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

January 23, 2025

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

Proposition 1. *Let assumption* **??** *holds, and let* $\Delta_i = \rho^{1,*} - \rho^{2,n}$ *be the reward gap where* $\rho^{1,*}$ *is true score function of* $\pi_i^\star$ *and* $\rho^{2,n}$ *is the score function of any suboptimal policy at episode* n. *The probability that agent* i *selects* $\pi_i^\star$ *by episode* n *satisfies,*

$$P(\pi_i^n \in \Pi_i^\star) \geq 1 - \prod_{k=1}^{n} \left[\frac{\beta_i^k |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|} + \left(1 - \beta_i^k\right) 2e^{-\frac{\Delta_i^2 k}{2}} \right].$$

Proof. At each episode n, the probability that agent i selects a suboptimal policy satisfies

$$P(\pi_i^{2,n} \notin \Pi_i^\star) \leq \frac{\beta_i^n |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|} + (1 - \beta_i^n)\, P\left(\rho^{2,n} \geq \max_{k \leq n} \rho^{1,k}\right).$$

The first term corresponds to random exploration, and the second term bounds the probability that a suboptimal policy is mistakenly evaluated as superior to the best-known policy due to overestimation. Decomposing the second term gives

$$\begin{aligned} P\left(\rho^{2,n} \geq \max_{k \leq n} \rho^{1,k}\right) \leq & P\left(\rho^{2,n} \geq \rho^{1,*} - \frac{\Delta_i}{2}\right) + \\ & P\left(\max_{k \leq n} \rho^{1,k} \leq \rho^{1,*} - \frac{\Delta_i}{2}\right). \end{aligned} \tag{1.1}$$

From assumption **??** and the Markov property, we apply Chernoff-Hoeffding bounds, we first consider the overestimation of a suboptimal policy, $P\left(\rho^{2,n} \geq \rho^{2,*} + \frac{\Delta_i}{2}\right) \leq e^{-\frac{\Delta_i^2(n)}{2}}$. Next, is the underestimation of the optimal policy, $P\left(\rho^{1,k} \leq \rho^{1,*} - \frac{\Delta_i}{2}\right) \leq e^{-\frac{\Delta_i^2(n)}{2}}$. Combining these bounds, we get, $P\left(\rho^{2,n} \geq \max_{k\leq n} \rho^{1,k}\right) \leq 2e^{-\frac{\Delta_i^2 n}{2}}$. Substituting this result into the original bound gives

$$P(\pi_i^{2,n} \notin \Pi_i^\star) \leq \frac{\beta_i^n |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|} + (1-\beta_i^n)\, 2e^{-\frac{\Delta_i^2 n}{2}}.$$

Modeling the sequence of probabilities through the learning process, the probability of selecting the optimal policy satisfies, $P(\pi_i^n \in \Pi_i^\star) = 1 - \prod_{k=1}^n P(\pi_i^{2,k} \notin \Pi_i^\star)$.

The independence assumption is justified as asynchronous Q-learning is a well-behaved stochastic iterative algorithm [?] with martingale-based Q-value updates and exponentially bounded errors [?]. Adaptive exploration rates ensure random resets, further weakening dependencies and enabling asymptotic independence across episodes.

Substituting the upper bound yields

$$P(\pi_i^n \in \Pi_i^\star) \geq 1 - \prod_{k=1}^{n} \left[\frac{\beta_i^k |\Pi_i \backslash \{\Pi_i^\star\}|}{|\Pi_i|} + \left(1-\beta_i^k\right) 2e^{-\frac{\Delta_i^2 k}{2}} \right].$$

□

2 Appendix B

We restate Theorem 1 for clarity,

Theorem 1. *Consider a discounted stochastic cooperative game. Suppose that each agent updates its policies by Algorithm 1. Let Assumptions 1 and 2 hold. Then, for any* $0 < \delta < 1$, *one has that for all* $k \geq T_k/T$, *we have,* $|Q_i^k - Q_i^\star| < \epsilon$ *with a probability at least* $1-\delta$, *where,*

$$T_k = \left(\frac{T^2 (\Delta_{\rho,i})^2}{2\sigma^2} \log\left(\frac{3}{\delta}\right) + \frac{T^2}{t_{mix}} \right) \cdot \max\left\{ \frac{32\sigma^2 \log(6/\delta)}{\epsilon^2}, \frac{16\gamma^2 (V^\star)^2 \log(6|\mathcal{S}||\mathcal{A}|N/\delta)}{\epsilon^2 (1-(1-\eta_i)^T)^2} \right\}. \tag{2.1}$$

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,

To simplify the notation and reduce clutter in the equations, we omit the explicit dependency on state-action pairs, assuming the context makes it clear. For instance, $Q_i^k(s_i^t, a_i^t)$ is abbreviated as Q_i^k, the transition probability $P(s^{t+1}|s^t, a^t)$ is denoted as P, and the value function $V_i^{k-1}(s^{t+1})$ is written as V_i^{k-1}. These changes streamline the presentation while maintaining clarity.

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

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.

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.

After t recursions, we obtain:

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

We can apply the triangle inequality to the error and get,

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

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

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

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

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

where $\eta_{min} = \min_{i\in\mathcal{N}}\min_{j\in[1,t]}\eta_i^j)$.

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, w.p. greater than $1-\delta$, that,

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

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 $\frac{443 t_{mix}}{\mu_{min}} \log \frac{4|\mathcal{S}||\mathcal{A}||\mathcal{N}|}{\delta} \leq t \leq T$, where μ_{min} is the stationary distribution of the Markov chain $(s_i^0, s_i^1, \dots)$.

Then, we have,

$$|e_0^t| \leq (1-\eta_{min})^{\frac{t\mu_{min}}{2}} |\Delta_i^0| \tag{2.8}$$

For learning rates $\eta_i^j \in (0,1)$ with $\eta_{\max} = \max_j \eta_i^j$, and $\eta_{\min} = \min_j \eta_i^j$, the following inequality holds,

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

To derive this, first bound the product term. $\prod_{j=l+1}^{t}(1-\eta_i^j) \leq (1-\eta_{\min})^{t-l}$, Substituting this into the summation, and using the definition of C^t we get,

$$\sum_{l=1}^{t} \left(\prod_{j=l+1}^{t} (1-\eta_i^j) \right) \eta_i^l \leq \frac{1-(1-\eta_{\min})^{C^t}}{\eta_{\min}}.$$

Applying Bernoulli's inequality, $(1-\eta_{\min})^{C^t} \geq 1 - C^t\eta_{\min}$, yields, $1-(1-\eta_{\min})^t \leq C^t \eta_{\min}$.
Substituting this back into the bound,

$$\sum_{l=1}^{t} \left(\prod_{j=l+1}^{t} (1-\eta_i^j) \right) \eta_i^l \leq \eta_{\max} \cdot \frac{C^t \eta_{\min}}{\eta_{\min}} = C^t \eta_{\max}. \tag{2.9}$$

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 reward under the joint optimal policy.

Lemma 2. *For any $\delta > 0$ and error tolerance $\epsilon > 0$, one has, $|e_1^t(s_i, a_i)| \leq \epsilon$ w.p. at least $1-\delta$, holds for all $i \in \mathcal{N}$, $s_i \in \mathcal{S}_i$, and $a_i \in \mathcal{A}_i$, for all*

$$t \geq t_{mix} + \frac{2\sigma^2}{\mu_{\min}\epsilon^2} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta} \right). \tag{2.11}$$

where, $\sigma^2 = Var[e_1^t]$

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 $\mathrm{Var}[e_1^t] = \sigma^2 \leq C^{t2}\eta_{\max}^2 (r_{\max} - r_{\min})^2$. Furthermore, $\{e_1^t\}$ forms a martingale difference sequence with $\mathbb{E}[e_1^t \mid \mathcal{H}_i^{k-1}] = 0$, satisfying the conditions for Bernstein's inequality,

$$P\left(|e_1^t| > \epsilon\right) \leq 2\exp\left(\frac{-C^t\epsilon^2}{2\sigma^2 + \frac{2}{3}(r_{\max} - r_{\min})\epsilon} \right). \tag{2.12}$$

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

$$P\left(\max_{i,s_i,a_i} |e_1^t| > \epsilon\right) \leq 2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i| \exp\left(\frac{-C^t\epsilon^2}{2\sigma^2 + \frac{2}{3}(r_{\max} - r_{\min})\epsilon}\right). \tag{2.13}$$

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:

$$2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i| \exp\left(-\frac{\mu_{\min}(t - t_{mix})\epsilon^2}{2\sigma^2 + \frac{2}{3}(r_{\max} - r_{\min})\epsilon}\right) \leq \delta. \tag{2.14}$$

Solving for t, we obtain:

$$t \geq t_{mix} + \frac{1}{\mu_{\min}} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right) \times \max\left\{\frac{2\sigma^2}{\epsilon^2}, \frac{3(r_{\max} - r_{\min})}{\epsilon}\right\}. \tag{2.15}$$

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

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

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

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

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

$$e_2^t = \sum_{l=1}^{C^t}(1 - \eta_i)^{C^t - l}\eta_i\gamma\left(P_i^{t_l} - P\right)V^\star \tag{2.16}$$

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

$$P\left(\left|\sum_{l=1}^{C^t}(1 - \eta_i)^{C^t - l}\eta_i\gamma\left(P_i^{t_l} - P\right)V^\star\right| \geq \epsilon\right) \leq \delta \tag{2.17}$$

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,

$$P\left(\left|\sum_{l=1}^{C^t}(1 - \eta_i)^{C^t - l}\eta_i\gamma\left(P_i^{t_l} - P\right)V^\star\right| \geq \epsilon\right) \leq \frac{\delta}{|\mathcal{S}_i||\mathcal{A}_i|N} \tag{2.18}$$

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,

$$P\left(\left|\sum_{l=1}^{C^t}(1 - \eta_i)^{C^t - l}\eta_i\gamma\left(P_i^{t_l} - P\right)V^\star\right| \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{\sigma^2}\right) \tag{2.19}$$

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

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 (2.19) and taking the complement, we get,

$$P\left(e_2^t \leq c\gamma\sqrt{\log\left(\frac{|\mathcal{S}||\mathcal{A}|N}{\delta}\right)}\right) \geq 1 - \frac{\delta}{|\mathcal{S}_i||\mathcal{A}_i|N} \tag{2.20}$$

□

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

$$t \geq t_{mix} + \frac{(\gamma V^\star \eta_{\max})^2 \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right)}{2\epsilon^2}. \tag{2.21}$$

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:

$$P\left(\|\xi_M\|_1 > \epsilon\right) \leq 2|\mathcal{A}_{-i}||\mathcal{S}_{-i}|\exp\left(-2C^t\epsilon^2\right). \tag{2.22}$$

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

$$P\left(\|\xi_T\|_1 > \epsilon\right) \leq 2|\mathcal{S}|\exp\left(-2C^t\epsilon^2\right). \tag{2.23}$$

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

$$P\left(\|\xi_{P,i}^l\|_1 > \epsilon\right) \leq 2\left(|\mathcal{A}_{-i}||\mathcal{S}_{-i}| + |\mathcal{S}|\right)\exp\left(-2C^t\epsilon^2\right). \tag{2.24}$$

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

$$P\left(\max_{i,s_i,a_i}\|\xi_{P,i}^l\|_1 > \epsilon\right) \leq 2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|\exp\left(-2C^t\epsilon^2\right). \tag{2.25}$$

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

$$|e_2^t| \leq \gamma V^\star \sum_{l=1}^{t}\left(\prod_{j=l+1}^{t}(1-\eta_i^j)\right)\eta_i^l\epsilon, \tag{2.26}$$

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:

$$\sum_{l=1}^{t}\left(\prod_{j=l+1}^{t}(1-\eta_i^j)\right)\eta_i^l \leq \eta_{\max}C^t, \tag{2.27}$$

we get:

$$|e_2^t| \leq \gamma V^\star \eta_{\max} C^t \epsilon. \tag{2.28}$$

Substituting ϵ:

$$|e_2^t| \leq \gamma V^\star \eta_{\max}\sqrt{C^t \cdot \frac{1}{2}\log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right)}. \tag{2.29}$$

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

$$\gamma V^\star \eta_{\max}\sqrt{C^t \cdot \frac{1}{2}\log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right)} \leq \epsilon. \tag{2.30}$$

Square both sides:

$$(\gamma V^\star \eta_{\max})^2 C^t \cdot \frac{1}{2}\log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right) \leq \epsilon^2. \tag{2.31}$$

Solve for C^t:

$$C^t \leq \frac{2\epsilon^2}{(\gamma V^\star \eta_{\max})^2 \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right)}. \tag{2.32}$$

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

$$t \geq \frac{(\gamma V^\star \eta_{\max})^2 \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right)}{2\epsilon^2}. \tag{2.33}$$

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

$$t_0 = t_{\mathrm{mix}} + \frac{(\gamma V^\star \eta_{\max})^2 \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right)}{2\epsilon^2}, \tag{2.34}$$

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

Lemma 4 (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:*

$$P\left(|e_3^t| \leq \epsilon\right) \geq 1-\delta, \tag{2.35}$$

for all

$$t \geq t_{mix} + \frac{1}{\mu_{\min}(1-(1-\eta_i)^T)^2} \cdot \frac{2\sigma^2}{\epsilon^2}\log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right). \tag{2.36}$$

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:

$$\Delta_{\rho,i} = \rho^\star - \rho(\pi_i, \boldsymbol{\pi}_{-i}), \tag{2.37}$$

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:

$$P\left(\phi^k = 0\right) \leq \exp\left(-\frac{(kT - t_{mix})(\Delta_{\rho,i})^2}{2\sigma^2}\right). \tag{2.38}$$

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT - t_{mix})(\Delta_{\rho,i})^2}{2\sigma^2}\right). \tag{2.39}$$

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:

$$P\left(\exists(i, s_i, a_i) : |e_3^t| > \epsilon\right) \leq \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \tag{2.40}$$

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

$$P\left(|e_3^t| > \epsilon\right) \leq \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}. \tag{2.41}$$

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

$$P\left(|e_3^t| \leq \epsilon\right) \geq 1 - \delta_{total}, \tag{2.42}$$

where

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

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

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

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2}\right). \tag{2.45}$$

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

$$\exp\left(-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2}\right) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|} \leq \delta. \tag{2.46}$$

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

$$\exp\left(-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2}\right) \leq \delta. \tag{2.47}$$

Taking the natural logarithm:

$$-\frac{(kT - t_{mix})\epsilon^2}{2\sigma^2} \leq \log(\delta). \tag{2.48}$$

Rearranging for $kT - t_{mix}$:

$$kT - t_{mix} \geq \frac{2\sigma^2}{\epsilon^2}|\log(\delta)|. \tag{2.49}$$

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:

$$(1-(1-\eta_i)^T)^2 t \geq t_{mix} + \frac{2\sigma^2}{\epsilon^2}|\log(\delta)|. \tag{2.50}$$

Rearranging:

$$t \geq t_{mix} + \frac{1}{\mu_{\min}(1-(1-\eta_i)^T)^2} \cdot \frac{2\sigma^2}{\epsilon^2}|\log(\delta)|. \tag{2.51}$$

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

$$t_0 = t_{mix} + \frac{1}{\mu_{\min}(1-(1-\eta_i)^T)^2} \cdot \frac{2\sigma^2}{\epsilon^2} \log\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta}\right), \tag{2.52}$$

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

For e_4^t one can easily notice that,

$$\left|P_i^l\left(V_i^{l-1} - V^\star\right)\right| \leq \left|V_i^{l-1} - V^\star\right| \leq \left|Q_i^{l-1} - Q^\star\right| = \left|\Delta_i^{l-1}\right| \tag{2.53}$$

Thus we can write,

$$|\Delta_i^t| \leq \left[(1-\eta)^{\frac{1}{2}t\mu_{\min}} + C^t\eta_{\max}\gamma\right]|\Delta_i^0| + |e_1^t| + |e_2^t| + |e_3^t|. \tag{2.54}$$

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

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

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:

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

□

Lemma 5 (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:*

$$P\left(|e_3^t| \leq \epsilon\right) \geq 1-\delta, \tag{2.55}$$

for all

$$t \geq t_{mix} + \frac{1}{2\mu_{\min}\epsilon^2}\ln\left(\frac{2|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}{\delta_{total}}\right), \tag{2.56}$$

where δ_{total} is given by:

$$\delta_{total} = \delta_{joint}(k) + \frac{\delta}{|\mathcal{N}||\mathcal{S}_i||\mathcal{A}_i|}, \tag{2.57}$$

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

$$\delta_{joint}(k) = \exp\left(-\frac{(kT-t_{mix})(\Delta_{\rho,i})^2}{2\sigma^2}\right). \tag{2.58}$$

—

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 $\Delta_{\rho,i}$:

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

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:

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

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