_{min} = \min_{s_i, a_i} \pi(s_i, a_i)$.

We require:

$$2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i| \exp \left(-\frac{\mu_{min}(t - t_{mix})\epsilon^2}{2\sigma^2 + \frac{2}{3}(r_{max} - r_{min})\epsilon} \right) \leq \delta. \tag{2.12}$$

Solving for t , we obtain:

$$t \geq t_{mix} + \frac{1}{\mu_{\min}} \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right) \times \max \left\{ \frac{2\sigma^2}{\epsilon^2}, \frac{3(r_{\max} - r_{\min})}{\epsilon} \right\}. \quad (2.13)$$

Thus, for any t satisfying the lower bound in (2.13), the error satisfies, $|e_1^t(s_i, a_i)| \leq \epsilon$, w.p. at least $1 - \delta$ for all i , s_i , and a_i . \square

Next, we want to analyse e_2^t . Given $V^*(s_i^t)$ for any $s_i^t \in \mathcal{S}_i$, there exist a constant $c \in [0, 1]$ such that

$$\sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i) \right) \eta_i \gamma (P_i^l - P) V^* \leq c\gamma \sqrt{\eta \log \left(\frac{|\mathcal{S}||\mathcal{A}|N}{\delta} \right)}$$

w.p. at least $1 - \delta$.

We can easily proof this inequality, first notice that from the update rule in (??), we can write e_2^t , as

$$e_2^t = \sum_{l=1}^{C^t} (1 - \eta_i)^{C^t-l} \eta_i \gamma (P_i^{t_l} - P) V^* \quad (2.14)$$

Given (s_i^t, a_i^t) , set,

$$P \left(\left| \sum_{l=1}^{C^t} (1 - \eta_i)^{C^t-l} \eta_i \gamma (P_i^{t_l} - P) V^* \right| \geq \epsilon \right) \leq \delta \quad (2.15)$$

By applying union bounds to any $(s_i^t, a_i^t) \in \mathcal{S}_i \times \mathcal{A}_i, \forall i \in \mathcal{N}$, we get,

$$P \left(\left| \sum_{l=1}^{C^t} (1 - \eta_i)^{C^t-l} \eta_i \gamma (P_i^{t_l} - P) V^* \right| \geq \epsilon \right) \leq \frac{\delta}{|\mathcal{S}_i||\mathcal{A}_i|N} \quad (2.16)$$

Using the Markov property and lemma 2 in [?], the state action pair (s_i^t, a_i^t) is independent for all t . Thus, by applying Hoeffding inequality, we have,

$$P \left(\left| \sum_{l=1}^{C^t} (1 - \eta_i)^{C^t-l} \eta_i \gamma (P_i^{t_l} - P) V^* \right| \geq \epsilon \right) \leq \exp \left(-\frac{\epsilon^2}{\sigma^2} \right) \quad (2.17)$$

where, $\sigma^2 \leq \left(\sum_{l=1}^{C^t} (1 - \eta_i)^{C^t-l} \eta_i \gamma V^* \right)^2$.

Setting, $\exp \left(-\frac{\epsilon^2}{\sigma^2} \right) = \frac{\delta}{|\mathcal{S}_i||\mathcal{A}_i|N}$ and solving for ϵ we get, $\epsilon = \sqrt{\frac{\sigma}{2} \log \left(\frac{|\mathcal{S}||\mathcal{A}|N}{\delta} \right)}$

Replacing t in (2.17) and taking the complement, we get,

$$P \left(e_2^t \leq c\gamma \sqrt{\log \left(\frac{|\mathcal{S}||\mathcal{A}|N}{\delta} \right)} \right) \geq 1 - \frac{\delta}{|\mathcal{S}_i||\mathcal{A}_i|N} \quad (2.18)$$

\square

where, $c = \frac{1-(1-\eta_i)^{2C^t}}{1-(1-\eta_i)^2} \eta^2$.

Then for a tight bound we have the following result.

Lemma 5. *For any $\delta > 0$, with probability at least $1 - \delta$, $|e_2^t| \leq \epsilon$, for all*

$$t \geq t_{mix} + \frac{(\gamma V^* \eta_{\max})^2 \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)}{2\epsilon^2}. \quad (2.19)$$

Proof. Let the transition probability error be $\xi_{P,i}^l = P_i^l - P$, which can be decomposed into marginal estimation error ξ_M and transition estimation error ξ_T conditioned on the marginals. The total error satisfies $\|\xi_{P,i}^l\|_1 \leq \|\xi_M\|_1 + \|\xi_T\|_1$.

Marginal Estimation Error (ξ_M): Applying concentration bounds for marginal estimation error ξ_M :

$$P(\|\xi_M\|_1 > \epsilon) \leq 2|\mathcal{A}_{-i}||\mathcal{S}_{-i}| \exp(-2C^t \epsilon^2). \quad (2.20)$$

Transition Estimation Error (ξ_T): Similarly, applying concentration bounds for transition estimation error ξ_T conditioned on the marginals:

$$P(\|\xi_T\|_1 > \epsilon) \leq 2|\mathcal{S}| \exp(-2C^t \epsilon^2). \quad (2.21)$$

Using the union bound, the total transition probability error satisfies:

$$P(\|\xi_{P,i}^l\|_1 > \epsilon) \leq 2(|\mathcal{A}_{-i}||\mathcal{S}_{-i}| + |\mathcal{S}|) \exp(-2C^t \epsilon^2). \quad (2.22)$$

Generalizing Across Agents, States, and Actions: Using the union bound again to generalize over all agents, states, and actions:

$$P\left(\max_{i,s_i,a_i} \|\xi_{P,i}^l\|_1 > \epsilon\right) \leq 2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i| \exp(-2C^t \epsilon^2). \quad (2.23)$$

Bounding $|e_2^t|$: From the above, substituting the bound into e_2^t , we have:

$$|e_2^t| \leq \gamma V^* \sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l \epsilon, \quad (2.24)$$

where $\epsilon = \sqrt{\frac{1}{2C^t} \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)}$.

Using the learning rate property:

$$\sum_{l=1}^t \left(\prod_{j=l+1}^t (1 - \eta_i^j) \right) \eta_i^l \leq \eta_{\max} C^t, \quad (2.25)$$

we get:

$$|e_2^t| \leq \gamma V^* \eta_{\max} C^t \epsilon. \quad (2.26)$$

Substituting ϵ :

$$|e_2^t| \leq \gamma V^* \eta_{\max} \sqrt{C^t \cdot \frac{1}{2} \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)}. \quad (2.27)$$

Solving for t : To ensure $|e_2^t| \leq \epsilon$, set:

$$\gamma V^* \eta_{\max} \sqrt{C^t \cdot \frac{1}{2} \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)} \leq \epsilon. \quad (2.28)$$

Square both sides:

$$(\gamma V^* \eta_{\max})^2 C^t \cdot \frac{1}{2} \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right) \leq \epsilon^2. \quad (2.29)$$

Solve for C^t :

$$C^t \leq \frac{2\epsilon^2}{(\gamma V^* \eta_{\max})^2 \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)}. \quad (2.30)$$

Finally, using $C^t = \frac{1}{t}$, we get:

$$t \geq \frac{(\gamma V^* \eta_{\max})^2 \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)}{2\epsilon^2}. \quad (2.31)$$

Thus, for any $t > t_0$, where:

$$t_0 = t_{\text{mix}} + \frac{(\gamma V^* \eta_{\max})^2 \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right)}{2\epsilon^2}, \quad (2.32)$$

we have $|e_2^t| \leq \epsilon$, w.p. at least $1 - \delta$. \square

Lemma 6 (Probabilistic Bound for e_3^t). *Let $\delta > 0$ and $\epsilon > 0$ be the confidence and error thresholds, respectively. Then, for any episode index $k > 0$, the error term e_3^t satisfies:*

$$P(|e_3^t| \leq \epsilon) \geq 1 - \delta, \quad (2.33)$$

for all

$$t \geq t_{\text{mix}} + \frac{1}{\mu_{\min}(1 - (1 - \eta_i)^T)^2} \cdot \frac{2\sigma^2}{\epsilon^2} \log \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right). \quad (2.34)$$

Proof. The key challenge is bounding the deviations of the exploration term e_3^t while accounting for the retention of policies due to global feedback ϕ^k .

Step 1: Global Feedback Failure Probability As noted in Remark ??, even when the optimal policy is selected (guaranteed by Proposition 1), retaining it depends on the global performance metric ρ . The global reward gap is defined as:

$$\Delta_{\rho,i} = \rho^* - \rho(\pi_i, \boldsymbol{\pi}_{-i}), \quad (2.35)$$

where ρ^* is the optimal global metric. Using Hoeffding's inequality, the probability of failing to retain the policy due to $\phi^k = 0$ satisfies:

$$P(\phi^k = 0) \leq \exp \left(-\frac{(kT - t_{\text{mix}})(\Delta_{\rho,i})^2}{2\sigma^2} \right). \quad (2.36)$$

This defines the joint failure probability $\delta_{joint}(k)$ as:

$$\delta_{joint}(k) = \exp \left(-\frac{(kT - t_{mix})(\Delta_{\rho,i})^2}{2\sigma^2} \right). \quad (2.37)$$

Step 2: Union Bound Across All Agents, States, and Actions To ensure the probabilistic bound holds uniformly across all agents, states, and actions, we apply a union bound. For any specific state-action pair, the probability of error satisfies:

$$P \left(\exists(i, s_i, a_i) : |e_3^t| > \epsilon \right) \leq \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \quad (2.38)$$

Combining the probabilities, the total failure probability is bounded by:

$$P \left(|e_3^t| > \epsilon \right) \leq \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \quad (2.39)$$

Taking the complement, the probability of satisfying $|e_3^t| \leq \epsilon$ is:

$$P \left(|e_3^t| \leq \epsilon \right) \geq 1 - \delta_{total}, \quad (2.40)$$

where

$$\delta_{total} = \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \quad (2.41)$$

Step 3: Relating $\delta_{joint}(k)$ to ϵ From Hoeffding's inequality, we have:

$$\delta_{joint}(k) = \exp \left(-\frac{(kT - t_{mix})(\Delta_{\rho,i})^2}{2\sigma^2} \right). \quad (2.42)$$

Assume the reward gap $\Delta_{\rho,i}$ is proportional to the error threshold ϵ , i.e., $\Delta_{\rho,i} \geq \epsilon$. Substituting this into $\delta_{joint}(k)$:

$$\delta_{joint}(k) = \exp \left(-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2} \right). \quad (2.43)$$

For the total failure probability to satisfy $\delta_{total} \leq \delta$, we need:

$$\exp \left(-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2} \right) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|} \leq \delta. \quad (2.44)$$

Neglecting the second term (for simplicity in scaling), we require:

$$\exp \left(-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2} \right) \leq \delta. \quad (2.45)$$

Taking the natural logarithm:

$$-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2} \leq \log(\delta). \quad (2.46)$$

Rearranging for $kT - t_{mix}$:

$$kT - t_{mix} \geq \frac{2\sigma^2}{\epsilon^2} |\log(\delta)|. \quad (2.47)$$

Step 4: Incorporating Learning Retention Factor The retention factor $(1 - (1 - \eta_i)^T)^2$ influences the effective bound on t . To account for this:

$$(1 - (1 - \eta_i)^T)^2 t \geq t_{mix} + \frac{2\sigma^2}{\epsilon^2} |\log(\delta)|. \quad (2.48)$$

Rearranging:

$$t \geq t_{mix} + \frac{1}{\mu_{\min}(1 - (1 - \eta_i)^T)^2} \cdot \frac{2\sigma^2}{\epsilon^2} |\log(\delta)|. \quad (2.49)$$

Conclusion Thus, for any $t \geq t_0$, where:

$$t_0 = t_{mix} + \frac{1}{\mu_{\min}(1 - (1 - \eta_i)^T)^2} \cdot \frac{2\sigma^2}{\epsilon^2} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right), \quad (2.50)$$

we have $|e_3^t| \leq \epsilon$ w.p. at least $1 - \delta$. \square

For e_4^t one can easily notice that,

$$|P_i^l (V_i^{l-1} - V^*)| \leq |V_i^{l-1} - V^*| \leq |Q_i^{l-1} - Q^*| = |\Delta_i^{l-1}| \quad (2.51)$$

Thus we can write,

$$|\Delta_i^t| \leq \left[(1 - \eta)^{\frac{1}{2}t\mu_{\min}} + C^t \eta_{i,\max} \gamma \right] |\Delta_i^0| + |e_1^t| + |e_2^t| + |e_3^t|. \quad (2.52)$$

To ensure $|\Delta_i^t| \leq \epsilon$, we bound each term in the inequality:

$$|\Delta_i^t| \leq \left[(1 - \eta)^{\frac{1}{2}t\mu_{\min}} + C^t \eta_{i,\max} \gamma \right] |\Delta_i^0| + |e_1^t| + |e_2^t| + |e_3^t|,$$

by $\frac{\epsilon}{4}$:

1. **Bounding e_0^t :** The term $e_0^t = \prod_{j=1}^t (1 - \eta_i^j) \Delta_i^0$ decays as $(1 - \eta)^{\frac{1}{2}t\mu_{\min}}$. For $e_0^t \leq \frac{\epsilon}{4}$, we require:

$$t \geq \frac{2}{\mu_{\min}} \log\left(\frac{4|\Delta_i^0|}{\epsilon}\right).$$

2. **Bounding e_1^t :** From Lemma 1, $|e_1^t| \leq \frac{\epsilon}{4}$ w.p. $1 - \frac{\delta}{3}$ for:

$$t \geq t_{mix} + \frac{2\sigma^2}{\mu_{\min}\epsilon^2} \log\left(\frac{6|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right).$$

3. **Bounding e_2^t :** From Lemma 2, $|e_2^t| \leq \frac{\epsilon}{4}$ w.p. $1 - \frac{\delta}{3}$ for:

$$t \geq t_{mix} + \frac{1}{\mu_{\min}} \frac{1}{(1 - (1 - \eta_i)^{C^t})^2} \log\left(\frac{6|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right).$$

4. **Bounding e_3^t :** From Lemma 3, $|e_3^t| \leq \frac{\epsilon}{4}$ w.p. $1 - \frac{\delta}{3}$ for:

$$t \geq t_{mix} + \frac{1}{\mu_{\min}} \frac{1}{(1 - (1 - \eta_i)^T)^2} \log \left(\frac{6|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right).$$

5. **Combining the Bounds:** Using the union bound, the total failure probability is δ . The overall sample complexity T_k is determined by the slowest-decaying term among e_0^t , e_1^t , e_2^t , and e_3^t . Thus:

$$T_k = \left(\frac{T^2(\Delta_{\rho,i})^2}{2\sigma^2} \log \left(\frac{3}{\delta} \right) + \frac{T^2}{t_{mix}} \right) \cdot \max \left\{ \frac{32\sigma^2 \log(6/\delta)}{\mu_{\min}\epsilon^2}, \frac{16\gamma^2(V^*)^2 \log(6|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|/\delta)}{\mu_{\min}\epsilon^2(1 - (1 - \eta_i)^T)^2} \right\}.$$

□

Lemma 7 (Probabilistic Bound for e_3^t). *Let $\delta > 0$ and $\epsilon > 0$ be the confidence and error thresholds, respectively. Then, for any episode index $k > 0$, the error term e_3^t satisfies:*

$$P(|e_3^t| \leq \epsilon) \geq 1 - \delta, \quad (2.53)$$

for all

$$t \geq t_{mix} + \frac{1}{2\mu_{\min}\epsilon^2} \ln \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta_{total}} \right), \quad (2.54)$$

where δ_{total} is given by:

$$\delta_{total} = \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}, \quad (2.55)$$

and $\delta_{joint}(k)$ represents the probability of failure due to global feedback, defined as:

$$\delta_{joint}(k) = \exp \left(-\frac{(kT - t_{mix})(\Delta_{\rho,i})^2}{2\sigma^2} \right). \quad (2.56)$$

—

Proof

The key steps in the proof are to correctly account for the failure probabilities due to: 1. The local exploration term. 2. The global feedback mechanism.

We need to bound the total failure probability $P(|e_3^t| > \epsilon)$ by considering both sources of deviation.

Step 1: Bounding the Local Failure Probability Using Hoeffding's inequality, the failure probability for any specific agent i , state s_i , and action a_i is:

$$P(|e_3^t| > \epsilon) \leq 2 \exp(-2\mu_{\min}t\epsilon^2).$$

By applying the union bound over all agents, states, and actions, we obtain:

$$P(\exists(i, s_i, a_i) : |e_3^t| > \epsilon) \leq \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}.$$

Step 2: Bounding the Global Failure Probability The global failure due to feedback (captured by ϕ^k) depends on the global reward gap $\Delta_{\rho,i}$:

$$\delta_{joint}(k) = \exp \left(-\frac{(kT - t_{mix})(\Delta_{\rho,i})^2}{2\sigma^2} \right).$$

Thus, the total failure probability becomes:

$$P(|e_3^t| > \epsilon) \leq \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}.$$

Step 3: Complementing the Probability Taking the complement, we get:

$$P(|e_3^t| \leq \epsilon) \geq 1 - \delta_{total},$$

where:

$$\delta_{total} = \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}.$$

Step 4: Time Bound for t For the bound to hold, we solve for t in the expression:

$$2 \exp(-2\mu_{\min} t \epsilon^2) = \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}.$$

Taking the natural logarithm and solving for t , we have:

$$t \geq \frac{1}{2\mu_{\min}\epsilon^2} \ln \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right).$$

Including the influence of $\delta_{joint}(k)$, the overall time bound becomes:

$$t \geq t_{mix} + \frac{1}{2\mu_{\min}\epsilon^2} \ln \left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta_{total}} \right),$$

where δ_{total} includes both local and global failure contributions.