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* $P(\lim_{k\to\infty} \pi_i^k \in \Pi_i^\star) \overset{a.s.}{=} 1$.

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

$$P(\pi_i^k \notin \Pi_i^\star) \leq \frac{\beta_i^k |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|} + \left(1 - \beta_i^k\right) P\left(\hat{r}_i^k(\pi_i^k) \geq \bar{r}_i^k(\pi_i^\star)\right), \tag{1.1}$$

where β_i^k is the exploration rate in (**??**), $|\Pi_i \backslash \{\Pi_i^\star\}|$ is the number of suboptimal policies, and $\bar{r}_i(\pi_i^k)$ is the long-term average reward of π_i^k under the limiting distribution of $\boldsymbol{\pi}_{-i}^k$, defined as

$$\bar{r}_i^k(\pi_i^k) = \mathbb{E}_{\boldsymbol{\pi}_{-i}^k \sim \boldsymbol{\pi}_{-i}^\infty} \left[\frac{1}{T} \sum_{t=1}^{T} r_i^t(\pi_i^k, \boldsymbol{\pi}_{-i}^k) \right]. \tag{1.2}$$

The inequality in (1.1) holds because $P\left(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^\star)\right)$ includes all cases where π_i^k appears better than $\pi_i^\star$, even if it is suboptimal. We now present the following lemma, which establishes the convergence of the empirical reward. First, we recall a fundamental result on martingales: Doob's forward convergence theorem [**?**, p. 109].

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

Lemma 1. *Under weak stabilization of $\boldsymbol{\pi}_{-i}^k$, for all $\pi_i^k \in \Pi_i$, $\lim_{k\to\infty} \hat{r}_i^k(\pi_i^k) \overset{a.s.}{=} \bar{r}_i(\pi_i^k)$.*

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 $\boldsymbol{\pi}_{-i}^k$ (i.e., $\boldsymbol{\pi}_{-i}^k \overset{d}{\to} \boldsymbol{\pi}_{-i}^\infty$ as $k \to \infty$) follows because the policy space of the other agents, Π_{-i}, is finite. Since $\{\boldsymbol{\pi}_{-i}^k\}$ lies in a compact subset of $\mathbb{R}^{N-1}$, it has a convergent subsequence by the Bolzano-Weierstrass Theorem [**?**, p.54]. By Assumption **??**, the joint policy process $(\pi_i^k, \boldsymbol{\pi}_{-i}^k)$ 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 $\boldsymbol{\pi}_{-i}^k$ converges weakly to the stationary distribution, i.e., $\boldsymbol{\pi}_{-i}^k \overset{d}{\to} \boldsymbol{\pi}_{-i}^\infty$. Next, under weak stabilization, the time-average reward $\hat{r}_i^k(\pi_i^k)$ converges to the ensemble average $\bar{r}_i(\pi_i^k)$. To establish almost sure convergence, define the sequence $M^k = \hat{r}_i^k(\pi_i^k) - \bar{r}_i(\pi_i^k)$. This sequence forms a martingale with respect to the history $\mathcal{H}^k$ since $\mathbb{E}[M^{k+1} \mid \mathcal{H}^k] = M^k$. Moreover, its increments are bounded. By Assumption 1, each reward is bounded by $r_i^t \leq r_{i,\max}$, which ensures that the cumulative reward within an episode satisfies $\hat{r}_i^k(\pi_i^k) \leq r_{i,\max}$. Consequently, the magnitude of the martingale satisfies $|M^k| \leq |r_{i,\max} - r_{i,\min}|$, ensuring that M^k is square-integrable since $\mathbb{E}[M^{k2}] < \infty$. By Theorem 1, $M^k \xrightarrow{\text{a.s.}} 0$, ensuring that $\hat{r}_i^k(\pi_i^k) \xrightarrow{\text{a.s.}} \bar{r}_i(\pi_i^k)$, completing the proof. □

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 $S_{\text{explore}}^k \sim \sum_{j=1}^k \beta_i^j \sim \sum_{j=1}^k \frac{1}{j} \sim \log k$, yielding $S_{\text{explore}}^k = O(\log k)$. 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, $\hat{r}_i^k(\pi_i^k)$ converges almost surely to $\bar{r}_i(\pi_i^k)$, and the convergence rate depends on the number of times a policy is played, denoted by $T_i^k(\pi_i^k)$. 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. Specifically, by Hoeffding's inequality, $P(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^\star)) \leq e^{-\theta T_i^k(\pi_i^k)}$, where $\theta > 0$ is a constant. Since the total number of plays for a policy scales as $T_i^k(\pi_i^k) \sim \frac{k}{|\Pi_i|}$ under uniform exploration, the likelihood of exploitation-driven switches decays rapidly. This ensures that the number of exploitation-driven switches satisfies $S_{\text{exploit}}^k = O(\log k)$. Since $S_{\text{explore}}^k = O(\log k)$ and $S_{\text{exploit}}^k = O(\log k)$, the total number of policy switches satisfies $S^k = S_{\text{explore}}^k + S_{\text{exploit}}^k = O(\log k)$. This result shows that S^k is sublinear, as $\frac{S^k}{k} \to 0$ as $k \to \infty$.

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 $S_k = O(\log k)$. Then $\lim_{k\to\infty} \beta_i^k \overset{a.s.}{=} 0$.*

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

Now, we are ready to finalize the proof of Proposition 1.Recall the starting inequality (1.1), the first term, $\frac{\beta_i^k |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|}$, represents the probability of selecting a suboptimal policy during exploration. By Lemma 1, $\beta_i^k \to 0$, thus

$$\lim_{k\to\infty} \frac{\beta_i^k |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|} = 0.$$

The second term, $\left(1 - \beta_i^k\right) P\left(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^\star)\right)$, represents the probability of selecting a suboptimal policy during exploitation. If $\pi_i^k \notin \Pi_i^\star$, then by Lemma 2, there exists a constant $\theta > 0$ such that

$$P\left(\hat{r}_i^k(\pi_i^k) \geq \hat{r}_i^k(\pi_i^\star)\right) \leq e^{-\theta T_i^k(\pi_i^k)}.$$

This exponential bound relies on the boundedness of rewards (Assumption **??**) and the convergence of empirical means (Lemma 2). Furthermore, because of uniform exploration, we have $T_i^k(\pi_i^\star) \geq \frac{k}{|\Pi_i|}$. Hence, $e^{-\theta T_i^k(\pi_i^k)} = O(e^{-\theta \frac{k}{|\Pi_i|}}) = O(\frac{1}{k})$, implying that the second term in (1.1) also vanishes as $k \to \infty$. Putting these two terms together, we have

$$P(\pi_i^k \notin \Pi_i^\star) \leq \frac{\beta_i^k |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|} + \left(1 - \beta_i^k\right) e^{-\theta T_i^k(\pi_i^\star)}.$$

As $k \to \infty$, Lemma 2 ensures $\beta_i^k \to 0$. Furthermore, since $T_i^k(\pi_i^\star) \to \infty$, it follows that $e^{-\theta T_i^k(\pi_i^\star)} \to 0$. Therefore, $\lim_{k\to\infty} P(\pi_i^k \in \Pi_i^\star) = 1$. □

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 $0 < \delta < 1$, $|Q_i^t - Q_i^\star| < \epsilon$ holds with probability exceeding $1 - \delta$ for all*

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

where $C = \max\left(\frac{443}{\mu_{\min}}, \frac{2\sigma_1^2}{\mu_{\min}}, \frac{(\gamma V_i^\star \eta_{i,\max})^2}{2}, \frac{2\sigma_2^2}{\mu_{\min}(1-(1-\eta_i^t)^T)^2}\right)$, and $\Delta_{\rho,i} = \rho^\star - \rho^{2,k}$. 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, $P\left(\lim_{t\to\infty} Q_i^t = Q_i^\star\right) \overset{a.s.}{=} 1$.*

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,

$$
\begin{aligned}
\Delta_i^t &:= Q_i^t - Q_i^\star \\
&= (1-\eta_i^t)Q_i^{t-1} + \eta_i^t\left(r_i + \gamma P^t V_i^{t-1}\right) - Q_i^\star \\
&= (1-\eta_i^t)(Q_i^{t-1} - Q_i^\star) + \eta_i^t\left(r_i + \gamma P^t V_i^{t-1} - Q_i^\star\right) \\
&\overset{(*)}{=} (1-\eta_i^t)\Delta_i^{t-1} + \eta_i^t\Big(\xi_r^t + \beta_i^k(2\mathbb{I}(t)-1) + \\
&\qquad\qquad\qquad\qquad \gamma\left(P^t V_i^{t-1} - PV^\star\right)\Big) \\
&= (1-\eta_i^t)\Delta_i^{t-1} + \eta_i^t\xi_r^t + \eta_i^t\beta(2\mathbb{I}(t)-2) + \eta_i^t\gamma\left(P^t - P\right)V^\star \\
&\qquad\qquad + \eta_i^t\gamma P^t\left(V_i^{t-1} - V^\star\right),
\end{aligned}
\tag{2.2}
$$

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

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:

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

By the triangle inequality, we have

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

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 [**?**].

For any two probability distributions μ and ν, let $d_{\mathrm{TV}}(\mu,\nu)$ denote their total variation distance. Consider a time-homogeneous, uniformly ergodic Markov chain $\{s_i^0, s_i^1, s_i^2, \dots\}$ on a finite state space $\mathcal{S}_i$ with transition kernel P and stationary distribution μ. We denote by $P^t(\cdot \mid s_i^0)$ the distribution of s_i^t given the initial state s_i^0. The mixing time t_{mix} of this Markov chain is defined by

$$
\begin{aligned}
t_{\mathrm{mix}}(\epsilon) &:= \min\left\{ t \;\middle|\; \max_{x\in\mathcal{X}} d_{\mathrm{TV}}(P^t(\cdot \mid s_i^0), \mu) \le \epsilon \right\}; \\
t_{\mathrm{mix}} &:= t_{\mathrm{mix}}(1/4).
\end{aligned}
\tag{2.5}
$$

In what follows, we restate Lemma 8 in [**?**], which addresses the concentration of the empirical distribution of a uniformly ergodic Markov chain, highlighting the importance of the mixing time.

Lemma 3. *Consider the above-mentioned Markov chain. For any $0 < \delta < 1$, if*

$$t \geq \frac{443 t_{mix}}{\mu_{\min}} \log \frac{4|\mathcal{X}|}{\delta},$$

then for any $s_i^1 \in \mathcal{S}_i$, one has

$$P_{s_i^1}\left\{\exists s_i^t \in \mathcal{S}_i : \left|\sum_{i=1}^{t} \mathbb{I}\{s_i^t\} - t\mu(s_i^t)\right| \geq \frac{1}{2} t\mu(s_i^t)\right\} \leq \delta.$$

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

$$|e_0^t| \leq (1 - \eta_i^t)^{\frac{1}{2} t\mu_{min}} |\Delta_i^0| \tag{2.6}$$

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

$$\begin{aligned} \left|\prod_{j=1}^{t} (1 - \eta_i^j)\Delta_i^0\right| &= \prod_{j=1}^{C^t(s_i,a_i)} (1 - \eta_i^j)|\Delta_i^0| \\ &\leq (1 - \eta_{i,min}^t)^{C^t(s_i,a_i)} |\Delta_i^0|, \end{aligned} \tag{2.7}$$

Now using Lemma 3, 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,

$$C^t(s_i, a_i) \geq t\mu_{min}/2 \tag{2.8}$$

Using the fact that, any aperiodic and irreducible Markov chain on a finite state space is uniformly ergodic [**?**]. Thus, (2.8) holds uniformly over all (s_i, a_i) and all agents $i \in \mathcal{N}$ and all $\frac{443 t_{\text{mix}}}{\mu_{min}} \log \frac{4|\mathcal{S}||\mathcal{A}||\mathcal{N}|}{\delta} \leq t \leq T$, 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

$$\begin{aligned} \sum_{l=1}^{t} \left(\prod_{j=l+1}^{t} (1 - \eta_i^j)\right) \eta_i^l &\leq \eta_i^l \frac{1 - (1 - \eta_{i,\min})^{C^t}}{\eta_{i,\min}} \\ &\overset{(*)}{\leq} \eta_{i,\max} \cdot \frac{C^t \eta_{i,\min}}{\eta_{i,\min}} \\ &= C^t \eta_{i,\max}. \end{aligned} \tag{2.9}$$

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

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

$$\xi_r^k = r_i(s_i^t, a_i^t, \boldsymbol{a}_{-i}^t) - r_i^\star(s_i^t, a_i^t, \boldsymbol{a}_{-i}^t), \tag{2.10}$$

where $r_i^\star(s_i^t, a_i^t, \boldsymbol{a}_{-i}^t)$ represents the optimal reward under the optimal policy which satisfies, $r_i^\star(s_i^t, a_i^t, \boldsymbol{a}_{-i}^t) = \bar{r}_i^k(\pi_i^k)$, with $\bar{r}_i^k(\pi_i^k)$ given in (1.2).

Lemma 5. *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 $t \geq t_{\text{mix}} + \frac{2\sigma_1^2}{\mu_{\min}\epsilon^2} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right)$, where $\sigma_1^2 = Var[e_1^t] \leq (C^t \eta_{i,max}(r_{\max} - r_{\min}))^2$.*

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 $\text{Var}[e_1^t] = \sigma_1^2 \leq (C^t \eta_{i,\max}(r_{\max} - r_{\min}))^2$.

From Lemma 1, the error term ξ_r^t can be viewed as nearly zero-mean once the environment is effectively stationary beyond t_{mix}. Informally, for large l, $\mathbb{E}[\xi_r^l | \mathcal{H}^{k-1}] \approx 0$. Therefore, $\{e_1^t\}$ forms a martingale difference sequence with $\mathbb{E}[e_1^t \mid \mathcal{H}_i^{k-1}] = 0$, hence the partial sum of these centered increments satisfies a classical martingale-based concentration principle for Bernstein's inequality, thus we have

$$P\left(|e_1^t| > \epsilon\right) \leq 2 \exp\left(\frac{-C^t \epsilon^2}{2\sigma_1^2 + \frac{2}{3}(r_{\max} - r_{\min})\epsilon}\right). \tag{2.11}$$

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

$$P\left(\max_{i, s_i, a_i} |e_1^t| > \epsilon\right) \leq 2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i| \exp\left(\frac{-C^t \epsilon^2}{2\sigma_1^2 + \frac{2}{3}(r_{\max} - r_{\min})\epsilon}\right). \tag{2.12}$$

From Lemma 1, the error term ξ_r^t behaves as a martingale difference sequence once the Markov chain has mixed sufficiently, i.e., once $t \geq t_{\text{mix}}$. Before t_{mix}, the chain may still be in a transient phase, meaning the empirical state-action visitation frequencies are not yet well-approximated by the stationary distribution. Thus, we cannot immediately assume that each state-action pair (s_i, a_i) is visited in proportion to $\mu(s_i, a_i)$.

However, beyond t_{mix}, Lemma 3 ensures that the empirical frequency of visits to any (s_i, a_i) is at least half of its expected stationary fraction. Applying this bound to the total visit count, we obtain

$$C^t(s_i, a_i) \geq \frac{1}{2}\mu(s_i, a_i)(t - t_{\text{mix}})$$

with high probability for sufficiently large t.

Since $\mu_{\min} = \min_{s_i, a_i} \mu(s_i, a_i)$, we generalize this bound over all state-action pairs, $C^t \approx \mu_{\min}(t - t_{\text{mix}})$ up to constant factors. This approximation is crucial in solving for t in the final probability bound.

Solving for t, given $C^t \approx \mu_{\min}(t - t_{\text{mix}})$, we require

$$2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i| \exp\left(-\frac{\mu_{\min}(t - t_{\text{mix}})\epsilon^2}{2\sigma_1^2 + \frac{2}{3}(r_{\max} - r_{\min})\epsilon}\right) \leq \delta. \tag{2.13}$$

Solving for t, we obtain

$$t \geq t_{\text{mix}}+\frac{1}{\mu_{\min}} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right) \times \max\left\{\frac{2\sigma_1^2}{\epsilon^2}, \frac{3(r_{\max}-r_{\min})}{\epsilon}\right\}. \tag{2.14}$$

Since the term $\frac{2\sigma_1^2}{\epsilon^2}$ dominates when ϵ is sufficiently small. Thus, to simplify the expression and ensure the bound holds in this regime, we express it as

$$t \geq t_{\text{mix}} + \frac{2\sigma_1^2}{\mu_{\min}\epsilon^2} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right). \tag{2.15}$$

Thus, for any t satisfying the lower bound in (2.15), the error satisfies, $|e_1^t(s_i, a_i)| \leq \epsilon$, with probability at least $1-\delta$ for all i, s_i, and a_i. $\square$

Next, we want to analyze e_2^t.

Lemma 6. *For any $\delta > 0$, with probability at least $1-\delta$, $|e_2^t| \leq \epsilon$, for all*

$$t \geq t_{\text{mix}} + \frac{2\,(\gamma\, V^\star)^2}{\mu_{\min}\,\epsilon^2} \ln\!\left(\frac{2\,|\mathcal{N}|\,|\mathcal{S}_i|\,|\mathcal{A}_i|}{\delta}\right) \tag{2.16}$$

where $V^\star = \frac{r_{max}}{1-\gamma}$.

Proof: We begin with the error bound for e_2^t:

$$\begin{aligned}
|e_2^t| &= \left|\sum_{l=1}^{t}\left(\prod_{j=l+1}^{t}(1-\eta_i^j)\right)\eta_i^l\gamma(P^l-P)V_i^\star\right| \\
&\leq \sum_{l=1}^{t}\left|\left(\prod_{j=l+1}^{t}(1-\eta_i^j)\right)\eta_i^l\gamma(P^l-P)V_i^\star\right| \\
&= \sum_{l=1}^{t}\left(\prod_{j=l+1}^{t}(1-\eta_i^j)\right)\eta_i^l\gamma\|(P^l-P)V_i^\star\| \\
&\leq \sum_{l=1}^{t}\left(\prod_{j=l+1}^{t}(1-\eta_i^j)\right)\eta_i^l\gamma\|V_i^\star\| \max_{i'\in\mathcal{N}, s_{i'}\in\mathcal{S}_{i'}, a_{i'}\in\mathcal{A}_{i'}} \|P^l-P\|_1 \\
&\leq \gamma V^\star\left(\max_{i'\in\mathcal{N}, s_{i'}\in\mathcal{S}_{i'}, a_{i'}\in\mathcal{A}_{i'}} \|P^l-P\|_1\right)\sum_{l=1}^{t}\left(\prod_{j=l+1}^{t}(1-\eta_i^j)\right)\eta_i^l
\end{aligned} \tag{2.17}$$

We use the known identity that for any sequence $\eta_i^j \in (0,1)$, $\sum_{l=1}^t \left(\prod_{j=l+1}^t (1-\eta_i^j)\right)\eta_i^l = 1-\prod_{j=1}^t(1-\eta_i^j) \leq 1$. Applying this bound to inequality (2.17):

$$|e_2^t| \leq \gamma\, V^\star \max_{i\in\mathcal{N}, s_i\in\mathcal{S}_i, a_i\in\mathcal{A}_i} \|P^l-P\|_1. \tag{2.18}$$

By Hoeffding's inequality and a union bound, with probability at least $1-\delta$, for all $l \geq t_{\text{mix}}$ and all (s_i, a_i), we have $\|P^l - P\|_1 \leq \epsilon_0$ if $C^l(s_i, a_i) \geq \frac{1}{2\epsilon_0^2} \ln\left(\frac{2\,|\mathcal{N}|\,|\mathcal{S}_i|\,|\mathcal{A}_i|}{\delta}\right)$. To ensure $|e_2^t| \leq \epsilon$, we set $\gamma\, V^\star\, \epsilon_0 = \epsilon$, so $\epsilon_0 = \frac{\epsilon}{\gamma\, V^\star}$. Substituting ϵ_0 into the visit count condition and using the approximation $C^t(s_i, a_i) \approx \mu_{\min}\,(t - t_{\text{mix}})$ for $t \geq t_{\text{mix}}$, we require:

$$\mu_{\min}\,(t - t_{\text{mix}}) \geq \frac{(\gamma\, V^\star)^2}{2\epsilon^2} \ln\left(\frac{2\,|\mathcal{N}|\,|\mathcal{S}_i|\,|\mathcal{A}_i|}{\delta}\right). \tag{2.19}$$

Solving for t yields the sample complexity bound:

$$t \geq t_{\text{mix}} + \frac{2\,(\gamma\, V^\star)^2}{\mu_{\min}\,\epsilon^2} \ln\left(\frac{2\,|\mathcal{N}|\,|\mathcal{S}_i|\,|\mathcal{A}_i|}{\delta}\right). \tag{2.20}$$

Thus, with this choice of t, we ensure that with probability at least $1-\delta$, $|e_2^t| \leq \epsilon$.

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:*

$$P\left(|e_3^t| \leq \epsilon\right) \geq 1 - \delta, \tag{2.21}$$

for all

$$t \geq t_{\text{mix}} + \frac{1}{\mu_{\min}(1-(1-\eta_i)^T)^2} \cdot \frac{2\sigma_2^2}{\epsilon^2} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right). \tag{2.22}$$

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:

$$\Delta_{\rho,i} = \rho^\star - \rho(\pi_i, \boldsymbol{\pi}_{-i}), \tag{2.23}$$

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

$$P\left(\phi^k = 0\right) \leq \exp\left(-\frac{(kT - t_{\text{mix}})(\Delta_{\rho,i})^2}{2\sigma_2^2}\right). \tag{2.24}$$

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT - t_{\text{mix}})(\Delta_{\rho,i})^2}{2\sigma_2^2}\right). \tag{2.25}$$

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:

$$P\left(\exists (i, s_i, a_i) : |e_3^t| > \epsilon\right) \leq \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \tag{2.26}$$

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

$$P\left(|e_3^t| > \epsilon\right) \leq \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \tag{2.27}$$

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

$$P\left(|e_3^t| \leq \epsilon\right) \geq 1 - \delta_{total}, \tag{2.28}$$

where

$$\delta_{total} = \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \tag{2.29}$$

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT - t_{\text{mix}})(\Delta_{\rho,i})^2}{2\sigma_2^2}\right). \tag{2.30}$$

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT - t_{\text{mix}})\epsilon^2}{2\sigma_2^2}\right). \tag{2.31}$$

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

$$\exp\left(-\frac{(kT - t_{\text{mix}})\epsilon^2}{2\sigma_2^2}\right) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|} \leq \delta. \tag{2.32}$$

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

$$\exp\left(-\frac{(kT - t_{\text{mix}})\epsilon^2}{2\sigma_2^2}\right) \leq \delta. \tag{2.33}$$

Taking the natural logarithm:

$$-\frac{(kT - t_{\text{mix}})\epsilon^2}{2\sigma_2^2} \leq \log(\delta). \tag{2.34}$$

Rearranging for $kT - t_{\text{mix}}$:

$$kT - t_{\text{mix}} \geq \frac{2\sigma_2^2}{\epsilon^2}|\log(\delta)|. \tag{2.35}$$

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:

$$(1 - (1 - \eta_i)^T)^2 t \geq t_{\text{mix}} + \frac{2\sigma_2^2}{\epsilon^2}|\log(\delta)|. \tag{2.36}$$

Rearranging:

$$t \geq t_{\text{mix}} + \frac{1}{\mu_{\min}(1 - (1 - \eta_i)^T)^2} \cdot \frac{2\sigma_2^2}{\epsilon^2}|\log(\delta)|. \tag{2.37}$$

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

$$t_0 = t_{\text{mix}} + \frac{1}{\mu_{\min}(1-(1-\eta_i)^T)^2} \cdot \frac{2\sigma_2^2}{\epsilon^2} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right), \tag{2.38}$$

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

For e_4^t one can easily notice that,

$$\left|P^l\left(V_i^{l-1} - V^\star\right)\right| \leq \left|V_i^{l-1} - V^\star\right| \leq \left|Q_i^{l-1} - Q^\star\right| = \left|\Delta_i^{l-1}\right| \tag{2.39}$$

Thus we can write,

$$|\Delta_i^t| \leq \left[(1-\eta)^{\frac{1}{2}t\mu_{\min}} + C^t\eta_{i,\max}\gamma\right]|\Delta_i^0| + |e_1^t| + |e_2^t| + |e_3^t|. \tag{2.40}$$

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:

$$t \geq t_{\text{mix}} + \frac{2\sigma_2^2}{\mu_{\min}\epsilon^2} \log\left(\frac{6|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right).$$

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

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

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

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

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:

$$T_k = \left(\frac{T^2(\Delta_{\rho,i})^2}{2\sigma_2^2} \log\left(\frac{3}{\delta}\right) + \frac{T^2}{t_{\text{mix}}}\right) \cdot \max\left\{\frac{32\sigma_2^2 \log(6/\delta)}{\mu_{\min}\epsilon^2}, \frac{16\gamma^2(V^\star)^2 \log(6|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|/\delta)}{\mu_{\min}\epsilon^2(1-(1-\eta_i)^T)^2}\right\}.$$

□

Lemma 8 (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:*

$$P\left(|e_3^t| \leq \epsilon\right) \geq 1 - \delta, \tag{2.41}$$

for all

$$t \geq t_{\mathrm{mix}} + \frac{1}{2\mu_{\min}\epsilon^2} \ln\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta_{total}}\right), \tag{2.42}$$

where δ_{total} is given by:

$$\delta_{total} = \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}, \tag{2.43}$$

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT - t_{\mathrm{mix}})(\Delta_{\rho,i})^2}{2\sigma_2^2}\right). \tag{2.44}$$

—

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:

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

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

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

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT - t_{\mathrm{mix}})(\Delta_{\rho,i})^2}{2\sigma_2^2}\right).$$

Thus, the total failure probability becomes:

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

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

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

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:

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

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:

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

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