The Blessing of Heterogeneity in Federated Q-Learning: Linear Speedup and Beyond

Jiin Woo Gauri Joshi Yuejie Chi
Carnegie Mellon University

May 2023; Revised December 2023

Abstract

When the data used for reinforcement learning (RL) are collected by multiple agents in a distributed manner, federated versions of RL algorithms allow collaborative learning without the need for agents to share their local data. In this paper, we consider federated Q-learning, which aims to learn an optimal Q-function by *periodically* aggregating local Q-estimates trained on local data alone. Focusing on infinite-horizon tabular Markov decision processes, we provide sample complexity guarantees for both the synchronous and asynchronous variants of federated Q-learning. In both cases, our bounds exhibit a linear speedup with respect to the number of agents and near-optimal dependencies on other salient problem parameters.

In the asynchronous setting, existing analyses of federated Q-learning, which adopt an equally weighted averaging of local Q-estimates, require that every agent covers the entire state-action space. In contrast, our improved sample complexity scales inverse proportionally to the minimum entry of the *average* stationary state-action occupancy distribution of all agents, thus only requiring the agents to *collectively cover* the entire state-action space, unveiling the *blessing of heterogeneity* in enabling collaborative learning by relaxing the coverage requirement of the single-agent case. However, its sample complexity still suffers when the local trajectories are highly heterogeneous. In response, we propose a novel federated Q-learning algorithm with importance averaging, giving larger weights to more frequently visited state-action pairs, which achieves a robust linear speedup as if all trajectories are centrally processed, regardless of the heterogeneity of local behavior policies.

Keywords: federated Q-learning, periodic averaging, sample complexity, linear speedup, blessing of heterogeneity

Contents

Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Emails: {jiinw,gaurij,yuejiec}@andrew.cmu.edu.

A preliminary version of this paper was presented at the 2023 International Conference on Machine Learning (ICML).

1 Introduction

Reinforcement Learning (RL) (Sutton and Barto, 2018) is an area of machine learning for sequential decision making, aiming to learn an optimal policy that maximizes the total rewards via interactions with an unknown environment. RL is widely used in many real-world applications, such as autonomous driving, games, clinical trials, and recommendation systems. However, due to the high dimensionality of the state-action space, training of RL agents typically requires a significant amount of computation and data to achieve desirable performance. Moreover, data collection can be extremely time-consuming with limited access in the wild, especially when performed by a single agent. On the other hand, it is possible to leverage multiple agents to collect data simultaneously, under the premise that they can learn a global policy collaboratively with the aid of a central server without the need of sharing local data. As a result, there is a growing need to conduct RL in a distributed or federated fashion.

Although there have been many studies analyzing federated learning (Kairouz et al., 2021) in other areas such as supervised machine learning (Bonawitz et al., 2019; McMahan et al., 2017; Wang et al., 2020b), there are only a few recent works focused on federated RL. They consider issues such as robustness to adversarial attacks (Fan et al., 2021; Wu et al., 2021), environment heterogeneity (Jin et al., 2022), as well as sample and communication complexities (Doan et al., 2021; Khodadadian et al., 2022; Shen et al., 2022). Encouragingly, some of these prior works offer non-asymptotic sample complexity analyses of federated RL algorithms that highlight a linear speedup of the required sample size in terms of the number of agents. However, the performance characterization of these federated algorithms is still far from complete.

1.1 Federated Q-learning: prior art and limitations

This paper focuses on Q-learning (Watkins and Dayan, 1992), one of the most celebrated model-free RL algorithms, which aims to learn the optimal Q-function directly without forming an estimate of the model. Two sampling protocols are typically studied: synchronous sampling and asynchronous sampling. With synchronous sampling, all state-action pairs are updated uniformly assuming access to a generative model or a simulator (Kearns and Singh, 1999). With asynchronous sampling, only the state-action pair that is visited by the behavior policy is updated at each time (Tsitsiklis, 1994). Despite its long history of theoretical investigation, the tight sample complexity of Q-learning in the single-agent setting has only recently been pinned down in Li et al. (2023). As we shall elucidate, there remains a large gap in terms of the sample complexity requirement between the federated setting and the single-agent setting in terms of dependencies on salient problem parameters.

To harness the power of multiple agents, Khodadadian et al. (2022) proposed and analyzed a federated variant of Q-learning with *asynchronous* sampling that periodically aggregates the local Q-estimates trained on local Markovian trajectories collected over K agents. To set the stage, consider an infinite-horizon tabular Markov decision process (MDP) with state space $\mathcal{S}$, action space $\mathcal{A}$, and a discount factor $\gamma \in [0, 1)$. To learn an ε-optimal Q-function estimate (in the ℓ_∞ sense), Khodadadian et al. (2022) requires a per-agent sample size on the order of

$$\widetilde{O}\left(\frac{|\mathcal{S}|^2}{K\mu_{\mathsf{min}}^5(1-\gamma)^9\varepsilon^2}\right) \tag{1}$$

for sufficiently small ε, where $\mu_{\mathsf{min}} := \min_{1\le k\le K}\min_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mu_{\mathsf{b}}^k(s,a)$ is the minimum entry of the stationary state-action occupancy distributions μ_{b}^k of the sample trajectories over all agents, and $\widetilde{O}$ hides logarithmic terms. On the other hand, the sample requirement of single-agent Q-learning (Li et al., 2023) for learning an ε-optimal Q-function is

$$\widetilde{O}\left(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2}\right) \tag{2}$$

for sufficiently small ε. Comparing the two sample complexity bounds reveals several drawbacks of existing analyses and raises the following natural questions.

- *Near-optimal sample size.* Despite the appealing linear speedup in terms of the number of agents K shown in Khodadadian et al. (2022), it has unfavorable dependencies on other salient problem parameters. In particular, since $1/\mu_{\mathsf{min}} \ge |\mathcal{S}||\mathcal{A}|$, the sample complexity in (1) will be better than that of the single-agent case in (2) only if K is at least above the order of $\frac{|\mathcal{S}|^6|\mathcal{A}|^4}{(1-\gamma)^5}$, which may not be practically feasible with large state-action space and long effective horizon. *Can we improve the dependency on the salient problem parameters for federated Q-learning while maintaining linear speedup?*

- *Benefits of heterogeneity.* Existing analyses in Khodadadian et al. (2022) require that each agent has a full coverage of the state-action space (i.e., $\mu_{\mathsf{min}} > 0$), which is as stringent as the single-agent setting. However, given that the insufficient coverage of individual agents can be complemented by each other when agents have heterogeneous local trajectories, it may not be necessary to require full coverage of the state-action space from every agent. *Can we exploit the heterogeneity in the agents' local trajectories and relax the coverage requirement on individual agents?*

1.2 Summary of our contributions

In this paper, we answer these questions in the affirmative, by providing a sample complexity analysis of federated Q-learning under both the synchronous and asynchronous settings. The main contributions are summarized as follows, with Table 1 providing a comparison with the prior art.

- *Sample complexity of federated synchronous Q-learning with equal averaging.* We show that with high probability, the sample complexity of federated synchronous Q-learning (FedSynQ) to learn an ε-optimal

sampling	reference	number of agents	coverage	sample complexity
synchronous	Chen et al. (2020); Wainwright (2019a)	1	full	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
	(Li et al., 2023)	1	full	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
	FedSynQ (Theorem 1)	K	full	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{K(1-\gamma)^5\varepsilon^2}$
asynchronous	Qu and Wierman (2020)	1	full	$\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$
	Li et al. (2021b)	1	full	$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$
	Li et al. (2023)	1	full	$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2}$
	FedAsynQ-EqAvg (Khodadadian et al., 2022)	K	full	$\frac{\lvert\mathcal{S}\rvert^2}{K\mu_{\mathsf{min}}^5(1-\gamma)^9\varepsilon^2}$
	FedAsynQ-EqAvg (Theorem 2)	K	partial	$\frac{C_{\mathsf{het}}}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}$
	FedAsynQ-ImAvg (Theorem 3)	K	partial	$\frac{1}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}$

Table 1: Comparison of sample complexity upper bounds of single-agent and federated Q-learning algorithms under synchronous and asynchronous sampling protocols to learn an ε-optimal Q-function in the ℓ_∞ sense, where logarithmic factors and burn-in costs are hidden. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, γ is the discount factor, K is the total number of agents, and t_{mix} is the mixing time of the behavior policy. In addition, $\mu_{\mathsf{min}} = \min_{k,s,a} \mu_{\mathsf{b}}^k(s,a)$ denotes the minimum entry of the stationary state-action occupancy distributions μ_{b}^k of all agents, $\mu_{\mathsf{avg}} := \min_{s,a} \frac{1}{K}\sum_{k=1}^K \mu_{\mathsf{b}}^k(s,a)$ denotes the minimum entry of the average stationary state-action occupancy distribution of all agents, and $C_{\mathsf{het}} := \max_{k,s,a} K\mu_{\mathsf{b}}^k(s,a)/\big(\sum_{k=1}^K \mu_{\mathsf{b}}^k(s,a)\big)$ captures the heterogeneity across the agents.

Q-function in the ℓ_∞ sense is (see Theorem 1)

$$\widetilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{K(1-\gamma)^5\varepsilon^2}\right), \tag{3}$$

which exhibits a linear speedup with respect to the number of agents K and nearly matches the tight sample complexity bound of single-agent synchronous Q-learning up to a factor of $1/(1-\gamma)$ in Li et al. (2023) for $K = 1$.

- *Sample complexity of federated asynchronous Q-learning with equal averaging.* We provide a sharpened sample complexity analysis of the algorithm developed in Khodadadian et al. (2022) for federated asynchronous Q-learning with equal averaging (FedAsynQ-EqAvg) that leads to new insights. To learn an ε-optimal Q-function in the ℓ_∞ sense, FedAsynQ-EqAvg requires at most (see Theorem 2)

$$\widetilde{O}\left(\frac{C_{\mathsf{het}}}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right) \tag{4}$$

samples per agent for sufficiently small ε (ignoring the burn-in cost that depends on the mixing times of the Markovian trajectories over all agents), where $\mu_{\mathsf{avg}} = \min_{s,a} \frac{1}{K}\sum_{k=1}^K \mu_{\mathsf{b}}^k(s,a) \geq \mu_{\mathsf{min}}$ is the minimum entry of the *average* stationary state-action occupancy distribution of all agents, and $C_{\mathsf{het}} = \max_{k,s,a} \frac{K\mu_{\mathsf{b}}^k(s,a)}{\sum_{k=1}^K \mu_{\mathsf{b}}^k(s,a)} \in [1, 1/\mu_{\mathsf{avg}}]$ captures the heterogeneity of the behavior policies across agents. This sample complexity not only proves a linear speedup with respect to the number of agents, but also greatly sharpens the dependency on all the salient problem parameters — including $1/(1-\gamma)$, $\lvert\mathcal{S}\rvert$, and $1/\mu_{\mathsf{min}}$ — by orders of magnitudes compared to the bound obtained in Khodadadian et al. (2022). More importantly, it uncovers that as long as the agents collectively cover the entire state-action space (i.e.,

$\mu_{\mathsf{avg}} > 0$), FedAsynQ-EqAvg still enables learning even when individual agents fail to cover the entire state-action space (i.e., $\mu_{\mathsf{min}} = 0$), unveiling the blessing of heterogeneity that was not elucidated in prior work (Khodadadian et al., 2022).

- *Sample complexity of federated asynchronous Q-learning with importance averaging.* Although heterogeneous behavior policies at agents may induce local trajectories covering different parts of the state-action space and relax the coverage requirement, equally weighting the local Q-estimates may hinder the convergence which is bottlenecked by the slowest converging agent. This is evident by the dependency on C_{het} in the sample complexity of FedAsynQ-EqAvg, which becomes larger when the local behavior policies are highly disparate. To address this issue, we propose a novel importance averaging scheme in federated Q-learning (FedAsynQ-ImAvg) that averages the local Q-estimates by assigning larger weights to more frequently updated local estimates. To learn an ε-optimal Q-function in the ℓ_∞ sense, FedAsynQ-ImAvg requires at most (see Theorem 3)

$$\widetilde{O}\left(\frac{1}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right) \tag{5}$$

  samples per agent for sufficiently small ε (ignoring the burn-in cost that depends on the mixing times of the Markovian trajectories over all agents). This improves over that of FedAsynQ-EqAvg by removing the dependency on C_{het}, which can be as large as $1/\mu_{\mathsf{avg}}$. More importantly, this suggests that FedAsynQ-ImAvg achieves stable linear speedup with respect to the profile of the local behavior policies while maintaining the blessing of heterogeneity that eases the burden of individual agents' coverage.

1.3 Related work

Analysis of single-agent Q-learning. There has been extensive research on the convergence guarantees of Q-learning, focusing on the single-agent case. Many initial studies have analyzed the asymptotic convergence of Q-learning (Borkar and Meyn, 2000; Jaakkola et al., 1994; Szepesvári, 1998; Tsitsiklis, 1994). Later, Beck and Srikant (2012); Chen et al. (2020); Even-Dar and Mansour (2003); Li et al. (2023); Wainwright (2019a) have studied the sample complexity of Q-learning under synchronous sampling, and Beck and Srikant (2012); Chen et al. (2021b); Even-Dar and Mansour (2003); Li et al. (2023, 2021b); Qu and Wierman (2020) have investigated the finite-time convergence of Q-learning under asynchronous sampling (also referred to as Markovian sampling). In addition, Bai et al. (2019); Jin et al. (2018); Li et al. (2021a); Yang et al. (2021); Zhang et al. (2020) studied Q-learning with optimism for online RL, and Shi et al. (2022); Yan et al. (2022) dealt with Q-learning with pessimism for offline RL.

Distributed and federated RL. Several recent works have developed distributed versions of RL algorithms to accelerate training (Assran et al., 2019; Espeholt et al., 2018; Mnih et al., 2016). Theoretical analysis of convergence and communication efficiency of these distributed RL algorithms have also been considered in recent works. For example, a collection of works (Chen et al., 2022a; Doan et al., 2019; Sun et al., 2020; Wai, 2020; Wang et al., 2020a; Zeng et al., 2021) have analyzed the convergence of decentralized temporal difference (TD) learning. Furthermore, Chen et al. (2022b); Shen et al. (2022) have analyzed the finite-time convergence of distributed actor-critic algorithms and Chen et al. (2021a) proposed a communication-efficient policy gradient algorithm with provable convergence guarantees.

Notation. Throughout this paper, we denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$, and $[K] \coloneqq \{1, \cdots, K\}$ for any positive integer $K > 0$. In addition, $f(\cdot) = \widetilde{O}(g(\cdot))$ or $f \lesssim g$ (resp. $f(\cdot) = \widetilde{\Omega}(g(\cdot))$ or $f \gtrsim g$) means that $f(\cdot)$ is orderwise no larger than (resp. no smaller than) $g(\cdot)$ modulo some logarithmic factors. The notation $f \asymp g$ means $f \lesssim g$ and $f \gtrsim g$ hold simultaneously.

2 Model and background

In this section, we introduce the mathematical model and background of Markov decision processes.

Infinite-horizon Markov decision process. We consider an infinite-horizon Markov decision process (MDP), which is represented by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the state space and the action space, respectively, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ indicates the transition kernel such that $P(s' \mid s, a)$ denotes the probability that action a in state s leads to state s', $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes a deterministic reward function, where $r(s, a)$ is the immediate reward for action a in state s, and $\gamma \in [0, 1)$ is the discount factor.

Policy, value function, and Q-function. A *policy* is an action-selection rule denoted by the mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, such that $\pi(a|s)$ is the probability of taking action a in state s. For a given policy π, the *value function* $V^\pi : \mathcal{S} \to \mathbb{R}$, which measures the expected discounted cumulative reward from an initial state s, is defined as

$$\forall s \in \mathcal{S} : \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]. \tag{6}$$

Here, the expectation is taken with respect to the randomness of the trajectory $\{s_t, a_t, r_t\}_{t=0}^{\infty}$, sampled based on the transition kernel (i.e., $s_{t+1} \sim P(\cdot|s_t, a_t)$) and the policy π (i.e., $a_t \sim \pi(\cdot|s_t)$) for any $t \geq 0$. Similarly, the state-action value function (i.e., *Q-function*) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which measures the expected discounted cumulative reward from an initial state-action pair (s, a), is defined as

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \qquad Q^\pi(s, a) := r(s, a) + \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a\right].$$

Again here, the expectation is taken with respect to the randomness of the trajectory $\{s_t, a_t, r_t\}_{t=1}^{\infty}$ generated similarly as above. Since the rewards lie within $[0, 1]$, it follows that for any policy π,

$$0 \leq V^\pi \leq \frac{1}{1-\gamma}, \qquad 0 \leq Q^\pi \leq \frac{1}{1-\gamma}. \tag{7}$$

Optimal policy and Bellman's principle of optimality. A policy that maximizes the value function uniformly over all states is called an *optimal policy* and denoted by $\pi^\star$. Note that the existence of such an optimal policy is always guaranteed (Puterman, 2014), which also maximizes the Q-function simultaneously. The corresponding optimal value function and Q-function are denoted by $V^\star := V^{\pi^\star}$ and $Q^\star := Q^{\pi^\star}$, respectively. It is well-known that the optimal Q-function $Q^\star$ can be determined as the unique fixed point of the Bellman operator $\mathcal{T}$, given by

$$\mathcal{T}(Q)(s, a) := r(s, a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q(s', a')\right]. \tag{8}$$

Q-learning (Watkins and Dayan, 1992), perhaps the most widely used model-free RL algorithm, seeks to learn the optimal Q-function based on samples collected from the underlying MDP without estimating the model.

3 Federated synchronous Q-learning: algorithm and theory

In this section, we begin with understanding federated synchronous Q-learning, where all the state-action pairs are updated simultaneously assuming access to a generative model or simulator at all the agents.

3.1 Problem setting

In the synchronous setting, each agent $k \in [K]$ has access to a generative model, and generates a new sample

$$s_t^k(s, a) \sim P(\cdot|s, a) \tag{9}$$

for every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ *independently* at every iteration t. Our goal is to learn the optimal Q-function $Q^\star$ collaboratively by aggregating the local Q-learning estimates *periodically*.

Review: synchronous Q-learning with a single agent. To facilitate algorithmic development, let us recall the synchronous Q-learning update rule with a single agent. Starting with certain initialization Q_0, at every iteration $t \geq 1$, the Q-function is updated according to

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_t(s,a) = (1-\eta) Q_{t-1}(s,a) + \eta \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t(s,a), a') \right), \tag{10}$$

where $s_t(s,a) \sim P(\cdot|s,a)$ is drawn independently for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, and η denotes the constant learning rate. The sample complexity of synchronous Q-learning has been recently investigated and sharpened in a number of works, e.g. Chen et al. (2020); Li et al. (2023); Wainwright (2019a).

3.2 Algorithm description

We propose a natural federated synchronous Q-learning algorithm called FedSynQ that alternates between local updates at agents and periodic averaging at a central server. The complete description is summarized in Algorithm 1. FedSynQ initializes a local Q-function as $Q_0^k = Q_0$ at each agent $k \in [K]$. Suppose at the beginning of each iteration $t \geq 1$, each agent maintains a local Q-function estimate Q_{t-1}^k and a local value function estimate V_{t-1}^k, which are related via

$$\forall s \in \mathcal{S}: \qquad V_t^k(s) := \max_{a \in \mathcal{A}} Q_t^k(s,a). \tag{11}$$

FedSynQ proceeds according to the following steps in the rest of the t-th iteration.

1. *Local updates:* Each agent first independently updates *all* entries of its Q-estimate Q_{t-1}^k to reach some *intermediate* estimate following the update rule:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_{t-\frac{1}{2}}^k(s,a) = (1-\eta) Q_{t-1}^k(s,a) + \eta \left(r(s,a) + \gamma V_{t-1}^k(s_t^k(s,a)) \right), \tag{12}$$

   where $s_t^k(s,a)$ is drawn according to (9), and $\eta \geq 0$ is the learning rate.

2. *Periodic averaging:* These intermediate estimates will be periodically averaged by the server to form the updated estimate Q_t^k at the end of the t-th iteration. Formally, denoting $\tau \geq 1$ as the synchronization period, it follows

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_t^k(s,a) = \begin{cases} \frac{1}{K} \sum_{k=1}^K Q_{t-\frac{1}{2}}^k(s,a) & \text{if } t \equiv 0 \pmod{\tau} \\ Q_{t-\frac{1}{2}}^k(s,a) & \text{otherwise} \end{cases}. \tag{13}$$

Denoting the number of total iterations by T, the algorithm outputs the final Q-estimate as the average of all local estimates, i.e. $Q_T = \frac{1}{K} \sum_k Q_T^k$. Without loss of generality, we assume the total number of iterations T is divisible by τ, where $C_{\mathsf{round}} = T/\tau$ is the rounds of communication.

Algorithm 1: Federated Synchronous Q-learning (FedSynQ)

```
inputs: learning rate $\eta$, discount factor $\gamma$, number of agents $K$, synchronization period $\tau$, number of iterations $T$.
initialization: $Q_0^k = Q_0$ for all $k$.
for $t = 1, \cdots, T$ do
  for $k \in [K]$ do
    Draw $s_t^k(s,a) \sim P(\cdot \mid s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
    Compute $Q_{t-\frac{1}{2}}^k$ according to (12).
    Compute $Q_t^k$ according to (13).
  end for
end for
return: $Q_T = \frac{1}{K} \sum_k Q_T^k$.
```

3.3 Performance guarantee

We are ready to provide the finite-time convergence analysis of Algorithm 1.

Theorem 1 (Finite-time convergence of FedSynQ)**.** *Consider any given $\delta \in (0,1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Suppose that the initialization of Algorithm 1 satisfies $0 \leq Q_0 \leq \frac{1}{1-\gamma}$, and the synchronization period τ obeys*

$$\tau \leq 1 + \frac{1}{\eta} \min \left\{ \frac{1-\gamma}{8\gamma}, \frac{1}{K} \right\}. \tag{14a}$$

There exist some sufficiently large constant $c_T > 0$ and sufficiently small constant $c_\eta > 0$, such that with probability at least $1-\delta$, the output of Algorithm 1 satisfies $\|Q_T - Q^\star\|_\infty \leq \varepsilon$, provided that the sample size per agent T and the learning rate η satisfy

$$T \geq \frac{c_T}{K(1-\gamma)^5\varepsilon^2} (\log((1-\gamma)^2\varepsilon))^2 \log \frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}, \tag{14b}$$

$$\eta = c_\eta K(1-\gamma)^4\varepsilon^2 \frac{1}{\log \frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}. \tag{14c}$$

Theorem 1 suggests that to achieve an ε-accurate Q-function estimate in an ℓ_∞ sense, the number of samples required at each agent is no more than

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{K(1-\gamma)^5\varepsilon^2}\right),$$

given that the agent collects $|\mathcal{S}||\mathcal{A}|$ samples at each iteration. A few implications are in order.

Linear speedup. The sample complexity exhibits an appealing linear speedup with respect to the number of agents K. In comparison, the sharpest upper bound known for single-agent Q-learning (Li et al., 2023) is $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4 \min\{\varepsilon,\varepsilon^2\}}\right)$, which matches with its algorithmic-dependent lower bound when $\varepsilon \in (0,1)$. Therefore, our federated setting enables faster learning as soon as the number of agents satisfies

$$K \gtrsim \frac{1}{(1-\gamma)\max\{1,\varepsilon\}}$$

up to logarithmic factors. When $K=1$, our bound nearly matches with the lower bound of single-agent Q-learning up to a factor of $1/(1-\gamma)$, indicating its near-optimality.

Communication efficiency. One key feature of our federated setting is the use of periodic averaging with the hope to improve communication efficiency. According to (14a), our theory requires that the synchronization period τ be inversely proportional to the learning rate η, which suggests that more frequent communication is needed to compensate the discrepancy of local updates when the learning rate is large. To provide insights, consider the parameter regime when $K \gtrsim \frac{1}{1-\gamma}$ and $\varepsilon \lesssim \frac{1}{K(1-\gamma)^2}$. Plugging the choice of the learning rate (14c) into the upper bound of τ in (14a), we can choose the synchronization period as $\tau \asymp \frac{1}{K^2(1-\gamma)^4\varepsilon^2}$ up to logarithmic factors, leading to a communication complexity no larger than $C_{\mathsf{round}} = \frac{T}{\tau} \lesssim \frac{K}{1-\gamma}$, which is almost independent of the final accuracy ε.

4 Federated asynchronous Q-learning: algorithm and theory

In this section, we study the sample complexity of federated asynchronous Q-learning, where K agents sample local trajectories using different behavior policies. In particular, we propose a novel aggregation algorithm FedAsynQ-ImAvg that leverages the heterogeneity of these policies and dramatically improves the sample complexity.

4.1 Problem setting

In the asynchronous setting, each agent $k \in [K]$ independently collects a sample trajectory $\{s_t^k, a_t^k, r_t^k\}_{t=0}^{\infty}$ from the same underlying MDP $\mathcal{M}$ following some stationary *local* behavior policy π_{b}^k such that

$$a_t^k \sim \pi_{\mathsf{b}}^k(\cdot|s_t^k), \quad r_t^k = r(s_t^k, a_t^k), \quad s_{t+1}^k \sim P(\cdot|s_t^k, a_t^k) \tag{15}$$

for all $t \geq 0$, where the initial state is initialized as s_0^k for each agent k. Note that the behavior policies $\{\pi_{\mathsf{b}}^k\}_{k\in[K]}$ are heterogeneous across agents and can be different from the optimal policy $\pi^\star$. Contrary to the generative model considered in the synchronous setting, the samples collected under the asynchronous setting are no longer independent across time but are Markovian, making the analysis significantly more challenging. The sample trajectory at each agent can be viewed as sampling a time-homogeneous Markov chain over the set of state-action pairs. Throughout this paper, we make the following standard uniform ergodicity assumption (Li et al., 2021b; Paulin, 2015).

Assumption 1 (Uniform ergodicity)**.** *For every agent $k \in [K]$, the Markov chain induced by the stationary behavior policy π_{b}^k is uniformly ergodic over the entire state-action space $\mathcal{S} \times \mathcal{A}$.*

Uniform ergodicity guarantees that the distribution of the state-action pair (s_t, a_t) of a trajectory converges to the stationary distribution of the Markov chain geometrically fast regardless of the initial state-action pair, and eventually, each state-action pair is visited in proportion to the stationary distribution.

Key parameters. Two important quantities concerning the resulting Markov chains will govern the performance guarantees. The first one is the stationary state-action distribution μ_{b}^k, which is the stationary distribution of the Markov chain induced by π_{b}^k over all state-action pairs; the second one is t_{mix}^k, which is the mixing time of the same Markov chain given by

$$t_{\mathsf{mix}}^k := \min\Big\{t \;\Big|\; \max_{(s_0,a_0)\in\mathcal{S}\times\mathcal{A}} d_{\mathsf{TV}}\big(P_t^k(\cdot\,|\,s_0,a_0),\, \mu_{\mathsf{b}}^k\big) \leq \frac{1}{4}\Big\}, \tag{16}$$

where $P_t^k(\cdot\,|\,s_0, a_0)$ denote the distribution of (s_t, a_t) conditioned on (s_0, a_0) for agent k, and $d_{\mathsf{TV}}(\cdot,\cdot)$ is the total variation distance. Further, let the largest mixing time of all the Markov chains induced by local behavior policies be

$$t_{\mathsf{mix}}^{\mathsf{max}} := \max_{k\in[K]} t_{\mathsf{mix}}^k. \tag{17}$$

In words, $t_{\mathsf{mix}}^{\mathsf{max}}$ approximately indicates the time that the transition of every agent starts to follow its stationary distribution regardless of its initial state.

Let us further define a few key parameters that measure the coverage and heterogeneity of the stationary state-action distribution μ_{b}^k across agents. First, define

$$\mu_{\mathsf{min}} := \min_{k\in[K]} \mu_{\mathsf{min}}^k, \qquad \text{where} \qquad \mu_{\mathsf{min}}^k := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_{\mathsf{b}}^k(s,a). \tag{18}$$

State-action pairs with small stationary probabilities are visited less frequently, and therefore can become bottlenecks in improving the quality of Q-function estimates. Clearly, $\mu_{\mathsf{min}} \leq \frac{1}{|\mathcal{S}||\mathcal{A}|}$. In addition, denote

$$\mu_{\mathsf{avg}} := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{1}{K}\sum_{k=1}^{K} \mu_{\mathsf{b}}^k(s,a). \tag{19}$$

In words, μ_{avg} is the minimum entry of the *average* stationary state-action distribution of all agents. The difference between μ_{avg} and μ_{min} stands out when an individual agent fails to cover the entire state-action space. While $\mu_{\mathsf{min}} = 0$ in such a case, μ_{avg} can still be positive as long as each state-action pair is explored by at least one of the agents, i.e., $\sum_{k=1}^{K} \mu_{\mathsf{b}}^k(s,a) > 0$. Note that μ_{avg} is always greater than or equal to μ_{min} since

$$\mu_{\mathsf{avg}} = \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{1}{K}\sum_{k=1}^{K} \mu_{\mathsf{b}}^k(s,a) \geq \min_{(s,a)\in\mathcal{S}\times\mathcal{A},k\in[K]} \mu_{\mathsf{b}}^k(s,a) = \mu_{\mathsf{min}}. \tag{20}$$

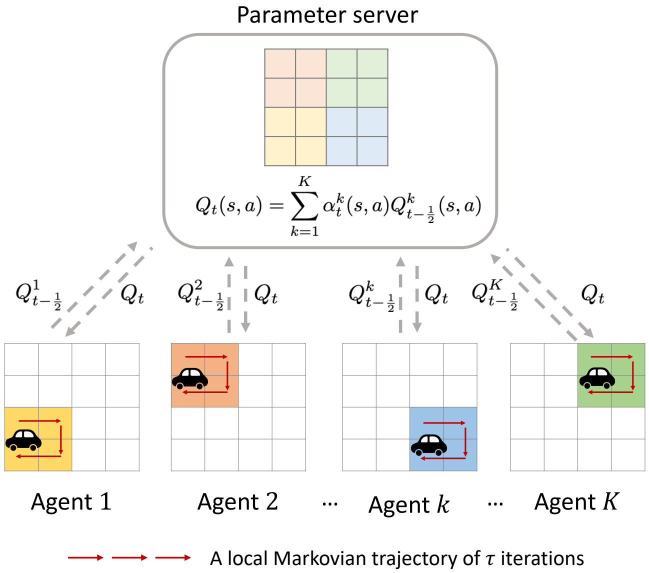

Figure 1: Federated asynchronous Q-learning with K agents and a parameter server. Each agent k performs τ local updates on its local Q-table along a Markovian trajectory induced by behavior policy π_{b}^k and sends the Q-table to the server. The server averages and synchronizes the local Q-tables every τ iterations. For importance averaging, the agents additionally send the number of visits over all the state-action pairs within each synchronization period, which is not pictured.

Last but not least, we measure the heterogeneity of the stationary state-action distributions across agents by

$$C_{\mathsf{het}} := \max_{k \in [K]} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\mu_{\mathsf{b}}^k(s,a)}{\frac{1}{K} \sum_{k=1}^{K} \mu_{\mathsf{b}}^k(s,a)}, \tag{21}$$

which satisfies $1 \leq C_{\mathsf{het}} \leq \min\{K, 1/\mu_{\mathsf{avg}}\}$, and in particular, $C_{\mathsf{het}} = 1$ when $\mu_b^k = \mu_b$ are all equal.

Review: asynchronous Q-learning with a single agent. Recall the update rule of asynchronous Q-learning with a single agent, where at each iteration $t \geq 1$, upon receiving a transition (s_{t-1}, a_{t-1}, s_t), the Q-estimate is updated via

$$Q_t(s,a) = \begin{cases} (1-\eta) Q_{t-1}(s,a) + \eta \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t, a') \right), & \text{if } (s,a) = (s_{t-1}, a_{t-1}), \\ Q_t(s,a) & \text{otherwise}, \end{cases} \tag{22}$$

where η denotes the learning rate and V_t is defined in (11). The sample complexity of asynchronous Q-learning has been recently investigated in Li et al. (2023, 2021b); Qu and Wierman (2020).

4.2 Algorithm description

Similar to the synchronous setting, we describe a federated asynchronous Q-learning algorithm, called FedAsynQ (see Algorithm 2), that learns the optimal Q-function by periodically averaging the local Q-estimates with the aid of a central server. See Figure 1 for an illustration. Inheriting the notation of Q_t^k and V_t^k from the synchronous setting (cf. (11)), FedAsynQ proceeds as follows in the rest of the t-th iteration.

1. *Local updates:* Each agent k samples a transition $(s_{t-1}^k, a_{t-1}^k, r_{t-1}^k, s_t^k)$ from its Markovian trajectory generated by the behavior policy π_{b}^k according to (15) and updates a *single* entry of its local Q-estimate Q_{t-1}^k:

$$Q_{t-\frac{1}{2}}^k(s,a) = \begin{cases} (1-\eta) Q_{t-1}^k(s,a) + \eta \big(r_{t-1}^k + \gamma V_{t-1}^k(s_t^k) \big) & \text{if } (s,a) = (s_{t-1}^k, a_{t-1}^k) \\ Q_{t-1}^k(s,a), & \text{otherwise} \end{cases}, \tag{23}$$

Algorithm 2: Federated Asynchronous Q-learning (FedAsynQ)

```
inputs: learning rate $\{\eta\}$, discount factor $\gamma$, number of agents $K$, synchronization period $\tau$, total number of iterations $T$.
initialization: $Q_0^k = Q_0$ for all $k \in [K]$.
for $t = 1, \cdots, T$ do
  for $k \in [K]$ do
    Draw action $a_{t-1}^k \sim \pi_{\mathsf{b}}^k(s_{t-1}^k)$, observe reward $r_{t-1}^k = r(s_{t-1}^k, a_{t-1}^k)$, and draw next state $s_t^k \sim P(\cdot \mid s_{t-1}^k, a_{t-1}^k)$.
    Compute $Q_{t-\frac{1}{2}}^k$ according to (23).
    Compute $Q_t^k$ according to (24).
  end for
end for
return: $Q_T(s,a) = \sum_{k=1}^K \alpha_T^k(s,a) Q_T^k(s,a)$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
```

where η denotes the learning rate.

2. *Periodic averaging:* The intermediate local estimates will be averaged every τ iterations, where $\tau \geq 1$ is the synchronization period. Here, we consider a more general weighted averaging scheme, where the updated estimate Q_t^k is:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_t^k(s,a) = \begin{cases} \sum_{k=1}^K \alpha_t^k(s,a) Q_{t-\frac{1}{2}}^k(s,a) & \text{if } t \equiv 0 \pmod{\tau} \\ Q_{t-\frac{1}{2}}^k(s,a) & \text{otherwise} \end{cases}, \tag{24}$$

where $\alpha_t^k = [\alpha_t^k(s,a)]_{s \in \mathcal{S}, a \in \mathcal{A}} \in [0,1]^{|\mathcal{S}||\mathcal{A}|}$ is an entry-wise weight assigned to agent k such that

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad \sum_{k=1}^K \alpha_t^k(s,a) = 1.$$

After a total of T iterations, FedAsynQ outputs a global Q-estimate $Q_T(s,a) = \sum_{k=1}^K \alpha_T^k(s,a) Q_T^k(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. In the subsections below, we provide two possible ways (equal and importance weighting) to choose α_t^k and their corresponding sample complexity analyses.

4.3 Performance guarantees with equal averaging

We begin with the most natural choice, which equally weights the local Q-estimates, that is,

$$\alpha_t^k(s,a) = \frac{1}{K}. \tag{25}$$

We call the resulting scheme FedAsynQ-EqAvg, which is also analyzed in Khodadadian et al. (2022). We have the following improved performance guarantee in the next theorem.

Theorem 2 (Finite-time convergence of FedAsynQ-EqAvg)**.** *Consider any given $\delta \in (0,1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Suppose that the initialization of* FedAsynQ-EqAvg *satisfies $0 \leq Q_0 \leq \frac{1}{1-\gamma}$. There exist some sufficiently large constant $c_T > 0$ and sufficiently small constant $c_\eta > 0$, such that with probability at least $1-\delta$, the output of* FedAsynQ-EqAvg *satisfies $\|Q_T - Q^\star\|_\infty \leq \varepsilon$, provided that the synchronization period τ, the sample size per agent T, and the learning rate η satisfy*

$$\tau_0 \leq \tau \leq \frac{1}{4\eta} \min\left\{\frac{1-\gamma}{4}, \frac{1}{K}\right\}, \tag{26a}$$

$$T \geq c_T \left(\frac{C_{\mathsf{het}}}{K \mu_{\mathsf{avg}} (1-\gamma)^5 \varepsilon^2} + T_0 \right) (\log((1-\gamma)^2 \varepsilon))^2 \log(TK) \log \frac{|\mathcal{S}||\mathcal{A}|T^2 K}{\delta}, \tag{26b}$$

$$\eta = c_\eta \min\left\{\frac{K(1-\gamma)^4\varepsilon^2}{C_{\mathsf{het}}}, \eta_0\right\} \frac{1}{\log(TK)\log\frac{|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}, \tag{26c}$$

where $\tau_0 = \frac{2176 t_{\mathsf{mix}}^{\mathsf{max}}}{\mu_{\mathsf{avg}}}\log 8K \log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}$, $T_0 = \frac{1}{\mu_{\mathsf{avg}}(1-\gamma)\eta_0}$, *and* $\eta_0 = \frac{\mu_{\mathsf{avg}}\min\{1-\gamma, K^{-1}\}}{t_{\mathsf{mix}}^{\mathsf{max}}}$, *independent of* ε.

Theorem 2 implies that to achieve an ε-accurate estimate (in the ℓ_∞ sense), the sample complexity per agent of FedAsynQ-EqAvg is no more than

$$\widetilde{O}\left(\frac{C_{\mathsf{het}}}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

for sufficiently small ε, when the burn-in cost T_0 — representing the impact of the mixing times — is amortized over time. A few implications are in order.

Linear speedup without full coverage. The sample complexity of FedAsynQ-EqAvg shows linear speedup with respect to the number of agents, which is especially pronounced when the local behavior policies are similar, i.e., $C_{\mathsf{het}} \approx 1$. Notably, the guarantee holds as long as all agents collectively cover the entire state-action space (i.e., $\mu_{\mathsf{avg}} > 0$), unveiling the benefit of heterogeneity in local behavior policies. This is surprising in view of the convergence guarantee provided in Khodadadian et al. (2022), which requires each agent visits the entire state-action space (i.e., $\mu_{\mathsf{min}} = 0$). Moreover, our sample complexity has sharpened dependency on nearly all problem-dependent parameters compared to the bound $\widetilde{O}\left(\frac{|\mathcal{S}|^2}{K\mu_{\mathsf{min}}^5(1-\gamma)^9\varepsilon^2}\right)$ obtained in Khodadadian et al. (2022) by at least a factor of

$$\frac{\mu_{\mathsf{avg}}|\mathcal{S}|^2}{C_{\mathsf{het}}\mu_{\mathsf{min}}^5(1-\gamma)^4} \geq \frac{|\mathcal{S}|^5|\mathcal{A}|^3}{(1-\gamma)^4}.$$

For $K = 1$, the bound nearly matches with the sharpest upper bound $\widetilde{O}\left(\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2}\right)$ for the single-agent case (Li et al., 2023) up to a factor of $1/(1-\gamma)$, when ignoring the burn-in cost.

Communication efficiency. To provide further insights on the communication complexity of FedAsynQ-EqAvg, consider the regime when ε is sufficiently small and the number of agents is sufficiently large such that $K \gtrsim \frac{1}{1-\gamma}$. By plugging the choice of the learning rate (26c) into the upper bound of τ in (26a), we can select the synchronization period as large as $\tau \asymp \frac{C_{\mathsf{het}}}{K^2(1-\gamma)^4\varepsilon^2}$ up to logarithmic factors, which ensures the communication complexity $C_{\mathsf{round}} = T/\tau$ is no more than $\widetilde{O}\left(\frac{K}{\mu_{\mathsf{avg}}(1-\gamma)}\right)$.

4.4 Performance guarantees with importance averaging

In the asynchronous setting, heterogeneous behavior policies induce local trajectories that cover the state-action space in a non-uniform manner. As a result, agents may update the Q-estimate for a state-action pair at different frequencies, resulting in noisier Q-estimates of state-action pairs that an agent rarely visits. Equally-weighted averaging of such local Q-estimates is not efficient, because the convergence speed to the optimal Q-function for each state-action pair is bottlenecked with the slowest converging agent that visits it least frequently. This is highlighted by the impact of the heterogeneity factor C_{het} in the sample complexity of FedAsynQ-EqAvg, which scales linearly with C_{het}, implying that increased heterogeneity among agents' trajectories may impede the convergence. For example, if only one agent exclusively visits a certain state-action pair (s, a) with probability one, while other agents never visit that particular state-action pair, the heterogeneity factor becomes $C_{\mathsf{het}} = K$ when $K \leq 1/\mu_{\mathsf{avg}}$, canceling out the linear speedup.

Our key idea to prevent such inefficiency is to increase the contribution of frequently updated local Q-estimates, which are likely to have smaller errors. By assigning a weight inversely proportional to the error of the corresponding local estimate, we can balance the heterogeneous training progress of the local estimates and obtain an average estimate with much lower error. Combining this idea with the property that the local error decreases exponentially with the number of local visits, we propose an importance averaging scheme FedAsynQ-ImAvg with weights given by

$$\alpha_t^k(s,a) = \frac{(1-\eta)^{-N_{t-\tau,t}^k(s,a)}}{\sum_{k'=1}^K (1-\eta)^{-N_{t-\tau,t}^{k'}(s,a)}} \tag{27}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $k \in [K]$, where $N_{t-\tau,t}^k(s,a)$ represents the number of iterations between $[t-\tau, t)$ when the agent k visits (s,a). The weights in (27) can be calculated at the server based on the number of visits to each state-action pair by the agents in one synchronization period. Therefore, each agent needs to send its $N_{t-\tau,t}^k(s,a)$ for each (s,a) along with its local Q-estimate, and FedAsynQ-ImAvg incurs twice the communication cost of FedAsynQ-EqAvg per iteration.

We have the following theorem on the finite-time convergence of FedAsynQ-ImAvg.

Theorem 3 (Finite-time convergence of FedAsynQ-ImAvg). *Consider any given $\delta \in (0,1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Suppose that the initialization of* FedAsynQ-ImAvg *satisfies $0 \le Q_0 \le \frac{1}{1-\gamma}$, and the synchronization period τ obeys*

$$\tau \le \frac{1}{4\eta} \min\left\{\frac{1-\gamma}{4}, \frac{1}{K}\right\}. \tag{28a}$$

There exist some sufficiently large constant $c_T > 0$ and sufficiently small constant $c_\eta > 0$, such that with probability at least $1-\delta$, the output of FedAsynQ-ImAvg *satisfies $\|Q_T - Q^\star\|_\infty \le \varepsilon$, provided that the sample size per agent T and the learning rate η satisfy*

$$T \ge c_T \left(\frac{1}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2} + \widetilde{T}_0\right) (\log((1-\gamma)^2\varepsilon))^2 \log(TK) \log\frac{|\mathcal{S}||\mathcal{A}|T^2K}{\delta}, \tag{28b}$$

$$\eta = c_\eta \min\left\{K(1-\gamma)^4\varepsilon^2, \widetilde{\eta}_0\right\} \frac{1}{\log(TK)\log\frac{|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}, \tag{28c}$$

where $\widetilde{T}_0 = \frac{1}{\mu_{\mathsf{avg}}(1-\gamma)\eta_0}$ and $\widetilde{\eta}_0 = \min\left\{\frac{1}{t_{\mathsf{mix}}^{\mathsf{max}}}, 1-\gamma, K^{-1}\right\}$, independent of ε.

Theorem 3 implies that to achieve an ε-accurate estimate (in the ℓ_∞ sense), the sample complexity per agent of FedAsynQ-ImAvg is no more than

$$\widetilde{O}\left(\frac{1}{K\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}\right)$$

for sufficiently small ε, when the burn-in cost $\widetilde{T}_0$ — representing the impact of the mixing times — is amortized over time. A few implications are in order.

Linear speedup without the curse of heterogeneity. The sample complexity of FedAsynQ-ImAvg is better than that of FedAsynQ-EqAvg, since it no longer depends on C_{het} which can be as large as $1/\mu_{\mathsf{avg}}$. FedAsynQ-ImAvg not only overcomes potential insufficient local coverage by exploiting the complementary coverage of agents' behavior policies, but also achieves linear speedup with respect to the number of agents without suffering from the potential performance degradation due to the associated statistical heterogeneity as in FedAsynQ-EqAvg. In fact, the performance of FedAsynQ-ImAvg matches with centralized Q-learning as if we collect and process all data trajectories at the central server, up to the burn-in cost and logarithmic factors.

Communication efficiency. To provide further insights on the communication complexity of FedAsynQ-ImAvg, consider again the regime when ε is sufficiently small and $K \gtrsim \frac{1}{1-\gamma}$. To minimize the communication frequency while preserving the sample efficiency, we again plug the choice of the learning rate (28c) into (28a) and select the synchronization period as large as $\tau \asymp \frac{1}{K^2(1-\gamma)^4\varepsilon^2}$ up to logarithmic factors. Then, this ensures the communication complexity $C_{\mathsf{round}} = T/\tau$ is no more than $\widetilde{O}\Big(\frac{K}{\mu_{\mathsf{avg}}(1-\gamma)}\Big)$.

5 Numerical experiments

In this section, we conduct numerical experiments to demonstrate the performance of the asynchronous Q-learning algorithms (FedAsynQ-EqAvg and FedAsynQ-ImAvg).

Experimental setup. Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ described in Figure 2, where $\mathcal{S} = \{0, 1\}$ and $\mathcal{A} = \{1, 2, \cdots, m\}$. The reward function r is set as $r(s = 1, a) = 1$ and $r(s = 0, a) = 0$ for any action $a \in \mathcal{A}$, and the discount factor is set as $\gamma = 0.9$. We now describe the transition kernel P. Here, we set the self-transitioning probabilities $p_a := P(0|0, a)$ and $q_a := P(1|1, a)$ uniformly at random from $[0.4, 0.6]$ for each $a \in \mathcal{A}$, and set the probability of transitioning to the other state as $P(1 - s|s, a) = 1 - P(s|s, a)$ for each $s \in \mathcal{S}$.

We evaluate the proposed federated asynchronous Q-learning algorithms on the above MDP with K agents selecting their behavior policies from $\Pi = \{\pi_1,\ \pi_2, \cdots,\ \pi_m\}$, where the i-th policy always chooses action i for any state, i.e., $\pi_i(i|s) = 1$ for all $s \in \mathcal{S}$. Here, we assign π_i to agent $k \in [K]$ if $i \equiv k \pmod m$. Note that if an agent has a behavior policy π_i, it can visit only two state-action pairs, $(s = 0, a = i)$ and $(s = 1, a = i)$, as described in Figure 2. Thus, each agent covers a subset of the state-action space, and at least $K = m$ agents are required to obtain local trajectories collectively covering the entire state-action space. Under this setting with $m = 20$, we run the algorithms for 100 simulations using samples randomly generated from the MDP and policies assigned to the agents. The Q-function is initialized with entries uniformly at random from $(0, \frac{1}{1-\gamma}]$ for each state-action pair.

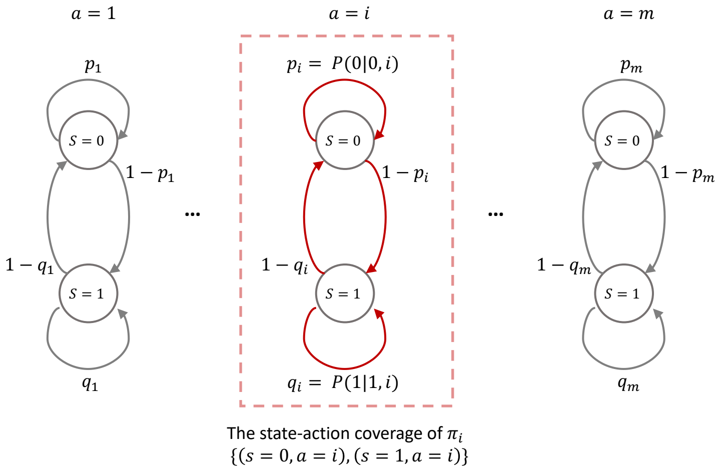

Figure 2: An illustration of the constructed synthetic MDP $\mathcal{M}$. The red arrows represent transitioning paths when action $a = i$ is taken in $s = 0$ and $s = 1$. A trajectory induced by π_i, which executes only action i for any state, can cover only two state-action pairs, $(s = 0, a = i)$ and $(s = 1, a = i)$.

Faster convergence of FedAsynQ-ImAvg. Figure 3 shows the normalized Q-estimate error $(1 - \gamma)\|Q_T - Q^\star\|_\infty$ with respect to the sample size T, with $K = 20$ and $\tau = 50$. Given the trajectories of agents collectively cover the entire state-action space, the global Q-estimates of both FedAsynQ-EqAvg and FedAsynQ-ImAvg converge to the optimal Q-function, yet at different speeds. Although FedAsynQ-EqAvg converges in the end, we can see that it converges much slower compared to FedAsynQ-ImAvg, because each entry of the Q-function is trained by only one agent while the other $m - 1$ agents never contribute useful information. However, the vacuous values of the $m - 1$ agents significantly slow down the global convergence under equal averaging.

Convergence speedup. Figure 4 demonstrates the impact of the number of agents on the convergence speed of FedAsynQ-EqAvg and FedAsynQ-ImAvg. It can be observed that there is indeed a speedup in terms of the number of agents K with respect to the squared ℓ_∞ error $\|Q_T - Q^\star\|_\infty^{-2}$, which is poised to scale linearly

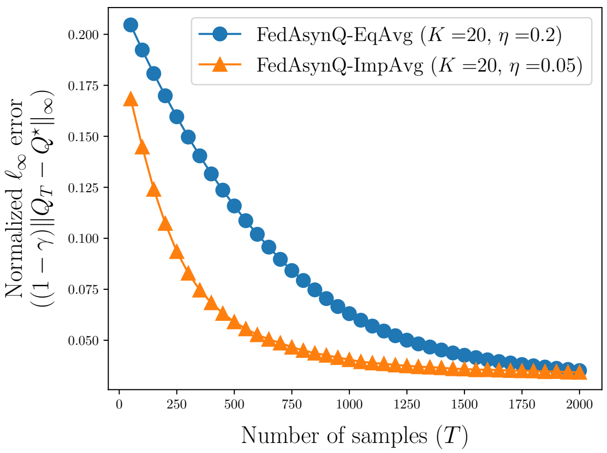

Figure 3: The normalized ℓ_∞ error of the Q-estimates $(1-\gamma)\|Q_T - Q^\star\|_\infty$ with respect to the number of samples T for both FedAsynQ-EqAvg and FedAsynQ-ImAvg, with $K = 20$ and $\tau = 50$. Here, the learning rates of FedAsynQ-ImAvg and FedAsynQ-EqAvg are set as $\eta = 0.05$ and $\eta = 0.2$, where each algorithm converges to the same error floor at the fastest speed, respectively.

with respect to the number of agents. In particular, the speedup is more rapid with FedAsynQ-ImAvg as K increases, while it increases much slower with FedAsynQ-EqAvg. This shows that FedAsynQ-ImAvg achieves much better convergence speedup in terms of the number of agents.

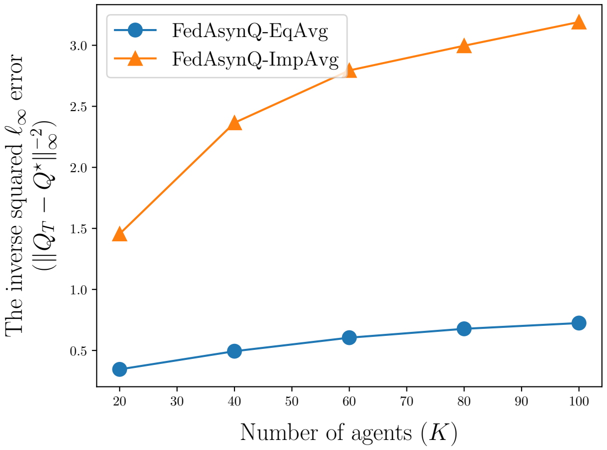

Figure 4: The inverse squared ℓ_∞ error $\|Q_T - Q^\star\|_\infty^{-2}$ with respect to the number of agents $K = 20, 40, 60, 80, 100$ for both FedAsynQ-EqAvg and FedAsynQ-ImAvg, with $T = 300$ and $\tau = 50$.

Communication efficiency. Figure 5 demonstrates the impact of the synchronization period τ on the convergence of FedAsynQ-ImAvg and FedAsynQ-EqAvg. With frequent averaging ($\tau = 1$), FedAsynQ-ImAvg slightly outperforms FedAsynQ-EqAvg, but there is no significant difference because the heterogeneity between local Q-functions after just one local update is very small. The performance of FedAsynQ-EqAvg degrades as we increase τ since FedAsynQ-EqAvg cannot cope with the increased heterogeneity between local Q-estimates as we increase the number of local steps. On the other end, the performance of FedAsynQ-ImAvg improves first (i.e., $\tau = 10,\ 25,\ 50$) as it balances the heterogeneity much better than FedAsynQ-EqAvg, but drops later if τ is too large (i.e., $\tau = 75,\ 100$) due to the high variance of the averaged Q-estimates.

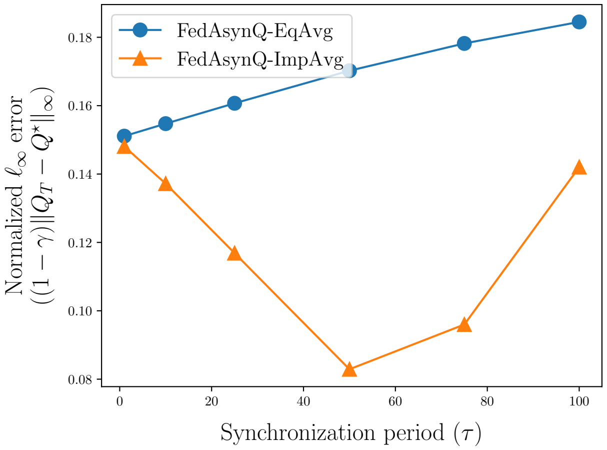

Figure 5: The normalized ℓ_∞ error of the Q-estimates $(1-\gamma)\|Q_T - Q^\star\|_\infty$ with respect to the synchronization period $\tau = 1, 10, 25, 50, 75, 100$ for both FedAsynQ-EqAvg and FedAsynQ-ImAvg, with $K = 20$ and $T = 300$.

6 Analysis outline

Let the matrix $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{A}|}$ represent the transition kernel of the underlying MDP, where $P(s,a) = P(\cdot|s,a)$ is the probability vector corresponding to the state transition at the state-action pair (s,a). For any vector $V \in \mathbb{R}^{|\mathcal{S}|}$, we define the variance parameter $\mathsf{Var}_{s,a}(V)$ with respect to the probability vector $P(s,a)$ as

$$\mathsf{Var}_{s,a}(V) \coloneqq \mathbb{E}_{s' \sim P(\cdot|s,a)} \big[V(s') - P(s,a)V\big]^2 = P(s,a)(V \circ V) - [P(s,a)V] \circ [P(s,a)V]. \tag{29}$$

Here, $\circ$ denotes the Hadamard product such that $a \circ b = [a_i b_i]_{i=1}^n$ for any vector $a = [a_i]_{i=1}^n, b = [b_i]_{i=1}^n \in \mathbb{R}^n$. With slight abuse of notation, we shall also assume $V^\star \in \mathbb{R}^{|\mathcal{S}|}$, $V_t^k \in \mathbb{R}^{|\mathcal{S}|}$, $Q^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $Q_t^k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $Q_{t+\frac{1}{2}}^k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $r \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ represent the corresponding functions in the matrix/vector form.

6.1 Basic facts

We first state a few basic facts that hold both for the synchronous and the asynchronous settings. It is easy to establish, by induction, that all iterates satisfy for all $1 \le k \le K$ and $t \ge 0$ that

$$0 \le Q_t^k \le \frac{1}{1-\gamma}, \qquad 0 \le V_t^k \le \frac{1}{1-\gamma}, \tag{30}$$

as long as $0 \le Q_0 = Q_0^k \le \frac{1}{1-\gamma}$; see a similar argument, e.g., in Li et al. (2023, Lemma 4). In addition, observe that

$$\|V_t^k - V^\star\|_\infty \le \|Q_t^k - Q^\star\|_\infty \tag{31}$$

since

$$\|V_t^k - V^\star\|_\infty = \max_{s \in \mathcal{S}} \Big| \max_{a \in \mathcal{A}} Q_t^k(s,a) - \max_{a \in \mathcal{A}} Q^\star(s,a) \Big| \le \max_{s \in \mathcal{S}, a \in \mathcal{A}} \big| Q_t^k(s,a) - Q^\star(s,a) \big| \le \|Q_t^k - Q^\star\|_\infty.$$

Letting Q_t be the average of the local Q-estimates at the end of the t-th iteration, i.e., $Q_t = \frac{1}{K}\sum_{k=1}^K Q_t^k$, it follows from (13) and (24) that for all $t \ge 0$ that

$$Q_t = \frac{1}{K}\sum_{k=1}^K Q_t^k = \frac{1}{K}\sum_{k=1}^K Q_{t-\frac{1}{2}}^k. \tag{32}$$

Denote the error between Q_t and $Q^\star$ by

$$\Delta_t = Q^\star - Q_t,$$

which is the quantity we aim to control. From (30), it holds immediately that for all $t \geq 0$,

$$\|\Delta_t\|_\infty \leq \frac{1}{1-\gamma}. \tag{33}$$

Next, we also introduce the following functions pertaining to periodic averaging. For any t,

- define $\iota(t) := \tau\lfloor \frac{t}{\tau} \rfloor$ as the most recent synchronization step until t;
- define $\phi(t) := \lfloor \frac{t}{\tau} \rfloor$ as the number of synchronization steps until t.

6.2 Proof outline of Theorem 1

Define the local empirical transition matrix at the t-th iteration $P_t^k \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ as

$$P_t^k((s,a),s') := \begin{cases} 1, & \text{if } s' = s_t^k(s,a) \\ 0, & \text{otherwise} \end{cases}, \tag{34}$$

then the local update rule (12) can be rewritten as

$$Q_{t-\frac{1}{2}}^k = (1-\eta)Q_{t-1}^k + \eta\left(r + \gamma P_t^k V_{t-1}^k\right). \tag{35}$$

The proof of Theorem 1 consists of the following steps.

Step 1: error decomposition. To analyze the error Δ_t, we first decompose the error into three terms, each of which can be bounded in a simple form. From (32), it follows that

$$\begin{aligned}\Delta_t = \frac{1}{K}\sum_{k=1}^K \left(Q^\star - Q_{t-\frac{1}{2}}^k\right) &\overset{\text{(i)}}{=} \frac{1}{K}\sum_{k=1}^K \left((1-\eta)(Q^\star - Q_{t-1}^k) + \eta(Q^\star - r - \gamma P_t^k V_{t-1}^k)\right) \\ &\overset{\text{(ii)}}{=} (1-\eta)\Delta_{t-1} + \eta\frac{\gamma}{K}\sum_{k=1}^K \left(PV^\star - P_t^k V_{t-1}^k\right) \\ &= (1-\eta)\Delta_{t-1} + \eta\frac{\gamma}{K}\sum_{k=1}^K \left(P - P_t^k\right)V_{t-1}^k + \eta\frac{\gamma}{K}\sum_{k=1}^K P\left(V^\star - V_{t-1}^k\right),\end{aligned}$$

where (i) follows from (35), and (ii) follows from Bellman's optimality equation $Q^\star = r + \gamma PV^\star$. By recursion over the above relation, we obtain

$$\Delta_t = \underbrace{(1-\eta)^t\Delta_0}_{=:E_t^1} + \underbrace{\eta\frac{\gamma}{K}\sum_{i=1}^t (1-\eta)^{t-i}\sum_{k=1}^K (P - P_i^k)V_{i-1}^k}_{=:E_t^2} + \underbrace{\eta\frac{\gamma}{K}\sum_{i=1}^t (1-\eta)^{t-i}\sum_{k=1}^K P(V^\star - V_{i-1}^k)}_{=:E_t^3}. \tag{36}$$

Here, the first term E_t^1 denotes the initialization error stemming from the disparity between the initial Q-values and the optimal Q-values, which diminishes exponentially throughout iterations. The second term, E_t^2, comprises a weighted sum accounting for the difference between the true transition probability and the realized transition in each iteration, where the difference arises from the randomness of transitions. Lastly, the final term, E_t^3, represents a weighted sum of value estimation errors from preceding iterations, which introduces a recursive relation.

Step 2: bounding the error terms. Now, we obtain a bound of each of the error terms in (36) separately.

- **Bounding $\|E_t^1\|_\infty$.** Using the fact that all agents start with the same initial Q-values, i.e., $Q_0^k = Q_0$, the first error term is bounded as follows:

$$\|E_t^1\|_\infty = (1-\eta)^t\|\Delta_0\|_\infty \leq \frac{(1-\eta)^t}{1-\gamma}, \tag{37}$$

where the last inequality follows from (33).

- **Bounding $\|E_t^2\|_\infty$.** Exploiting conditional independence across transitions in different iterations and applying Freedman's inequality (Freedman, 1975), the second error term is bounded using Lemma 1 below, whose proof is provided in Appendix B.1.

  Lemma 1. *For any given $\delta \in (0,1)$, the following holds*

$$\left\|E_t^2\right\|_\infty \leq \frac{8\gamma}{1-\gamma}\sqrt{\frac{\eta}{K}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} \tag{38}$$

  for all $0 \leq t \leq T$ with probability at least $1-\delta$, as long as η satisfies $\eta \leq \frac{K}{2}(\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})^{-1}$.

- **Bounding $\|E_t^3\|_\infty$.** For E_t^3, we obtain the following recursive relation using Lemma 2 below, whose proof is provided in Appendix B.2.

  Lemma 2. *Let β be any integer that satisfies $0 \leq \beta \leq \phi(T)$. For any given $\delta \in (0,1)$, the following holds*

$$\|E_t^3\|_\infty \leq \frac{2\gamma}{1-\gamma}(1-\eta)^{\beta\tau} + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \gamma(1+4\eta(\tau-1))\max_{\iota(t)-\beta\tau\leq i<t}\|\Delta_i\|_\infty$$

  for all $\beta\tau \leq t \leq T$ with probability at least $1-\delta$, as long as η satisfies $\tau\eta < 1/2$.

Step 3: solving a recursive relation. By putting all the bounds derived in the previous step together, for any $\beta\tau \leq t \leq T$, the total error bound can be written in a simple recursive form as follows:

$$\|\Delta_t\|_\infty \leq \zeta + \gamma(1+4\eta(\tau-1))\max_{\iota(t)-\beta\tau\leq i<t}\|\Delta_i\|_\infty \leq \zeta + \left(\frac{1+\gamma}{2}\right)\max_{\iota(t)-\beta\tau\leq i<t}\|\Delta_i\|_\infty, \tag{39}$$

where in the first inequality we introduce the short-hand notation

$$\zeta := \frac{4(1-\eta)^{\beta\tau}}{1-\gamma} + \frac{8\gamma}{1-\gamma}\sqrt{\frac{\eta}{K}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|KT}{\delta}}, \tag{40}$$

and the second inequality follows from the assumption $\tau - 1 \leq \frac{1-\gamma}{8\gamma\eta}$.

By invoking the recursive relation in (39) L times, where the choices of β and L will be made momentarily, it follows that for any $L\beta\tau \leq t \leq T$,

$$\begin{aligned}\|\Delta_t\|_\infty &\leq \sum_{i=0}^{L-1}\left(\frac{1+\gamma}{2}\right)^i\zeta + \left(\frac{1+\gamma}{2}\right)^L\max_{\iota(t)-L\beta\tau\leq i<t}\|\Delta_i\|_\infty \\ &\leq \frac{2}{1-\gamma}\zeta + \left(\frac{1+\gamma}{2}\right)^L\left(\frac{1}{1-\gamma}\right),\end{aligned} \tag{41}$$

where the second line uses the crude bound in (33).

Setting $\beta = \left\lfloor \frac{1}{\tau}\sqrt{\frac{(1-\gamma)T}{2\eta}} \right\rfloor$ and $L = \left\lceil \sqrt{\frac{\eta T}{1-\gamma}} \right\rceil$, which ensures $L\beta\tau \leq T$, and plugging their choices into (40) and (41) at $t = T$, we obtain that

$$\begin{aligned}\|\Delta_T\|_\infty &\leq \frac{8(1-\eta)^{\beta\tau}}{(1-\gamma)^2} + \frac{16\gamma}{(1-\gamma)^2}\sqrt{\frac{\eta}{K}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{32\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \left(\frac{1+\gamma}{2}\right)^L\left(\frac{1}{1-\gamma}\right) \\ &\leq \frac{32}{(1-\gamma)^2}\left(\exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right) + \gamma\sqrt{\frac{\eta}{K}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \gamma\eta\sqrt{\tau-1}\sqrt{\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}\right) \\ &\leq \frac{64}{(1-\gamma)^2}\left(\exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right) + \gamma\sqrt{\frac{\eta}{K}\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}\right),\end{aligned} \tag{42}$$

where the second line follows from

$$(1-\eta)^{\beta\tau} \leq \exp(-\eta\beta\tau) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right),$$

$$\left(\frac{1+\gamma}{2}\right)^L = \left(1-\frac{1-\gamma}{2}\right)^L \leq \exp\left(-\frac{(1-\gamma)}{2}L\right) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right),$$

and the third line follows from the choice of the synchronization period such that

$$\tau - 1 \leq \frac{1}{\eta}\min\left\{\frac{1-\gamma}{8\gamma}, \frac{1}{K}\right\}. \tag{43}$$

Thus, for any given $\varepsilon \in (0, \frac{1}{1-\gamma})$, we can guarantee that $\|\Delta_T\|_\infty \leq \varepsilon$ if

$$\begin{aligned} T &\geq c_T \frac{1}{K(1-\gamma)^5\varepsilon^2}(\log((1-\gamma)^2\varepsilon))^2 \log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}, \\ \eta &= c_\eta K(1-\gamma)^4\varepsilon^2 \frac{1}{\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \end{aligned} \tag{44}$$

for some sufficiently large c_T and sufficiently small c_η.

6.3 Proof outline of Theorem 2

For simplicity, we introduce the following notation. Let $\mathcal{U}^k_{v_1,v_2}(s,a)$ represent a set of iteration indices between $[v_1, v_2)$ for some $0 \leq v_1 \leq v_2 \leq T$ where agent k visits (s,a), i.e.,

$$\mathcal{U}^k_{v_1,v_2}(s,a) \coloneqq \left\{u \in [v_1, v_2) : (s^k_u, a^k_u) = (s,a)\right\},$$

and $N^k_{v_1,v_2}(s,a)$ denotes the number of visits of agent k on (s,a) during iterations between $[v_1, v_2)$, i.e.,

$$N^k_{v_1,v_2}(s,a) = |\mathcal{U}^k_{v_1,v_2}(s,a)|.$$

Define the local empirical transition matrix at the t-th iteration $P^k_t \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ as

$$P^k_t((s,a),s') \coloneqq \begin{cases} 1 & \text{if } (s,a,s') = (s^k_{t-1}, a^k_{t-1}, s^k_t) \\ 0 & \text{otherwise} \end{cases}. \tag{45}$$

Then the local update rule (23) can be rewritten as

$$Q^k_{t-\frac{1}{2}}(s,a) = \begin{cases} (1-\eta)Q^k_{t-1}(s,a) + \eta(r^k_{t-1} + \gamma P^k_t(s,a)V^k_{t-1}) & \text{if } (s,a) = (s^k_{t-1}, a^k_{t-1}) \\ Q^k_{t-1}(s,a), & \text{otherwise} \end{cases}. \tag{46}$$

The proof of Theorem 2 consists of the following steps.

Step 1: error decomposition. Consider any $0 \leq t \leq T$ such that $t \equiv 0 \pmod{\tau}$, i.e., t is a synchronization step. To analyze Δ_t, we first decompose the error for each $(s,a) \in \mathcal{S}\times\mathcal{A}$ as follows:

$$\begin{aligned} \Delta_t(s,a) &= \frac{1}{K}\sum_{k=1}^K (Q^\star(s,a) - Q^k_{t-\frac{1}{2}}(s,a)) \\ &= \left(\frac{1}{K}\sum_{k=1}^K (1-\eta)^{N^k_{t-\tau,t}(s,a)}\right)\Delta_{t-\tau}(s,a) \\ &\quad + \frac{\gamma}{K}\sum_{k=1}^K \sum_{u\in\mathcal{U}^k_{t-\tau,t}(s,a)} \eta(1-\eta)^{N^k_{u+1,t}(s,a)}(P(s,a) - P^k_{u+1}(s,a))V^k_u \end{aligned}$$

$$+\frac{\gamma}{K}\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{t-\tau,t}(s,a)}\eta(1-\eta)^{N^k_{u+1,t}(s,a)}P(s,a)(V^\star-V^k_u), \tag{47}$$

where we invoke the following recursive relation of the local error at iteration u such that $(s_{u-1},a_{u-1})=(s,a)$:

$$\begin{aligned}
&Q^\star(s,a)-Q^k_{u-\frac{1}{2}}(s,a)\\
&=(1-\eta)(Q^\star(s,a)-Q^k_{u-1}(s,a))+\eta(Q^\star(s,a)-r^k_{u-1}-\gamma P^k_u(s,a)V^k_{u-1})\\
&=(1-\eta)(Q^\star(s,a)-Q^k_{u-1}(s,a))+\eta(\gamma P(s,a)V^\star-\gamma P^k_u(s,a)V^k_{u-1})\\
&=(1-\eta)(Q^\star(s,a)-Q^k_{u-1}(s,a))+\gamma\eta(P(s,a)-P^k_u(s,a))V^k_{u-1}+\gamma P(s,a)(V^\star-V^k_{u-1}).
\end{aligned} \tag{48}$$

Here, the second equality follows from Bellman's optimality equation. Denoting

$$\lambda_{v_1,v_2}(s,a)\coloneqq\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{N^k_{v_1,v_2}(s,a)} \tag{49}$$

for any integer $0\le v_1\le v_2\le T$, we apply recursion to the relation (47) over the synchronization periods, and obtain

$$\begin{aligned}
&\Delta_t(s,a)\\
&=\left(\prod_{h=0}^{\phi(t)-1}\lambda_{h\tau,(h+1)\tau}(s,a)\right)\Delta_0(s,a)\\
&\quad+\sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right)\frac{\gamma}{K}\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\eta(1-\eta)^{N^k_{u+1,(h+1)\tau}(s,a)}(P(s,a)-P^k_{u+1}(s,a))V^k_u\\
&\quad+\sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right)\frac{\gamma}{K}\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\eta(1-\eta)^{N^k_{u+1,(h+1)\tau}(s,a)}P(s,a)(V^\star-V^k_u)\\
&=\underbrace{\omega_{0,t}(s,a)\Delta_0(s,a)}_{=:E^1_t(s,a)}+\underbrace{\gamma\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{0,t}(s,a)}\omega^k_{u,t}(s,a)(P(s,a)-P^k_{u+1}(s,a))V^k_u}_{=:E^2_t(s,a)}\\
&\quad+\underbrace{\gamma\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{0,t}(s,a)}\omega^k_{u,t}(s,a)P(s,a)(V^\star-V^k_u)}_{=:E^3_t(s,a)},
\end{aligned} \tag{50}$$

which is decomposed in a similar manner as (36). Here, we define

$$\omega_{0,t}(s,a)\coloneqq\prod_{h=0}^{\phi(t)-1}\lambda_{h\tau,(h+1)\tau}(s,a), \tag{51a}$$

$$\omega^k_{u,t}(s,a)\coloneqq\frac{1}{K}\eta(1-\eta)^{N^k_{u+1,(\phi(u)+1)\tau}(s,a)}\prod_{l=\phi(u)+1}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a). \tag{51b}$$

We record the following useful lemma whose proof is provided in Appendix C.2.

Lemma 3. *Consider integers v_1 and v_2 such that $0\le v_1\le v_2\le t\le T$, where $t\equiv 0$ (mod τ), and a state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$. Suppose that $\eta\tau\le 1$. The parameters defined in (51) satisfy*

$$\lambda_{v_1,v_2}(s,a)\le\exp\left(-\frac{\eta}{2K}\sum_{k=1}^{K}N^k_{v_1,v_2}(s,a)\right), \tag{52a}$$

$$\omega_{0,t}(s,a) + \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{0,t}^k(s,a)} \omega_{u,t}^k(s,a) = 1, \tag{52b}$$

$$\sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{0,h'\tau}^k(s,a)} \omega_{u,t}^k(s,a) \le \exp\left(-\frac{\eta}{2K}\sum_{k=1}^{K} N_{h'\tau,t}^k(s,a)\right), \quad \forall 0 \le h' \le \phi(t), \tag{52c}$$

$$\sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{0,t}^k(s,a)} (\omega_{u,t}^k(s,a))^2 \le \frac{2\eta}{K}. \tag{52d}$$

Step 2: bounding the error terms. Here, we derive the bound of the error terms in (50) separately for all the state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$.

- **Bounding $|E_t^1(s,a)|$.** Using the initialization condition that $Q_0(s,a) = Q_0^k(s,a)$ for every agent $k \in [K]$, we bound the first term for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows:

  $$|E_t^1(s,a)| \le \omega_{0,t}(s,a)(\|Q_0\|_\infty + \|Q^\star\|_\infty) \overset{\text{(i)}}{\le} \frac{2\omega_{0,t}(s,a)}{1-\gamma} \overset{\text{(ii)}}{\le} \frac{2}{1-\gamma}\exp\left(-\frac{\eta\mu_{\mathsf{avg}}t}{8}\right), \tag{53}$$

  where (i) holds because $\|Q_0\|_\infty, \ \|Q^\star\|_\infty \le \frac{1}{1-\gamma}$ (cf. (30)) and (ii) follows from the fact that

  $$\omega_{0,t}(s,a) \le \exp\left(-\frac{\eta}{2K}\sum_{k=1}^{K} N_{0,t}^k(s,a)\right) \le \exp\left(-\frac{\eta\mu_{\mathsf{avg}}t}{8}\right), \tag{54}$$

  where the first inequality holds according to (52a) of Lemma 3, and the last inequality follows from the fact that $\sum_{k=1}^{K} N_{0,t}^k(s,a) \ge \frac{K\mu_{\mathsf{avg}}t}{4}$ for all $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [T]$ at least with probability $1-\delta$ according to Lemma 10 and the union bound, as long as $t \ge t_{\mathsf{th}}$.

- **Bounding $|E_t^2(s,a)|$.** By carefully treating the statistical dependency via a decoupling argument and applying Freedman's inequality, we can obtain the following bound, whose proof is provided in Appendix C.3.

  Lemma 4. *For any given $\delta \in (0,1)$, the following holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $1 \le t \le T$:*

  $$\left|E_t^2(s,a)\right| \le \frac{7241\gamma}{(1-\gamma)}\sqrt{\frac{C_{\mathsf{het}}\eta}{K}\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \tag{55}$$

  with probability at least $1-4\delta$, as long as $\tau \ge t_{\mathsf{th}}$ and $\frac{3}{T} \le \eta \le \min\left\{\frac{1}{16\tau}, \frac{1}{4\tau K}, \frac{1}{128KC_{\mathsf{het}}\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}\right\}$.

- **Bounding $|E_t^3(s,a)|$.** For E_t^3, we can obtain the following recursive relation, whose proof is provided in Appendix C.4.

  Lemma 5. *Let β be any integer that satisfies $0 < \beta \le \phi(T)$. For any given $\delta \in (0,1)$, the following holds*

  $$|E_t^3(s,a)| \le \frac{2\gamma}{1-\gamma}\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right) + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}} + \frac{1+\gamma}{2}\max_{\phi(t)-\beta\le h\le\phi(t)-1}\|\Delta_{h\tau}\|_\infty, \tag{56}$$

  for all $\beta\tau \le t \le T$ with probability at least $1-\delta$, as long as $\beta\tau \ge t_{\mathsf{th}}$ and $\eta \le \min\{\frac{1-\gamma}{4\gamma\tau}, \frac{1}{2\tau}\}$.

Step 3: solving a recursive relation. By putting all the bounds derived in the previous step together, for any $\beta\tau \le t \le T$, the total error bound can be written in a simple recursive form as follows:

$$\|\Delta_t\|_\infty \le \theta + \frac{1+\gamma}{2} \max_{\phi(t)-\beta \le h \le \phi(t)-1} \|\Delta_{h\tau}\|_\infty, \tag{57}$$

where we define

$$\theta := \frac{4}{1-\gamma} \exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right) + \frac{7241\gamma}{(1-\gamma)} \sqrt{\frac{C_{\mathsf{het}}\eta}{K} \log(TK) \log \frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma} \sqrt{\log \frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}}. \tag{58}$$

Then, by invoking the recursive relation for L_1 times, where the choices of β and L_1 will be made momentarily, it follows that for any $L_1\beta\tau \le t \le T$,

$$\|\Delta_t\|_\infty \le \sum_{l=0}^{L_1-1} \left(\frac{1+\gamma}{2}\right)^l \theta + \left(\frac{1+\gamma}{2}\right)^{L_1} \max_{\phi(t)-\beta L \le i \le \phi(t)-1} \|\Delta_{i\tau}\|_\infty \le \frac{2}{1-\gamma} \left(\theta + \left(\frac{1+\gamma}{2}\right)^{L_1}\right), \tag{59}$$

where the last inequality follows from (33).

Setting $\beta = \left\lfloor \frac{1}{\tau}\sqrt{\frac{2(1-\gamma)T}{\mu_{\mathsf{avg}}\eta}} \right\rfloor$ and $L_1 = \left\lceil \frac{1}{2}\sqrt{\frac{\mu_{\mathsf{avg}}\eta T}{(1-\gamma)}} \right\rceil$, which ensures $L_1\beta\tau \le T$, and plugging the choices into (58) and (59) at $t = T$, we obtain

$$\begin{aligned}
\|\Delta_T\|_\infty \; &\le \frac{8\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right)}{(1-\gamma)^2} + \frac{14481\gamma}{(1-\gamma)^2} \sqrt{\frac{C_{\mathsf{het}}\eta}{K} \log(TK) \log \frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2} \sqrt{\log \frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}} + \frac{2}{1-\gamma}\left(\frac{1+\gamma}{2}\right)^L \\
&\le \frac{16}{(1-\gamma)^2} \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{8}\right) + \frac{14481\gamma}{(1-\gamma)^2} \sqrt{\frac{C_{\mathsf{het}}\eta}{K} \log(TK) \log \frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2} \sqrt{\log \frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}} \\
&\le \frac{14497}{(1-\gamma)^2} \left(\exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{8}\right) + \gamma\sqrt{\frac{C_{\mathsf{het}}\eta}{K} \log(TK) \log \frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}\right),
\end{aligned} \tag{60}$$

where the second line follows from

$$\begin{aligned}
\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right) &\le \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{8}\right), \\
\left(\frac{1+\gamma}{2}\right)^{L_1} &= \left(1 - \frac{1-\gamma}{2}\right)^{L_1} \le \exp\left(-\frac{1-\gamma}{2}L_1\right) \le \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right),
\end{aligned}$$

and the third line follows from the choice of the synchronization period such that

$$t_{\mathsf{th}} \le \tau \le \frac{1}{4\eta} \min\left\{\frac{1-\gamma}{4}, \frac{1}{K}\right\}. \tag{61}$$

Thus, for any given $\varepsilon \in (0, \frac{1}{1-\gamma}]$, we can guarantee that $\|\Delta_T\|_\infty \le \varepsilon$ if

$$\begin{aligned}
T &\ge c_T(\log((1-\gamma)^2\varepsilon))^2 \log(TK) \log \frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta} \frac{1}{\mu_{\mathsf{avg}}} \max\left\{\frac{C_{\mathsf{het}}}{K(1-\gamma)^5\varepsilon^2}, \frac{t_{\mathsf{mix}}^{\mathsf{max}}}{\mu_{\mathsf{avg}}(1-\gamma)\min\{1-\gamma, K^{-1}\}}\right\}, \\
\eta &= c_\eta \left(\log(TK) \log \frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}\right)^{-1} \min\left\{\frac{K(1-\gamma)^4\varepsilon^2}{C_{\mathsf{het}}}, \frac{\mu_{\mathsf{avg}}\min\{1-\gamma, K^{-1}\}}{t_{\mathsf{mix}}^{\mathsf{max}}}\right\}
\end{aligned}$$

for some sufficiently large c_T and sufficiently small c_η.

6.4 Proof outline of Theorem 3

The proof of Theorem 3 consists of the following steps.

Step 1: error decomposition. Consider any $0 \le t \le T$ such that $t \equiv 0 \pmod{\tau}$, i.e., t is a synchronization step. To analyze Δ_t, invoking the recursive relation of the local error (cf. (48)), we first decompose the error for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows:

$$\begin{aligned}
\Delta_t(s,a) &= \sum_{k=1}^{K} \alpha_t^k(s,a)(Q^\star(s,a) - Q_{t-\frac{1}{2}}^k(s,a)) \\
&= \left(\sum_{k=1}^{K} \alpha_t^k(s,a)(1-\eta)^{N_{t-\tau,t}^k(s,a)} \right) \Delta_{t-\tau}(s,a) \\
&\quad + \gamma \sum_{k=1}^{K} \alpha_t^k(s,a) \sum_{u \in \mathcal{U}_{t-\tau,t}^k(s,a)} \eta(1-\eta)^{N_{u+1,t}^k(s,a)} (P(s,a) - P_{u+1}^k(s,a)) V_u^k \\
&\quad + \gamma \sum_{k=1}^{K} \alpha_t^k(s,a) \sum_{u \in \mathcal{U}_{t-\tau,t}^k(s,a)} \eta(1-\eta)^{N_{u+1,t}^k(s,a)} P(s,a)(V^\star - V_u^k) \\
&= \left(\frac{K}{\sum_{k'=1}^{K} (1-\eta)^{-N_{t-\tau,t}^{k'}(s,a)}} \right) \Delta_{t-\tau}(s,a) \\
&\quad + \gamma \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{t-\tau,t}^k(s,a)} \frac{\eta(1-\eta)^{-N_{t-\tau,u+1}^k(s,a)}}{\sum_{k'=1}^{K} (1-\eta)^{-N_{t-\tau,t}^{k'}(s,a)}} (P(s,a) - P_{u+1}^k(s,a)) V_u^k \\
&\quad + \gamma \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{t-\tau,t}^k(s,a)} \frac{\eta(1-\eta)^{-N_{t-\tau,u+1}^k(s,a)}}{\sum_{k'=1}^{K} (1-\eta)^{-N_{t-\tau,t}^{k'}(s,a)}} P(s,a)(V^\star - V_u^k),
\end{aligned} \tag{62}$$

where the last line uses the definition of $\alpha_t^k(s,a)$ in (27). Denoting

$$\widetilde{\lambda}_{v_1,v_2}(s,a) := \frac{K}{\sum_{k=1}^{K} (1-\eta)^{N_{v_1,v_2}^k(s,a)}} \tag{63}$$

for any integer $0 \le v_1 \le v_2 \le T$, we apply recursion to the relation (62) over the synchronization period, and obtain

$$\begin{aligned}
&\Delta_t(s,a) \\
&= \left(\prod_{h=0}^{\phi(t)-1} \widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a) \right) \Delta_0(s,a) \\
&\quad + \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a) \right) \gamma \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \frac{\eta(1-\eta)^{-N_{h\tau,u+1}^k(s,a)}}{\sum_{k'=1}^{K} (1-\eta)^{-N_{h\tau,(h+1)\tau}^{k'}(s,a)}} (P(s,a) - P_{u+1}^k(s,a)) V_u^k \\
&\quad + \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a) \right) \gamma \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \frac{\eta(1-\eta)^{-N_{h\tau,u+1}^k(s,a)}}{\sum_{k'=1}^{K} (1-\eta)^{-N_{h\tau,(h+1)\tau}^{k'}(s,a)}} P(s,a)(V^\star - V_u^k) \\
&= \underbrace{\widetilde{\omega}_{0,t}(s,a)\Delta_0(s,a)}_{=:E_t^1(s,a)} + \underbrace{\gamma \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{0,t}^k(s,a)} \widetilde{\omega}_{u,t}^k(s,a)(P(s,a) - P_{u+1}^k(s,a)) V_u^k}_{=:E_t^2(s,a)}
\end{aligned}$$

$$+\gamma \underbrace{\sum_{k=1}^{K} \sum_{u\in\mathcal{U}_{0,t}^k(s,a)} \widetilde{\omega}_{u,t}^k(s,a) P(s,a)(V^\star - V_u^k)}_{=:E_t^3(s,a)}, \tag{64}$$

which is again decomposed similarly as (36). Here, we define

$$\widetilde{\omega}_{0,t}(s,a) := \prod_{h=0}^{\phi(t)-1} \widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a), \tag{65a}$$

$$\widetilde{\omega}_{u,t}^k(s,a) := \frac{\eta(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)}}{\sum_{k'=1}^K (1-\eta)^{-N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}} \left(\prod_{l=\phi(u)+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right). \tag{65b}$$

We record the following useful lemma whose proof is provided in Appendix C.5.

Lemma 6. *Consider any integers $0 \le v_1 \le v_2 \le t \le T$ where $t \equiv 0 \pmod{\tau}$ and any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. Suppose that $\eta\tau \le 1$, then the parameters defined in (65) satisfy*

$$\frac{1}{3K} \le \alpha_t^k(s,a) \le \frac{3}{K}, \tag{66a}$$

$$\widetilde{\omega}_{0,t}(s,a) \le (1-\eta)^{\frac{1}{K}\sum_{k=1}^K N_{0,t}^k(s,a)}, \tag{66b}$$

$$\widetilde{\omega}_{0,t}(s,a) + \sum_{k=1}^{K} \sum_{u\in\mathcal{U}_{0,t}^k(s,a)} \widetilde{\omega}_{u,t}^k(s,a) = 1, \tag{66c}$$

$$\sum_{k=1}^{K} \sum_{u\in\mathcal{U}_{0,h'\tau}^k(s,a)} \widetilde{\omega}_{u,t}^k(s,a) \le (1-\eta)^{\frac{1}{K}\sum_{k=1}^K N_{h'\tau,t}^k(s,a)}, \quad \forall 0 \le h' \le \phi(t), \tag{66d}$$

$$\sum_{k=1}^{K} \sum_{u\in\mathcal{U}_{0,t}^k(s,a)} (\widetilde{\omega}_{u,t}^k(s,a))^2 \le \frac{6\eta}{K}. \tag{66e}$$

Step 2: bounding the error terms. Here, we derive the bound of each error term in (64) separately for all the state-action pairs $(s,a) \in \mathcal{S}\times\mathcal{A}$.

- **Bounding $|E_t^1(s,a)|$.** Using the initialization condition that $Q_0(s,a) = Q_0^k(s,a)$ for every client $k \in [K]$, we bound the first term for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ as follows:

  $$|E_t^1(s,a)| \le \widetilde{\omega}_{0,t}(\|Q_0\|_\infty + \|Q^\star\|_\infty) \overset{\text{(i)}}{\le} \frac{2\widetilde{\omega}_{0,t}}{1-\gamma} \overset{\text{(ii)}}{\le} \frac{2}{1-\gamma}(1-\eta)^{\frac{1}{K}\sum_{k=1}^K N_{0,t}^k(s,a)} \overset{\text{(iii)}}{\le} \frac{2}{1-\gamma}(1-\eta)^{\frac{1}{4}\mu_{\mathsf{avg}}t}, \tag{67}$$

  where (i) holds because $\|Q_0\|_\infty,\ \|Q^\star\|_\infty \le \frac{1}{1-\gamma}$ (cf. (30)), (ii) follows from (66b) of Lemma 6, and (iii) holds for all $(s,a,t) \in \mathcal{S}\times\mathcal{A}\times[T]$ with probability at least $1-\delta$ according to Lemma 10, as long as $t \ge t_{\mathsf{th}}$.

- **Bounding $|E_t^2(s,a)|$.** By carefully treating the statistical dependency via a decoupling argument and applying Freedman's inequality, we can obtain the following bound, whose proof is provided in Appendix C.6.

  Lemma 7. *For any given $\delta \in (0,1)$, the following holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $1 \le t \le T$:*

  $$\left|E_t^2(s,a)\right| \le \frac{2064\gamma}{(1-\gamma)}\sqrt{\frac{\eta}{K}\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \tag{68}$$

  with probability at least $1-2\delta$, as long as

  $$\frac{3}{T} < \eta \le \min\left\{\frac{1}{16\tau}, \frac{K}{256\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}, \frac{1}{34816 t_{\mathsf{mix}}^{\mathsf{max}}\log(8K)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}\right\}.$$

- **Bounding $|E_t^3(s,a)|$.** For E_t^3, similarly to Lemma 5, we can obtain the following recursive relation, whose proof is provided in Appendix C.7.

  Lemma 8. *Let β be any integer that satisfies $\frac{t_{\mathsf{th}}}{\tau} \leq \beta \leq \phi(T)$. For any given $\delta \in (0,1)$, the following holds*

  $$|E_t^3(s,a)| \leq \frac{2(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}}}{1-\gamma} + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}} + \frac{1+\gamma}{2}\max_{\phi(t)-\beta\leq h\leq\phi(t)-1}\|\Delta_{h\tau}\|_\infty, \tag{69}$$

  for all $\beta\tau \leq t \leq T$ with probability at least $1-\delta$, as long as $\eta \leq \min\{\frac{1-\gamma}{4\gamma\tau}, \frac{1}{2\tau}\}$.

Step 3: solving a recursive relation. By putting all the bounds derived in the previous step together, for any $\beta\tau \leq t \leq T$, the total error bound can be written in a simple recursive form as follows:

$$\|\Delta_t\|_\infty \leq \theta + \frac{1+\gamma}{2}\max_{\phi(t)-\beta\leq h\leq\phi(t)-1}\|\Delta_{h\tau}\|_\infty, \tag{70}$$

where we define

$$\widetilde{\theta} := \frac{4}{1-\gamma}(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}} + \frac{2064\gamma}{(1-\gamma)}\sqrt{\frac{\eta}{K}\log{(TK)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}}. \tag{71}$$

Then, by invoking the recursive relation for L_2 times, where the choices of β and L_2 will be made momentarily, it follows that for any $L_2\beta\tau \leq t \leq T$,

$$\|\Delta_t\|_\infty \leq \sum_{l=0}^{L_2-1}\left(\frac{1+\gamma}{2}\right)^l\widetilde{\theta} + \left(\frac{1+\gamma}{2}\right)^{L_2}\max_{\phi(t)-\beta L\leq i\leq\phi(t)-1}\|\Delta_{i\tau}\|_\infty \leq \frac{2}{1-\gamma}\left(\theta + \left(\frac{1+\gamma}{2}\right)^{L_2}\right), \tag{72}$$

where the last inequality follows from (33).

Setting $L_2 = \left\lceil \frac{1}{2}\sqrt{\frac{\mu_{\mathsf{avg}}\eta T}{(1-\gamma)}}\right\rceil$ and $\beta = \left\lfloor \frac{1}{\tau}\sqrt{\frac{2(1-\gamma)T}{\mu_{\mathsf{avg}}\eta}}\right\rfloor$, which ensures $L_2\beta\tau \leq T$, and plugging the choices into (71) and (72) at $t = T$, we obtain

$$\begin{aligned}
\|\Delta_T\|_\infty &\leq \frac{8(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}}}{(1-\gamma)^2} + \frac{4128\gamma}{(1-\gamma)^2}\sqrt{\frac{\eta}{K}\log{(TK)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}} + \frac{2}{1-\gamma}\left(\frac{1+\gamma}{2}\right)^{L_2} \\
&\leq \frac{16}{(1-\gamma)^2}\exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right) + \frac{4128\gamma}{(1-\gamma)^2}\sqrt{\frac{\eta}{K}\log{(TK)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}} \\
&\leq \frac{4144}{(1-\gamma)^2}\left(\exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right) + \gamma\sqrt{\frac{\eta}{K}\log{(TK)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}\right),
\end{aligned} \tag{73}$$

where the second line follows from

$$(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}} \leq \exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{4}\right) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right),$$

$$\left(\frac{1+\gamma}{2}\right)^{L_2} = \left(1-\frac{1-\gamma}{2}\right)^{L_2} \leq \exp\left(-\frac{1-\gamma}{2}L_2\right) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right),$$

and the third line follows from the choice of the synchronization period such that

$$\tau \leq \frac{1}{4\eta}\min\left\{\frac{1-\gamma}{4}, \frac{1}{K}\right\}. \tag{74}$$

Thus, for any given $\varepsilon \in (0, \frac{1}{1-\gamma})$, optimizing η and T to make (73) bounded by ε and recalling $\beta\tau \geq t_{\mathsf{th}}$, we can guarantee that $\|\Delta_T\|_\infty \leq \varepsilon$ if

$$
\begin{aligned}
&T \geq c_T(\log((1-\gamma)^2\varepsilon))^2 \log(TK) \log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta} \frac{1}{\mu_{\mathsf{avg}}} \max\left\{\frac{1}{K(1-\gamma)^5\varepsilon^2}, \frac{t_{\mathsf{mix}}^{\mathsf{max}}}{(1-\gamma)}, \frac{1}{(1-\gamma)\min\{1-\gamma, K^{-1}\}}\right\}, \\
&\eta = c_\eta \min\left\{K(1-\gamma)^4\varepsilon^2 \frac{1}{\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}, \frac{1}{\mu_{\mathsf{avg}} t_{\mathsf{th}}}, \frac{1}{t_{\mathsf{mix}}^{\mathsf{max}} \log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}\right\} \\
&\quad = c_\eta \left(\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}\right)^{-1} \min\left\{K(1-\gamma)^4\varepsilon^2, \frac{1}{t_{\mathsf{mix}}^{\mathsf{max}}}, \min\left\{1-\gamma, K^{-1}\right\}\right\}
\end{aligned}
$$

for some sufficiently large c_T and sufficiently small c_η.

7 Discussions

We presented a sample complexity analysis of federated Q-learning in both synchronous and asynchronous settings. Our sample complexity not only leads to linear speedup with respect to the number of agents, but also significantly improves the dependencies on other salient problem parameters over the prior art. For federated asynchronous Q-learning, we proposed a novel importance averaging scheme that weighs the agents' local Q-estimates according to the number of visits to each state-action pair. This allows agents to leverage the blessing of heterogeneity of their local behavior policies and collaboratively learn the optimal Q-function that otherwise would not be possible, without requiring each individual agent to cover the entire state-action space. Looking ahead, this work opens up many exciting future directions, some outlined below.

- *Improved sample complexity.* While our sample complexity bounds are near-optimal with respect to the size of the state-action space, it is still sub-optimal with respect to the effective horizon length as well as the mixing time when benchmarking with the sample complexity in the single-agent setting (Li et al., 2023). It will be interesting to close this gap, and further improve the sample complexity with variance reduction techniques (Li et al., 2021b; Wainwright, 2019b) in the federated setting.
- *Understanding communication asynchrony across agents.* As a starting point, our work assumes that all agents communicate with the server in a synchronous manner to perform periodic averaging. However, in practical federated networks, some agents might be stragglers due to communication slowdowns, which warrants further investigation (Kairouz et al., 2021).
- *Other RL settings and function approximation.* Besides the infinite-horizon tabular MDPs, it will be of great interest to extend our analysis framework to other RL settings including but not limited to the finite-horizon setting, the average reward setting, heterogeneous environments across the agents (Yang et al., 2023), as well as incorporating function approximation.
- *Federated offline RL.* In many applications, offline RL is attracting a growing amount of interest, which aims to explore history datasets to improve the learned policy without exploration, e.g. via pessimistic variants of Q-learning (Shi et al., 2022). It will be appealing to develop federated offline Q-learning algorithms to enable learning from geographically distributed history datasets.

Acknowledgements

This work is supported in part by the grants NSF CCF-2007911, CCF-2106778, CNS-2148212, and ONR N00014-19-1-2404 to Y. Chi, NSF-CCF 2007834, CCF-2045694, CNS-2112471, and ONR N00014-23-1-2149 to G. Joshi, and the Hsu Chang Memorial Fellowship at Carnegie Mellon University to J. Woo.

References

Assran, M., Romoff, J., Ballas, N., Pineau, J., and Rabbat, M. (2019). Gossip-based actor-learner architectures for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, volume 32.

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al. (2019). Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, pages 374–388.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469.

Chen, T., Zhang, K., Giannakis, G. B., and Başar, T. (2021a). Communication-efficient policy gradient methods for distributed reinforcement learning. *IEEE Transactions on Control of Network Systems*, 9(2):917–929.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pages 8223–8234.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2021b). A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*.

Chen, Z., Zhou, Y., and Chen, R. (2022a). Multi-agent off-policy TDC with near-optimal sample and communication complexities. *Transactions on Machine Learning Research*.

Chen, Z., Zhou, Y., Chen, R.-R., and Zou, S. (2022b). Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis. In *International Conference on Machine Learning*, volume 162, pages 3794–3834. PMLR.

Doan, T., Maguluri, S., and Romberg, J. (2019). Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1626–1635.

Doan, T. T., Maguluri, S. T., and Romberg, J. (2021). Finite-time performance of distributed temporal-difference learning with linear function approximation. *SIAM Journal on Mathematics of Data Science*, 3(1):298–320.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. (2018). IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1406–1415.

Even-Dar, E. and Mansour, Y. (2003). Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25.

Fan, X., Ma, Y., Dai, Z., Jing, W., Tan, C., and Low, B. K. H. (2021). Fault-tolerant federated reinforcement learning with theoretical guarantee. In *Advances in Neural Information Processing Systems*, volume 34, pages 1007–1021.

Freedman, D. A. (1975). On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, pages 703–710.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Jin, H., Peng, Y., Yang, W., Wang, S., and Zhang, Z. (2022). Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 18–37.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Kearns, M. J. and Singh, S. P. (1999). Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems*, pages 996–1002.

Khodadadian, S., Sharma, P., Joshi, G., and Maguluri, S. T. (2022). Federated reinforcement learning: Linear speedup under Markovian sampling. In *International Conference on Machine Learning*, pages 10997–11057.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023). Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research.*

Li, G., Shi, L., Chen, Y., and Chi, Y. (2021a). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021b). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, volume 48, pages 1928–1937. PMLR.

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons.

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR.

Shen, H., Zhang, K., Hong, M., and Chen, T. (2022). Towards understanding asynchronous advantage actor-critic: convergence and linear speedup. *arXiv preprint arXiv:2012.15511.*

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, volume 162, pages 19967–20025. PMLR.

Sun, J., Wang, G., Giannakis, G. B., Yang, Q., and Yang, Z. (2020). Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 4485–4495. PMLR.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Szepesvári, C. (1998). The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pages 1064–1070.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202.

Wai, H.-T. (2020). On the convergence of consensus algorithms with Markovian noise and gradient bias. In *Conference on Decision and Control*, pages 4897–4902. IEEE.

Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*.

Wainwright, M. J. (2019b). Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*.

Wang, G., Lu, S., Giannakis, G., Tesauro, G., and Sun, J. (2020a). Decentralized TD tracking with linear function approximation and its finite-time analysis. In *Advances in Neural Information Processing Systems*, volume 33.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. (2020b). Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Wu, Z., Shen, H., Chen, T., and Ling, Q. (2021). Byzantine-resilient decentralized policy evaluation with linear function approximation. *IEEE Transactions on Signal Processing*, 69:3839–3853.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022). The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*.

Yang, K., Yang, L., and Du, S. (2021). Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR.

Yang, T., Cen, S., Wei, Y., Chen, Y., and Chi, Y. (2023). Federated natural policy gradient methods for multi-task reinforcement learning. *arXiv preprint arXiv:2311.00201*.

Zeng, S., Doan, T. T., and Romberg, J. (2021). Finite-time analysis of decentralized stochastic approximation with applications in multi-agent and multi-task learning. In *Conference on Decision and Control*, pages 2641–2646. IEEE.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33.

A Preliminaries

We record a few useful inequalities that will be used throughout our analysis. To start with, our analysis leverages Freedman's inequality (Freedman, 1975), which we record a user-friendly version as follows.

Theorem 4 (Theorem 6 in Li et al. (2023))**.** *Suppose that $Y_n = \sum_{k=1}^{n} X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \leq R \qquad \text{and} \qquad \mathbb{E}\left[X_k \mid \{X_j\}_{j:j<k}\right] = 0 \qquad \text{for all } k \geq 1.$$

Define

$$W_n \coloneqq \sum_{k=1}^{n} \mathbb{E}_{k-1}\left[X_k^2\right],$$

where we write $\mathbb{E}_{k-1}$ for the expectation conditional on $\{X_j\}_{j:j<k}$. Then for any given $\sigma^2 \geq 0$, one has

$$\mathbb{P}\left\{|Y_n| \geq \tau \text{ and } W_n \leq \sigma^2\right\} \leq 2\exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right). \tag{75}$$

In addition, suppose that $W_n \leq \sigma^2$ holds deterministically. For any positive integer $m \geq 1$, with probability at least $1-\delta$ one has

$$|Y_n| \leq \sqrt{8 \max\left\{W_n, \frac{\sigma^2}{2^m}\right\} \log \frac{2m}{\delta}} + \frac{4}{3} R \log \frac{2m}{\delta}. \tag{76}$$

Another useful relation concerns the concentration of empirical distributions of uniformly ergodic Markov chains, which is rephrased from Li et al. (2021b).

Lemma 9 ((Li et al., 2021b, Lemma 8))**.** *Consider any time homogeneous and uniformly ergodic Markov chain $(X_0, X_1, X_2, \ldots)$ with transition kernel P, finite state space $\mathcal{X}$, and stationary distribution μ. Let t_{mix} be the mixing time of the Markov chain and μ_{min} be the minimum entry of the stationary distribution μ. Consider any $0 < \delta < 1$. For any $x \in \mathcal{X}$, if $t \geq \frac{443 t_{\mathsf{mix}}}{\nu} \log \frac{4|\mathcal{X}|}{\delta}$ for $\nu \geq \mu(x)$, then*

$$\forall y \in \mathcal{X}: \quad \mathbb{P}_{X_1 = y}\left\{\left|\sum_{i=1}^{t} \mathbb{1}\{X_i = x\} - t\mu(x)\right| \geq \frac{1}{2} t\nu\right\} \leq \frac{\delta}{|\mathcal{X}|}.$$

Remark 1. Lemma 9 is a slightly generalized version of in Li et al. (2021b, Lemma 8), where the concentration bound is characterized in terms of any given threshold $\nu \geq \mu(x)$, not scaling with the stationary distribution $\mu(x)$. It can be shown using the Bernstein's inequality for Markov chains (Paulin, 2015, Theorem 3.11) in the same manner as Li et al. (2021b, Lemma 8), except that the threshold is set to $\frac{\nu t}{2}$ instead of $\frac{\mu(x) t}{2}$. We omit further details for conciseness and refer interested readers to the proof in Li et al. (2021b).

In addition, we provide the concentration bound of the total number of visits of multiple agents agents with independent uniformly ergodic Markov chains, whose proof is provided in Appendix C.1. Denote

$$t_{\mathsf{th}}(s,a) := \frac{2176 t_{\mathsf{mix}}^{\mathsf{max}} \log 8K \log \frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}{\mu_{\mathsf{avg}}(s,a)} \quad \text{and} \quad t_{\mathsf{th}} := \frac{2176 t_{\mathsf{mix}}^{\mathsf{max}} \log 8K \log \frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}{\mu_{\mathsf{avg}}}. \tag{77}$$

Here, $\mu_{\mathsf{avg}}(s,a) := \frac{1}{K}\sum_{k=1}^{K} \mu_{\mathsf{b}}^k(s,a)$ is the average behavior policy over all agents.

Lemma 10. *Consider any $\delta \in (0,1)$. Under the asynchronous sampling, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $0 \leq u < v \leq T$ such that $v - u \geq t_{\mathsf{th}}(s,a)$, the following holds :*

$$\frac{1}{4}(v-u)K\mu_{\mathsf{avg}}(s,a) \leq \sum_{k=1}^{K} N_{u,v}^k(s,a) \leq 2(v-u)K\mu_{\mathsf{avg}}(s,a) \tag{78}$$

with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T^2}$.

B Proofs for federated synchronous Q-learning (Section 3)

Define the following actions

$$a^\star(s) = \arg\max_{a \in \mathcal{A}} Q^\star(s,a), \quad a_i^k(s) = \arg\max_{a \in \mathcal{A}} Q_i^k(s,a), \quad a_i(s) = \arg\max_{a \in \mathcal{A}} \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a) \tag{79}$$

for any state $s \in \mathcal{S}$, which will be useful throughout the proof.

B.1 Proof of Lemma 1

For notation simplicity, let $z_i^k(s,a) := \eta(1-\eta)^{t-i}(P(s,a) - P_i^k(s,a))V_{i-1}^k$, then the entries of $E_t^2 = [E_t^2(s,a)]$ can be written as

$$E_t^2(s,a) = \eta\frac{\gamma}{K}\sum_{i=1}^{t}(1-\eta)^{t-i}\sum_{k=1}^{K}(P(s,a) - P_i^k(s,a))V_{i-1}^k = \frac{\gamma}{K}\sum_{i=1}^{t}\sum_{k=1}^{K} z_i^k(s,a), \tag{80}$$

which we plan to bound by invoking Freedman's inequality (cf. Theorem 4) using the fact $z_i^k(s,a)$ is independent of the transition events of other agents $k' \neq k$ at i and has zero mean conditioned on the events before iteration i, i.e.,

$$\mathbb{E}[z_i^k(s,a)|V_{i-1}^K,\dots,V_{i-1}^1,\dots,V_0^K,\dots,V_0^1] = 0, \qquad \forall k \in [K],\ 1 \le i \le t. \tag{81}$$

Before applying Freedman's inequality, we first derive the following properties of the variable $z_i^k(s,a)$.

- First, we can bound

$$B_t(s,a) \coloneqq \max_{k\in[K],1\le i\le t} |z_i^k(s,a)| \le \max_{k\in[K],1\le i\le t} \eta\big(\|P(s,a)\|_1 + \|P_i^k(s,a)\|_1\big)\|V_{i-1}^k\|_\infty \le \frac{2\eta}{1-\gamma}, \tag{82}$$

  where the first inequality uses $(1-\eta)^{t-i} \le 1$, and the last inequality follows from $\|P(s,a)\|_1 \le 1$, $\|P_i^k(s,a)\|_1 \le 1$, and $\|V_{i-1}^k\|_\infty \le \frac{1}{1-\gamma}$ (cf. (30)).

- Next, we have

$$\begin{aligned} W_t(s,a) &\coloneqq \sum_{i=1}^t \sum_{k=1}^K \mathbb{E}\big[(z_i^k(s,a))^2 | V_{i-1}^K,\dots,V_{i-1}^1,\dots,V_0^K,\dots,V_0^1\big] \\ &= \sum_{i=1}^t \sum_{k=1}^K \mathsf{Var}\big(z_i^k(s,a)|V_{i-1}^K,\dots,V_{i-1}^1,\dots,V_0^K,\dots,V_0^1\big) \\ &= \sum_{i=1}^t \sum_{k=1}^K \eta^2(1-\eta)^{2(t-i)} \mathsf{Var}_{s,a}(V_{i-1}^k) \\ &\le \frac{2K}{(1-\gamma)^2} \sum_{i=1}^t \eta^2(1-\eta)^{2(t-i)} \le \frac{2\eta K}{(1-\gamma)^2} \coloneqq \sigma^2, \end{aligned} \tag{83}$$

  where we recall the definition of $\mathsf{Var}_{s,a}$ in (29). Here, the first inequality holds since

$$\mathsf{Var}_{s,a}(V_{i-1}^k) \le \|P(s,a)\|_1(\|V_{i-1}^k\|_\infty)^2 + (\|P(s,a)\|_1\|V_{i-1}^k\|_\infty)^2 \le \frac{2}{(1-\gamma)^2}$$

  and the last inequality follows from

$$\sum_{i=1}^t \eta^2(1-\eta)^{2(t-i)} \le \frac{\eta^2(1-(1-\eta)^{2t})}{1-(1-\eta)^2} \le \eta. \tag{84}$$

By substituting the above bounds (cf. (82) and (83)) and $m = 1$ into Freedman's inequality (see Theorem 4), it follows that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in [T]$,

$$\begin{aligned} \left|\sum_{i=1}^t \sum_{k=1}^K z_i^k(s,a)\right| &\le \sqrt{8 \max\{W_t(s,a), \frac{\sigma^2}{2^m}\} \log \frac{2m|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{4}{3} B_t(s,a) \log \frac{2m|\mathcal{S}||\mathcal{A}|T}{\delta} \\ &\le \sqrt{\frac{32\eta K}{(1-\gamma)^2} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{6\eta}{1-\gamma} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta} \\ &\le \frac{8\gamma}{1-\gamma} \sqrt{\frac{\eta}{K} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} \end{aligned} \tag{85}$$

with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$, where the last inequality holds under the assumption $\eta \le \frac{K}{2}(\log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta})^{-1}$. Applying the union bound over all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in [T]$ then completes the proof.

B.2 Proof of Lemma 2

For any $\beta\tau \le t \le T$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can decompose the entries of $E_t^3 = [E_t^3(s,a)]$ as

$$\begin{aligned}
|E_t^3(s,a)| &= \left| \frac{\eta\gamma}{K} \sum_{i=0}^{t-1} \sum_{k=1}^{K} (1-\eta)^{t-i-1} P(s,a)(V^\star - V_i^k) \right| \\
&\le \underbrace{\left| \frac{\eta\gamma}{K} \sum_{i=0}^{\iota(t)-\beta\tau-1} \sum_{k=1}^{K} (1-\eta)^{t-i-1} P(s,a)(V^\star - V_i^k) \right|}_{=:E_t^{3a}(s,a)} + \underbrace{\left| \frac{\eta\gamma}{K} \sum_{i=\iota(t)-\beta\tau}^{t-1} \sum_{k=1}^{K} (1-\eta)^{t-i-1} P(s,a)(V^\star - V_i^k) \right|}_{=:E_t^{3b}(s,a)}. \qquad (86)
\end{aligned}$$

We shall bound these two terms separately.

Step 1: bounding $E_t^{3a}(s,a)$. First, the bound of E_t^{3a} is obtained as follows:

$$\begin{aligned}
E_t^{3a}(s,a) &\le \eta \frac{\gamma}{K} \sum_{k=1}^{K} \sum_{i=0}^{\iota(t)-\beta\tau-1} (1-\eta)^{t-i} \|P(s,a)\|_1 (\|V^\star\|_\infty + \|V_i^k\|_\infty) \\
&\le \frac{2\eta\gamma}{1-\gamma} \sum_{i=0}^{\iota(t)-\beta\tau-1} (1-\eta)^{t-i-1} \le \frac{2\gamma}{1-\gamma}(1-\eta)^{\beta\tau}, \qquad (87)
\end{aligned}$$

where the second inequality holds due to the fact that $\|P(s,a)\|_1 \le 1$ and $\|V^\star\|_\infty \le \frac{1}{1-\gamma}$, $\|V_i^k\|_\infty \le \frac{1}{1-\gamma}$, and the last inequality follows from

$$\sum_{i=0}^{\iota(t)-\beta\tau-1} (1-\eta)^{t-i-1} \le (1-\eta)^{\beta\tau} + (1-\eta)^{\beta\tau+1} + \ldots + (1-\eta)^{t-1} \le \frac{(1-\eta)^{\beta\tau}}{1-(1-\eta)} \le \frac{(1-\eta)^{\beta\tau}}{\eta}.$$

Step 2: decomposing the bound on $E_t^{3b}(s,a)$. Next, $E_t^{3b}(s,a)$ can be bounded as follows

$$\begin{aligned}
E_t^{3b}(s,a) &= \left| \frac{\eta\gamma}{K} \sum_{i=\iota(t)-\beta\tau}^{t-1} \sum_{k=1}^{K} (1-\eta)^{t-i-1} P(s,a)(V^\star - V_i^k) \right| \\
&\le \gamma \sum_{i=\iota(t)-\beta\tau}^{t-1} \eta(1-\eta)^{t-i-1} \left| \frac{1}{K} \sum_{k=1}^{K} P(s,a)(V^\star - V_i^k) \right| \\
&\le \gamma \sum_{i=\iota(t)-\beta\tau}^{t-1} \eta(1-\eta)^{t-i-1} \left\| \frac{1}{K} \sum_{k=1}^{K} (V^\star - V_i^k) \right\|_\infty, \qquad (88)
\end{aligned}$$

where the second inequality holds since $\|P(s,a)\|_1 \le 1$. To continue, denoting

$$d_{v,w}^k(s,a) \coloneqq Q_w^k(s,a) - Q_v^k(s,a), \qquad (89)$$

we claim the following bound for any $0 \le i < T$, which will be shown in Appendix B.2.1:

$$\left\| \frac{1}{K} \sum_{k=1}^{K} (V^\star - V_i^k) \right\|_\infty \le \|\Delta_i\|_\infty + 2 \max_k \left\| d_{\iota(i),i}^k \right\|_\infty. \qquad (90)$$

In view of (90), it boils down to control $\max_k \left\| d_{\iota(i),i}^k \right\|_\infty$. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $k \in [K]$, and $0 \le i < T$, by the definition (89), it follows that

$$\left| d_{\iota(i),i}^k(s,a) \right| = \left| \sum_{j=\iota(i)}^{i-1} d_{j,j+1}^k(s,a) \right| \le \underbrace{2\eta \sum_{j=\iota(i)}^{i-1} \|\Delta_j^k\|_\infty}_{:=B_1} + \underbrace{\gamma\eta \left| \sum_{j=\iota(i)}^{i-1} (P_{j+1}^k(s,a) - P(s,a))V^\star \right|}_{:=B_2}, \qquad (91)$$

where

$$\Delta_j^k = Q^\star - Q_j^k. \tag{92}$$

The inequality (91) holds by the local update rule:

$$\begin{aligned}
d_{j,j+1}^k(s,a) &= Q_{j+1}^k(s,a) - Q_j^k(s,a)\\
&= \eta(r(s,a) + \gamma P_{j+1}^k(s,a)V_j^k - Q_j^k(s,a))\\
&\overset{\text{(i)}}{=} \eta(r(s,a) + \gamma P_{j+1}^k(s,a)V_j^k - r(s,a) - \gamma P(s,a)V^\star + Q^\star(s,a) - Q_j^k(s,a))\\
&= \eta(\gamma P_{j+1}^k(s,a)V_j^k - \gamma P(s,a)V^\star + Q^\star(s,a) - Q_j^k(s,a))\\
&= \gamma\eta P_{j+1}^k(s,a)(V_j^k - V^\star) + \gamma\eta(P_{j+1}^k(s,a) - P(s,a))V^\star + \eta\Delta_j^k(s,a)\\
&\le 2\eta\|\Delta_j^k\|_\infty + \gamma\eta(P_{j+1}^k(s,a) - P(s,a))V^\star, \qquad (93)
\end{aligned}$$

where (i) follows from Bellman's optimality equation, and the last inequality follows from $\|P_{j+1}^k(s,a)\|_1 \le 1$ and $\|V_j^k - V^\star\|_\infty \le \|\Delta_j^k\|_\infty$ (cf. (31)).

Next, we bound each term in (91) separately.

- **Bounding B_1.** The local error $\|\Delta_j^k\|_\infty$ is bounded as stated in the following lemma, whose proof is provided in Appendix B.2.2.

  Lemma 11. *Assume $\tau\eta \le \frac{1}{2}$. For any given $\delta \in (0,1)$, the following bound holds for any $1 \le i \le T$ and $k \in [K]$:*

  $$\|\Delta_i^k\|_\infty \le \|\Delta_{\iota(i)}\|_\infty + \frac{2}{1-\gamma}\sqrt{\eta \log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \tag{94}$$

  with at least probability $1-\delta$, where $\iota(i)$ is the most recent synchronization step until i.

  Using the fact that $i - \iota(i) \le \tau - 1$, we can claim that

  $$2\eta\sum_{j=\iota(i)}^{i-1}\|\Delta_j^k\|_\infty \le 2\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty + \frac{4\eta(\tau-1)}{1-\gamma}\sqrt{\eta\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}. \tag{95}$$

- **Bounding B_2.** Using the fact that the empirical transitions are independent and centered on the true transition probability, by invoking Hoeffding's inequality and the union bound, we can claim that the following holds for all $(s,a,k,t) \in \mathcal{S}\times\mathcal{A}\times[K]\times[T]$,

  $$\gamma\eta\left|\sum_{j=\iota(i)}^{i-1}(P_{j+1}^k(s,a) - P(s,a))V^\star\right| \le \frac{\gamma\eta}{1-\gamma}\sqrt{\frac{1}{2}\sum_{j=\iota(i)}^{i-1}\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \le \frac{\gamma\eta}{1-\gamma}\sqrt{(\tau-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \tag{96}$$

  with probability at least $1-\delta$ for any given $\delta\in(0,1)$, where τ is the synchronization period.

By substituting the bound of B_1 and B_2 into (91), and applying the union bound, we obtain that: for any given $\delta \in (0,1)$, the following holds for any $0 \le i \le T$ and $k \in [K]$:

$$\begin{aligned}
\|d_{\iota(i),i}^k\|_\infty &\le 2\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty + \frac{4\eta((\tau-1)\sqrt{\eta} + \sqrt{\tau-1})}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|KT}{\delta}}\\
&\le 2\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty + \frac{8\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|KT}{\delta}} \qquad (97)
\end{aligned}$$

with at least probability $1-\delta$, where $\iota(i)$ is the most recent synchronization step until i. Here, the second line uses the fact $\eta\tau < 1$.

By combining (97) and (90) and substituting it into (88) and using the fact that $\sum_{i=\iota(t)-\beta\tau}^{t-1} \eta(1-\eta)^{t-i-1} \leq 1$, we can obtain the bound $E_t^{3b}(s,a)$ as follows:

$$|E_t^{3b}(s,a)| \leq \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \gamma \sum_{i=\iota(t)-\beta\tau}^{t-1} \eta(1-\eta)^{t-i-1}\left(\|\Delta_i\|_\infty + 4\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty\right)$$
$$\leq \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \gamma(1+4\eta(\tau-1)) \max_{\iota(t)-\beta\tau\leq i<t} \|\Delta_i\|_\infty. \tag{98}$$

Step 3: putting all together. Now, we have the bounds of E_t^{3a} and E_t^{3b} separately derived above. By combining the bounds in (86), we can finally claim the advertised bound and this completes the proof.

B.2.1 Proof of (90)

On one end, it follows that for any $s \in \mathcal{S}$,

$$\frac{1}{K}\sum_{k=1}^{K}\left(V^\star(s) - V_i^k(s)\right) = Q^\star(s,a^\star(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_i^k(s))$$
$$\leq Q^\star(s,a^\star(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a^\star(s)) = \Delta_i(s,a^\star(s)), \tag{99}$$

where we use the definitions in (79). On the other end, it follows that

$$\frac{1}{K}\sum_{k=1}^{K}\left(V^\star(s) - V_i^k(s)\right) = Q^\star(s,a^\star(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_{\iota(i)}(s)) + \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_{\iota(i)}(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_i^k(s))$$
$$\geq Q^\star(s,a_{\iota(i)}(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_{\iota(i)}(s)) + \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_{\iota(i)}(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_i^k(s))$$
$$= \Delta_i(s,a_{\iota(i)}(s)) + \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_{\iota(i)}(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_i^k(s)), \tag{100}$$

where the inequality follows from the fact that $a^\star(s)$ is the optimal action for state s. Notice that the latter terms can be further lower bounded as

$$\frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_{\iota(i)}(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_i^k(s))$$
$$= \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_{\iota(i)}(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_{\iota(i)}^k(s,a_{\iota(i)}(s)) + \frac{1}{K}\sum_{k=1}^{K} Q_{\iota(i)}^k(s,a_{\iota(i)}(s))$$
$$- \frac{1}{K}\sum_{k=1}^{K} Q_{\iota(i)}^k(s,a_i^k(s)) + \frac{1}{K}\sum_{k=1}^{K} Q_{\iota(i)}^k(s,a_i^k(s)) - \frac{1}{K}\sum_{k=1}^{K} Q_i^k(s,a_i^k(s))$$
$$\geq \frac{1}{K}\sum_{k=1}^{K}\left(d_{\iota(i),i}^k(s,a_{\iota(i)}(s)) - d_{\iota(i),i}^k(s,a_i^k(s))\right), \tag{101}$$

where the inequality follows from the definition (89) and the fact that

$$Q_{\iota(i)}^k(s,a_{\iota(i)}(s)) - Q_{\iota(i)}^k(s,a_i^k(s)) \geq 0.$$

The above holds, since $Q_{\iota(i)}^k = Q_{\iota(i)}$ for all $k \in [K]$ agents after periodic averaging at $\iota(i)$, and $a_{\iota(i)}(s)$ is the optimal action at state s at time $\iota(i)$ for every agent.

Combining (99), (100) and (101), we obtain

$$\Delta_i(s, a_{\iota(i)}(s)) + \frac{1}{K}\sum_{k=1}^{K}\left(d^k_{\iota(i),i}(s, a_{\iota(i)}(s)) - d^k_{\iota(i),i}(s, a^k_i(s))\right) \leq \frac{1}{K}\sum_{k=1}^{K}\left(V^\star(s) - V^k_i(s)\right) \leq \Delta_i(s, a^\star(s)),$$

which immediately implies (90).

B.2.2 Proof of Lemma 11

By applying the decomposition in (36) to the local error for agent k, we decompose Δ^k_i as follows:

$$\Delta^k_i(s,a) = \underbrace{(1-\eta)^{i-\iota(i)}\Delta^k_{\iota(i)}(s,a)}_{:=D_1} + \underbrace{\gamma\sum_{j=\iota(i)+1}^{i}\eta(1-\eta)^{i-j}(P(s,a) - P^k_j(s,a))V^\star}_{:=D_2}$$
$$+ \underbrace{\gamma\sum_{j=\iota(i)+1}^{i}\eta(1-\eta)^{i-j}P^k_j(s,a)(V^\star - V^k_{j-1})}_{:=D_3}. \tag{102}$$

We shall bound each term separately.

- **Bounding D_1.** Since $\Delta^k_{\iota(i)} = \Delta_{\iota(i)}$ for every agent k at the synchronization step $\iota(i)$,
$$|D_1| \leq (1-\eta)^{i-\iota(i)}\|\Delta_{\iota(i)}\|_\infty. \tag{103}$$
- **Bounding D_2.** In a similar manner to (96), by invoking Hoeffding inequality and using the fact that $\sum_{j=\iota(i)+1}^{i}(\eta(1-\eta)^{i-j})^2 \leq \eta$ (cf. (84)), we can claim that the following holds for all $(s,a,k,t) \in \mathcal{S}\times\mathcal{A}\times[K]\times[T]$,
$$|D_2| \leq \gamma\sqrt{\sum_{j=\iota(i)+1}^{i}(\eta(1-\eta)^{i-j})^2\|V^\star\|^2_\infty \log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \leq \frac{\gamma}{1-\gamma}\sqrt{\eta\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \tag{104}$$
with probability at least $1-\delta$ for any given $\delta \in (0,1)$.
- **Bounding D_3.** By bounding $\|V^\star - V^k_{j-1}\|_\infty$ with the local error $\|\Delta^k_{j-1}\|_\infty$ (cf. (31)) and using $\|P^k_j(s,a)\|_1 \leq 1$, we have
$$|D_3| \leq \gamma\sum_{j=\iota(i)+1}^{i}\eta(1-\eta)^{i-j}\|P^k_j(s,a)\|_1\|V^\star - V^k_{j-1}\|_\infty \leq \gamma\sum_{j=\iota(i)+1}^{i}\eta(1-\eta)^{i-j}\|\Delta^k_{j-1}\|_\infty. \tag{105}$$

By combining the bounds obtained above in (102), we obtain the following recursive relation

$$\|\Delta^k_i\|_\infty \leq (1-\eta)^{i-\iota(i)}\|\Delta_{\iota(i)}\|_\infty + \underbrace{\frac{\gamma}{1-\gamma}\sqrt{\eta\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}}_{:=\rho} + \gamma\sum_{j=\iota(i)+1}^{i}\eta(1-\eta)^{i-j}\|\Delta^k_{j-1}\|_\infty. \tag{106}$$

By invoking the recursive relation with some algebraic calculations, we obtain the following bound

$$\|\Delta^k_i\|_\infty \leq (1-\eta)^{i-\iota(i)}\|\Delta_{\iota(i)}\|_\infty + \rho$$
$$+ \gamma\sum_{j_1=\iota(i)+1}^{i}\eta(1-\eta)^{i-j_1}\left((1-\eta)^{j_1-1-\iota(i)}\|\Delta_{\iota(i)}\|_\infty + \rho + \gamma\sum_{j_2=\iota(i)+1}^{j_1-1}\eta(1-\eta)^{j_1-1-j_2}\|\Delta^k_{j_2-1}\|_\infty\right)$$

$$
\begin{aligned}
&= \left((1-\eta)^{i-\iota(i)} + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-1-\iota(i)} \right) \|\Delta_{\iota(i)}\|_\infty + \left(1 + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-j_1} \right) \rho \\
&\quad + \gamma^2 \sum_{j_1=\iota(i)+1}^{i} \sum_{j_2=\iota(i)+1}^{j_1-1} \eta^2 (1-\eta)^{i-1-j_2} \|\Delta^k_{j_2-1}\|_\infty \\
&\le \left((1-\eta)^{i-\iota(i)} + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-1-\iota(i)} \right) \|\Delta_{\iota(i)}\|_\infty + \left(1 + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-j_1} \right) \rho \\
&\quad + \gamma^2 \sum_{j_1=\iota(i)+1}^{i} \sum_{j_2=\iota(i)+1}^{j_1-1} \eta^2 (1-\eta)^{i-1-j_2} \left((1-\eta)^{j_2-1-\iota(i)} \|\Delta_{\iota(i)}\|_\infty + \rho + \cdots \right) \\
&\le \left((1-\eta)^{i-\iota(i)} + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-1-\iota(i)} + \cdots + \gamma^l \sum_{j_1=\iota(i)+1}^{i} \cdots \sum_{j_l=\iota(i)+1}^{j_{l-1}-1} \eta^l (1-\eta)^{i-l-\iota(i)} \right) \|\Delta_{\iota(i)}\|_\infty \\
&\quad + \left(1 + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-j_1} + \cdots + \gamma^l \sum_{j_1=\iota(i)+1}^{i} \cdots \sum_{j_l=\iota(i)+1}^{j_{l-1}-1} \eta^l (1-\eta)^{i-l+1-j_l} \right) \rho \\
&\quad + \gamma^{l+1} \sum_{j_1=\iota(i)+1}^{i} \cdots \sum_{j_{l+1}=\iota(i)+1}^{j_l-1} \eta^{l+1} (1-\eta)^{i-l-j_{l+1}} \left(\|\Delta^k_{j_{l+1}-1}\| \right) \\
&\overset{\text{(i)}}{\le} \sum_{l=0}^{i-\iota(i)} \gamma^l \binom{i-\iota(i)}{l} \eta^l (1-\eta)^{i-\iota(i)-l} \|\Delta^k_{\iota(i)}\|_\infty + \sum_{l=0}^{i-\iota(i)-1} \gamma^l \binom{i-\iota(i)}{l} \eta^l \rho \\
&\le ((1-\eta)+\gamma\eta)^{i-\iota(i)} \|\Delta^k_{\iota(i)}\|_\infty + (1+\gamma\eta)^{i-\iota(i)} \rho \\
&\overset{\text{(ii)}}{\le} \|\Delta^k_{\iota(i)}\|_\infty + 2\rho,
\end{aligned}
\tag{107}
$$

where (i) follows from $\Delta^k_{j_{i-\iota(i)}-1} = \Delta^k_{\iota(i)}$ since $j_l \le i-l+1$,

$$
\sum_{j_1=\iota(i)+1}^{i} \sum_{j_2=\iota(i)+1}^{j_1-1} \cdots \sum_{j_l=\iota(i)+1}^{j_{l-1}-1} \eta^l (1-\eta)^{i-l-\iota(i)} = \binom{i-\iota(i)}{l} \eta^l (1-\eta)^{i-l-\iota(i)},
$$

$$
\sum_{j_1=\iota(i)+1}^{i} \cdots \sum_{j_l=\iota(i)+1}^{j_{l-1}-1} \eta^l (1-\eta)^{i-l+1-j_l} \le \sum_{j_1=\iota(i)+1}^{i} \cdots \sum_{j_l=\iota(i)+1}^{j_{l-1}-1} \eta^l \le \binom{i-\iota(i)}{l} \eta^l,
$$

and (ii) follows from $(1+\gamma\eta)^{i-\iota(i)} \le (1+\gamma\eta)^\tau \le e^{\tau\eta} \le 2$ since $i-\iota(i) \le \tau$ and $\tau\eta \le \frac{1}{2}$. This completes the proof.

C Proofs for federated asynchronous Q-learning (Section 4)

C.1 Proof of Lemma 10

To describe the joint probabilistic transitions of K agents formally, we first introduce the following Markov chain $X_t = (X_t^1, \ldots, X_t^K)$, $t = 0, 1, \ldots$, where $X_t^k \in \mathcal{S} \times \mathcal{A}$ is the state-action pair visited by agent k at time t. The joint transition kernel P of K agents is given by

$$
P := \begin{pmatrix} P^1 & & & \\ & P^2 & & \\ & & \ddots & \\ & & & P^K \end{pmatrix},
\tag{108}
$$

where P^k is the transition kernel of agent k, $k = 1, \dots, K$. Since the agents are independent, the stationary distribution of the joint Markov chain is μ, given by

$$\mu(x) \coloneqq \prod_{k=1}^{K} \mu_{\mathsf{b}}^{k}(x^k), \qquad \forall x = (x^1, x^2, \cdots, x^K) \in (\mathcal{S} \times \mathcal{A})^K, \tag{109}$$

where μ_{b}^{k} denotes the stationary distribution of agent k, which are induced by its behavior policy π_{b}^{k}. Next, we define the mixing time of the joint Markov chain as follows:

$$t_{\mathsf{mix}}(\epsilon) \coloneqq \min\left\{ t \;\middle|\; \sup_{x_0 \in (\mathcal{S}\times\mathcal{A})^K} d_{\mathsf{TV}}(P_t(\cdot|x_0),\, \mu) \le \epsilon \right\} \quad \text{and} \quad t_{\mathsf{mix}} \coloneqq t_{\mathsf{mix}}\left(\frac{1}{4}\right), \tag{110}$$

where

$$P_t(\cdot|x_0) = \prod_{k=1}^{K} P_t^k(\cdot|x_0^k) \tag{111}$$

denotes the distribution of the joint state-action pairs of all agents after t transitions starting from $x_0 = (x_0^1, \dots, x_0^K)$. The mixing time of the joint Markov chain can be connected to those of the individual chains via the following relation

$$t_{\mathsf{mix}}(\epsilon) \le \max_{k} t_{\mathsf{mix}}^{k}(\epsilon/K), \qquad t_{\mathsf{mix}} \le 4 \log 8K \max_{k\in[K]} t_{\mathsf{mix}}^{k}, \tag{112}$$

which will be proven at the end of the proof.

We now turn to the proof of Lemma 10. Define the event

$$\mathcal{B}_{u,v}(s,a) \coloneqq \left\{ \left| \sum_{k=1}^{K} N_{u,v}^{k}(s,a) - (v-u) \sum_{k=1}^{K} \mu_{\mathsf{b}}^{k}(s,a) \right| \ge \frac{1}{2}(v-u) \sum_{k=1}^{K} \mu_{\mathsf{b}}^{k}(s,a) \right\}. \tag{113}$$

We first establish that

$$\max_{x_0 \in (\mathcal{S}\times\mathcal{A})^K} \mathbb{P}\left\{ \mathcal{B}_{u,v}(s,a) \middle| \{(s_0^k, a_0^k)\}_{k=1}^{K} = x_0 \right\} \le \frac{\delta}{|\mathcal{S}||\mathcal{A}|T^2} \tag{114}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $1 \le u < v \le T$ provided that $u \ge t_{\mathsf{th}}(s,a)/2$ and $v - u \ge t_{\mathsf{th}}(s,a)/2$. To this end, we decompose the probability into two terms as follows:

$$\begin{aligned}
\mathbb{P}\left\{ \mathcal{B}_{u,v}(s,a) \middle| \{(s_0^k, a_0^k)\}_{k=1}^{K} = x_0 \right\} &= \underbrace{\mathbb{P}\left\{ \mathcal{B}_{u,v}(s,a) \middle| \{(s_0^k, a_0^k)\}_{k=1}^{K} \sim \mu \right\}}_{=:G_1} \\
&\quad + \underbrace{\mathbb{P}\left\{ \mathcal{B}_{u,v}(s,a) \middle| \{(s_0^k, a_0^k)\}_{k=1}^{K} = x_0 \right\} - \mathbb{P}\left\{ \mathcal{B}_{u,v}(s,a) \middle| \{(s_0^k, a_0^k)\}_{k=1}^{K} \sim \mu \right\}}_{=:G_2},
\end{aligned}$$

and show each of the terms is bounded by $\frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2}$ for any $x_0 \in (\mathcal{S} \times \mathcal{A})^K$. We shall derive the bounds of these two terms separately.

Step 1: bounding G_1. This is for the case that the distribution of the initial state follows the joint stationary distribution. Since the total number of visits can be written as

$$\sum_{k=1}^{K} N_{u,v}^{k}(s,a) = \sum_{k=1}^{K} \sum_{i=u+1}^{v} Z_i^k(s,a) = \sum_{i=u+1}^{v} \bar{Z}_i(s,a),$$

where

$$Z_i^k(s,a) = \begin{cases} 1, & \text{if } (s,a) \in (s_{i-1}^k, a_{i-1}^k) \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \bar{Z}_i(s,a) = \sum_{k=1}^{K} Z_i^k(s,a),$$

and

$$\nu_{u,v}(s,a) \coloneqq \mathbb{E}_{(s_0^k,a_0^k)\sim\mu^k \forall k\in[K]}\left[\sum_{i=u+1}^{v} \bar{Z}_i(s,a)\right] = (v-u)\sum_{k=1}^{K}\mu_{\mathsf{b}}^k(s,a),$$

we can invoke Bernstein's inequality for Markov chains (Paulin, 2015, Theorem 3.11) and obtain

$$\begin{aligned} G_1 &= \mathbb{P}_{\{(s_0^k,a_0^k)\}_{k=1}^K\sim\mu}\left[\left|\sum_{i=u+1}^{v}\bar{Z}_i(s,a)-\nu_{u,v}(s,a)\right| \geq \frac{1}{2}\nu_{u,v}(s,a)\right] \\ &\leq 2\exp\left(-\frac{(\nu_{u,v}(s,a)/2)^2\gamma_{\mathrm{ps}}}{8((v-u)+1/\gamma_{\mathrm{ps}})V_f+20C(\nu_{u,v}(s,a)/2)}\right). \end{aligned} \tag{115}$$

Here, γ_{ps} is the pseudo spectral gap satisfying

$$\gamma_{\mathrm{ps}} \geq \frac{1}{2t_{\mathsf{mix}}} \tag{116a}$$

for uniformly ergodic Markov chains according to Paulin (2015, Proposition 3.4). The parameters C and V_f are defined and bounded as follows

$$C \coloneqq \max_{u<i\leq v}\left|\bar{Z}_i(s,a)-\mathbb{E}[\bar{Z}_i(s,a)]\right| \leq K, \tag{116b}$$

$$V_f \coloneqq \mathsf{Var}(\bar{Z}_i(s,a)) = \sum_{k=1}^{K}(1-\mu_{\mathsf{b}}^k(s,a))\mu_{\mathsf{b}}^k(s,a) \leq \sum_{k=1}^{K}\mu_{\mathsf{b}}^k(s,a). \tag{116c}$$

Plugging (116) into (115), we have

$$\begin{aligned} G_1 &\leq 2\exp\left(-\frac{(\nu_{u,v}(s,a))^2}{8t_{\mathsf{mix}}(24(v-u)(\sum_{k=1}^{K}\mu_{\mathsf{b}}^k(s,a))+10K\nu_{u,v}(s,a))}\right) \\ &\leq 2\exp\left(-\frac{(v-u)(\sum_{k=1}^{K}\mu_{\mathsf{b}}^k(s,a))}{8t_{\mathsf{mix}}(24+10K)}\right) \leq \frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2}, \end{aligned} \tag{117}$$

where the last inequality holds since $(v-u)$ is large enough to satisfy the following condition:

$$v-u \geq \frac{t_{\mathsf{th}}(s,a)}{2} \geq \frac{1088(\max_{k\in[K]} t_{\mathsf{mix}}^k)\log 8K \log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}{\frac{1}{K}\sum_{k=1}^{K}\mu_{\mathsf{b}}^k(s,a)} \geq \frac{272t_{\mathsf{mix}}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}{\frac{1}{K}\sum_{k=1}^{K}\mu_{\mathsf{b}}^k(s,a)}.$$

Step 2: bounding G_2. By the same argument of Li et al. (2021b, Section A.1), using the fact that the difference caused by the initial state becomes very small after sufficiently long time, we have we have

$$\begin{aligned} G_2 &\coloneqq \mathbb{P}\Big\{\mathcal{B}_{u,v}(s,a)\big|\{(s_0^k,a_0^k)\}_{k=1}^K = x_0\Big\} - \mathbb{P}\Big\{\mathcal{B}_{u,v}(s,a)\big|\{(s_0^k,a_0^k)\}_{k=1}^K \sim \mu\Big\} \\ &\leq d_{\mathsf{TV}}(P_u(\cdot|x_0),\mu) \leq \frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2}, \end{aligned} \tag{118}$$

where the last inequality holds due to

$$u \geq \frac{t_{\mathsf{th}}(s,a)}{2} \geq 4\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}\max_{k\in[K]}t_{\mathsf{mix}}^k \geq \max_{k\in[K]}t_{\mathsf{mix}}^k\left(\frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2K}\right) \geq t_{\mathsf{mix}}\left(\frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2}\right). \tag{119}$$

Here, the second inequality follows from the fact that $t_{\mathsf{mix}}^k(\epsilon) \leq 2t_{\mathsf{mix}}^k\log_2\frac{2}{\epsilon}$ (Paulin, 2015), and the last inequality follows from (112).

Step 3: summing things up. By combining the above bound, we complete the proof of (114), provided that $u \geq t_{\mathsf{th}}(s,a)/2$ and $v - u \geq t_{\mathsf{th}}(s,a)$. Then, we can obtain the following bound for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $0 \leq u < v \leq T$:

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{4}(v-u)\sum_{k=1}^{K}\mu_{\mathsf{b}}^{k}(s,a) \leq \sum_{k=1}^{K}N_{u,v}^{k}(s,a) \leq 2(v-u)\sum_{k=1}^{K}\mu_{\mathsf{b}}^{k}(s,a)\right\} \\
&\leq \mathbb{P}\left\{\left|\sum_{k=1}^{K}N^{k}_{u+\frac{t_{\mathsf{th}}(s,a)}{2},v}(s,a) - \left(v-u-\frac{t_{\mathsf{th}}(s,a)}{2}\right)\sum_{k=1}^{K}\mu_{\mathsf{b}}^{k}(s,a)\right| \geq \frac{1}{2}\left(v-u-\frac{t_{\mathsf{th}}(s,a)}{2}\right)\sum_{k=1}^{K}\mu_{\mathsf{b}}^{k}(s,a)\right\} \\
&= \max_{x_0 \in (\mathcal{S}\times\mathcal{A})^K}\mathbb{P}\left\{\mathcal{B}_{u+\frac{t_{\mathsf{th}}(s,a)}{2},v}(s,a)\,\middle|\,\{(s_0^k,a_0^k)\}_{k=1}^{K} = x_0\right\} \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|T^2}.
\end{aligned} \tag{120}$$

Proof of (112). Notice that by the definition of d_{TV} and (111), we have

$$d_{\mathsf{TV}}(P_t(\cdot|x_0),\,\mu) \leq \sum_{k=1}^{K} d_{\mathsf{TV}}(P_t^k(\cdot|x_0^k),\,\mu_{\mathsf{b}}^k)$$

for any $x_0 \in (\mathcal{S}\times\mathcal{A})^K$. Hence, setting $t = \max_{k\in[K]} t_{\mathsf{mix}}^k\left(\frac{\epsilon}{K}\right)$, we have

$$\max_{x_0\in(\mathcal{S}\times\mathcal{A})^K} d_{\mathsf{TV}}(P_t(\cdot|x_0),\,\mu) \leq \sum_{k=1}^{K}\frac{\epsilon}{K} = \epsilon,$$

which immediately implies

$$t_{\mathsf{mix}}(\epsilon) \leq \max_k t_{\mathsf{mix}}^k(\epsilon/K).$$

The proof is complete by using the fact that $t_{\mathsf{mix}}(\epsilon) \leq 2t_{\mathsf{mix}}\log_2\frac{2}{\epsilon}$ (Paulin, 2015), which leads to

$$t_{\mathsf{mix}} \leq \max_{k\in[K]} t_{\mathsf{mix}}^k\left(\frac{1}{4K}\right) \leq 4\log 8K \max_{k\in[K]} t_{\mathsf{mix}}^k.$$

C.2 Proof of Lemma 3

First, (52a) is derived as follows:

$$\begin{aligned}
\lambda_{v_1,v_2}(s,a) = \frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{N_{v_1,v_2}^k(s,a)} \leq \frac{1}{K}\sum_{k=1}^{K}\exp(-\eta N_{v_1,v_2}^k(s,a)) \leq 1 - \frac{1}{2}\frac{1}{K}\sum_{k=1}^{K}\eta N_{v_1,v_2}^k(s,a) \\
\leq \exp\left(-\frac{\eta}{2K}\sum_{k=1}^{K}N_{v_1,v_2}^k(s,a)\right)
\end{aligned} \tag{121}$$

using the fact that $1 - x \leq \exp(-x) \leq 1 - \frac{x}{2}$ holds for any $0 \leq x < 1$, and $\eta N_{h\tau,(h+1)\tau}^{k'}(s,a) \leq \eta\tau \leq 1$.

Next, we obtain (52b) through the following derivation:

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{u\in\mathcal{U}_{0,t}^k(s,a)}\omega_{u,t}^k(s,a) &= \sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)}\omega_{u,t}^k(s,a) \\
&= \sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right)\sum_{k=1}^{K}\frac{1}{K}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)}\left(\eta(1-\eta)^{N_{u+1,(h+1)\tau}^k(s,a)}\right) \\
&\overset{\text{(i)}}{=} \sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right)\sum_{k=1}^{K}\frac{1}{K}(1-(1-\eta)^{N_{h\tau,(h+1)\tau}^k(s,a)})
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(ii)}}{=} \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a) \right) (1-\lambda_{h\tau,(h+1)\tau}(s,a)) \\
&\overset{\text{(iii)}}{=} 1-\lambda_{0,\tau}\lambda_{\tau,2\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} = 1-\omega_{0,t}(s,a),
\end{aligned} \tag{122}$$

where (i) follows from the geometric sum

$$\begin{aligned}
\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \eta(1-\eta)^{N^k_{u+1,(h+1)\tau}(s,a)} &= \eta+\eta(1-\eta)+\cdots+\eta(1-\eta)^{N^k_{h\tau,(h+1)\tau}(s,a)-1} \\
&= 1-(1-\eta)^{N^k_{h\tau,(h+1)\tau}(s,a)},
\end{aligned} \tag{123}$$

(ii) follows from the definition (49), and (iii) follows by cancellation.

Similarly, (52c) can be obtained with some algebraic calculations as follows:

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{u\in\mathcal{U}^k_{0,h'\tau}(s,a)} \omega^k_{u,t}(s,a) &= \sum_{k=1}^{K}\sum_{h=0}^{h'-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \omega^k_{u,t}(s,a) \\
&\overset{\text{(i)}}{=} \sum_{h=0}^{h'-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a) \right) (1-\lambda_{h\tau,(h+1)\tau}(s,a)) \\
&\overset{\text{(ii)}}{\le} \lambda_{h'\tau,(h'+1)\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} - \lambda_{0,\tau}\lambda_{\tau,2\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} \\
&\le \lambda_{h'\tau,(h'+1)\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} \overset{\text{(iii)}}{\le} \prod_{h=h'}^{\phi(t)-1} \exp\left(-\frac{\eta}{2K}\sum_{k=1}^{K} N^k_{h\tau,(h+1)\tau}(s,a) \right),
\end{aligned} \tag{124}$$

where (i) follows from similar derivations as above, (ii) follows by cancellation, and (iii) follows from (52a).

Finally, (52d) is derived as follows:

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{u\in\mathcal{U}^k_{0,t}(s,a)} (\omega^k_{u,t}(s,a))^2 &= \sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} (\omega^k_{u,t}(s,a))^2 \\
&= \frac{1}{K}\sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a) \right)^2 \sum_{k=1}^{K}\frac{1}{K} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \left(\eta(1-\eta)^{N^k_{u+1,(h+1)\tau}(s,a)}\right)^2 \\
&\overset{\text{(i)}}{\le} \frac{2\eta}{K}\sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a) \right) \sum_{k=1}^{K}\frac{1}{K} \left(1-(1-\eta)^{(N^k_{h\tau,(h+1)\tau}(s,a))}\right) \\
&= \frac{2\eta}{K}\sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a) \right) (1-\lambda_{h\tau,(h+1)\tau}(s,a)) \\
&\overset{\text{(ii)}}{\le} \frac{2\eta}{K},
\end{aligned}$$

where (i) holds since

$$\begin{aligned}
\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \left(\eta(1-\eta)^{N^k_{u+1,(h+1)\tau}(s,a)}\right)^2 &= \eta^2+\eta^2(1-\eta)^2+\cdots+\eta(1-\eta)^{2(N^k_{u+1,(h+1)\tau}(s,a)-1)} \\
&\le \eta\left(1-(1-\eta)^{2N^k_{u+1,(h+1)\tau}(s,a)}\right) \\
&\le 2\eta\left(1-(1-\eta)^{N^k_{u+1,(h+1)\tau}(s,a)}\right)
\end{aligned} \tag{125}$$

and (ii) can be similarly derived to the proof of (52c) (cf. (124)).

C.3 Proof of Lemma 4

Without loss of generality, we prove the claim for some fixed $1 \leq t \leq T$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. For notation simplicity, let

$$y^k_{u,t}(s,a) = \begin{cases} \omega^k_{u,t}(s,a)(P(s,a) - P^k_{u+1}(s,a))V^k_u & \text{if } (s^k_u, a^k_u) = (s,a) \\ 0 & \text{otherwise} \end{cases}, \tag{126}$$

where

$$\omega^k_{u,t}(s,a) = \frac{\eta}{K}(1-\eta)^{N^k_{u+1,(\phi(u)+1)\tau}(s,a)} \prod_{h=\phi(u)+1}^{\phi(t)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right), \tag{127}$$

then $E^2_t(s,a) = \gamma \sum_{k=1}^K \sum_{u=0}^{t-1} y^k_{u,t}(s,a)$. However, due to the dependency between $P^k_{u+1}(s,a)$ and $\omega^k_{u,t}(s,a)$ arising from the Markovian sampling, it is difficult to track the sum of $y := \{y^k_{u,t}(s,a)\}$ directly. To address this issue, we will first analyze the sum using a collection of approximate random variables $\widehat{y} = \{\widehat{y}^k_{u,t}(s,a)\}$ drawn from a carefully constructed set $\widehat{\mathcal{Y}}$, which is closely coupled with the target $\{y^k_{u,t}(s,a)\}_{0 \leq u < t}$, i.e.,

$$D(y, \widehat{y}) := \left| \sum_{k=1}^{K} \sum_{u=0}^{t-1} \left(y^k_{u,t}(s,a) - \widehat{y}^k_{u,t}(s,a) \right) \right| \tag{128}$$

is sufficiently small. In addition, $\widehat{y}$ shall exhibit some useful statistical independence and thus easier to control its sum; we shall control this over the entire set $\widehat{\mathcal{Y}}$. Finally, leveraging the proximity above, we can obtain the desired bound on y via triangle inequality. We now provide details on executing this proof outline, where the crust is in designing the set $\widehat{\mathcal{Y}}$ with a controlled size.

Before describing our construction, let's introduce the following useful event:

$$\mathcal{B}_M(s,a) := \bigcap_{u=0}^{t-M\tau} \left\{ \frac{1}{4}\mu_{\mathsf{avg}}(s,a)KM\tau \leq \sum_{k=1}^{K} N^k_{u,u+M\tau}(s,a) \leq 2\mu_{\mathsf{avg}}(s,a)KM\tau \right\}, \tag{129}$$

where $M = M(s,a) := \lfloor \frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau} \rfloor$. Note that $M\tau \geq \tau \geq t_{\mathsf{th}}$ (see (77) for the definition of $t_{\mathsf{th}}(s,a)$), and $1 \leq 1/(16\eta\mu_{\mathsf{avg}}(s,a)\tau) \leq M(s,a) \leq 1/(8\eta\mu_{\mathsf{avg}}(s,a)\tau)$ if $\eta\tau \leq 1/16$. Then, $\mathcal{B}_M(s,a)$ holds with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$ according to Lemma 10. The rest of the proof shall be carried out under the event $\mathcal{B}_M(s,a)$.

Step 1: constructing $\widehat{\mathcal{Y}}$. To decouple dependency between $P^k_{u+1}(s,a)$ and $\omega^k_{u,t}(s,a)$, we will introduce approximates of $\omega^k_{u,t}(s,a)$ that only depend on history until u by replacing a factor dependent on future with some constant. To gain insight, we first decompose $\omega^k_{u,t}(s,a)$ as follows:

$$\begin{aligned} \omega^k_{u,t}(s,a) &= \frac{\eta}{K}(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)} \frac{(1-\eta)^{N^k_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{k'=1}^K (1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}} \prod_{h=\phi(u)}^{\phi(t)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right) \\ &= \underbrace{\frac{\eta}{K}(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)} \prod_{h=\phi(u)}^{\phi(t)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right)}_{:=\bar{\omega}^k_{u,t}(s,a)} \\ &\quad + \underbrace{\frac{\eta}{K}(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)} \left(\frac{(1-\eta)^{N^k_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{k'=1}^K (1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}} - 1 \right) \prod_{h=\phi(u)}^{\phi(t)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right)}_{:=\chi^k_{u,t}(s,a)}. \end{aligned}$$

Considering that $\chi^k_{u,t}(s,a)$ can be made small enough, which will be shown in the following step, we analyze the dominant factor $\bar{\omega}^k_{u,t}(s,a)$ in detail as follows:

$$
\begin{aligned}
\bar{\omega}^k_{u,t}(s,a) &= \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right) \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right)^{-1} \right) \\
&\quad \times \frac{\eta}{K} (1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)} \prod_{h=\phi(u)}^{\phi(t)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right) \\
&= \underbrace{\frac{\eta}{K} (1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)} \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right)^{-1}}_{\text{dependent on history until } u} \\
&\quad \times \underbrace{\prod_{h=h_0(u,t)}^{\phi(t)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right)}_{\text{dependent on history and future until } t} \\
&= \underbrace{\frac{\eta}{K} (1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)} \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right)^{-1}}_{\coloneqq x^k_u(s,a)} \\
&\quad \times \prod_{l=1}^{l(u,t)} \underbrace{\prod_{h=\max\{0,\phi(t)-lM\}}^{\phi(t)-(l-1)M-1} \left(\frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \right)}_{\coloneqq z_l(s,a)},
\end{aligned}
\tag{130}
$$

where we denote $h_0(u,t) = \max\{0, \phi(t) - l(u,t)M\}$, with $l(u,t) \coloneqq \lceil \frac{(t-u)}{M\tau} \rceil$.

Motivated by the above decomposition, we will construct $\widehat{\mathcal{Y}}$ by approximating the future-dependent parameter $z_l(s,a)$ for $1 \le l \le L$, where we define

$$
L \coloneqq \min\left\{ \left\lceil \frac{t}{M\tau} \right\rceil, \lceil 128 \log\left(K/\eta\right) \rceil \right\}. \tag{131}
$$

We note that $L \le 128 \log(TK)$ for $\eta \ge 3/T$. Using the fact that $1 - x \le \exp(-x) \le 1 - \frac{x}{2}$ holds for any $0 \le x < 1$, and $\eta N^{k'}_{h\tau,(h+1)\tau}(s,a) \le \eta\tau \le \frac{1}{2}$,

$$
\begin{aligned}
\exp\left(-\frac{2\eta}{K} \sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a) \right) &\le 1 - \frac{\eta}{K} \sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a) \\
&\le \frac{1}{K} \sum_{k'=1}^{K} (1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)} \\
&\le \frac{1}{K} \sum_{k'=1}^{K} \exp(-\eta N^{k'}_{h\tau,(h+1)\tau}(s,a)) \\
&\le 1 - \frac{1}{2}\frac{1}{K} \sum_{k'=1}^{K} \eta N^{k'}_{h\tau,(h+1)\tau}(s,a) \\
&\le \exp\left(-\frac{\eta}{2K} \sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a) \right).
\end{aligned}
\tag{132}
$$

Therefore, for $1 \le l < L$, under $\mathcal{B}_M(s,a)$, the range of $z_l(s,a)$ is bounded as follows:

$$z_l(s,a) \in \left[\exp(-4\eta\mu_{\mathsf{avg}}(s,a)M\tau),\;\; \exp(-\frac{1}{8}\eta\mu_{\mathsf{avg}}(s,a)M\tau)\right].$$

Using this property, we construct a set of values that can cover possible realizations of $z_l(s,a)$ in a fine-grained manner as follows:

$$\mathcal{Z} := \left\{\exp\left(-\frac{1}{8}\eta\mu_{\mathsf{avg}}(s,a)M\tau - \frac{i\eta}{K}\right) \;\;\middle|\, i \in \mathbb{Z}: \;\; 0 \le i < 4K\mu_{\mathsf{avg}}(s,a)M\tau\right\}. \tag{133}$$

Note that the distance of adjacent elements of $\mathcal{Z}$ is bounded by $\eta/Ke^{-1/8\eta\mu_{\mathsf{avg}}(s,a)M\tau}$, and the size of the set is bounded by $4K\mu_{\mathsf{avg}}(s,a)M\tau$. For $l = L$, because the number of iterations involved in $z_L(s,a)$ can be less than $M\tau$, it follows that $z_L(s,a) \in [\exp(-4\eta\mu_{\mathsf{avg}}(s,a)M\tau), 1]$. Hence, we construct the set

$$\mathcal{Z}_0 := \left\{\exp\left(-\frac{i\eta}{K}\right) \;\;\middle|\, i \in \mathbb{Z}: \;\; 0 \le i < 4K\mu_{\mathsf{avg}}(s,a)M\tau\right\}. \tag{134}$$

In sum, we can always find $(\widehat{z}_1, \cdots, \widehat{z}_l, \cdots, \widehat{z}_L) \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0$ where its entry-wise distance to $(z_l(s,a))_{l\in[L-1]}$ (resp. $z_L(s,a)$) is at most $\eta/Ke^{-1/8\eta\mu_{\mathsf{avg}}(s,a)M\tau}$ (resp. η/K).

Moreover, we approximate $x_u^k(s,a)$ by clipping it when the accumulated number of visits of all agents is not too large as follows:

$$\widehat{x}_u^k(s,a) = \begin{cases} x_u^k(s,a) & \text{if } \sum_{k=1}^K N^k_{h_0(u,t)\tau,\phi(u)\tau}(s,a) \le 2K\mu_{\mathsf{avg}}(s,a)M\tau \\ 0 & \text{otherwise} \end{cases}. \tag{135}$$

Note that the clipping never occurs and $\widehat{x}_u^k(s,a) = x_u^k(s,a)$ for all u as long as $\mathcal{B}_M(s,a)$ holds. To provide useful properties of $\widehat{x}_u^k(s,a)$ that will be useful later, we record the following lemma whose proof is provided in Appendix C.3.1.

Lemma 12. *For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, consider any integers $1 \le t \le T$ and $1 \le l \le \lceil \frac{t}{M\tau} \rceil$, where $M = \lfloor \frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau} \rfloor$. Suppose that $4\eta\tau \le 1$, then $\widehat{x}_u^k(s,a)$ defined in (135) satisfy*

$$\forall u \in [h_0, \phi(t) - (l-1)M) \quad : \quad \widehat{x}_u^k(s,a) \le \frac{9\eta}{K}, \tag{136a}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^K \widehat{x}_u^k(s,a) \le 16\eta\mu_{\mathsf{avg}}(s,a)M\tau, \tag{136b}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^K (\widehat{x}_u^k(s,a))^2 \le \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K}, \tag{136c}$$

where $h_0 = \max\{0, \phi(t) - lM\}$.

Finally, for each $\boldsymbol{z} = (\widehat{z}_1, \cdots, \widehat{z}_L) \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0$, setting

$$\widehat{\omega}^k_{u,t}(s,a;\boldsymbol{z}) = \widehat{x}_u^k(s,a) \prod_{l=1}^{l(u,t)} \widehat{z}_l, \tag{137}$$

an approximate random sequence $\widehat{y}_{\boldsymbol{z}} = \{\widehat{y}^k_{u,t}(s,a;\boldsymbol{z})\}_{0\le u<t}$ can be constructed as follows:

$$\widehat{y}^k_{u,t}(s,a;\boldsymbol{z}) = \begin{cases} \widehat{\omega}^k_{u,t}(s,a;\boldsymbol{z})(P(s,a) - P^k_{u+1}(s,a))V^k_u & \text{if } (s^k_u, a^k_u) = (s,a) \text{ and } l(u,t) \le L \\ 0 & \text{otherwise} \end{cases}. \tag{138}$$

If $t > LM\tau$, for any $u < t - LM\tau$, i.e., $l(u,t) > L$, we set $\widehat{y}^k_{u,t}(s,a;\boldsymbol{z}) = 0$ since the magnitude of $\omega^k_{u,t}(s,a)$ becomes negligible when the time difference between u and t is large enough, and the fine-grained approximation using $\mathcal{Z}$ is no longer needed, as shall be seen momentarily. Finally, denote a collection of the approximates induced by $\mathcal{Z}^{L-1} \times \mathcal{Z}_0$ as

$$\widehat{\mathcal{Y}} = \{\widehat{y}_{\boldsymbol{z}} : \quad \boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0\}.$$

Step 2: bounding the approximation error $D(y, \widehat{y}_{\boldsymbol{z}})$. We now show that under $\mathcal{B}_M(s,a)$, there exists $\widehat{y}_{\boldsymbol{z}} := \widehat{y}_{\boldsymbol{z}(y)} \in \widehat{\mathcal{Y}}$ such that

$$D(y, \widehat{y}_{\boldsymbol{z}}) < \frac{525}{1-\gamma}\sqrt{\frac{C_{\mathsf{het}}\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}} \tag{139}$$

with at least probability $1-2\delta$. To this end, we first decompose the approximation error as follows:

$$\begin{aligned}
&\min_{\widehat{y}_{\boldsymbol{z}}\in\widehat{\mathcal{Y}}} D(y, \widehat{y}_{\boldsymbol{z}}) \\
&= \min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{k=1}^{K}\sum_{u=0}^{t-1}\left(y_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right)\right| \\
&\le \max_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{k=1}^{K}\sum_{u=0}^{t-LM\tau-1} y_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right| + \min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{k=1}^{K}\sum_{u=t-LM\tau}^{t-1} y_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right| \\
&\le \underbrace{\max_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{k=1}^{K}\sum_{u=0}^{t-LM\tau-1} y_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right|}_{=:D_1} \\
&\quad + \underbrace{\min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{k=1}^{K}\sum_{u=t-LM\tau}^{t-1} (\bar{\omega}_{u,t}^k(s,a) - \widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z}))(P(s,a) - P_{u+1}^k(s,a))V_u^k\right|}_{=:D_2} \\
&\quad + \underbrace{\left|\sum_{k=1}^{K}\sum_{u=t-LM\tau}^{t-1} \chi_{u,t}^k(s,a)(P(s,a) - P_{u+1}^k(s,a))V_u^k\right|}_{=:D_3},
\end{aligned}$$

and will bound each term separately.

- **Bounding D_1.** This term appears only when $t > LM\tau$. Since $\widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) = 0$ for all $u < t - LM\tau$ regardless of $\boldsymbol{z}$ by construction,

$$\begin{aligned}
\left|\sum_{k=1}^{K}\sum_{u=0}^{t-LM\tau-1} y_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right| &\le \sum_{k=1}^{K}\sum_{u\in\mathcal{U}_{0,t-LM\tau}^k(s,a)} \omega_{u,t}^k(s,a)\|P(s,a) - P_{u+1}^k(s,a)\|_1\|V_u^k\|_\infty \\
&\overset{\text{(i)}}{\le} \frac{2}{1-\gamma}\sum_{k=1}^{K}\sum_{u\in\mathcal{U}_{0,t-LM\tau}^k(s,a)} \omega_{u,t}^k(s,a) \\
&\le \frac{2}{1-\gamma}\prod_{h=\phi(t)-LM}^{\phi(t)-1}\left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N_{h\tau,(h+1)\tau}^{k'}(s,a)}\right) \\
&\overset{\text{(ii)}}{\le} \frac{2}{1-\gamma}\exp\left(-\frac{\eta}{2K}\sum_{k'=1}^{K} N_{t-LM\tau,t}^{k'}(s,a)\right) \\
&\overset{\text{(iii)}}{\le} \frac{2}{1-\gamma}\exp\left(-\frac{1}{8}\eta\mu_{\mathsf{avg}}(s,a)LM\tau\right) \\
&\overset{\text{(iv)}}{\le} \frac{2\eta}{(1-\gamma)K},
\end{aligned}$$

  where (i) holds since $\|P(s,a)\|_1$, $\|P_u^k(s,a)\|_1 \le 1$ and $\|V_{u-1}^k\|_\infty \le \frac{1}{1-\gamma}$ (cf. (30)), (ii) follows from (132), (iii) holds due to $\mathcal{B}_M(s,a)$, and (iv) holds because $L \ge 128\log\frac{K}{\eta} \ge \frac{8}{\eta\mu_{\mathsf{avg}}(s,a)M\tau}\log\frac{K}{\eta}$ given that $\eta\mu_{\mathsf{avg}}(s,a)M\tau \ge 1/16$.

- **Bounding D_2.** Since $\widehat{x}_u^k(s,a) = x_u^k(s,a)$ when $\mathcal{B}_M(s,a)$ holds, in view of (138), we have

$$\begin{aligned}
&\min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{k=1}^{K}\sum_{u=t-LM\tau}^{t-1}(\bar{\omega}_{u,t}^k(s,a)-\widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z}))(P(s,a)-P_{u+1}^k(s,a))V_u^k\right| \\
&\le \min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{t-LM\tau,t}(s,a)}\left|\bar{\omega}_{u,t}^k(s,a)-\widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z})\right|\|P(s,a)-P_{u+1}^k(s,a)\|_1\|V_u^k\|_\infty \\
&\le \frac{2}{1-\gamma}\min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left(\sum_{l=1}^{L}\sum_{h=\phi(t)-lM}^{\phi(t)-(l-1)M-1}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\sum_{k=1}^{K}\widehat{x}_u^k(s,a)\left|\prod_{l'=1}^{l}z_{l'}(s,a)-\prod_{l'=1}^{l}\widehat{z}_{l'}\right|\right),
\end{aligned}$$

where the last inequality holds since $\|P(s,a)\|_1$, $\|P_u^k(s,a)\|_1 \le 1$ and $\|V_{u-1}^k\|_\infty \le \frac{1}{1-\gamma}$ (cf. (30)) and the definition of $\widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z})$ defined in (137).

Note that for any given $\{z_l(s,a)\}_{l\in[L]}$, under $\mathcal{B}_M(s,a)$, there exists $\widehat{\boldsymbol{z}}^\star = (\widehat{z}_1^\star,\dots,\widehat{z}_l^\star,\dots,\widehat{z}_L^\star)\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0$ such that $|\widehat{z}_l^\star - z_l(s,a)| \le \frac{\eta}{K}\exp(-1/8\eta\mu_{\mathsf{avg}}(s,a)M\tau)$ for $l<L$ and $|\widehat{z}_L^\star - z_L(s,a)| \le \frac{\eta}{K}$. Also, recall that $z_l(s,a)$, $\widehat{z}_l^\star \le \exp(-1/8\eta\mu_{\mathsf{avg}}(s,a)M\tau)$ for $l<L$ and $z_L(s,a)$, $\widehat{z}_L^\star \le 1$. Then, for any $l \le L$ it follows that:

$$\begin{aligned}
\left|\prod_{l'=1}^{l}z_{l'}(s,a)-\prod_{l'=1}^{l}\widehat{z}_{l'}^\star\right| &\le \Big(\Big|\prod_{l'=1}^{l}z_{l'}(s,a)-\widehat{z}_1^\star\prod_{l'=2}^{l}z_{l'}(s,a)\Big|+\cdots+\Big|z_l\prod_{l'=1}^{l-1}\widehat{z}_{l'}^\star-\prod_{l'=1}^{l}\widehat{z}_{l'}^\star\Big|\Big) \\
&\le \exp\Big(-\frac{1}{8}(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\Big)\sum_{l'=1}^{l}\frac{\eta}{K} \\
&\le \exp\Big(-\frac{1}{8}(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\Big)\frac{L\eta}{K}.
\end{aligned}$$

Then, applying the above bound and (136b) in Lemma 12,

$$\begin{aligned}
D_2 &\le \frac{2}{1-\gamma}\sum_{l=1}^{L}\sum_{h=\phi(t)-lM}^{\phi(t)-(l-1)M-1}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\sum_{k=1}^{K}\widehat{x}_u^k(s,a)\left|\prod_{l'=1}^{l}z_{l'}(s,a)-\prod_{l'=1}^{l}\widehat{z}_{l'}^\star\right| \\
&\le \frac{2}{1-\gamma}\frac{L\eta}{K}\sum_{l=1}^{L}\exp\Big(-\frac{1}{8}(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\Big)\sum_{h=\phi(t)-lM}^{\phi(t)-(l-1)M-1}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\sum_{k=1}^{K}\widehat{x}_u^k(s,a) \\
&\le \frac{2}{1-\gamma}\frac{L\eta}{K}\frac{1}{1-\exp(-1/8\eta\mu_{\mathsf{avg}}(s,a)M\tau)}(16\eta\mu_{\mathsf{avg}}(s,a)M\tau) \\
&\overset{\text{(i)}}{\le} \frac{2}{1-\gamma}\frac{L\eta}{K}\frac{16}{\eta\mu_{\mathsf{avg}}(s,a)M\tau}16\eta\mu_{\mathsf{avg}}(s,a)M\tau \le \frac{512\eta L}{(1-\gamma)K},
\end{aligned}$$

where (i) holds since $\eta\mu_{\mathsf{avg}}(s,a)M\tau/8 \le 1$ and $e^{-x} \le 1-\frac{1}{2}x$ for any $0 \le x \le 1$.

- **Bounding D_3.** Applying Freedman's inequality, we can obtain the following bound, whose proof is provided in Appendix C.3.2.

  Lemma 13. *Consider any $\delta\in(0,1)$ and L defined in (131). For any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $1\le t\le T$, the following holds:*

$$D_3 \le \frac{9}{1-\gamma}\sqrt{\frac{C_{\mathsf{het}}\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}} \tag{140}$$

  with probability at least $1-2\delta$, as long as $\tau \ge t_{\mathsf{th}}$, and $\eta \le \min\{\frac{1}{4\tau K}, \frac{1}{KC_{\mathsf{het}}L\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}\}$.

By combining the bounds obtained above,

$$\begin{aligned}\min_{\widehat{y}_{\boldsymbol{z}}\in\widehat{\mathcal{Y}}} D(y,\widehat{y}_{\boldsymbol{z}}) &\le \frac{2\eta}{(1-\gamma)K}+\frac{512\eta L}{(1-\gamma)K}+\frac{9}{1-\gamma}\sqrt{\frac{C_{\mathsf{het}}\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}\\ &\le \frac{525}{1-\gamma}\sqrt{\frac{C_{\mathsf{het}}\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}\end{aligned}$$

since $\eta\le\frac{K}{128\log(TK)}\le K/L$ due to $L\le128\log(TK)$.

Step 3: concentration bound over $\mathcal{Y}$. We now show that for all elements in $\widehat{\mathcal{Y}}=\{\widehat{y}_{\boldsymbol{z}}:\ \boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0\}$ satisfy

$$\left|\sum_{k=1}^{K}\sum_{u=0}^{t-1}\widehat{y}_{u,t}^{k}(s,a;\boldsymbol{z})\right|<\frac{115}{(1-\gamma)}\sqrt{\frac{\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \tag{141}$$

with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$. It suffices to establish (141) for a fixed $\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0$ with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\mathcal{Y}|}$, where

$$|\widehat{\mathcal{Y}}|=|\mathcal{Z}^{L-1}\times\mathcal{Z}_0|\le(4K\mu_{\mathsf{avg}}(s,a)M\tau)^L\le(K/\eta)^L\le(TK)^L \tag{142}$$

because $\eta\mu_{\mathsf{avg}}(s,a)M\tau\le1/4$ and $\eta\ge1/T$.

For any fixed $\boldsymbol{z}=(\widehat{z}_1,\cdots,\widehat{z}_L)\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0$, since $\widehat{\omega}_{u,t}^{k}(s,a;\boldsymbol{z})=\widehat{x}_u^k(s,a)\prod_{l=1}^{l(u,t)}\widehat{z}_l$ only depends on the events happened until u, which is independent to a transition at $u+1$. Thus, we can apply Freedman's inequality to bound the sum of $\widehat{y}_{u,t}^{k}(s,a;\boldsymbol{z})$ since

$$\mathbb{E}[\widehat{y}_{u,t}^{k}(s,a;\boldsymbol{z})|\mathcal{Y}_u]=0, \tag{143}$$

where $\mathcal{Y}_u$ denotes the history of visited state-action pairs and updated values of all agents until u, i.e., $\mathcal{Y}_u=\{(s_v^k,a_v^k),V_v^k\}_{k\in[K],v\le u}$. Before applying Freedman's inequality, we need to calculate the following quantities. First,

$$B_t(s,a):=\max_{k\in[K],0\le u<t}|\widehat{y}_{u,t}^{k}(s,a;\boldsymbol{z})|\le\widehat{x}_u^k(s,a)\prod_{l=1}^{l(u,t)}\widehat{z}_l\|P(s,a)-P_{u+1}^k(s,a)\|_1\|V_u^k\|_\infty\le\frac{18\eta}{(1-\gamma)K}, \tag{144}$$

where the last inequality follows from $\|P(s,a)\|_1,\ \|P_u^k(s,a)\|_1\le1$, $\|V_{u-1}^k\|_\infty\le\frac{1}{1-\gamma}$ (cf. (30)), $\hat{z}_l\le1$, and (136a) in Lemma 12. Next, we can bound the variance as

$$\begin{aligned}W_t(s,a)&:=\sum_{u=0}^{t}\sum_{k=1}^{K}\mathbb{E}[(\widehat{y}_{u,t}^{k}(s,a;\boldsymbol{z}))^2|\mathcal{Y}_u]\\ &=\sum_{l=1}^{L}\sum_{h=\max\{0,\phi(t)-lM\}}^{\phi(t)-(l-1)M-1}\sum_{k=1}^{K}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^{k}(s,a)}(\widehat{x}_u^k(s,a)\prod_{l'=1}^{l}\widehat{z}_{l'})^2\mathsf{Var}_{P(s,a)}(V_u^k)\\ &\overset{\text{(i)}}{\le}\frac{2}{(1-\gamma)^2}\sum_{l=1}^{L}\left(\prod_{l'=1}^{l}\widehat{z}_{l'}^2\right)\sum_{h=max\{0,\phi(t)-lM\}}^{\phi(t)-(l-1)M-1}\sum_{k=1}^{K}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^{k}(s,a)}(\widehat{x}_u^k(s,a))^2\\ &\overset{\text{(ii)}}{\le}\frac{2}{(1-\gamma)^2}\sum_{l=1}^{L}\left(\prod_{l'=1}^{l}\widehat{z}_{l'}^2\right)\frac{64\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K}\\ &\overset{\text{(iii)}}{\le}\frac{128\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K(1-\gamma)^2}\sum_{l=1}^{L}\exp\left(-1/4(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\right)\end{aligned}$$

$$\leq \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K(1-\gamma)^2}\frac{1}{1-\exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau)}$$
$$\overset{\text{(iv)}}{\leq} \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K(1-\gamma)^2}\frac{8}{\eta\mu_{\mathsf{avg}}(s,a)M\tau} = \frac{1024\eta}{K(1-\gamma)^2} =: \sigma^2, \tag{145}$$

where (i) holds due to the fact that $\|\mathsf{Var}_P(V)\|_\infty \leq \|P\|_1(\|V\|_\infty)^2 + (\|P\|_1\|V\|_\infty)^2 \leq \frac{2}{(1-\gamma)^2}$ because $\|V\|_\infty \leq \frac{1}{1-\gamma}$ (cf. (30)) and $\|P\|_1 \leq 1$, (ii) follows from (136c) in Lemma 12, (iii) holds due to the range of $\mathcal{Z}$ and $\mathcal{Z}_0$ is bounded by $\exp(-1/8\eta\mu_{\mathsf{avg}}(s,a)M\tau)$ and 1, respectively, and (iv) holds since $e^{-x} \leq 1 - \frac{1}{2}x$ for any $0 \leq x \leq 1$ and $\eta\mu_{\mathsf{avg}}(s,a)M\tau/4 \leq 1$.

Now, by substituting the above bounds of W_t and B_t into Freedman's inequality (see Theorem 4) and setting $m = 1$, it follows that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $t \in [T]$ and $\widehat{y}_{\boldsymbol{z}} \in \widehat{\mathcal{Y}}$,

$$\left|\sum_{k=1}^{K}\sum_{u=0}^{t-1} \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right| \leq \sqrt{8\max\{W_t(s,a), \frac{\sigma^2}{2^m}\}\log\frac{4m|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}} + \frac{4}{3}B_t(s,a)\log\frac{4m|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}$$
$$\leq \sqrt{8192\frac{\eta}{K(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}} + \frac{24\eta}{K(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}$$
$$\overset{\text{(i)}}{\leq} \frac{115}{(1-\gamma)}\sqrt{\frac{\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}, \tag{146}$$

with at least probability $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}$, where (i) holds because $|\widehat{\mathcal{Y}}| \leq (TK)^L$ (cf. (142)), and $\frac{\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta} \leq 1$ when $L \leq 128\log(TK)$ and $\eta \leq \frac{K}{128\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}$. Therefore, it follows that (141) holds.

Step 4: putting things together. We now putting all the results obtained in the previous steps together to achieve the claimed bound. Under $\mathcal{B}_M(s,a)$, there exists $\widehat{y}_{\boldsymbol{z}} := \widehat{y}_{\boldsymbol{z}(y)} \in \widehat{\mathcal{Y}}$ such that (139) holds. Hence,

$$\sum_{k=1}^{K}\sum_{u=0}^{t-1} y_{u,t}^k(s,a) \leq \left|\sum_{k=1}^{K}\sum_{u=0}^{t-1} \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right| + D(y,\widehat{y}_{\boldsymbol{z}})$$
$$\leq \frac{115}{(1-\gamma)}\sqrt{\frac{\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} + \frac{525}{1-\gamma}\sqrt{\frac{C_{\mathsf{het}}\eta L}{K}\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}$$
$$\leq \frac{7241}{(1-\gamma)}\sqrt{\frac{C_{\mathsf{het}}\eta}{K}\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}},$$

where the second line holds due to (141) and (139), and the last line holds because $L \leq 128\log(TK)$. By taking a union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $t \in [T]$, we complete the proof.

C.3.1 Proof of Lemma 12

For notational simplicity, let $\overline{h}$ be the largest integer among $h \in (h_0, \phi(t) - (l-1)M)$ such that

$$\sum_{k=1}^{K} N_{h_0\tau,(h-1)\tau}^k(s,a) \leq 2K\mu_{\mathsf{avg}}(s,a)M\tau. \tag{147}$$

Then, the following holds:

$$\sum_{k=1}^{K} N_{h_0\tau,\overline{h}\tau}^k(s,a) = \sum_{k=1}^{K} N_{(\overline{h}-1)\tau,\overline{h}\tau}^k(s,a) + \sum_{k=1}^{K} N_{h_0\tau,(\overline{h}-1)\tau}^k(s,a)$$
$$\leq K\tau + 2K\mu_{\mathsf{avg}}(s,a)M\tau. \tag{148}$$

Also, for the following proofs, we provide an useful bound as follows:

$$\sum_{k'=1}^{K} \frac{(1-\eta)^{-N_{h\tau,(h+1)\tau}^{k'}(s,a)}}{K} \leq \frac{\sum_{k'=1}^{K} e^{\eta N_{h\tau,(h+1)\tau}^{k'}(s,a)}}{K} \leq 1 + 2\eta \frac{\sum_{k'=1}^{K} N_{h\tau,(h+1)\tau}^{k'}(s,a)}{K}$$
$$\leq \exp\left(2\eta \frac{\sum_{k'=1}^{K} N_{h\tau,(h+1)\tau}^{k'}(s,a)}{K}\right), \tag{149}$$

which holds since $1 + x \leq e^x \leq 1 + 2x$ for any $x \in [0,1]$ and $\eta N_{h\tau,(h+1)\tau}^{k'}(s,a) \leq \eta\tau \leq 1$.

According to (135), for any integer $u \in [\overline{h}\tau, t-(l-1)M\tau)$, $\widehat{x}_u^k(s,a)$ is clipped to zero. Now, we prove the bounds in Lemma 12 respectively.

Proof of (136a)**.** For $u \in [h_0\tau, \overline{h}\tau)$,

$$\begin{aligned}
\widehat{x}_u^k(s,a) &= \frac{\eta}{K}(1-\eta)^{-N_{\phi(u)\tau,u+1}^k(s,a)} \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N_{h\tau,(h+1)\tau}^{k'}(s,a)}\right)^{-1} \\
&\overset{\text{(i)}}{\leq} \frac{3\eta}{K} \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N_{h\tau,(h+1)\tau}^{k'}(s,a)}\right)^{-1} \\
&\overset{\text{(ii)}}{\leq} \frac{3\eta}{K} \exp\left(\frac{2\eta}{K}\sum_{k'=1}^{K} N_{h_0\tau,(\overline{h}-1)\tau}^{k'}(s,a)\right) \\
&\overset{\text{(iii)}}{\leq} \frac{3\eta}{K} \exp\left(4\eta\mu_{\mathsf{avg}}(s,a)M\tau\right) \overset{\text{(iv)}}{\leq} \frac{9\eta}{K},
\end{aligned}$$

where (i) holds since $(1+\eta)^x \leq e^{\eta x}$ and $\eta N_{\phi(u)\tau,u+1}^k(s,a) \leq \eta\tau \leq 1$, (ii) holds due to (132) and the fact that $\phi(u) \leq \overline{h} - 1$, (iii) follows from the condition of $\overline{h}$ in (147), and (iv) holds because $4\eta\mu_{\mathsf{avg}}(s,a)M\tau \leq 1$.

Proof of (136b)**.** By the definition of $\overline{h}$, it follows that

$$\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \sum_{k=1}^{K} \widehat{x}_u^k(s,a) = \sum_{h=h_0}^{\overline{h}-1} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \sum_{k=1}^{K} x_u^k(s,a).$$

Using the following relation for each h:

$$\begin{aligned}
&\sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \sum_{k=1}^{K} x_u^k(s,a) \\
&= \frac{1}{K}\left(\sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \eta(1-\eta)^{-N_{\phi(u)\tau,u+1}^k(s,a)}\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N_{h'\tau,(h'+1)\tau}^{k'}(s,a)}\right)^{-1} \\
&= \left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-N_{h\tau,(h+1)\tau}^k(s,a)} - 1\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N_{h'\tau,(h'+1)\tau}^{k'}(s,a)}\right)^{-1} \\
&\leq \left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-N_{h\tau,(h+1)\tau}^k(s,a)} - 1\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-N_{h'\tau,(h'+1)\tau}^{k}(s,a)}\right),
\end{aligned}$$

where the last inequality follows from Jensen's inequality, and applying (149), we can complete the proof as follows:

$$\sum_{h=h_0}^{\overline{h}-1} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \sum_{k=1}^{K} x_u^k(s,a) \leq \prod_{h'=h_0}^{\overline{h}-1} \left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-N_{h'\tau,(h'+1)\tau}^{k}(s,a)}\right) - 1$$

$$
\begin{aligned}
&\le \exp\left(\frac{2\eta\sum_{k'=1}^{K} N^{k'}_{h_0\tau,\overline{h}\tau}(s,a)}{K}\right) - 1 \\
&\overset{\text{(i)}}{\le} \exp\left(4\eta\mu_{\mathsf{avg}}(s,a)M\tau + 2\eta\tau\right) - 1 \\
&\overset{\text{(ii)}}{\le} 16\eta\mu_{\mathsf{avg}}(s,a)M\tau,
\end{aligned}
$$

where (i) follows from (148), and (ii) holds because $e^x \le 1 + 2x$ for any $x \in [0,1]$, $2\eta\tau \le 1/2$, and $4\eta\mu_{\mathsf{avg}}(s,a)M\tau \le 1/2$.

Proof of (136c). Similarly,

$$
\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} (\widehat{x}^k_u(s,a))^2 = \sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} (x^k_u(s,a))^2.
$$

Using the following relation for each h:

$$
\begin{aligned}
&\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} (x^k_u(s,a))^2 \\
&= \frac{1}{K^2}\left(\sum_{k=1}^{K} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \eta^2(1-\eta)^{-2N^k_{\phi(u)\tau,u+1}(s,a)}\right) \prod_{h'=h_0}^{h-1}\left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}\right)^{-2} \\
&\le \frac{\eta}{K}\left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-2N^k_{h\tau,(h+1)\tau}(s,a)} - 1\right) \prod_{h'=h_0}^{h-1}\left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}\right)^{-2} \\
&\le \frac{\eta}{K}\left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-2N^k_{h\tau,(h+1)\tau}(s,a)} - 1\right) \prod_{h'=h_0}^{h-1}\left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-2N^k_{h'\tau,(h'+1)\tau}(s,a)}\right),
\end{aligned}
$$

where the last inequality follows from Jensen's inequality, and applying (149) under the condition $2\eta\tau \le 1$, we can complete the proof as follows:

$$
\begin{aligned}
\sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} (x^k_u(s,a))^2 &\le \frac{\eta}{K}\prod_{h'=h_0}^{\overline{h}-1}\left(\frac{1}{K}\sum_{k=1}^{K}(1-\eta)^{-2N^k_{h'\tau,(h'+1)\tau}(s,a)}\right) - 1 \\
&\le \frac{\eta}{K}\left(\exp\left(4\eta\frac{\sum_{k'=1}^{K} N^{k'}_{h_0\tau,\overline{h}\tau}(s,a)}{K}\right) - 1\right) \\
&\overset{\text{(i)}}{\le} \frac{\eta}{K}\left(\exp\left(8\eta\mu_{\mathsf{avg}}(s,a)M\tau + 4\eta\tau\right) - 1\right) \\
&\overset{\text{(ii)}}{\le} \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K},
\end{aligned}
$$

where (i) follows from (148), and (ii) holds because $e^x \le 1 + 4x$ for any $x \in [0,2]$, $4\eta\tau \le 1$, and $8\eta\mu_{\mathsf{avg}}(s,a)M\tau \le 1$.

C.3.2 Proof of Lemma 13

Recall that

$$
\chi^k_{u,t}(s,a) = \frac{\eta}{K}(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)}\left(\frac{(1-\eta)^{N^k_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{k'=1}^{K}(1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}} - 1\right)\prod_{h=\phi(u)}^{\phi(t)-1}\left(\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N^{k'}_{h\tau,(h+1)\tau}(s,a)}\right)
$$

$$= \left(\frac{(1-\eta)^{N^k_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{k'=1}^K (1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}} - 1 \right) \omega^k_{u,t}(s,a).$$

We can observe that $\chi^k_{u,t}(s,a)$ and $\omega^k_{u,t}(s,a)$ are solely determined by the number of visits of agents during local steps, i.e., $(N^k_{h\tau,(h+1)\tau}(s,a))_{k\in[K],h\in[\phi(t)-LM,\phi(t)-1]}$. It thus suffice to consider $\{\chi^k_{u,t}(s,a;\boldsymbol{N})\}_{0\le u<t,k\in[K]}$ and $\{\omega^k_{u,t}(s,a;\boldsymbol{N})\}_{0\le u<t,k\in[K]}$ constructed with each of the possible combinations of number of visits for all $k\in[K]$ and $h\in[\phi(t)-LM,\phi(t)-1]$, i.e., $\boldsymbol{N}\in[0,\tau]^{KLM}$. Then, by setting $X = 9\sqrt{\frac{C_{\mathsf{het}}\eta L}{K(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}$ and taking an union bound,

$$\begin{aligned}
&\mathbb{P}\left[\left|\sum_{k=1}^K\sum_{u=t-LM\tau}^{t-1}\chi^k_{u,t}(s,a)(P(s,a)-P^k_{u+1}(s,a))V^k_u\right|\ge X\right]\\
&= \sum_{\boldsymbol{N}\in[0,\tau]^{KLM}}\mathbb{P}\left[\left|\sum_{k=1}^K\sum_{u=t-LM\tau}^{t-1}\chi^k_{u,t}(s,a)(P(s,a)-P^k_{u+1}(s,a))V^k_u\right|\ge X, \chi^k_{u,t}(s,a)=\chi^k_{u,t}(s,a;\boldsymbol{N})\right]\\
&\le \sum_{\boldsymbol{N}\in[0,\tau]^{KLM}}\mathbb{P}\left[\left|\sum_{k=1}^K\sum_{u=t-LM\tau}^{t-1}\chi^k_{u,t}(s,a;\boldsymbol{N})(P(s,a)-P^k_{u+1}(s,a))V^k_u\right|\ge X\right],
\end{aligned}$$

and it suffices to show that

$$\mathbb{P}\left[\left|\sum_{k=1}^K\sum_{u=t-LM\tau}^{t-1}\chi^k_{u,t}(s,a;\boldsymbol{N})(P(s,a)-P^k_{u+1}(s,a))V^k_u\right|\ge X\right]\le\frac{\delta}{|\mathcal{S}||\mathcal{A}|T(1+\tau)^{KLM}}.$$

Since $\chi^k_{u,t}(s,a;\boldsymbol{N})$ is a constant, which does not depend on $P^k_{u+1}(s,a)$,

$$\mathbb{E}[\chi^k_{u,t}(s,a;\boldsymbol{N})(P(s,a)-P^k_{u+1}(s,a))V^k_u|\mathcal{Y}_u]=0, \tag{150}$$

where $\mathcal{Y}_u$ denotes the history of visited state-action pairs and updated values of all agents until u, i.e., $\mathcal{Y}_u=\{(s^k_v,a^k_v),V^k_v\}_{k\in[K],v\le u}$, and thus, we can apply Freedman's inequality to bound the sum.

Before applying Freedman's inequality, we need to calculate the following quantities. First,

$$\begin{aligned}
B_t(s,a) &:= \max_{k\in[K],t-LM\tau\le u<t}|\chi^k_{u,t}(s,a;\boldsymbol{N})(P(s,a)-P^k_{u+1}(s,a))V^k_u|\\
&\le \max_{k\in[K],t-LM\tau\le u<t}\left|1-\frac{\frac{1}{K}\sum_{k'=1}^K(1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{(1-\eta)^{N^k_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}\right|\omega^k_{u,t}(s,a;\boldsymbol{N})\|P(s,a)-P^k_{u+1}(s,a)\|_1\|V^k_u\|_\infty\\
&\overset{\text{(i)}}{\le}\frac{2}{1-\gamma}\max_{k\in[K],t-LM\tau\le u<t}\left|1-\frac{\frac{1}{K}\sum_{k'=1}^K(1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{(1-\eta)^{N^k_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}\right|\omega^k_{u,t}(s,a;\boldsymbol{N})\\
&\overset{\text{(ii)}}{\le}\frac{8\eta\mu_{\mathsf{max}}(s,a)\tau}{1-\gamma}\max_{k\in[K],t-LM\tau\le u<t}\omega^k_{u,t}(s,a;\boldsymbol{N})\overset{\text{(iii)}}{\le}\frac{8\eta^2\mu_{\mathsf{max}}(s,a)\tau}{(1-\gamma)K},
\end{aligned}$$

where (i) holds because $\|P(s,a)\|_1$, $\|P^k_u(s,a)\|_1\le1$, $\|V^k_{u-1}\|_\infty\le\frac{1}{1-\gamma}$ (cf. (30)), (ii) follows from the fact that (which will be shown at the end of the proof)

$$\left|1-\frac{\frac{1}{K}\sum_{k'=1}^K(1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{(1-\eta)^{N^k_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}\right|\le4\eta\mu_{\mathsf{max}}(s,a)\tau, \tag{151}$$

with $\mu_{\mathsf{max}}(s,a):=\max_k\mu^k_{\mathsf{b}}(s,a)$, and (iii) holds due to the fact that $\omega^k_{u,t}(s,a;\boldsymbol{N})\le\frac{\eta}{K}$.

Next, we can bound the variance as

$$W_t(s,a):=\sum_{u=\max\{0,t-LM\tau\}}^{t-1}\sum_{k=1}^K\mathbb{E}\Big[\Big(\chi^k_{u,t}(s,a;\boldsymbol{N})(P(s,a)-P^k_{u+1}(s,a))V^k_u\Big)^2|\mathcal{Y}_u\Big]$$

$$\begin{aligned}
&\overset{\text{(i)}}{\leq} (4\eta\mu_{\mathsf{max}}(s,a)\tau)^2 \sum_{h=\max\{0,\phi(t)-LM\}}^{\phi(t)-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} \Big(\omega^k_{u,t}(s,a;\boldsymbol{N})\Big)^2 \mathsf{Var}_{P(s,a)}(V^k_u)\\
&\overset{\text{(ii)}}{\leq} \frac{2(4\eta\mu_{\mathsf{max}}(s,a)\tau)^2}{(1-\gamma)^2} \sum_{h=\max\{0,\phi(t)-LM\}}^{\phi(t)-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} \Big(\omega^k_{u,t}(s,a;\boldsymbol{N})\Big)^2\\
&\overset{\text{(iii)}}{\leq} \frac{2(4\eta\mu_{\mathsf{max}}(s,a)\tau)^2}{(1-\gamma)^2}\frac{6\eta}{K} =: \sigma^2,
\end{aligned}$$

where (i) follows from (151), (ii) holds due to the fact that $\|\mathsf{Var}_P(V)\|_\infty \leq \|P\|_1(\|V\|_\infty)^2 + (\|P\|_1\|V\|_\infty)^2 \leq \frac{2}{(1-\gamma)^2}$ because $\|V\|_\infty \leq \frac{1}{1-\gamma}$ (cf. (30)) and $\|P\|_1 \leq 1$, (iii) follows from (52d) in Lemma 3.

Now, by substituting the above bounds of W_t and B_t into Freedman's inequality (see Theorem 4) and setting $m=1$, it follows that for any $s\in\mathcal{S}$, $a\in\mathcal{A}$, $t\in[T]$ and $\boldsymbol{N} = (N^k_{h\tau,(h+1)\tau}(s,a))_{k\in[K],h\in[\phi(t)-LM,\phi(t)-1]} \in [0,\tau]^{KLM}$,

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\sum_{u=t-LM\tau}^{t-1} \chi^k_{u,t}(s,a;\boldsymbol{N})(P(s,a)-P^k_{u+1}(s,a))V^k_u\right|\\
&\leq \sqrt{8\max\{W_t(s,a),\frac{\sigma^2}{2^m}\}\log\frac{4m|\mathcal{S}||\mathcal{A}|T(1+\tau)^{KLM}}{\delta}} + \frac{4}{3}B_t(s,a)\log\frac{4m|\mathcal{S}||\mathcal{A}|T(1+\tau)^{KLM}}{\delta}\\
&\leq \sqrt{96\frac{(4\eta\mu_{\mathsf{max}}(s,a)\tau)^2\eta}{K(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)^{KLM}}{\delta}} + \frac{12\eta^2\mu_{\mathsf{max}}(s,a)\tau}{K(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)^{KLM}}{\delta}\\
&\leq \sqrt{384\frac{(4\eta\tau K)(\mu_{\mathsf{max}}(s,a)^2\eta M\tau)L\eta}{K(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta}} + \frac{12L\eta(\mu_{\mathsf{max}}(s,a)\eta M\tau)}{(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta}\\
&\overset{\text{(i)}}{\leq} \sqrt{48\frac{C_{\mathsf{het}}L\eta}{K(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta}} + \frac{2C_{\mathsf{het}}L\eta}{(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta}\\
&\overset{\text{(ii)}}{\leq} 9\sqrt{\frac{C_{\mathsf{het}}\eta L}{K(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}} \qquad (152)
\end{aligned}$$

with at least probability $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T(1+\tau)^{KLM}}$, where we invoke the definition of C_{het} (cf. (21)). Here, (i) holds because $\eta\tau K \leq 1/4$ and $\mu_{\mathsf{max}}(s,a)\eta M\tau \leq C_{\mathsf{het}}\mu_{\mathsf{avg}}(s,a)\eta M\tau \leq \frac{C_{\mathsf{het}}}{8}$, and (ii) follows from the fact that $\eta \leq \frac{1}{128KC_{\mathsf{het}}\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}} \leq \frac{1}{KC_{\mathsf{het}}L\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}$.

Proof of (151). Using the fact that for $0<\eta<1$,

$$(1-\eta)^{-n} \leq e^{\eta n} \leq 1+2\eta n \;\text{ if }\; n\geq 0 \;\text{ and }\; \eta n \leq 1, \;\text{ and }\; (1-\eta)^n \geq 1-\eta n \;\text{ if }\; n\leq 0 \text{ or } n\geq 1,$$

we can obtain the bounds as follows:

$$\begin{aligned}
1-\frac{\eta}{K}\sum_{k'=1}^{K} N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a) \leq \frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)} &\leq \frac{\frac{1}{K}\sum_{k'=1}^{K}(1-\eta)^{N^{k'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{(1-\eta)^{N^{k}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}\\
&\leq (1-\eta)^{-N^{k}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}\\
&\leq 1+2\eta N^{k}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a).
\end{aligned}$$

Thus, recalling $\mu_{\mathsf{max}}(s,a) := \max_k \mu^k_{\mathsf{b}}(s,a)$, and using the fact that for any $(s,a,k,u)\in\mathcal{S}\times\mathcal{A}\times[K]\times[T]$:

$$N^{k}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a) \leq 2\mu_{\mathsf{max}}(s,a)\tau$$

at least with probability $1-\delta$, as long as $\tau \geq 443\left(\frac{t_{\mathsf{mix}}^k}{\mu_{\mathsf{max}}(s,a)}\right)\log\frac{4|\mathcal{S}||\mathcal{A}|TK}{\delta}$, which naturally holds if $\tau \geq t_{\mathsf{th}}$ (see (77) for the definition of t_{th}), according to Lemma 9,

$$\left|1-\frac{\frac{1}{K}\sum_{k'=1}^K(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{k'}(s,a)}}{(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{k}(s,a)}}\right| \leq 2\eta\max\left\{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{k}(s,a), \frac{1}{K}\sum_{k'=1}^K N_{\phi(u)\tau,(\phi(u)+1)\tau}^{k'}(s,a)\right\}$$
$$\leq 4\eta\mu_{\mathsf{max}}(s,a)\tau.$$

C.4 Proof of Lemma 5

For any $t \geq \beta\tau$, the error term can be decomposed as follows:

$$\begin{aligned} E_t^3(s,a) &= \gamma\sum_{k=1}^K\sum_{u\in\mathcal{U}_{0,t}^k(s,a)}\omega_{u,t}^k(s,a)P(s,a)(V^\star - V_u^k) \\ &= \underbrace{\gamma\sum_{k=1}^K\sum_{u\in\mathcal{U}_{0,(\phi(t)-\beta)\tau}^k(s,a)}\omega_{u,t}^k(s,a)P(s,a)(V^\star - V_u^k)}_{=:E_t^{3a}(s,a)} \\ &\qquad + \underbrace{\gamma\sum_{k=1}^K\sum_{u\in\mathcal{U}_{(\phi(t)-\beta)\tau,t}^k(s,a)}\omega_{u,t}^k(s,a)P(s,a)(V^\star - V_u^k)}_{=:E_t^{3b}(s,a)}. \end{aligned} \tag{153}$$

We shall these two terms separately.

- **Bounding $E_t^{3a}(s,a)$.** First, the bound on $E_t^{3a}(s,a)$ is derived as follows:

$$\begin{aligned} |E_t^{3a}(s,a)| &\leq \gamma\sum_{k=1}^K\sum_{u\in\mathcal{U}_{0,(\phi(t)-\beta)\tau}^k(s,a)}\omega_{u,t}^k(s,a)\|P(s,a)\|_1\|(V^\star - V_u^k)\|_\infty \\ &\overset{\text{(i)}}{\leq} \frac{2\gamma}{1-\gamma}\sum_{k=1}^K\sum_{u\in\mathcal{U}_{0,(\phi(t)-\beta)\tau}^k(s,a)}\omega_{u,t}^k(s,a) \\ &\overset{\text{(ii)}}{\leq} \frac{2\gamma}{1-\gamma}\exp\left(-\frac{\eta}{2K}\sum_{k=1}^K N_{(\phi(t)-\beta)\tau,t}^k(s,a)\right) \\ &\overset{\text{(iii)}}{\leq} \frac{2\gamma}{1-\gamma}\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right), \end{aligned} \tag{154}$$

where (i) holds because $\|V_u^k\|_\infty, \|V^\star\|_\infty \leq \frac{1}{1-\gamma}$ (cf. (30)) and $\|P(s,a)\|_1 \leq 1$, (ii) holds due to (52c) in Lemma 3, and (iii) follows from the fact that $\sum_{k=1}^K N_{(\phi(t)-\beta)\tau,t}^k(s,a) \geq \frac{K\mu_{\mathsf{avg}}\beta\tau}{4}$ according to Lemma 10 as long as $\beta\tau \geq t_{\mathsf{th}}$.

- **Bounding $E_t^{3b}(s,a)$.** Next, we bound $E_t^{3b}(s,a)$ as follows:

$$\begin{aligned} |E_t^{3b}(s,a)| &\leq \gamma\sum_{k=1}^K\sum_{u\in\mathcal{U}_{(\phi(t)-\beta)\tau,t}^k(s,a)}\omega_{u,t}^k(s,a)\left\|V^\star - V_u^k\right\|_\infty \\ &\overset{\text{(i)}}{\leq} \gamma\sum_{k=1}^K\sum_{h=\phi(t)-\beta}^{\phi(t)-1}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)}\omega_{u,t}^k(s,a)(\|\Delta_{h\tau}\|_\infty + \|Q_u^k - Q_{h\tau}^k\|_\infty) \end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \gamma \sum_{k=1}^{K} \sum_{h=\phi(t)-\beta}^{\phi(t)-1} \sum_{u \in \mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \omega^k_{u,t}(s,a)((1+2\eta\tau)\|\Delta_{h\tau}\|_\infty + \sigma_{\mathsf{local}}) \tag{155}$$

where (i) follows from the following bound, which will be shown in Appendix C.4.1,

$$\|V^\star - V^k_u\|_\infty \leq \|\Delta^k_{\iota(u)}\|_\infty + \|Q^k_u - Q^k_{\iota(u)}\|_\infty, \tag{156}$$

and (ii) holds due to the following lemma.

Lemma 14. *Assume $\eta\tau \leq \frac{1}{2}$. For any given $\delta \in (0,1)$, the following holds for any $k \in [K]$ and $0 \leq u < T$:*

$$\|Q^k_u - Q^k_{\iota(u)}\|_\infty \leq 2\eta\tau\|\Delta^k_{\iota(u)}\|_\infty + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}} \tag{157}$$

with probability at least $1-\delta$.

Here, for notation simplicity, we denote $\sigma_{\mathsf{local}} := \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}}$.

Then, with some algebraic calculations, we can obtain the bound on $E^{3b}_t(s,a)$ as follows:

$$\begin{aligned} |E^{3b}_t(s,a)| &\overset{\text{(i)}}{\leq} \sigma_{\mathsf{local}} + \gamma \sum_{h=\phi(t)-\beta}^{\phi(t)-1} (1+2\eta\tau)\|\Delta_{h\tau}\|_\infty \sum_{k=1}^{K} \sum_{u \in \mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \omega^k_{u,t}(s,a) \\ &\overset{\text{(ii)}}{\leq} \sigma_{\mathsf{local}} + \frac{1+\gamma}{2} \max_{\phi(t)-\beta \leq h < \phi(t)} \|\Delta_{h\tau}\|_\infty \sum_{k=1}^{K} \sum_{h=\phi(t)-\beta}^{\phi(t)-1} \sum_{u \in \mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \omega^k_{u,t}(s,a) \\ &\overset{\text{(iii)}}{\leq} \sigma_{\mathsf{local}} + \frac{1+\gamma}{2} \max_{\phi(t)-\beta \leq h < \phi(t)} \|\Delta_{h\tau}\|_\infty, \end{aligned} \tag{158}$$

where (i) holds according to (52b) of Lemma 3, (ii) holds when η is small enough that $\eta \leq \frac{1-\gamma}{4\gamma\tau}$, and (iii) follows from (52b) of Lemma 3.

Now we have the bounds of $E^{3a}_t(s,a)$ and $E^{3b}_t(s,a)$ separately obtained above. By combining the bounds in (153), we can claim the advertised bound, which completes the proof.

C.4.1 Proof of (156)

We prove the claim by showing

$$\Delta^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) - d^k_{\iota(u),u}(s, a^\star(s)) \leq V^\star(s) - V^k_u(s) \leq \Delta^k_{\iota(u)}(s, a^\star(s)) - d^k_{\iota(u),u}(s, a^\star(s)) \tag{159}$$

for any $s \in \mathcal{S}$. The upper bound is derived as follows:

$$\begin{aligned} V^\star(s) - V^k_u(s) &= Q^\star(s, a^\star(s)) - Q^k_u(s, a^k_u(s)) \\ &\leq Q^\star(s, a^\star(s)) - Q^k_u(s, a^\star(s)) \\ &= Q^\star(s, a^\star(s)) - Q^k_{\iota(u)}(s, a^\star(s)) - \underbrace{(Q^k_u(s, a^\star(s)) - Q^k_{\iota(u)}(s, a^\star(s)))}_{d^k_{\iota(u),u}(s,a^\star(s))} \end{aligned} \tag{160}$$

using the fact that $Q^k_u(s, a^k_u(s)) \geq Q^k_u(s, a^\star(s))$. Similarly, the lower bound is obtained as follows:

$$\begin{aligned} V^\star(s) - V^k_u(s) &= Q^\star(s, a^\star(s)) - Q^k_u(s, a^k_u(s)) \\ &= Q^\star(s, a^\star(s)) - Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) + Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) - Q^k_u(s, a^k_u(s)) \\ &= Q^\star(s, a^\star(s)) - Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) + Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) - Q^k_{\iota(u)}(s, a^k_u(s)) - d^k_{\iota(u),u}(s, a^k_u(s)) \\ &\geq Q^\star(s, a^k_{\iota(u)}(s)) - Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) + Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) - Q^k_{\iota(u)}(s, a^k_u(s)) - d^k_{\iota(u),u}(s, a^k_u(s)) \\ &\geq Q^\star(s, a^k_{\iota(u)}(s)) - Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) - d^k_{\iota(u),u}(s, a^k_u(s)) \end{aligned} \tag{161}$$

using the fact that $Q^\star(s, a^k_{\iota(u)}(s)) \leq Q^\star(s, a^\star(s))$ and $Q^k_{\iota(u)}(s, a^k_{\iota(u)}(s)) \geq Q^k_{\iota(u)}(s, a^k_u(s))$.

C.4.2 Proof of Lemma 14

For any $0 \leq u < T$, $k \in [K]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can write the bound as

$$|Q_u^k(s,a) - Q_{\iota(u)}^k(s,a)| \leq \underbrace{2\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} \|\Delta_v^k\|_\infty}_{:=B_1} + \underbrace{\left| \gamma\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} (P_{v+1}^k(s,a) - P(s,a))V^\star \right|}_{:=B_2}. \tag{162}$$

The inequality holds by the local update rule:

$$\begin{aligned} Q_{v+1}^k(s,a) - Q_v^k(s,a) &= (1-\eta)Q_v^k(s,a) + \eta(r(s,a) + \gamma P_{v+1}^k(s,a)V_v^k) - Q_v^k(s,a) \\ &= \eta(r(s,a) + \gamma P_{v+1}^k(s,a)V_v^k - Q_v^k(s,a)) \\ &= \eta(\gamma P_{v+1}^k(s,a)V_v^k - \gamma P(s,a)V^\star + Q^\star(s,a) - Q_v^k(s,a)) \\ &= \gamma\eta P_{v+1}^k(s,a)(V_v^k - V^\star) + \gamma\eta(P_{v+1}^k(s,a) - P(s,a))V^\star + \eta\Delta_v^k(s,a), \end{aligned} \tag{163}$$

and

$$\begin{aligned} |Q_u^k(s,a) - Q_{\iota(u)}^k(s,a)| &\leq \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} |Q_{v+1}^k(s,a) - Q_v^k(s,a)| \\ &\leq \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} \big(\eta|\Delta_v^k(s,a)| + \gamma\eta|P_{v+1}^k(s,a)(V_v^k - V^\star)|\big) \\ &\qquad + \left| \gamma\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} (P_{v+1}^k(s,a) - P(s,a))V^\star \right| \\ &\leq \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} 2\eta\|\Delta_v^k\|_\infty + \left| \gamma\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} (P_{v+1}^k(s,a) - P(s,a))V^\star \right|, \end{aligned} \tag{164}$$

where the last inequality holds since $\|P_{v+1}^k(s,a)\|_1 \leq 1$ and $\|V_v^k - V^\star\|_\infty \leq \|Q_v^k - Q^\star\|_\infty$ (cf. (31)).

Now, we shall bound each term separately.

- **Bounding B_1.** The local error $\|\Delta_v^k\|_\infty$ is bounded as follows.

  Lemma 15. *Assume $\eta\tau \leq \frac{1}{2}$. For any given $\delta \in (0,1)$, the following holds for any $k \in [K]$ and $0 \leq u < T$:*

  $$\|\Delta_u^k\|_\infty \leq \|\Delta_{\iota(u)}^k\|_\infty + \frac{2\gamma}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} \tag{165}$$

  with probability at least $1-\delta$.

  Then, combining the fact that the number of local updates before the periodic averaging is at most $\tau - 1$, we can conclude that

  $$\begin{aligned} 2\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} \|\Delta_v^k\|_\infty &\leq 2\eta|\mathcal{U}_{\iota(u),u}^k(s,a)| \max_{v \in \mathcal{U}_{\iota(u),u}^k(s,a)} \|\Delta_v^k\|_\infty \\ &\leq 2\eta(\tau-1)\left(\|\Delta_{\iota(u)}^k\|_\infty + \frac{2}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} \right). \end{aligned} \tag{166}$$

- **Bounding B_2.** Exploiting the independence of the transitions and applying the Hoeffding inequality and using the fact that $|\mathcal{U}^k_{\iota(u),u}(s,a)| \leq \tau - 1$, B_2 is bounded as follows:

$$
\begin{aligned}
B_2 &\leq \gamma\eta \sqrt{\sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} |(P^k_{v+1}(s,a) - P(s,a))V^\star| \log \frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} \\
&\leq \frac{2\gamma\eta}{1-\gamma}\sqrt{(\tau-1)\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} \qquad (167)
\end{aligned}
$$

  for any $k \in [K]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $0 \leq u < T$ with probability at least $1-\delta$, where the last inequality follows from $\|V^\star\|_\infty \leq \frac{1}{1-\gamma}$, $\|P^k_{v+1}(s,a)\|_1$, and $\|P(s,a)\|_1 \leq 1$.

By substituting the bound on B_1 and B_2 into (162) and using the condition that $\eta\tau < 1$, we can claim the stated bound holds and this completes the proof.

C.4.3 Proof of Lemma 15

For each state-action $(s,a) \in \mathcal{S} \times \mathcal{A}$ and agent k, by invoking the recursive relation (48) derived from the local Q-learning update in (23), Δ^k_u is decomposed as follows:

$$
\begin{aligned}
\Delta^k_u(s,a) = \underbrace{(1-\eta)^{N^k_{\iota(u),u}(s,a)}\Delta^k_{\iota(u)}(s,a)}_{=:D_1} + \underbrace{\gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)}(P(s,a) - P^k_{v+1}(s,a))V^\star}_{=:D_2} \\
\underbrace{+\gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)} P^k_{v+1}(s,a)(V^\star - V^k_v)}_{=:D_3}. \qquad (168)
\end{aligned}
$$

Now, we obtain the bound on the three decomposed terms separately.

- **Bounding D_1.** The term D_1 can be bounded by

$$
|D_1| \leq (1-\eta)^{N^k_{\iota(u),u}(s,a)}\|\Delta^k_{\iota(u)}\|_\infty. \qquad (169)
$$

- **Bounding D_2.** By applying the Hoeffding bound using the independence of transitions, the second term is bounded as follows:

$$
\begin{aligned}
|D_2| &\leq \gamma \sqrt{\sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} (\eta(1-\eta)^{N^k_{v+1,u}(s,a)})^2(\|V^\star\|_\infty)^2 \log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} \\
&\leq \frac{\gamma}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} := \rho \qquad (170)
\end{aligned}
$$

  with probability at least $1-\delta$, where the last inequality holds due to the fact that $\|V^\star\|_\infty \leq \frac{1}{1-\gamma}$ and

$$
\sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} (\eta(1-\eta)^{N^k_{v+1,u}(s,a)})^2 \leq \eta^2(1 + (1-\eta)^2 + (1-\eta)^4 + \cdots) \leq \eta.
$$

  See (Li et al., 2021b, Lemma 1) for the detailed explanation of the bound.

- **Bounding D_3.** Lastly, we bound the third term as follows:

$$
|D_3| \leq \gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)}\|P^k_{v+1}(s,a)\|_1\|V^\star - V^k_v\|_\infty
$$

$$\leq \gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)} \|\Delta^k_v\|_\infty, \tag{171}$$

where the last inequality follows from the fact that $\|P^k_{v+1}(s,a)\|_1 = 1$ and

$$Q^k_v(s, a^\star(s)) - Q^\star(s, a^\star(s)) \leq V^k_v(s) - V^\star(s) \leq Q^k_v(s, a^k_v(s)) - Q^\star(s, a^k_v(s))$$

for any $s \in \mathcal{S}$, where we denote $a^\star(s) = \arg\max_a Q^\star(s,a)$, $a^k_v(s) = \arg\max_a Q^k_v(s,a)$.

By combining the bounds of the above three terms, we obtain the following recursive relation:

$$|\Delta^k_u(s,a)| \leq (1-\eta)^{N^k_{\iota(u),u}(s,a)} \|\Delta^k_{\iota(u)}\|_\infty + \rho + \gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)} \|\Delta^k_v\|_\infty. \tag{172}$$

Using the recursive relation, we will prove that the following claim holds for any $0 \leq m < \tau$ by induction:

$$\|\Delta^k_{\iota(u)+m}\|_\infty \leq \|\Delta^k_{\iota(u)}\|_\infty + 2\rho, \tag{173}$$

which completes the proof of Lemma 15. First, if $m = 0$, the claim is obviously true. Suppose the claim holds for $\iota(u), \iota(u)+1, \cdots, \iota(u)+m-1$. Then, for $u = \iota(u)+m$, by invoking the recursive relation (172), we can show that the claim (173) holds for m as follows:

$$\begin{aligned}
&|\Delta^k_{\iota(u)+m}(s,a)| \\
&\leq (1-\eta)^{N^k_{\iota(u),u}(s,a)} \|\Delta^k_{\iota(u)}\|_\infty + \rho + \gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)} (\|\Delta^k_{\iota(u)}\|_\infty + 2\rho) \\
&= ((1-\eta)^{N^k_{\iota(u),u}(s,a)} + \gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)}) \|\Delta^k_{\iota(u)}\|_\infty + (1 + 2\gamma \sum_{v \in \mathcal{U}^k_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^k_{v+1,u}(s,a)})\rho \\
&= ((1-\eta)^{N^k_{\iota(u),u}(s,a)} + \gamma(1 - (1-\eta)^{N^k_{\iota(u),u}(s,a)})) \|\Delta^k_{\iota(u)}\|_\infty + (1 + 2\gamma(1 - (1-\eta)^{N^k_{\iota(u),u}(s,a)}))\rho \\
&\leq \|\Delta^k_{\iota(u)}\|_\infty + 2\rho,
\end{aligned} \tag{174}$$

where the last inequality holds since

$$(1-\eta)^{N^k_{\iota(u),u}(s,a)} \geq (1-\eta)^\tau \geq (\frac{1}{4})^{\eta\tau} \geq \frac{1}{2}$$

provided that $\eta\tau \leq \frac{1}{2}$.

C.5 Proof of Lemma 6

First, using the fact that

$$1 \leq (1-\eta)^{-N^k_{t-\tau,t}(s,a)} \leq e^{\eta\tau} \leq 3$$

given that $\eta\tau \leq 1$, by the definition of α^k_t (cf. (27)), we derive (66a) as follows:

$$\frac{1}{3K} \leq \frac{1}{K \max_{k' \in [K]} (1-\eta)^{-N^{k'}_{t-\tau,t}(s,a)}} \leq \alpha^k_t(s,a) = \frac{(1-\eta)^{-N^k_{t-\tau,t}(s,a)}}{\sum_{k'=1}^K (1-\eta)^{-N^{k'}_{t-\tau,t}(s,a)}} \leq \frac{(1-\eta)^{-N^k_{t-\tau,t}(s,a)}}{K} \leq \frac{3}{K}.$$

Moving onto (66b), it follows that

$$\begin{aligned}
\widetilde{\omega}_{0,t}(s,a) &= \prod_{h=0}^{\phi(t)-1} \widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a) \\
&= \prod_{h=0}^{\phi(t)-1} \sum_{k=1}^K \alpha^k_{(h+1)\tau}(s,a)(1-\eta)^{N^k_{h\tau,(h+1)\tau}(s,a)}
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} \prod_{h=0}^{\phi(t)-1} \frac{K}{\sum_{k=1}^{K}(1-\eta)^{-N^k_{h\tau,(h+1)\tau}(s,a)}} \\
&\overset{\text{(ii)}}{\leq} \prod_{h=0}^{\phi(t)-1} \frac{1}{(1-\eta)^{-\frac{1}{K}\sum_{k=1}^{K} N^k_{h\tau,(h+1)\tau}(s,a)}} \\
&= (1-\eta)^{\sum_{h=0}^{\phi(t)-1}\frac{1}{K}\sum_{k=1}^{K} N^k_{h\tau,(h+1)\tau}(s,a)} = (1-\eta)^{\frac{1}{K}\sum_{k=1}^{K} N^k_{0,t}(s,a)},
\end{aligned}
$$

where (i) follows from the definition of α_t^k (cf. (27)), (ii) follows from Jensen's inequality.

Next, we obtain (66c) through the following derivation:

$$
\begin{aligned}
&\sum_{k=1}^{K} \sum_{u\in\mathcal{U}^k_{0,t}(s,a)} \widetilde{\omega}^k_{u,t}(s,a) = \sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \widetilde{\omega}^k_{u,t}(s,a) \\
&= \sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1} \alpha^k_{(h+1)\tau}(s,a) \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \eta(1-\eta)^{N^k_{u+1,(h+1)\tau}(s,a)} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \\
&= \sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1} \alpha^k_{(h+1)\tau}(s,a) \left(1-(1-\eta)^{N^k_{h\tau,(h+1)\tau}(s,a)}\right) \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \\
&\overset{\text{(i)}}{=} \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \sum_{k=1}^{K} \alpha^k_{(h+1)\tau}(s,a) \left(1-(1-\eta)^{N^k_{h\tau,(h+1)\tau}(s,a)}\right) \\
&\overset{\text{(ii)}}{=} \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \left(1-\sum_{k=1}^{K} \alpha^k_{(h+1)\tau}(s,a)(1-\eta)^{N^k_{h\tau,(h+1)\tau}(s,a)}\right) \\
&= \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \left(1-\widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a)\right) \\
&\overset{\text{(iii)}}{=} 1-\widetilde{\lambda}_{0,\tau}(s,a)\widetilde{\lambda}_{\tau,2\tau}(s,a)\cdots\widetilde{\lambda}_{(\phi(t)-1)\tau,t}(s,a) = 1-\widetilde{\omega}_{0,t}(s,a),
\end{aligned}
\tag{175}
$$

where (i) follows by reordering the summation, (ii) follows by $\sum_{k=1}^{K}\alpha^k_t(s,a)=1$, and (iii) holds by cancellation.

In a similar manner, (66d) is derived as follows:

$$
\begin{aligned}
\sum_{k=1}^{K} \sum_{u\in\mathcal{U}^k_{0,h'\tau}(s,a)} \widetilde{\omega}^k_{u,t}(s,a) &= \sum_{k=1}^{K}\sum_{h=0}^{h'-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \widetilde{\omega}^k_{u,t}(s,a) \\
&= \sum_{h=0}^{h'-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \left(1-\widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a)\right) \\
&\leq \prod_{l=h'}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a) \\
&\leq (1-\eta)^{\frac{1}{K}\sum_{k=1}^{k} N^k_{h'\tau,t}(s,a)},
\end{aligned}
$$

where the last inequality follows from

$$
\prod_{l=h'}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a) = \prod_{h=h'}^{\phi(t)-1} \frac{K}{\sum_{k=1}^{K}(1-\eta)^{-N^k_{h\tau,(h+1)\tau}(s,a)}} \leq \prod_{h=h'}^{\phi(t)-1} \frac{1}{(1-\eta)^{-\frac{1}{K}\sum_{k=1}^{K} N^k_{h\tau,(h+1)\tau}(s,a)}}
$$

due to Jensen's inequality.

Finally, with basic algebraic calculations, (66e) is derived as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{u\in\mathcal{U}_{0,t}^{k}(s,a)}(\widetilde{\omega}_{u,t}^{k}(s,a))^2 = \sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1}\sum_{u\in\mathcal{U}^{k}_{h\tau,(h+1)\tau}(s,a)}(\widetilde{\omega}_{u,t}^{k}(s,a))^2\\
&= \sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1}(\alpha^{k}_{(h+1)\tau}(s,a))^2\left(\prod_{l=h+1}^{\phi(t)-1}\widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right)^2\sum_{u\in\mathcal{U}^{k}_{h\tau,(h+1)\tau}(s,a)}\left(\eta(1-\eta)^{N^{k}_{u+1,(h+1)\tau}(s,a)}\right)^2\\
&\overset{\text{(i)}}{\le} 2\sum_{k=1}^{K}\sum_{h=0}^{\phi(t)-1}(\alpha^{k}_{(h+1)\tau}(s,a))^2\left(\prod_{l=h+1}^{\phi(t)-1}\widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right)^2\eta\left(1-(1-\eta)^{N^{k}_{h\tau,(h+1)\tau}(s,a)}\right)\\
&\overset{\text{(ii)}}{\le} \frac{6\eta}{K}\sum_{h=0}^{\phi(t)-1}\left(\prod_{l=h+1}^{\phi(t)-1}\widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right)^2\sum_{k=1}^{K}\alpha^{k}_{(h+1)\tau}(s,a)\left(1-(1-\eta)^{N^{k}_{h\tau,(h+1)\tau}(s,a)}\right)\\
&\overset{\text{(iii)}}{\le} \frac{6\eta}{K},
\end{aligned}
$$

where (i) holds because

$$
\begin{aligned}
\sum_{u\in\mathcal{U}^{k}_{h\tau,(h+1)\tau}(s,a)}(\eta(1-\eta)^{N^{k}_{u+1,(h+1)\tau}(s,a)})^2 &= \eta^2\frac{1-(1-\eta)^{2(N^{k}_{h\tau,(h+1)\tau}(s,a))}}{1-(1-\eta)^2}\\
&\le \eta(1-(1-\eta)^{2(N^{k}_{h\tau,(h+1)\tau}(s,a))})\\
&\le 2\eta(1-(1-\eta)^{(N^{k}_{h\tau,(h+1)\tau}(s,a))})
\end{aligned}
\tag{176}
$$

given that $2x-x^2\ge x$ for $x\le 1$ and $(1-x^2)\le 2(1-x)$, (ii) follows from (66a), and (iii) follows from the same reasoning of (175).

C.6 Proof of Lemma 7

Without loss of generality, we prove the claim for some fixed $1\le t\le T$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$. For notation simplicity, let

$$
\widetilde{y}_{u,t}^{k}(s,a) = \begin{cases}\widetilde{\omega}_{u,t}^{k}(s,a)(P(s,a)-P_{u+1}^{k}(s,a))V_u^k & \text{if } (s_u^k,a_u^k)=(s,a)\\ 0 & \text{otherwise}\end{cases}, \tag{177}
$$

where

$$
\widetilde{\omega}_{u,t}^{k}(s,a) = \frac{\eta(1-\eta)^{-N^{k}_{\phi(u)\tau,u+1}(s,a)}}{K}\prod_{h=\phi(u)}^{\phi(t)-1}\frac{K}{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}}, \tag{178}
$$

then $E_t^2(s,a)=\gamma\sum_{k=1}^{K}\sum_{u=0}^{t-1}\widetilde{y}_{u,t}^{k}(s,a)$. However, due to the dependency between $P_{u+1}^{k}(s,a)$ and $\widetilde{\omega}_{u,t}^{k}(s,a)$ arising from the Markovian sampling, it is difficult to track the sum of $\widetilde{y}:=\{\widetilde{y}_{u,t}^{k}(s,a)\}$ directly. To address this issue, we will first analyze the sum using a collection of approximate random variables $\widehat{y}=\{\widehat{y}_{u,t}^{k}(s,a)\}$ drawn from a carefully constructed set $\widehat{\mathcal{Y}}$, which is closely coupled with the target $\{\widetilde{y}_{u,t}^{k}(s,a)\}_{0\le u<t}$, i.e.,

$$
D(\widetilde{y},\widehat{y}) := \left|\sum_{k=1}^{K}\sum_{u=0}^{t-1}(\widetilde{y}_{u,t}^{k}(s,a)-\widehat{y}_{u,t}^{k}(s,a))\right| \tag{179}
$$

is sufficiently small. In addition, $\widehat{y}$ shall exhibit some useful statistical independence and thus easier to control its sum; we shall control this over the entire set $\widehat{\mathcal{Y}}$. Finally, leveraging the proximity above, we can obtain the desired bound on $\widetilde{y}$ via triangle inequality. We now provide details on executing this proof outline, where the crust is in designing the set $\widehat{\mathcal{Y}}$ with a controlled size.

Before describing our construction, let's introduce the following useful event:

$$\mathcal{B}_M := \bigcap_{u=0}^{t-M\tau} \left\{ \frac{1}{4}\mu_{\mathsf{avg}}(s,a)KM\tau \le \sum_{k=1}^{K} N^k_{u,u+M\tau}(s,a) \le 2\mu_{\mathsf{avg}}(s,a)KM\tau \right\}, \tag{180}$$

where $M = M(s,a) := \lfloor \frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau} \rfloor$. Note that $M \ge \frac{1}{16\eta\mu_{\mathsf{avg}}(s,a)\tau}$ since $\eta\tau \le 1/16$. Combining this with the assumption $\eta \le \frac{1}{16t_{\mathsf{th}}(s,a)\mu_{\mathsf{avg}}(s,a)}$ (see (77) for the definition of $t_{\mathsf{th}}(s,a)$), it follows that $M\tau \ge t_{\mathsf{th}}(s,a)$ always holds. Then, $\mathcal{B}_M$ holds with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$ according to Lemma 10. The rest of the proof shall be carried out under the event $\mathcal{B}_M$.

Step 1: constructing $\widehat{\mathcal{Y}}$. To decouple dependency between $P^k_{u+1}(s,a)$ and $\widetilde{\omega}^k_{u,t}(s,a)$, we will introduce approximates of $\widetilde{\omega}^k_{u,t}(s,a)$ that only depend on history until u by replacing a factor dependent on future with some constant. To gain insight, we factorize $\widetilde{\omega}^k_{u,t}(s,a)$ into two components as follows:

$$\begin{aligned}
\widetilde{\omega}^k_{u,t}(s,a) &= \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{K}{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}} \frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}}{K} \right) \\
&\quad \times \frac{\eta(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)}}{K} \prod_{h=\phi(u)}^{\phi(t)-1} \frac{K}{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}} \\
&= \underbrace{\left(\prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}}{K} \right) \frac{\eta(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)}}{K} \right)}_{\text{dependent on history until } u} \\
&\quad \times \underbrace{\left(\prod_{h=h_0(u,t)}^{\phi(t)-1} \frac{K}{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}} \right)}_{\text{dependent on history and future until } t} \\
&= \underbrace{\left(\prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}}{K} \right) \frac{\eta(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)}}{K} \right)}_{:=x^k_u(s,a)} \\
&\quad \times \prod_{l=1}^{l(u,t)} \underbrace{\left(\prod_{h=\max\{0,\phi(t)-lM\}}^{\phi(t)-(l-1)M-1} \frac{K}{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}} \right)}_{:=z_l(s,a)}.
\end{aligned} \tag{181}$$

where we denote $l(u,t) := \lceil \frac{(t-u)}{M\tau} \rceil$ and $h_0(u,t) = \max\{0, \phi(t) - l(u,t)M\}$.

Motivated by the above decomposition, we will construct $\widehat{\mathcal{Y}}$ by approximating the future-dependent parameter $z_l(s,a)$ for $1 \le l \le L$, where $L := \min\{\lceil \frac{t}{M\tau} \rceil, \lceil 64 \log(K/\eta) \rceil\}$. Using the fact that $1 + x \le \exp(x) \le 1 + 2x$ holds for any $0 \le x < 1$, and $\eta\frac{\sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a)}{K} \le \eta\tau \le 1$, and applying Jensen's inequality,

$$\begin{aligned}
\exp\left(-\eta \frac{\sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a)}{K} \right) \ge \frac{K}{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}} &\ge \frac{K}{\sum_{k'=1}^{K} e^{\eta N^{k'}_{h\tau,(h+1)\tau}(s,a)}} \\
&\ge \frac{1}{1 + 2\eta \sum_{k'=1}^{K} \frac{\sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a)}{K}}
\end{aligned}$$

$$\geq \exp\left(-2\eta\frac{\sum_{k'=1}^{K} N_{h\tau,(h+1)\tau}^{k'}(s,a)}{K}\right).$$

Therefore, for $1 \leq l < L$, under $\mathcal{B}_M$, the range of $z_l(s,a)$ is bounded as follows:

$$z_l(s,a) \in \left[\exp(-4\eta\mu_{\mathsf{avg}}(s,a)M\tau),\;\; \exp(-\frac{1}{4}\eta\mu_{\mathsf{avg}}(s,a)M\tau)\right].$$

Using this property, we construct a set of values that can cover possible realizations of $z_l(s,a)$ in a fine-grained manner as follows:

$$\mathcal{Z} := \left\{\exp\left(-\frac{1}{4}\eta\mu_{\mathsf{avg}}(s,a)M\tau - \frac{i\eta}{K}\right) \;\middle|\, i \in \mathbb{Z}:\;\; 0 \leq i < 4K\mu_{\mathsf{avg}}(s,a)M\tau\right\}. \tag{182}$$

Note that the distance of adjacent elements of $\mathcal{Z}$ is bounded by $\eta/Ke^{-1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau}$, and the size of the set is bounded by $4K\mu_{\mathsf{avg}}(s,a)M\tau$. For $l = L$, because the number of iterations involved in $z_L(s,a)$ can be less than $M\tau$, it follows that $z_L(s,a) \in [\exp(-4\eta\mu_{\mathsf{avg}}(s,a)M\tau), 1]$. Hence, we construct the set

$$\mathcal{Z}_0 := \left\{\exp\left(-\frac{i\eta}{K}\right) \;\middle|\, i \in \mathbb{Z}:\;\; 0 \leq i < 4K\mu_{\mathsf{avg}}(s,a)M\tau\right\}. \tag{183}$$

In sum, we can always find $(\widehat{z}_1, \cdots, \widehat{z}_l, \cdots, \widehat{z}_L) \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0$ where its entry-wise distance to $(z_l(s,a))_{l\in[L-1]}$ (resp. $z_L(s,a)$) is at most $\eta/Ke^{-1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau}$ (resp. η/K).

Moreover, we approximate $x_u^k(s,a)$ by clipping it when the accumulated number of visits of all agents is not too large as follows:

$$\widehat{x}_u^k(s,a) = \begin{cases} x_u^k(s,a) & \text{if } \sum_{k=1}^{K} N_{h_0(u,t)\tau,\phi(u)\tau}^{k}(s,a) \leq 2K\mu_{\mathsf{avg}}(s,a)M\tau \\ 0 & \text{otherwise} \end{cases}. \tag{184}$$

Note that the clipping never occurs and $\widehat{x}_u^k(s,a) = x_u^k(s,a)$ for all u as long as $\mathcal{B}_M$ holds. To provide useful properties of $\widehat{x}_u^k(s,a)$ that will be useful later, we record the following lemma whose proof is provided in Appendix C.6.1.

Lemma 16. *For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, consider any integers $1 \leq t \leq T$ and $1 \leq l \leq \lceil\frac{t}{M\tau}\rceil$, where $M = \lfloor\frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau}\rfloor$. Suppose that $4\eta\tau \leq 1$, then $\widehat{x}_u^k(s,a)$ defined in (184) satisfy*

$$\forall u \in [h_0, \phi(t) - (l-1)M) \;\;:\;\; \widehat{x}_u^k(s,a) \leq \frac{9\eta}{K}, \tag{185a}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^{k}(s,a)} \sum_{k=1}^{K} \widehat{x}_u^k(s,a) \leq 16\eta\mu_{\mathsf{avg}}(s,a)M\tau, \tag{185b}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^{k}(s,a)} \sum_{k=1}^{K} (\widehat{x}_u^k(s,a))^2 \leq \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K}, \tag{185c}$$

where $h_0 = \max\{0, \phi(t) - lM\}$.

Finally, for each $\boldsymbol{z} = (\widehat{z}_1, \cdots, \widehat{z}_L) \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0$, setting $\widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z}) = \widehat{x}_u^k(s,a)\prod_{l=1}^{l(u,t)} \widehat{z}_l$, an approximate random sequence $\widehat{y}_{\boldsymbol{z}} = \{\widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\}_{0\leq u<t}$ can be constructed as follows:

$$\widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) = \begin{cases} \widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z})(P(s,a) - P_{u+1}^k(s,a))V_u^k & \text{if } (s_u^k, a_u^k) = (s,a) \text{ and } l(u,t) \leq L \\ 0 & \text{otherwise} \end{cases}. \tag{186}$$

If $t > LM\tau$, for any $u < t - LM\tau$, i.e., $l(u,t) > L$, we set $\widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) = 0$ since the magnitude of $\widetilde{\omega}_{u,t}^k(s,a)$ becomes negligible when the time difference between u and t is large enough, and the fine-grained approximation using $\mathcal{Z}$ is no longer needed, as shall be seen momentarily. Finally, denote a collection of the approximates induced by $\mathcal{Z}^{L-1} \times \mathcal{Z}_0$ as

$$\widehat{\mathcal{Y}} = \{\widehat{y}_{\boldsymbol{z}} : \quad \boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0\}.$$

Step 2: bounding the approximation error $D(\widetilde{y}, \widehat{y}_{\boldsymbol{z}})$. We now show that under $\mathcal{B}_M$, there always exists $\widehat{y}_{\boldsymbol{z}} := \widehat{y}_{\boldsymbol{z}(\widetilde{y})} \in \widehat{\mathcal{Y}}$ such that

$$D(\widetilde{y}, \widehat{y}_{\boldsymbol{z}}) < \frac{129}{1-\gamma}\sqrt{\frac{L\eta}{K}}. \tag{187}$$

To this end, we first decompose the approximation error as follows:

$$\begin{aligned}
&\min_{\widehat{y}_{\boldsymbol{z}} \in \widehat{\mathcal{Y}}} D(\widetilde{y}, \widehat{y}_{\boldsymbol{z}}) \\
&= \min_{\boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0} \left| \sum_{k=1}^{K} \sum_{u=0}^{t-1} \left(\widetilde{y}_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) \right) \right| \\
&\le \underbrace{\max_{\boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0} \left| \sum_{k=1}^{K} \sum_{u=0}^{t-LM\tau-1} \widetilde{y}_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) \right|}_{=:D_1} + \underbrace{\min_{\boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0} \left| \sum_{k=1}^{K} \sum_{u=t-LM\tau}^{t-1} \widetilde{y}_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) \right|}_{=:D_2}
\end{aligned}$$

- **Bounding D_1.** This term appears only when $t > LM\tau$. Since $\widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) = 0$ for all $u < t - LM\tau$ regardless of $\boldsymbol{z}$ by construction,

$$\begin{aligned}
\left| \sum_{k=1}^{K} \sum_{u=0}^{t-LM\tau-1} \widetilde{y}_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) \right| &\le \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{0,t-LM\tau}^k(s,a)} \widetilde{\omega}_{u,t}^k(s,a) \| P(s,a) - P_{u+1}^k(s,a) \|_1 \| V_u^k \|_\infty \\
&\overset{\text{(i)}}{\le} \frac{2}{1-\gamma} \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{0,t-LM\tau}^k(s,a)} \widetilde{\omega}_{u,t}^k(s,a) \\
&\overset{\text{(ii)}}{\le} \frac{2}{1-\gamma} (1-\eta)^{\frac{1}{K}\sum_{k=1}^{K} N_{t-LM\tau,t}^k(s,a)} \\
&\overset{\text{(iii)}}{\le} \frac{2}{1-\gamma} e^{-\eta \frac{1}{4} \mu_{\mathsf{avg}}(s,a) LM\tau} \\
&\overset{\text{(iv)}}{\le} \frac{2\eta}{(1-\gamma)K},
\end{aligned}$$

where (i) holds since $\|P(s,a)\|_1,\ \|P_u^k(s,a)\|_1 \le 1$ and $\|V_{u-1}^k\|_\infty \le \frac{1}{1-\gamma}$ (cf. (30)), (ii) follows from (66d) in Lemma 6, (iii) holds due to $\mathcal{B}_M$, and (iv) holds because $L \ge 64 \log \frac{K}{\eta} \ge \frac{4}{\eta \mu_{\mathsf{avg}}(s,a) M\tau} \log \frac{K}{\eta}$ given that $\eta \mu_{\mathsf{avg}}(s,a) M\tau \ge 1/16$.

- **Bounding D_2.** Since $\widehat{x}_u^k(s,a) = x_u^k(s,a)$ when $\mathcal{B}_M$ holds, in view of (186), we have

$$\begin{aligned}
&\min_{\boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0} \left| \sum_{k=1}^{K} \sum_{u=t-LM\tau}^{t-1} \widetilde{y}_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z}) \right| \\
&\le \min_{\boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0} \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_{t-LM\tau,t}^k(s,a)} \left| \widetilde{\omega}_{u,t}^k(s,a) - \widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z}) \right| \| P(s,a) - P_{u+1}^k(s,a) \|_1 \| V_u^k \|_\infty \\
&\le \frac{2}{1-\gamma} \min_{\boldsymbol{z} \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0} \left(\sum_{l=1}^{L} \sum_{h=\phi(t)-lM}^{\phi(t)-(l-1)M-1} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} \sum_{k=1}^{K} \widehat{x}_u^k(s,a) \left| \prod_{l'=1}^{l} z_{l'}(s,a) - \prod_{l'=1}^{l} \widehat{z}_{l'} \right| \right),
\end{aligned}$$

where the last inequality holds since $\|P(s,a)\|_1,\ \|P_u^k(s,a)\|_1 \le 1$ and $\|V_{u-1}^k\|_\infty \le \frac{1}{1-\gamma}$ (cf. (30)).

Note that for any given $\{z_l(s,a)\}_{l \in [L]}$, under $\mathcal{B}_M$, there exists $\widehat{\boldsymbol{z}}^\star = (\widehat{z}_1^\star, \dots, \widehat{z}_l^\star, \dots, \widehat{z}_L^\star) \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0$ such that $|\widehat{z}_l^\star - z_l(s,a)| \le \frac{\eta}{K} \exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau)$ for $l < L$ and $|\widehat{z}_L^\star - z_L(s,a)| \le \frac{\eta}{K}$. Also, recall that $z_l(s,a),\ \widehat{z}_l^\star \leq \exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau)$ for $l < L$ and $z_L(s,a),\ \widehat{z}_L^\star \leq 1$. Then, for any $l \leq L$ it follows that:

$$\left|\prod_{l'=1}^{l} z_{l'}(s,a) - \prod_{l'=1}^{l} \widehat{z}_{l'}^\star\right| \leq \Big(\Big|\prod_{l'=1}^{l} z_{l'}(s,a) - \widehat{z}_1^\star \prod_{l'=2}^{l} z_{l'}(s,a)\Big| + \cdots + \Big|z_l \prod_{l'=1}^{l-1} \widehat{z}_{l'}^\star - \prod_{l'=1}^{l} \widehat{z}_{l'}^\star\Big|\Big)$$
$$\leq \exp\Big(-\frac{1}{4}(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\Big)\sum_{l'=1}^{l}\frac{\eta}{K}$$
$$\leq \exp\Big(-\frac{1}{4}(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\Big)\frac{L\eta}{K}.$$

Then, applying the above bound and (185b) in Lemma 16,

$$\min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{k=1}^{K}\sum_{u=t-LM\tau}^{t-1} \widetilde{y}_{u,t}^k(s,a) - \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right|$$
$$\leq \frac{2}{1-\gamma}\sum_{l=1}^{L}\sum_{h=\phi(t)-lM}^{\phi(t)-(l-1)M-1}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\sum_{k=1}^{K}\widehat{x}_u^k(s,a)\left|\prod_{l'=1}^{l} z_{l'}(s,a) - \prod_{l'=1}^{l}\widehat{z}_{l'}^\star\right|$$
$$\leq \frac{2}{1-\gamma}\frac{L\eta}{K}\sum_{l=1}^{L}\exp\Big(-\frac{1}{4}(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\Big)\sum_{h=\phi(t)-lM}^{\phi(t)-(l-1)M-1}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\sum_{k=1}^{K}\widehat{x}_u^k(s,a)$$
$$\leq \frac{2}{1-\gamma}\frac{L\eta}{K}\frac{1}{1-\exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau)}(16\eta\mu_{\mathsf{avg}}(s,a)M\tau)$$
$$\overset{\text{(i)}}{\leq} \frac{2}{1-\gamma}\frac{L\eta}{K}\frac{8}{\eta\mu_{\mathsf{avg}}(s,a)M\tau}16\eta\mu_{\mathsf{avg}}(s,a)M\tau \leq \frac{256L\eta}{(1-\gamma)K},$$

where (i) holds since $1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau \leq 1$ and $e^{-x} \leq 1 - \frac{1}{2}x$ for any $0 \leq x \leq 1$.

By combining the bounds obtained above and using the fact that $\frac{4\eta L}{K} \leq 1$ and $L \leq 64\log(TK)$, we can conclude that

$$\min_{\widehat{y}_{\boldsymbol{z}}\in\widehat{\mathcal{Y}}} D(\widetilde{y},\widehat{y}_{\boldsymbol{z}}) \leq \frac{2\eta}{(1-\gamma)K} + \frac{256L\eta}{(1-\gamma)K} \leq \frac{129}{1-\gamma}\sqrt{\frac{L\eta}{K}}.$$

Step 3: concentration bound over $\mathcal{Y}$. We now show that for all elements in $\widehat{\mathcal{Y}} = \{\widehat{y}_{\boldsymbol{z}} :\ \boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0\}$ satisfy

$$\left|\sum_{k=1}^{K}\sum_{u=0}^{t-1}\widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right| < \frac{624}{(1-\gamma)}\sqrt{\frac{\eta}{K}\log(TK)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} \tag{188}$$

with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$. It suffices to establish (188) for a fixed $\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0$ with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\mathcal{Y}|}$, where

$$|\widehat{\mathcal{Y}}| = |\mathcal{Z}^{L-1}\times\mathcal{Z}_0| \leq (4K\mu_{\mathsf{avg}}(s,a)M\tau)^L \leq (K/\eta)^L \leq (TK)^L. \tag{189}$$

For any fixed $\boldsymbol{z} = (\widehat{z}_1,\cdots,\widehat{z}_L)\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0$, since $\widehat{\omega}_{u,t}^k(s,a;\boldsymbol{z}) = \widehat{x}_u^k(s,a)\prod_{l=1}^{l(u,t)}\widehat{z}_l$ only depends on the events happened until u, which is independent to a transition at $u+1$. Thus, we can apply Freedman's inequality to bound the sum of $\widehat{y}_{u,t}^k(s,a;\boldsymbol{z})$ since

$$\mathbb{E}[\widehat{y}_{u,t}^k(s,a;\boldsymbol{z})|\mathcal{Y}_u] = 0, \tag{190}$$

where $\mathcal{Y}_u$ denotes the history of visited state-action pairs and updated values of all agents until u, i.e., $\mathcal{Y}_u = \{(s_v^k, a_v^k), V_v^k\}_{k\in[K],v\leq u}$. Before applying Freedman's inequality, we need to calculate the following quantities. First,

$$B_t(s,a) := \max_{k\in[K],0\leq u<t} |\widehat{y}_{u,t}^k(s,a;\boldsymbol{z})| \leq \widehat{x}_u^k(s,a) \prod_{l=1}^{l(u,t)} \widehat{z}_l \|P(s,a) - P_{u+1}^k(s,a)\|_1 \|V_u^k\|_\infty \leq \frac{18\eta}{(1-\gamma)K}, \tag{191}$$

where the last inequality follows from $\|P(s,a)\|_1$, $\|P_u^k(s,a)\|_1 \leq 1$, $\|V_{u-1}^k\|_\infty \leq \frac{1}{1-\gamma}$ (cf. (30)), $\hat{z}_l \leq 1$, and (185a) in Lemma 16. Next, we can bound the variance as

$$\begin{aligned}
W_t(s,a) &:= \sum_{u=t-LM\tau}^{t-1} \sum_{k=1}^{K} \mathbb{E}[(\widehat{y}_{u,t}^k(s,a;\boldsymbol{z}))^2 | \mathcal{Y}_u] \\
&= \sum_{l=1}^{L} \sum_{h=\max\{0,\phi(t)-lM\}}^{\phi(t)-(l-1)M-1} \sum_{k=1}^{K} \sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} (\widehat{x}_u^k(s,a) \prod_{l'=1}^{l} \widehat{z}_{l'})^2 \mathsf{Var}_{P(s,a)}(V_u^k) \\
&\overset{\text{(i)}}{\leq} \frac{2}{(1-\gamma)^2} \sum_{l=1}^{L} \left(\prod_{l'=1}^{l} \widehat{z}_{l'}^2\right) \sum_{h=max\{0,\phi(t)-lM\}}^{\phi(t)-(l-1)M-1} \sum_{k=1}^{K} \sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^k(s,a)} (\widehat{x}_u^k(s,a))^2 \\
&\overset{\text{(ii)}}{\leq} \frac{2}{(1-\gamma)^2} \sum_{l=1}^{L} \left(\prod_{l'=1}^{l} \widehat{z}_{l'}^2\right) \frac{64\eta^2 \mu_{\mathsf{avg}}(s,a) M\tau}{K} \\
&\overset{\text{(iii)}}{\leq} \frac{128\eta^2 \mu_{\mathsf{avg}}(s,a) M\tau}{K(1-\gamma)^2} \sum_{l=1}^{L} \exp\left(-1/2(l-1)\eta\mu_{\mathsf{avg}}(s,a)M\tau\right) \\
&\leq \frac{128\eta^2 \mu_{\mathsf{avg}}(s,a) M\tau}{K(1-\gamma)^2} \frac{1}{1-\exp(-1/2\eta\mu_{\mathsf{avg}}(s,a)M\tau)} \\
&\overset{\text{(iv)}}{\leq} \frac{128\eta^2 \mu_{\mathsf{avg}}(s,a) M\tau}{K(1-\gamma)^2} \frac{4}{\eta\mu_{\mathsf{avg}}(s,a)M\tau} = \frac{512\eta}{K(1-\gamma)^2} := \sigma^2,
\end{aligned} \tag{192}$$

where (i) holds due to the fact that $\|\mathsf{Var}_P(V)\|_\infty \leq \|P\|_1(\|V\|_\infty)^2 + (\|P\|_1\|V\|_\infty)^2 \leq \frac{2}{(1-\gamma)^2}$ because $\|V\|_\infty \leq \frac{1}{1-\gamma}$ (cf. (30)) and $\|P\|_1 \leq 1$, (ii) follows from (185c) in Lemma 16, (iii) holds due to the range of $\mathcal{Z}$ and $\mathcal{Z}_0$ is bounded by $\exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)M\tau)$ and 1, respectively, and (iv) holds since $e^{-x} \leq 1 - \frac{1}{2}x$ for any $0 \leq x \leq 1$ and $1/2\eta\mu_{\mathsf{avg}}(s,a)M\tau \leq 1$.

Now, by substituting the above bounds of W_t and B_t into Freedman's inequality (see Theorem 4) and setting $m = 1$, it follows that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $t \in [T]$ and $\widehat{y}_{\boldsymbol{z}} \in \widehat{\mathcal{Y}}$,

$$\begin{aligned}
\left|\sum_{k=1}^{K}\sum_{u=0}^{t-1} \widehat{y}_{u,t}^k(s,a;\boldsymbol{z})\right| &\leq \sqrt{8\max\{W_t(s,a), \frac{\sigma^2}{2^m}\}\log\frac{4m|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}} + \frac{4}{3}B_t(s,a)\log\frac{4m|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta} \\
&\leq \sqrt{4096\frac{\eta}{K(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}} + \frac{24\eta}{K(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta} \\
&\overset{\text{(i)}}{\leq} \frac{78}{(1-\gamma)}\sqrt{\frac{\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}},
\end{aligned} \tag{193}$$

with at least probability $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}$, where (i) holds because $|\widehat{\mathcal{Y}}| \leq (TK)^L$ given that $\eta\mu_{\mathsf{avg}}(s,a)M\tau \leq 1/4$, and $\frac{4\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta} \leq 1$. Therefore, it follows that (188) holds.

Step 4: putting things together. We now putting all the results obtained in the previous steps together to achieve the claimed bound. Under $\mathcal{B}_M$, there always exists $\widehat{y}_{\boldsymbol{z}} := \widehat{y}_{\boldsymbol{z}(\widehat{y})} \in \widehat{\mathcal{Y}}$ such that (187) holds. Hence, setting $q = \frac{2064}{(1-\gamma)}\sqrt{\frac{\eta}{K}\log{(TK)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}$,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{u=0}^{t-1}\widetilde{y}^k_{u,t}(s,a) &\leq \left|\sum_{k=1}^{K}\sum_{u=0}^{t-1}\widehat{y}^k_{u,t}(s,a;\boldsymbol{z})\right| + D(\widetilde{y},\widehat{y}_{\boldsymbol{z}}) \\ &\leq \frac{78}{(1-\gamma)}\sqrt{\frac{\eta L}{K}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}} + \frac{129}{1-\gamma}\sqrt{\frac{L\eta}{K}} \\ &\leq \frac{2064}{(1-\gamma)}\sqrt{\frac{\eta}{K}\log{(TK)}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2K}{\delta}}, \qquad (194)\end{aligned}$$

where the second line holds due to (188) and (187), and the last line holds due to $L \leq 64\log{(TK)}$. By taking a union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $t \in [T]$, we complete the proof.

C.6.1 Proof of Lemma 16

For notational simplicity, let $\overline{h}$ be the largest integer among $h \in (h_0, \phi(t)-(l-1)M)$ such that

$$\sum_{k=1}^{K} N^k_{h_0\tau,(h-1)\tau}(s,a) \leq 2K\mu_{\mathsf{avg}}(s,a)M\tau. \qquad (195)$$

Then, the following holds:

$$\begin{aligned}\sum_{k=1}^{K} N^k_{h_0\tau,\overline{h}\tau}(s,a) &= \sum_{k=1}^{K} N^k_{(\overline{h}-1)\tau,\overline{h}\tau}(s,a) + \sum_{k=1}^{K} N^k_{h_0\tau,(\overline{h}-1)\tau}(s,a) \\ &\leq K\tau + 2K\mu_{\mathsf{avg}}(s,a)M\tau. \qquad (196)\end{aligned}$$

Also, for the following proofs, we provide an useful bound as follows:

$$\begin{aligned}\sum_{k'=1}^{K}\frac{(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}}{K} \leq \frac{\sum_{k'=1}^{K} e^{\eta N^{k'}_{h\tau,(h+1)\tau}(s,a)}}{K} &\leq 1 + 2\eta\frac{\sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a)}{K} \\ &\leq \exp\left(2\eta\frac{\sum_{k'=1}^{K} N^{k'}_{h\tau,(h+1)\tau}(s,a)}{K}\right), \qquad (197)\end{aligned}$$

which holds since $1+x \leq e^x \leq 1+2x$ for any $x \in [0,1]$ and $\eta N^{k'}_{h\tau,(h+1)\tau}(s,a) \leq \eta\tau \leq 1$.

According to (184), for any integer $u \in [\overline{h}\tau, t-(l-1)M\tau)$, $\widehat{x}^k_u(s,a)$ is clipped to zero. Now, we prove the bounds in Lemma 16 respectively.

Proof of (185a). For $u \in [h_0\tau, \overline{h}\tau)$,

$$\begin{aligned}\widehat{x}^k_u(s,a) &= \prod_{h=h_0}^{\phi(u)-1}\left(\frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}}{K}\right)\frac{\eta(1-\eta)^{-N^k_{\phi(u)\tau,u+1}(s,a)}}{K} \\ &\overset{\text{(i)}}{\leq} \prod_{h=h_0}^{\phi(u)-1}\left(\frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h\tau,(h+1)\tau}(s,a)}}{K}\right)\frac{3\eta}{K} \\ &\overset{\text{(ii)}}{\leq} \exp\left(\frac{2\eta}{K}\sum_{k'=1}^{K} N^{k'}_{h_0\tau,(\overline{h}-1)\tau}(s,a)\right)\frac{3\eta}{K} \\ &\overset{\text{(iii)}}{\leq} \exp(4\eta\mu_{\mathsf{avg}}(s,a)M\tau)\frac{3\eta}{K} \overset{\text{(iv)}}{\leq} \frac{9\eta}{K}, \qquad (198)\end{aligned}$$

where (i) holds since $(1+\eta)^x \leq e^{\eta x}$ and $\eta N^k_{\phi(u)\tau,u+1}(s,a) \leq \eta\tau \leq 1$, (ii) holds due to (197) and the fact that $\phi(u) \leq \overline{h}-1$, (iii) follows from the definition of $\overline{h}$ in (195), and (iv) holds because $4\eta\mu_{\mathsf{avg}}(s,a)M\tau \leq 1$.

Proof of (185b)**.** By the definition of $\overline{h}$, it follows that

$$\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} \widehat{x}^k_u(s,a) = \sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} x^k_u(s,a).$$

Using the following relation for each h:

$$\begin{aligned}
&\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} x^k_u(s,a)\\
&= \left(\prod_{h'=h_0}^{h-1} \frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}}{K}\right) \sum_{k=1}^{K} \frac{\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \eta(1-\eta)^{-N^k_{h\tau,u+1}(s,a)}}{K}\\
&= \left(\prod_{h'=h_0}^{h-1} \frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}}{K}\right) \sum_{k=1}^{K} \frac{(1-\eta)^{-N^k_{h\tau,(h+1)\tau}(s,a)}-1}{K}\\
&= \left(\prod_{h'=h_0}^{h} \frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}}{K}\right) - \left(\prod_{h'=h_0}^{h-1} \frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}}{K}\right),
\end{aligned}$$

and applying (197), we can complete the proof as follows:

$$\begin{aligned}
\sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} x^k_u(s,a) &\le \prod_{h'=h_0}^{\overline{h}-1} \exp\left(\frac{2\eta\sum_{k'=1}^{K} N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}{K}\right) - 1\\
&\le \exp\left(\frac{2\eta\sum_{k'=1}^{K} N^{k'}_{h_0\tau,\overline{h}\tau}(s,a)}{K}\right) - 1\\
&\overset{\text{(i)}}{\le} \exp\left(4\eta\mu_{\mathsf{avg}}(s,a)M\tau + 2\eta\tau\right) - 1\\
&\overset{\text{(ii)}}{\le} 16\eta\mu_{\mathsf{avg}}(s,a)M\tau,
\end{aligned}$$

where (i) follows from (196), and (ii) holds because $e^x \le 1+2x$ for any $x\in[0,1]$ and $2\eta\tau \le 4\eta\mu_{\mathsf{avg}}(s,a)M\tau \le 1/2$.

Proof of (185c)**.** Similarly,

$$\sum_{h=h_0}^{\phi(t)-(l-1)M-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} (\widehat{x}^k_u(s,a))^2 = \sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} (x^k_u(s,a))^2.$$

Using the following relation for each h:

$$\begin{aligned}
&\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \sum_{k=1}^{K} (x^k_u(s,a))^2\\
&= \left(\prod_{h'=h_0}^{h-1} \frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}}{K}\right)^2 \sum_{k=1}^{K} \frac{\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \eta^2(1-\eta)^{-2N^k_{h\tau,u+1}(s,a)}}{K^2}\\
&\le \left(\prod_{h'=h_0}^{h-1} \frac{\sum_{k'=1}^{K}(1-\eta)^{-N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}}{K}\right)^2 \sum_{k=1}^{K} \frac{\eta((1-\eta)^{-2N^k_{h\tau,(h+1)\tau}(s,a)}-1)}{K^2}
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{\eta}{K}\left(\prod_{h'=h_0}^{h-1}\exp\left(2\eta\frac{\sum_{k'=1}^{K}N^{k'}_{h'\tau,(h'+1)\tau}(s,a)}{K}\right)\right)^2\left(\exp\left(4\eta\frac{\sum_{k'=1}^{K}N^{k'}_{h\tau,(h+1)\tau}(s,a)}{K}\right)-1\right)\\
&= \frac{\eta}{K}\exp\left(4\eta\frac{\sum_{k'=1}^{K}N^{k'}_{h_0\tau,h\tau}(s,a)}{K}\right)\left(\exp\left(4\eta\frac{\sum_{k'=1}^{K}N^{k'}_{h\tau,(h+1)\tau}(s,a)}{K}\right)-1\right)\\
&= \frac{\eta}{K}\left(\exp\left(4\eta\frac{\sum_{k'=1}^{K}N^{k'}_{h_0\tau,(h+1)\tau}(s,a)}{K}\right)-\exp\left(4\eta\frac{\sum_{k'=1}^{K}N^{k'}_{h_0\tau,h\tau}(s,a)}{K}\right)\right),
\end{aligned}
\tag{199}
$$

where the inequality is derived similarly to (197) under the condition $2\eta\tau \le 1$, we can complete the proof as follows:

$$
\begin{aligned}
\sum_{h=h_0}^{\overline{h}-1}\sum_{u\in\mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)}\sum_{k=1}^{K}(x_u^k(s,a))^2 &\le \frac{\eta}{K}\left(\exp\left(4\eta\frac{\sum_{k'=1}^{K}N^{k'}_{h_0\tau,\overline{h}\tau}(s,a)}{K}\right)-1\right)\\
&\overset{\text{(i)}}{\le} \frac{\eta}{K}\left(\exp\left(8\eta\mu_{\mathsf{avg}}(s,a)M\tau+4\eta\tau\right)-1\right)\\
&\overset{\text{(ii)}}{\le} \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)M\tau}{K},
\end{aligned}
\tag{200}
$$

where (i) follows from (196), and (ii) holds because $e^x \le 1+4x$ for any $x\in[0,2]$ and $4\eta\tau \le 8\eta\mu_{\mathsf{avg}}(s,a)M\tau \le 1$.

C.7 Proof of Lemma 8

The proof follows a similar structure to that of Lemma 5. We omit common parts of the proofs and refer to Appendix C.4 to check the detailed derivations. First, we decompose the error term as follows:

$$
\begin{aligned}
E_t^3(s,a) = {}&\underbrace{\gamma\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{0,(\phi(t)-\beta)\tau}(s,a)}\widetilde{\omega}^k_{u,t}(s,a)P(s,a)(V^\star-V_u^k)}_{=:E_t^{3a}(s,a)}\\
&+\underbrace{\gamma\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{(\phi(t)-\beta)\tau,t}(s,a)}\widetilde{\omega}^k_{u,t}(s,a)P(s,a)(V^\star-V_u^k)}_{=:E_t^{3b}(s,a)}.
\end{aligned}
\tag{201}
$$

We shall bound these two terms separately.

- **Bounding $E_t^{3a}(s,a)$.** First, the bound of $E_t^{3a}(s,a)$ is derived as follows:

$$
\begin{aligned}
|E_t^{3a}(s,a)| &\le \gamma\sum_{k=1}^{K}\sum_{u\in\mathcal{U}^k_{0,(\phi(t)-\beta)\tau}(s,a)}\widetilde{\omega}^k_{u,t}(s,a)\|P(s,a)\|_1\|V^\star-V_u^k\|_\infty\\
&\overset{\text{(i)}}{\le} \frac{2}{1-\gamma}(1-\eta)^{\frac{1}{K}\sum_{k=1}^{K}N^k_{(\phi(t)-\beta)\tau,t}(s,a)}\\
&\overset{\text{(ii)}}{\le} \frac{2}{1-\gamma}(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}},
\end{aligned}
\tag{202}
$$

where (i) holds due to Lemma 6 (cf. (66d)), and (ii) follows fromapplying Lemma 10 that with probability at least $1-\delta$,

$$
\sum_{k=1}^{K}N^k_{(\phi(t)-\beta)\tau,t}(s,a) \ge \frac{K\beta\tau\mu_{\mathsf{avg}}}{4}
$$

holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $0\le u<v\le T$ as long as $\beta\tau \ge t_{\mathsf{th}}$.

- **Bounding $E_t^{3b}(s,a)$.** Combining (156) and Lemma 14 to bound $\|V^\star - V_u^k\|_\infty$, we bound $E_t^{3b}(s,a)$ as follows:

$$\begin{aligned}
|E_t^{3b}(s,a)| &\le \gamma \sum_{k=1}^{K} \sum_{u \in \mathcal{U}^k_{(\phi(t)-\beta)\tau,t}(s,a)} \widetilde{\omega}^k_{u,t}(s,a) \left\| V^\star - V_u^k \right\|_\infty \\
&\le \gamma \sum_{k=1}^{K} \sum_{h=\phi(t)-\beta}^{\phi(t)-1} \sum_{u \in \mathcal{U}^k_{h\tau,(h+1)\tau}(s,a)} \widetilde{\omega}^k_{u,t}(s,a)((1+2\eta\tau)\|\Delta_{h\tau}\|_\infty + \sigma_{\mathsf{local}}) \\
&\le \sigma_{\mathsf{local}} + \frac{1+\gamma}{2} \max_{\phi(t)-\beta \le h < \phi(t)} \|\Delta_{h\tau}\|_\infty \qquad (203)
\end{aligned}$$

where we denote $\sigma_{\mathsf{local}} := \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TK}{\delta}}$ for notational simplicity, and the last inequality follows from Lemma 6 (cf. (66c)) and the assumption that $\eta \le \frac{1-\gamma}{4\gamma\tau}$.

Now we have the bounds of $E_t^{3a}(s,a)$ and $E_t^{3b}(s,a)$ separately obtained above. By combining the bounds in (201), we can claim the advertised bound, which completes the proof.