Federated Offline Reinforcement Learning: Collaborative Single-Policy Coverage Suffices

Jiin Woo [1] **Laixi Shi** [2] **Gauri Joshi** [1] **Yuejie Chi** [1]

Abstract

Offline reinforcement learning (RL), which seeks to learn an optimal policy using offline data, has garnered significant interest due to its potential in critical applications where online data collection is infeasible or expensive. This work explores the benefit of federated learning for offline RL, aiming at collaboratively leveraging offline datasets at multiple agents. Focusing on finite-horizon episodic tabular Markov decision processes (MDPs), we design FedLCB-Q, a variant of the popular model-free Q-learning algorithm tailored for federated offline RL. FedLCB-Q updates local Q-functions at agents with novel learning rate schedules and aggregates them at a central server using importance averaging and a carefully designed pessimistic penalty term. Our sample complexity analysis reveals that, with appropriately chosen parameters and synchronization schedules, FedLCB-Q achieves linear speedup in terms of the number of agents without requiring high-quality datasets at individual agents, as long as the local datasets collectively cover the state-action space visited by the optimal policy, highlighting the power of collaboration in the federated setting. In fact, the sample complexity almost matches that of the single-agent counterpart, as if all the data are stored at a central location, up to polynomial factors of the horizon length. Furthermore, FedLCB-Q is communication-efficient, where the number of communication rounds is only linear with respect to the horizon length up to logarithmic factors.

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA [2]Department of Computing Mathematical Sciences, California Institute of Technology, CA 91125, USA. Correspondence to: Jiin Woo <jiinw@andrew.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Offline RL (Levine et al., 2020), also known as batch RL, addresses the challenge of learning a near-optimal policy using offline datasets collected a priori, without further interactions with an environment. Fueled by the cost-effectiveness of utilizing pre-collected datasets compared to real-time explorations, offline RL has received increasing attention. However, the performance of offline RL crucially depends on the quality of offline datasets due to the lack of additional interactions with the environment, where the quality is determined by how thoroughly the state-action space is explored during data collection.

Encouragingly, recent research (Rashidinejad et al., 2021; Shi et al., 2022; Xie et al., 2021b; Li et al., 2024b) indicates that being more conservative on unseen state-action pairs, known as the principle of pessimism, enables learning of a near-optimal policy even with partial coverage of the state-action space, as long as the distribution of datasets encompasses the trajectory of the optimal policy. However, acquiring high-quality datasets that have good coverage of the optimal policy poses challenges because it requires the state-action visitation distribution induced by a behavior policy employed for data collection to be very close to the optimal policy. Alternatively, multiple datasets can be merged into one dataset to supplement insufficient coverage of one other, but this may be impractical when offline datasets are scattered and cannot be easily shared due to privacy and communication constraints.

Federated offline RL. Driven by the need to harvest multiple datasets to address insufficient coverage, there is a growing interest in implementing offline RL in a federated manner without the need to share datasets (Zhou et al., 2024; Woo et al., 2023; Khodadadian et al., 2022). For model-based RL, a study has proposed a federated variant of pessimistic value iteration (Zhou et al., 2024), which requires sharing of model estimates. On the other hand, for model-free RL, while Woo et al. (2023) introduced a federated Q-learning algorithm that achieves linear speedup with collaborative coverage of agents, due to the absence of pessimism, it still carries the risk of overestimation on state-action pairs that are insufficiently covered by the agents. Indeed, it remains unknown whether the principle of pessimism can be

type	reference	number of agents	coverage	sample complexity	communication rounds
model-based	VI-LCB (Xie et al., 2021b)	1	single	$\frac{H^6SC^\star}{\varepsilon^2}$	-
	PEVI-Adv (Xie et al., 2021b)	1	single	$\frac{H^4SC^\star}{\varepsilon^2}$	-
	VI-LCB (Li et al., 2024b)	1	single	$\frac{H^4SC^\star}{\varepsilon^2}$	-
model-free	LCB-Q (Shi et al., 2022)	1	single	$\frac{H^6SC^\star}{\varepsilon^2}$	-
	LCB-Q-Adv (Shi et al., 2022)	1	single	$\frac{H^4SC^\star}{\varepsilon^2}$	-
	FedAsynQ (Woo et al., 2023)	M	collaborative	$\frac{H^6}{Md_{\mathsf{avg}}\varepsilon^2}$	$\frac{HM}{d_{\mathsf{avg}}}$
	FedLCB-Q (Theorem 3.1)	M	collaborative	$\frac{H^7SC^\star_{\mathsf{avg}}}{M\varepsilon^2}$	H

Table 1: Comparison of sample complexity upper bounds of model-based and model-free algorithms for offline RL to learn an ε-optimal policy in finite-horizon non-stationary MDPs, where logarithmic factors and burn-in costs are hidden. Here, S is the size of state space, A is the size of action space, H is the horizon length, M is the number of agents, $C^\star$ and $C^\star_{\mathsf{avg}}$ denote the single-policy concentrability and the average single-policy concentrability, respectively (cf. (8) and (9)), and d_{avg} is the minimum entry of the average stationary state-action occupancy distribution of all agents. We follow standard conversion to translate the best sample complexity in Woo et al. (2023) to the finite-horizon setting for comparison.

implemented in federated offline RL to eliminate the risk of overestimation, while fully utilizing the collaborative coverage provided by agents, and without sharing datasets or model estimates.

Our goal in this paper is to develop a federated variant of Q-learning (Watkins & Dayan, 1992) for offline RL, which allows agents to learn a near-optimal Q-function with improved sample efficiency and relaxed coverage assumption. In the single-agent case, pessimism is implemented by penalizing the value estimates by subtracting a penalty term measuring the uncertainty of the estimates (Yan et al., 2023; Shi et al., 2022). However, federated settings are communication-constrained, implying that agents only have a limited chance of synchronization and they perform multiple local updates without knowing other agents' training progress. Allowing multiple local updates leads to higher uncertainty of local Q-estimates beyond the control of the pessimism penalty, potentially impacting both sample complexity and communication efficiency. This underscores the technical challenge of incorporating pessimism while managing local updates and raises the question:

How to judiciously incorporate the principle of pessimism in federated RL without hurting its sample and communication efficiency?

1.1. Our contribution

This work presents a federated Q-learning algorithm with pessimism for offline RL, which achieves *linear speedup* and *low communication cost*, while requiring only *collaborative coverage of the optimal policy*. Formally, we consider episodic finite-horizon tabular Markov decision processes (MDPs) with S states, A actions, and horizon length H. A total number of M agents, each with K trajectories (collected using its local behavior policy), collaborate in a federated setting with the help of a central server to learn the optimal policy. Our main contributions are summarized as below; see also Table 1 for a detailed comparison.

- **Federated Q-learning for offline RL.** We propose a federated offline Q-learning algorithm named FedLCB-Q, which involves iterative local updates at agents and global aggregation at a central server with scheduled synchronizations. We introduce essential components that implement pessimism compensating for the uncertainty in both local and global Q-function updates. First, to address the uncertainty arising from independent local updates, we employ *learning rate rescaling* at local agents and *importance averaging* at server aggregation. The former restricts the drifts of local Q-estimates by rapidly decreasing the learning rates during local updates, and the latter reduces uncertainty of the aggregated Q-estimates by assigning smaller weights to rarely updated local val-

ues. Additionally, for every global aggregation, a *global penalty* calculated based on aggregated visitation counts is subtracted from the aggregated global Q-estimate. These design choices play a crucial role in achieving both sample and communication efficiency while preventing the overestimation of the Q-function.

- **Linear speedup with collaborative single-policy coverage.** Our analysis of sample complexity of FedLCB-Q (see Theorem 3.1) demonstrates that FedLCB-Q finds an ε-optimal policy, as long as the total number of samples per agent $T = KH$ exceeds

$$\widetilde{O}\left(\frac{H^7 S C^\star_{\mathsf{avg}}}{M\varepsilon^2}\right),$$

  where $C^\star_{\mathsf{avg}}$ denotes the average single-policy concentrability coefficient of all agents (see (10) for the formal definition). This shows linear speedup in terms of number agents M, which is achieved with a significantly weaker data requirement at individual agents than prior art. In truth, each agent affords to have a non-expert dataset collected by a sub-optimal behavior policy, as long as all agents collectively cover the state-action pairs visited by the optimal policy, even they don't cover the entire state-action space as in Woo et al. (2023). The bound nearly matches the sample complexity obtained for a single-agent pessimistic Q-learning algorithm (Shi et al., 2022) with a similar Hoeffding-style penalty, up to a factor of H, as if all the datasets are processed at a central location.

- **Low communication cost.** Under appropriate choices of synchronization schedules, FedLCB-Q requires approximately $\widetilde{O}(H)$ rounds of synchronizations to achieve the targeted accuracy (see Corollary 3.2), which is almost independent with the size of the state-action space and the number of agents. The analysis suggests that frequent synchronizations are not necessary, outperforming prior art (Woo et al., 2023).

1.2. Related work

Offline RL. Offline RL addresses the problem of learning improved policies from a logged static dataset. The main challenge of offline RL is how to reliably estimate the values of unseen or rarely visited state-action pairs. To tackle this challenge, most offline RL algorithms prevent agents from taking uncertain actions by regularizing the policy to be close to the behavior policy (Fujimoto et al., 2019; Siegel et al., 2020; Fujimoto & Gu, 2021) or penalizing value estimates on out-of-distribution state-action pairs (Kumar et al., 2020; Liu et al., 2020; Kostrikov et al., 2022; Wu et al., 2019), which is also known as the principle of pessimism. Recently, the pessimistic approach has been developed and theoretically studied for various RL settings, such as model-based approaches (Xie et al., 2021b; Rashidinejad et al., 2021; Kidambi et al., 2020; Yu et al., 2020; Jin et al., 2021; Li et al., 2024b; Yin & Wang, 2021; Kim & Oh, 2023; Shi & Chi, 2022), policy-based approaches (Xie et al., 2021a; Zanette et al., 2021), and model-free approaches (Shi et al., 2022; Yan et al., 2023; Uehara et al., 2023). Most of these works have focused on the single-agent case and suggested that the state-action visitation distribution induced by the behavior policy should cover that of the optimal policy (Rashidinejad et al., 2021; Shi et al., 2022; Yan et al., 2023), and the distribution mismatch among the two visitation distributions governs the hardness of offline RL (Li et al., 2024b). Another interesting work (Shi et al., 2023) considered offline RL from multiple perturbed data sources, requiring a centralized setting in which an agent has full access to all the datasets.

Federated RL. There has been an increasing interest in federated and distributed RL, driven by the need to address more realistic constraints, including privacy, communication efficiency, and data heterogeneity, as well as training speedup. Recent works have investigated federated RL from various perspectives, such as robustness to adversarial attacks (Wu et al., 2021; Fan et al., 2021), environment or task heterogeneity (Yang et al., 2023; Jin et al., 2022; Wang et al., 2023; Zhou et al., 2024), as well as sample and communication complexities under asynchronous sampling (Khodadadian et al., 2022; Woo et al., 2023) and online sampling (Zheng et al., 2024; Zhang et al., 2024). In addition, to address unreliable estimation on unseen state-action pairs in local batch datasets under the federated setting, Shen et al. (2023) proposed a federated offline policy gradient algorithm that regularizes the distribution of an estimated policy to be close to the averaged visitation distributions of agents with regularization loss, and Zhou et al. (2024) studied a federated variant of pessimistic value iteration. However, for model-free RL, although Woo et al. (2023) provided a federated Q-learning algorithm that achieves linear speedup in terms of the number of agents with relaxed coverage assumption for individual agents, it still requires agents to cover the entire state-action space uniformly due to the lack of pessimism.

Q-learning. Characterizing the finite-sample complexity of single-agent Q-learning has been examined extensively under various data collection and function approximation schemes, including but not limited the synchronous setting (Even-Dar & Mansour, 2003; Beck & Srikant, 2012; Li et al., 2024a; Wainwright, 2019), the asynchronous and offline setting (Li et al., 2021; 2024a; Qu & Wierman, 2020; Yan et al., 2023; Shi et al., 2022), the online setting (Jin et al., 2018; Bai et al., 2019; Wang et al., 2019), under function approximation (Fan et al., 2020; Chen et al., 2019; Xu & Gu, 2020), to mention just a few.

Notation. In this paper, we use $\Delta(\mathcal{S})$ to refer to the probability simplex over a set $\mathcal{S}$, and $[K] := \{1, \cdots, K\}$ 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 $f \gtrsim g$) indicates that $f(\cdot)$ is order-wise not larger than (resp. not smaller than) $g(\cdot)$ up to some logarithmic factors. The notation $f \asymp g$ signifies that both $f \lesssim g$ and $f \gtrsim g$ simultaneously hold.

2. Background and problem formulation

2.1. Background

Basics of episodic finite-horizon MDPs. Consider an episodic finite-horizon MDP represented by

$$\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H\big),$$

where $\mathcal{S}$ is the state space of size S, $\mathcal{A}$ is the action space of size A, H is the horizon length, $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denote the probability transition kernel and the reward function at the h-th time step $(1 \le h \le H)$, respectively.

A policy is denoted by $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the probability distribution over the action space at time step h in state s. With slight abuse of notation, we use $\pi_h(s)$ to denote the selected action when the policy π_h is deterministic. For $h = 1, \ldots, H$, the value function $V_h^\pi(s)$ of policy π is defined as the expected cumulative rewards starting from state s at step h by following π, i.e.,

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^H r_t\big(s_t, a_t\big) \,\Big|\, s_h = s\right], \tag{1}$$

where the expectation is taken over the randomness of the trajectory $\{s_t, a_t, r_t\}_{t=h}^H$ induced by the policy π as well as the MDP transitions according to $a_t \sim \pi_t(\cdot \,|\, s_t)$ and $s_{t+1} \sim P_t(\cdot \,|\, s_t, a_t)$. Similarly, the Q-function $Q_h^\pi(s, a)$ of a policy π at step h in state-action pair (s, a) is defined as

$$Q_h^\pi(s, a) := r_h(s, a) + \mathbb{E}\left[\sum_{t=h+1}^H r_t(s_t, a_t) \,\Big|\, s_h = s, a_h = a\right], \tag{2}$$

where the expectation is again over the randomness induced by π and the MDP transitions.

It is well-known (Puterman, 2014) that one can always find a deterministic *optimal* policy $\pi^\star = \{\pi_h^\star\}_{h=1}^H$, which maximizes the value function (resp. the Q-function) *simultaneously* over all states (resp. state-action pairs) among all policies. The resulting optimal value function $V^\star = \{V_h^\star\}_{h=1}^H$ and optimal Q-functions $Q^\star = \{Q_h^\star\}_{h=1}^H$ are denoted respectively by

$$V_h^\star(s) := V_h^{\pi^\star}(s) = \max_\pi V_h^\pi(s),$$

$$Q_h^\star(s, a) := Q_h^{\pi^\star}(s, a) = \max_\pi Q_h^\pi(s, a)$$

for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Given an initial state distribution $\rho \in \Delta(\mathcal{S})$, the expected value of a given policy π and that of the optimal policy $\pi^\star$ at the initial step are defined respectively by

$$V_1^\pi(\rho) := \mathbb{E}_{s_1 \sim \rho}\big[V_1^\pi(s_1)\big], \quad V_1^\star(\rho) := \mathbb{E}_{s_1 \sim \rho}\big[V_1^\star(s_1)\big]. \tag{3}$$

Bellman equations. Of crucial importance are the Bellman equations that connect the value functions across different time steps (Bertsekas, 2017). For any policy π, it follows that

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_{h,s,a}}\big[V_{h+1}^\pi(s')\big] \tag{4}$$

for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, where $V_{H+1}^\pi(s) = 0$ for any $s \in \mathcal{S}$. Moreover, Bellman's optimality equation says that

$$Q_h^\star(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_{h,s,a}}\big[V_{h+1}^\star(s')\big] \tag{5}$$

for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, and the optimal policy satisfies $\pi_h^\star(s) = \arg\max_{a \in \mathcal{A}} Q_h^\star(s, a)$.

2.2. Problem formulation: federated offline RL

In offline RL, one has access to a offline dataset containing episodes collected by following some behavior policy. Here, we formulate a federated version of the offline RL problem with M agents, where each agent has access to a local offline dataset. For $1 \le m \le M$, the offline dataset $\mathcal{D}^m$ at agent m is composed of K episodes,[1] each generated independently according to a behavior policy $\mu^m = \{\mu_h^m\}_{h=1}^H$, resulting in

$$\mathcal{D}^m := \Big\{\big(s_{k,1}^m,\, a_{k,1}^m,\, r_{k,1}^m,\, \ldots,\, s_{k,H}^m,\, a_{k,H}^m,\, r_{k,H}^m\big)\Big\}_{k=1}^K,$$

where the initial state $s_{k,1}^m \sim \rho$ is drawn from some initial state distribution $\rho \in \Delta(\mathcal{S})$, $s_{k,h}^m\, a_{k,h}^m,\, r_{k,h}^m$ are the state, action and reward at step h in the k-th episode, $a_{k,h}^m \sim \mu_h^m(\cdot \,|\, s_{k,h}^m)$ and $r_{k,h}^m = r_h(s_{k,h}^m\, a_{k,h}^m)$.

Goal. The goal of federated offline RL is to learn an ε-optimal policy $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ satisfying

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon$$

using the history dataset $\mathcal{D} = \big\{\mathcal{D}^m\big\}_{1 \le m \le M}$ without sharing the local offline datasets, with the help of a parameter server. Furthermore, it is greatly desirable to achieve as high

[1] For simplicity, we assume all the agents have the same number of episodes. It is straightforward to generalize to the scenario when the local offline datasets have different sizes.

accuracy as possible, in a memory- and communication-efficient manner.

Metric. Obviously, the success of offline RL highly relies on the quality of the history dataset. In order to define the metric, let us first introduce the occupancy distributions $d_h^\pi(s)$ and $d_h^\pi(s,a)$ induced by policy π at step h, given by

$$d_h^\pi(s) := \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi), \tag{6}$$

$$d_h^\pi(s,a) := \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi)\, \pi_h(a \,|\, s). \tag{7}$$

Recent works (Rashidinejad et al., 2021; Xie et al., 2021b; Shi et al., 2022) have advocated the notion of *single-policy concentrability*, which measures the mismatch between the occupancy distributions induced by the optimal policy $\pi^\star$ and the behavior policy μ, with the benefit that this assumes away the need for the offline dataset to cover the entire state-action space, which is often impractical. Li et al. (2024b) offered a more refined notion called *single-policy clipped concentrability*, defined as follows.

Definition 2.1 (single-policy clipped concentrability)**.** The single-policy clipped concentrability coefficient $C^\star \in [1/S, \infty)$ of a behavior policy μ is defined to be the smallest quantity that satisfies

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a),\, 1/S\}}{d_h^\mu(s,a)} \le C^\star, \tag{8}$$

where we adopt the convention $0/0 = 0$.

The single-policy clipped concentrability coefficient $C^\star < \infty$ is finite whenever the behavior policy covers the state-action pairs visited by the *optimal* policy, rather than having to cover the entire state-action space. Recall that since $\pi^\star$ is deterministic, $d_h^{\pi^\star}(s,a) = d_h^{\pi^\star}(s)\mathbb{I}(a = \pi_h^\star(s))$, that is, $d_h^{\pi^\star}(s,a)$ is non-zero only for the optimal action $a = \pi_h^\star(s)$. Compared with the unclipped counterpart introduced in Rashidinejad et al. (2021), the clipping of the occupancy distribution $d_h^{\pi^\star}(s,a)$ by the threshold $1/S$ ensures that $C^\star$ will not be excessively large when $d_h^{\pi^\star}(s)$ is highly concentrated in a small number of states in state space.

In the federated setting, we further introduce a tailored notion that highlights the potential benefit of collaborative learning in the presence of multiple agents. For ease of notation, denote

$$d_h^m(s) = d_h^{\mu^m}(s) \quad \text{and} \quad d_h^m(s,a) = d_h^{\mu^m}(s,a)$$

as the occupancy distributions induced by the behavior policy μ^m at agent m. Based on these, we define the average occupancy distributions as

$$d_h^{\mathsf{avg}}(s) = \frac{1}{M}\sum_{m=1}^{M} d_h^m(s), \quad d_h^{\mathsf{avg}}(s,a) = \frac{1}{M}\sum_{m=1}^{M} d_h^m(s,a). \tag{9}$$

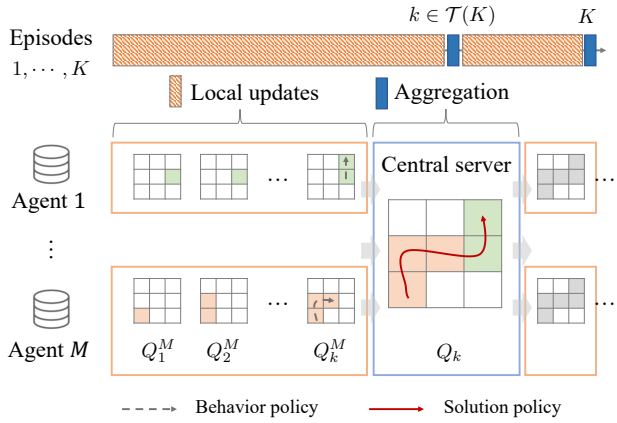


Figure 1: FedLCB-Q with M agents and a central server. Each agent m performs local updates on its local Q-table Q_k^m for each episode k in a local dataset $\mathcal{D}^m$. When synchronization is scheduled, $k \in \mathcal{T}(K)$, the agents send their local Q-tables to the server and the server aggregates the Q-tables into a global Q-table and synchronizes local Q-tables.

Definition 2.2 (average single-policy clipped concentrability)**.** The average single-policy concentrability coefficient $C_{\mathsf{avg}}^\star \in [1/S, \infty)$ of multiple behavior policies $\{\mu^m\}_{m\in[M]}$ is defined to be the smallest quantity that satisfies

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a), 1/S\}}{d_h^{\mathsf{avg}}(s,a)} \le C_{\mathsf{avg}}^\star, \tag{10}$$

where we adopt the convention $0/0 = 0$.

An important implication of the above definition is that, as long as the agents *collaboratively* cover the state-action pairs visited by the optimal policy, the average single-policy concentrability coefficient $C_{\mathsf{avg}}^\star < \infty$ is finite. Therefore, this is much weaker than the coverage requirement in the single-agent case.

3. Algorithm and theoretical guarantees

In this section, we first introduce the proposed model-free federated offline RL algorithm called FedLCB-Q, followed by its theoretical performance guarantees.

3.1. Algorithm description

We introduce a federated variant of Q-learning algorithm for offline RL, called FedLCB-Q, that learns a near-optimal Q-function without overestimation on unseen components of the state-action space. The complete description of FedLCB-Q is provided in Appendix A. On a high level, FedLCB-Q performs local Q-function updates at all the agents using its own local offline dataset, and occasionally, globally aggregates the local estimates in a pessimistic

fashion at a central server. To facilitate flexible communication patterns, we follow a synchronization schedule $\mathcal{T}(K)$, which contains the indices of episodes where communication occurs between the agents and the server.

To begin, FedLCB-Q initializes the local estimate ($Q^m_{0,h}$ and $V^m_{0,h}$) at each agent $m \in [M]$ and the global estimates ($Q_{0,h}$ and $V_{0,h}$) for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H+1]$ at the server as follows:

$$Q^m_{0,h}(s,a) = 0, \qquad V^m_{0,h}(s,a) = 0, \tag{11a}$$

$$Q_{0,h}(s,a) = 0, \qquad V_{0,h}(s,a) = 0. \tag{11b}$$

Then, FedLCB-Q proceeds the following steps for each episode $k \in [K]$.

1. **Local updates:** Each agent m samples the kth trajectory $\{(s^m_{k,h}, a^m_{k,h}, r^m_{k,h})\}_{h=1}^H$ from its local offline datasets $\mathcal{D}^m$. For each step $h \in [H]$, agent m updates its local Q-estimate $Q^m_{k,h}$ for $(s,a) = (s^m_{k,h}, a^m_{k,h})$ as follows:

$$\begin{aligned} &Q^m_{k,h}(s,a) \\ &= (1-\eta^m_{k,h}(s,a))Q^m_{k-1,h}(s,a) \\ &\quad + \eta^m_{k,h}(s,a)(r^m_{k,h} + V^m_{k-1,h+1}(s^m_{k,h+1})), \end{aligned} \tag{12}$$

   where $\eta^m_{k,h}(s,a)$ is the learning rate, whose schedule will be specified later, and $V^m_{k-1,h}(s)$ is set as

$$V^m_{k-1,h}(s) = V^m_{\iota(k),h}(s) = V_{\iota(k),h}(s),$$

   where $\iota(k)$ denotes the most recent episode where aggregation occurs before the kth episode, i.e.,

$$\iota(k) := \max_{k'} \{1 \le k' < k : k' \in \mathcal{T}(K)\}.$$

2. **Pessimistic aggregation:** If synchronization is scheduled at episode k, i.e., $k \in \mathcal{T}(K)$, each agent sends its local Q-estimate to a central server for aggregation after finishing the local update for the kth episode. Then, the server updates the global Q-estimate $Q_{k,h}$ by averaging the local Q-estimates and subtracting a penalty for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows:

$$Q_{k,h}(s,a) = \left(\sum_{m=1}^M \alpha^m_{k,h}(s,a) Q^m_{k,h}(s,a)\right) - B_{k,h}(s,a), \tag{13}$$

   where $\alpha^m_{k,h} = [\alpha^m_{k,h}(s,a)]_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in [0,1]^{SA}$ is an entry-wise weight matrix assigned to agent m for each $h \in [H]$, and $B_{k,h}(s,a)$ is a penalty term (to be specified later below) that introduces the pessimism preventing the overestimation of unseen state-action pairs. Accordingly, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, the global value estimate is updated as

$$V_{k,h}(s) = \max\left\{V_{\iota(k),h}(s),\ \max_{a\in\mathcal{A}} Q_{k,h}(s,a)\right\}, \tag{14}$$

   where the outer maximum ensures a monotonic update, as we explain later in the analysis. If $V_{k,h}(s) = \max_{a\in\mathcal{A}} Q_{k,h}(s,a)$, the global policy is updated as $\pi_{k,h}(s) = \arg\max_{a\in\mathcal{A}} Q_{k,h}(s,a)$, otherwise $\pi_{k,h}(s) = \pi_{\iota(k),h}(s)$. After aggregation, the server sends the global Q-function and value estimates to every agent, where $Q^m_{k,h} = Q_{k,h}$, $V^m_{k,h} = V_{k,h}$ for all $(k,m) \in \mathcal{T}(K) \times [M]$.

At the end of K episodes, FedLCB-Q outputs a global Q-estimate $\widehat{Q}_h(s,a) = Q_{K,h}(s,a)$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and a solution policy $\widehat{\pi}_h(s) = \pi_{K,h}(s)$ for all $(s,h) \in \mathcal{S}\times[H]$. For simplicity, we assume that the aggregation step always occurs after the last episode K, i.e., $K \in \mathcal{T}(K)$.

3.2. Choices of key parameters

The success of FedLCB-Q relies on careful and judicious selections of key algorithmic parameters, in a data-driven manner, which we detail below. To begin, let us introduce the following useful notation, which pertains to the counters for visits of agents on each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. For any $(m,k,h) \in [M]\times[K]\times[H]$,

- $n^m_{k,h}(s,a)$: the number of episodes in the interval $(\iota(k), k]$ during which agent m visits (s,a) at step h, i.e., $n^m_{k,h}(s,a) := |\{\iota(k) < i \le k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.
- $N^m_{k,h}(s,a)$: the number of episodes in the interval $[1,k]$ during which agent m visits (s,a) at step h, i.e., $N^m_{k,h}(s,a) := |\{1 \le i \le k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.
- $n_{k,h}(s,a)$: the cumulative count of local episodes across all agents within the interval $(\iota(k), k]$, wherein each agent visits (s,a) at step h, i.e., $n_{k,h}(s,a) := \sum_{m=1}^M n^m_{k,h}(s,a) = \sum_{m=1}^M |\{\iota(k) < i \le k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.
- $N_{k,h}(s,a)$: the cumulative count of local episodes across all agents within the interval $[1,k]$, wherein each agent visits (s,a) at step h, i.e., $N_{k,h}(s,a) := \sum_{m=1}^M N^m_{k,h}(s,a) = \sum_{m=1}^M |\{1 \le i \le k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.

Pessimism in the federated RL. In offline RL, pessimism is key to preventing the overestimation of Q-function on unseen state-action space. For a single-agent case, the pessimism is implemented by subtracting a penalty term computed based on the visiting counter of an agent for each state-action pair, which makes the estimation highly dependent on the quality of agents' datasets (Rashidinejad et al., 2021). For example, when an agent has non-expert data collected using a highly sub-optimal behavior policy, it is inevitable to subtract a large penalty for optimal actions that cannot be reached with the agent's behavior policy, and this

leads to slow convergence or convergence to a sub-optimal policy close to the behavior policy. In the federated setting, from the perspective of a server, as the aggregated information from multiple agents increases confidence, it is natural to be less pessimistic compared to an individual agent. Based on this intuition, given some prescribed probability $\delta \in (0,1)$, we suggest a global penalty computed with the aggregated counters of agents at $k \in \mathcal{T}(K)$:

$$B_{k,h}(s,a) \coloneqq \frac{(H+1)n_{k,h}(s,a)}{N_{k,h}(s,a) + Hn_{k,h}(s,a)} \sqrt{\frac{c_B \zeta_1^2 H^4}{N_{k,h}(s,a)}}, \tag{15}$$

where $B_{k,h}(s,a) = 0$ if $N_{k,h}(s,a) = 0$, and $\zeta_1 = \log\left(\frac{SAMK^2H}{\delta}\right)$ and c_B is some positive constant. Here, the penalty for each state-action pair decreases as long as the agents collectively explore the state-action pair enough. This relaxes the dependency on an individual agent and prevents the estimated policy from being restricted to a local behavior policy.

Local update uncertainty. To guarantee that the pessimism introduced by the global penalty is enough to prevent overestimation on rarely seen state-action pairs, the penalty should dominate the uncertainty of the Q-estimates. However, when agents independently update their own local Q-estimates without frequent communication, the global penalty, which is subtracted only at the aggregation step, may fail to cover the increasing uncertainty of the local Q-estimates during local updates. To handle this, we propose a choice of key parameters (learning rates $\eta_{k,h}^m$ and averaging weights $\alpha_{k,h}^m$) that effectively controls the uncertainty arising from the local updates as follows.

- **Importance averaging.** In the federated setting, agents have offline datasets with heterogeneous distributions induced by different behavior policies, leading to imbalanced uncertainty of local Q-estimates.To minimize the uncertainty of the averaged estimate, we propose the following entrywise weighting scheme for averaging:

$$\alpha_{k,h}^m(s,a) \coloneqq \frac{N_{\iota(k),h}(s,a) + (H+1)Mn_{k,h}^m(s,a)}{M(N_{k,h}(s,a) + Hn_{k,h}(s,a))}, \tag{16}$$

  where $\alpha_{k,h}^m(s,a) = \frac{1}{M}$ if $n_{k,h}(s,a) = 0$. By assigning smaller weights to less frequently updated local Q-estimates with smaller $n_{k,h}^m(s,a)$, which has high uncertainty, the averaged Q-estimate can always maintain an uncertainty level low enough to be dominated by the global penalty, regardless of the heterogeneity in local data distributions. The idea aligns with the notion of importance averaging introduced by Woo et al. (2023), which favors frequently updated local Q-values. Nevertheless, our approach differs in that, unlike Woo et al. (2023), where the assigned weights are determined solely based on local counters $n_{k,h}^m$ in a myopic manner, our weights, factoring in the global counter $N_{\iota(k),h}$, limit bias towards specific agents as the training of local Q-estimates stabilizes. The weighting scheme, mindful of the entire training progress, prevents some local values that have undergone intense updates recently from dominating the global learning of the Q-function, preserving the information accumulated through old updates.

- **Learning rates rescaling.** Local updates without synchronization increase the deviation of local Q-estimates, and this increases the variance of the global Q-estimate at aggregation. However, requiring agents to communicate frequently may be too stringent for many applications in the federated setting. To address this issue, we propose a novel choice of learning rate that exhibits slower decay based on a global counter $N_{\iota(k),h}$, and faster decay during local updates according to the local counter $n_{k,h}^m$:

$$\eta_{k,h}^m(s,a) \coloneqq \frac{M(H+1)}{N_{\iota(k),h}(s,a) + M(H+1)n_{k,h}^m(s,a)}. \tag{17}$$

  The rescaling of the learning rate is crucial to obtain linear speedup without frequent synchronizations. The gradual decay with a global counter allows more aggressive updates of the Q-estimates once collective information from all agents is aggregated, which enables convergence speedup. On the other hand, the fast decrease in learning rates during local updates ensures that agents adaptively slow down their drifts and maintain low variance of their local Q-estimates, without overly restricting the length of local updates. We will further discuss how this effectively reduces the variance of local estimates in Appendix B.1.

The computation of the global penalty (15) and importance averaging (16) at a server requires local counters $n_{k,h}^m(s,a)$ from every agent, and determining the learning rates (17) at each agent requires access to recently aggregated global counters $N_{\iota(k),h}(s,a)$. Therefore, for FedLCB-Q with the specified parameters choices, agents and a server additionally exchange the updated local and global counters at every aggregation step.

3.3. Theoretical guarantees

Given the parameters described above, we now give sample complexity guarantees on the performance of the proposed FedLCB-Q algorithm.

Theorem 3.1. *Consider $\delta \in (0,1)$ and let $\widehat{\pi}$ be the solution policy of* FedLCB-Q. *If a synchronization schedule $\mathcal{T}(K)$ is independent of trajectories in datasets $\mathcal{D}$ and satisfies*

$$\tau_1 \le \sqrt{\frac{H^2 S C_{\mathsf{avg}}^{\star} K}{M}} \quad \textit{and} \quad \frac{\tau_{u+1}}{\tau_u} \le 1 + \frac{2}{H} \tag{18}$$

for any $u \geq 1$, where τ_u is the number of episodes between the $(u-1)$-th and the u-th aggregations. Denoting the total number of samples per agent $T = KH$, the following holds:

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \leq c\left(\sqrt{\frac{H^7 SC_{\mathsf{avg}}^\star \zeta_1^2}{MT}} + \frac{H^4 SC_{\mathsf{avg}}^\star \zeta_1}{MT}\right) \tag{19}$$

at least with probability $1-\delta$, where $\zeta_1 = \log\left(\frac{SAMK^2H}{\delta}\right)$ and $c > 0$ is some universal constant.

Theorem 3.1 implies that as long as the initial synchronization occurs early and the synchronization intervals do not increase too rapidly (cf. (18)), FedLCB-Q is guaranteed to find an ε-optimal policy, i.e., $V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \leq \varepsilon$, for any target accuracy $\varepsilon \in (0, H]$, if the total number of samples per agent T exceeds

$$\widetilde{O}\left(\frac{H^7 SC_{\mathsf{avg}}^\star}{M\varepsilon^2}\right).$$

A few implications are in order.

Linear speedup without expert datasets. The value function gap shows linear speedup with respect to the number of agents M, highlighting the benefit of collaboration. Notably, the guarantee holds even when every agent has low-quality datasets collected by some sub-optimal behavior policy, as long as agents' local data distributions collectively cover the distribution of the optimal policy, where the average single-policy concentrability $C_{\mathsf{avg}}^\star$ (cf. (10)) is finite. On the other end, when performing offline RL using a single agent, it requires that the behavior policy of the single agent individually cover the optimal policy, i.e., $C^\star < \infty$ (cf. (8)), which is much more stringent. Therefore, federated offline RL enables policy learning that otherwise will not be possible in the single-agent setting. Specializing to the case $M = 1$, our bound nearly matches the sample complexity bound $\widetilde{O}\left(\frac{H^6 SC^\star}{\varepsilon^2}\right)$ obtained for a single-agent pessimistic Q-learning algorithm with a similar Hoeffding-style penalty (Shi et al., 2022), up to a factor of H.

Comparison with offline RL using shared datasets. To benchmark the tightness of our bound, let us consider the minimax lower bound of the sample complexity for single-agent offline RL (Li et al., 2024b), as if we collect all the agents' datasets at a central location. Note that the effective single-policy concentrability coefficient (cf. (8)) for the combined datasets $\mathcal{D}_{\mathsf{all}} = \cup_{m=1}^M \mathcal{D}^m$ becomes

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a),\, 1/S\}}{\sum_{m=1}^M d_h^m(s,a)}$$

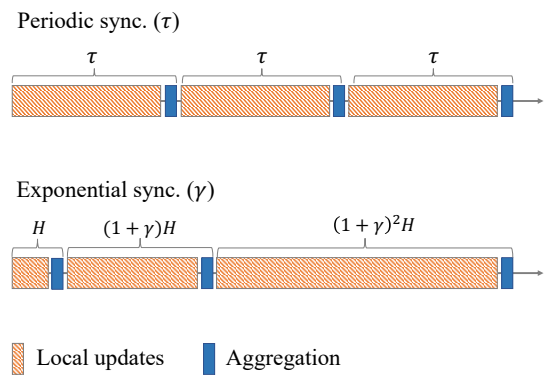


Figure 2: Illustration of the periodic synchronization with constant period τ and the exponential synchronization with a rate γ.

$$= \max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a),\, 1/S\}}{M d_h^{\mathsf{avg}}(s,a)} = \frac{C_{\mathsf{avg}}^\star}{M},$$

leading to the minimax lower bound (Li et al., 2024b)

$$\widetilde{\Omega}\left(\frac{H^4 SC_{\mathsf{avg}}^\star}{M\varepsilon^2}\right).$$

Comparing with the sample complexity bound of FedLCB-Q, obtained as $\widetilde{O}\left(\frac{H^7 SC_{\mathsf{avg}}^\star}{M\varepsilon^2}\right)$, this suggests that the performance of FedLCB-Q is near-optimal up to polynomial factors of H^3 even when compared with the single-agent counterpart assuming shared access to all agents' datasets.

Communication efficiency. Theorem 3.1 suggests initiating the first synchronization early and avoiding rapid increases in synchronization intervals (cf. (18)) to ensure fast convergence. This is attributed to large deviations among agents in the early stages, arising due to coarse Q-estimates and large learning rates, which diminish as training proceeds. For communication efficiency, it is essential to design a synchronization schedule that meets the constraints with the least number of synchronizations. We investigate the following two specific synchronization schedules for FedLCB-Q:

(a) **Periodic synchronization:** For a fixed period $\tau \geq 1$, communication between agents and a server occurs after every τ episodes, i.e., $\tau_i = \tau$ for all $i \geq 1$, and we denote the synchronization schedule as $\mathcal{T}_{\mathsf{period}}(K, \tau)$.

(b) **Exponential synchronization:** For a fixed ratio $\gamma > 0$, initializing $\tau_1 = H$, set $\tau_i = \lfloor (1+\gamma)\tau_{i-1} \rfloor$ for each $i \geq 2$. Under this scheduling, agents communicate frequently at initial iterations, but the period between aggregation steps increases exponentially with the rate of $(1+\gamma)$ and synchronization occurs rarely as training proceeds enough. We denote the synchronization schedule as $\mathcal{T}_{\mathsf{exp}}(K, \gamma)$.

We now analyze the number of communication rounds required to achieve a target accuracy for the above schedules.

Corollary 3.2. *For any given $\delta \in (0,1)$ and target error $\varepsilon \in (0, \min\{H, \frac{H^3 SC^\star_{\mathsf{avg}}}{M}\}]$, suppose the total number of samples per agent $T = KH$ satisfies*

$$T \asymp \frac{H^7 SC^\star_{\mathsf{avg}}}{M\varepsilon^2},$$

and FedLCB-Q *performs under the periodic synchronization scheduling, i.e., $\mathcal{T}(K) = \mathcal{T}_{\mathsf{period}}(K,\tau)$, with $\tau \asymp \sqrt{\frac{HSC^\star_{\mathsf{avg}} T}{M}}$, or the exponential synchronization scheduling, i.e., $\mathcal{T}(K) = \mathcal{T}_{\mathsf{exp}}(K,\gamma)$, with $\gamma = \frac{2}{H}$. Then, each schedule requires the number of synchronizations at most*

$$(\mathsf{Periodic}) \quad |\mathcal{T}_{\mathsf{period}}(K,\tau)| \lesssim \sqrt{\frac{MK}{H^2 SC^\star_{\mathsf{avg}}}}, \tag{20a}$$

$$(\mathsf{Exponential}) \quad |\mathcal{T}_{\mathsf{exp}}(K,\gamma)| \lesssim H, \tag{20b}$$

respectively, and the solution policy $\widehat{\pi}$ of FedLCB-Q *is an ε-optimal policy at least with probability $1-\delta$.*

Corollary 3.2 implies that FedLCB-Q requires only $\widetilde{O}(H)$ aggregations to achieve the target accuracy under appropriate synchronization schedules, such as the exponential synchronization schedule. Notably, the number of communication rounds is nearly independent of the size of the state-action space, the total number of episodes, or the number of agents, and this outperforms prior art (Woo et al., 2023). Furthermore, analysis suggests that exponential synchronization with a modest rate $\gamma = 2/H$ is a key to achieving such communication efficiency. With our strategic choices of learning rates, local Q-estimates stabilize as training proceeds, and thus agents can perform more local updates than previous rounds without increasing uncertainty beyond the control of the global pessimism penalty. Exponential synchronization reduces the number of synchronizations by capturing the additional room for local updates arising from the stabilization of Q-estimates. On the other hand, periodic synchronization does not exploit this benefit, even if we set the period τ maximally under (18) due to which it necessitates more communication rounds, which increase with K and M.

4. Discussions

We investigated federated offline RL, which enables multiple agents with history datasets to collaboratively learn an optimal policy, without sharing datasets. We proposed a federated offline Q-learning algorithm called FedLCB-Q, which iteratively performs local updates with rescaled learning rates at agents, and global aggregation with weighted averaging and global penalty at a server, which effectively controls the uncertainty in both local and global Q-estimates. Our sample complexity analysis demonstrates that FedLCB-Q achieves linear speedup in terms of the number of agents requiring only collective coverage of agents' datasets over the distribution of the optimal policy, not restricted to the quality of individual datasets. Furthermore, we showed that FedLCB-Q is communication-efficient, requiring only $\widetilde{O}(H)$ synchronizations under the exponential synchronization scheduling. For future exploration, this work paves the way for many interesting directions, some of which are outlined below.

- *Tightening H dependency.* Although our sample complexity bound is nearly optimal with respect to most salient problem parameters, such as state space size and single-policy concentrability coefficient, it falls short of optimality in terms of horizon length compared to the minimax sample complexity lower bound in the single-agent setting (Xie et al., 2021b). Closing this gap and improving sample complexity with variance reduction techniques, as proposed by Shi et al. (2022), will be an interesting avenue for future exploration.

- *Beyond episodic tabular MDPs.* Extending episodic tabular MDPs, it would be interesting to broaden our analysis framework to encompass other RL settings, including, the infinite-horizon setting (Woo et al., 2023; Yan et al., 2023), infinite state-action space setting (Bose et al., 2024), and the integration of function approximation.

- *Improving robustness.* Our work focuses on a scenario in which agents collect datasets from a common MDP without any disturbances. Yet, in real-world scenarios, some agents may possess datasets collected from perturbed MDPs. This introduces the need for additional considerations regarding robustness, as discussed in Shi et al. (2023). Therefore, enhancing our work to effectively handle the variability or noisiness of MDPs would be a compelling avenue for improvement.

- *Multi-task RL.* In many applications where various clients pursue different objectives, multi-task reinforcement learning holds a significant interest. It will be of great interest to extend our work to the multi-task RL setting (Yang et al., 2023; Jin et al., 2022; Zhou et al., 2024), which enables agents to learn their own optimal policies for their personalized goals while benefiting from collaboration by sharing common features of tasks.

Acknowledgements

This work is supported in part by the grants NSF CCF-2007911, CCF-2106778, CNS-2148212, ONR N00014-19-1-2404 to Y. Chi, and NSF-CCF 2007834, CCF-2045694, CNS-2112471, ONR N00014-23-1-2149 to G. Joshi.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

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

Bertsekas, D. P. *Dynamic programming and optimal control (4th edition)*. Athena Scientific, 2017.

Bose, A., Du, S. S., and Fazel, M. Offline multi-task transfer rl with representational penalization. *arXiv preprint arXiv:2402.12570*, 2024.

Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T., and Clarke, J.-P. Performance of Q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, 2019.

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

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

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

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

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 20132–20145, 2021.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, pp. 4863–4873, 2018.

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

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096, 2021.

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

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 21810–21823, 2020.

Kim, B. and Oh, M.-H. Model-based offline reinforcement learning with count-based conservatism. In *International Conference on Machine Learning*, pp. 16728–16746, 2023.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1179–1191, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

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

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024a.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233 – 260, 2024b. doi: 10.1214/23-AOS2342.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems*, 2020.

Mitzenmacher, M. and Upfal, E. *Probability and computing*. Cambridge University Press, 2005.

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

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

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 2021.

Shen, H., Lu, S., Cui, X., and Chen, T. Distributed offline policy optimization over batch data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 4443–4472, 2023.

Shi, C., Xiong, W., Shen, C., and Yang, J. Provably efficient offline reinforcement learning with perturbed data sources. In *International Conference on Machine Learning*, pp. 31353–31388, 2023.

Shi, L. and Chi, Y. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

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

Siegel, N., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., Heess, N., and Riedmiller, M. Keep doing what worked: Behavior modelling priors for offline reinforcement learning. In *International Conference on Learning Representations*, 2020.

Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Offline minimax soft-q-learning under realizability and partial coverage. In *Advances in Neural Information Processing Systems*, volume 37, 2023.

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

Wang, H., Mitra, A., Hassani, H., Pappas, G. J., and Anderson, J. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023.

Wang, Y., Dong, K., Chen, X., and Wang, L. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations*, 2019.

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

Woo, J., Joshi, G., and Chi, Y. The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. In *International Conference on Machine Learning*, pp. 37157–37216, 2023.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

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

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems*, 2021b.

Xu, P. and Gu, Q. A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, pp. 10555–10565. PMLR, 2020.

Yan, Y., Li, G., Chen, Y., and Fan, J. The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*, 69(11):7185–7219, 2023.

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

Yin, M. and Wang, Y.-X. Towards instance-optimal offline reinforcement learning with pessimism. In *Advances in Neural Information Processing Systems*, 2021.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142, 2020.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2021.

Zhang, C., Wang, H., Mitra, A., and Anderson, J. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *International Conference on Learning Representations*, 2024.

Zheng, Z., Gao, F., Xue, L., and Yang, J. Federated Q-learning: Linear regret speedup with low communication cost. *International Conference on Learning Representations*, 2024.

Zhou, D., Zhang, Y., Sonabend-W, A., Wang, Z., Lu, J., and Cai, T. Federated offline reinforcement learning. *Journal of the American Statistical Association*, 0(0):1–12, 2024. doi: 10.1080/01621459.2024.2310287.

A. Complete description of FedLCB-Q

We provide the complete description of FedLCB-Q in Algorithm 1, with its agent-end and server-end subroutines described in Algorithm 2 and Algorithm 3 respectively.

Algorithm 1 Federated pessimistic Q-learning (FedLCB-Q)

Parameters: horizon length H, number of agents M, total number of episodes per agent K, synchronization schedule $\mathcal{T}(K)$, target error $\delta \in (0,1)$, $\zeta_1 = \log\left(\frac{SAK^2MH}{\delta}\right)$, $c_B > 0$.
Initialization: set $Q^m_{0,h}(s,a) = 0$, $V^m_{0,h}(s) = 0$, $N^m_{0,h}(s,a) = 0$, $n^m_{0,h}(s,a) = 0$, $N_{0,h}(s,a) = 0$, $n_{0,h}(s,a) = 0$ for all $(m,s,a,h) \in [M] \times \mathcal{S} \times \mathcal{A} \times [H+1]$.
for $k = 1, \cdots, K$ **do**
  // Update the local Q-estimate and visitation counts at each agent
  $(Q^m_{k,h}, n^m_{k,h})$ = `Local-Q-learning()`;
  if $k \in \mathcal{T}(K)$ **then**
    // Agent-to-server communication
    Agents communicate $Q^m_{k,h}$ and $n^m_{k,h}$ to the server;
    // Global pessimistic averaging in a server
    $(Q_{k,h}, V_{k,h}, \pi_{k,h})$ = `Global-pessimistic-averaging()`;
    // Server-to-agent communication
    Server communication $Q_{k,h}$, $V_{k,h}$ and $N_{k,h}$ to agents;
    // Synchronize local Q-estimates
    for $(m,s,a,h) \in [M] \times \mathcal{S} \times \mathcal{A} \times [H]$ **do**
      $Q^m_{k,h}(s,a) = Q_{k,h}(s,a)$, $V^m_{k,h}(s) = V_{k,h}(s)$
    end for
  end if
end for
return: $\widehat{Q} = \{Q_{K,h}\}_{h\in[H]}$ and $\widehat{\pi} = \{\pi_{K,h}\}_{h\in[H]}$.

Algorithm 2 `Local-Q-learning` (agents)

for $m = 1, \cdots, M$ **do**
  Sample the k-th trajectory $\{(s^m_{k,h}, a^m_{k,h}, r^m_{k,h}, s^m_{k,h+1})\}_{h=1}^H$ from $\mathcal{D}^m$
  for $h = 1, \cdots, H$ **do**
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      $Q^m_{k,h}(s,a) = Q^m_{k-1,h}(s,a)$, $V^m_{k,h}(s) = V^m_{k-1,h}(s)$
    end for
    // Update the local counters and learning rates
    $n^m_{k,h}(s^m_{k,h}, a^m_{k,h}) = n^m_{k-1,h}(s^m_{k,h}, a^m_{k,h}) + 1$
    $\eta^m_{k,h}(s^m_{k,h}, a^m_{k,h}) = \frac{M(H+1)}{N_{\iota(k),h}(s^m_{k,h}, a^m_{k,h}) + M(H+1)n^m_{k,h}(s^m_{k,h}, a^m_{k,h})}$
    // Update local Q-estimates
    $Q^m_{k,h}(s^m_{k,h}, a^m_{k,h}) = \big(1 - \eta^m_{k,h}(s^m_{k,h}, a^m_{k,h})\big) Q^m_{k-1,h}(s^m_{k,h}, a^m_{k,h}) + \eta^m_{k,h}(s,a)(r^m_{k,h} + V^m_{k-1,h+1}(s^m_{k,h+1}))$
  end for
end for

Algorithm 3 `Global-pessimistic-averaging` (server)

for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ **do**
  // Update the average counter
  $n_{k,h}(s,a) = \sum_{m=1}^{M} n_{k,h}^m(s,a)$, $N_{k,h}(s,a) = N_{\iota(k),h}(s,a) + n_{k,h}(s,a)$
  // Compute global penalty and averaging weights
  $B_{k,h}(s,a) = \frac{(H+1)n_{k,h}(s,a)}{N_{k,h}(s,a)+Hn_{k,h}(s,a)} \sqrt{\frac{c_B \zeta_1^2 H^4}{N_{k,h}(s,a)}}$ if $N_{k,h}(s,a) > 0$, otherwise, $B_{k,h}(s,a) = 0$
  for $m = 1 \cdots M$ **do**
    $\alpha_{k,h}^m(s,a) = \frac{1}{M} \frac{N_{\iota(k),h}(s,a)+M(H+1)n_{k,h}^m(s,a)}{N_{k,h}(s,a)+Hn_{k,h}(s,a)}$ if $n_{k,h}^m(s,a) > 0$, otherwise, $\alpha_{k,h}^m(s,a) = \frac{1}{M}$
  end for
  // Update global Q-estimates
  $Q_{k,h}(s,a) = \sum_{m=1}^{M} \alpha_{k,h}^m(s,a) Q_{k,h}^m(s,a) - B_{k,h}(s,a)$
  $V_{k,h}(s) = \max\left\{V_{\iota(k),h}(s), \max_{a\in\mathcal{A}} Q_{k,h}(s,a)\right\}$
  $\pi_{k,h}(s) = \arg\max_{a\in\mathcal{A}} Q_{k,h}(s,a)$ if $V_{k,h}(s) = \max_{a\in\mathcal{A}} Q_{k,h}(s,a)$, otherwise, $\pi_{k,h}(s) = \pi_{\iota(k),h}(s)$
end for

B. Analysis

In this section, we will outline useful properties of FedLCB-Q and the key steps of the proof of Theorem 3.1, deferring the details, such as proofs of supporting lemmas, to Appendix C and D.

Throughout the paper, we adopt the following shorthand notation

$$P_{h,s,a} \coloneqq P_h(\cdot \mid s,a) \in [0,1]^{1\times S}, \tag{21}$$

which represents the transition probability vector given the current state-action pair (s,a) at step h. In addition, define $P_{k,h}^m \in \{0,1\}^{1\times S}$ as the empirical transition vector at step h of the k-th episode at agent m, namely

$$P_{k,h}^m(s) = \mathbb{I}(s = s_{k,h+1}^m), \quad \text{for all } s \in \mathcal{S}. \tag{22}$$

These are the notations pertaining to the counters for visits of agents on each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. For any $(m,k,h) \in [M] \times [K] \times [H]$,

- $l_{k,h}^m(s,a)$: a set of episodes in the interval $(\iota(k), k]$ during which agent m visits (s,a) at step h, i.e., $l_{k,h}^m(s,a) \coloneqq \{\iota(k) < i \le k : (s_{i,h}^m, a_{i,h}^m) = (s,a)\}$.
- $L_{k,h}^m(s,a)$: a set of episodes in the interval $[1,k]$ during which agent m visits (s,a) at step h, i.e. $L_{k,h}^m(s,a) \coloneqq \{1 \le i \le k : (s_{i,h}^m, a_{i,h}^m) = (s,a)\}$.

We also introduce the following notation related to the synchronization schedule $\mathcal{T}(K)$. For any positive integer k and u,

- t_u: the index of episodes, after which the uth synchronization occurs.
- τ_u: the number of local updates (episodes) taken between the $(u-1)$th and the uth synchronizations.
- $\iota(k)$: the most recent episode where the aggregation occurs before the kth episode.
- $\phi(k)$: the minimum index of aggregation occurring after k-th episode.

B.1. Basic facts

Error recursion of Q-estimates. We begin with the following key error decomposition of the Q-estimate at each synchronization, whose proof is provided in Appendix D.1.

Lemma B.1 (Q-estimation error decomposition)**.** *Consider a Q-function $Q^\pi = \{Q_h^\pi(s,a)\}_{[H]\times\mathcal{S}\times\mathcal{A}}$ and value function $V^\pi = \{V_h^\pi(s)\}_{[H]\times\mathcal{S}}$ induced by a policy π. Then, for any $[H]\times\mathcal{S}\times\mathcal{A}$ and $k\in\mathcal{T}(K)$, the error between Q_h^π and $Q_{k,h}$ is decomposed as follows:*

$$\begin{aligned} Q_h^\pi(s,a) - Q_{k,h}(s,a) = &\underbrace{\omega_{0,k,h}(s,a)(Q_h^\pi(s,a) - Q_{0,h}(s,a))}_{=:D_1^\pi(s,a,k,h)\text{: initialization error}} \\ &+ \underbrace{\sum_{m=1}^{M}\sum_{i\in L_{k,h}^m(s,a)} \omega_{i,k,h}^m(s,a)(P_{h,s,a} - P_{i,h}^m)V_{i-1,h+1}^m}_{=:D_2(s,a,k,h)\text{: transition variance}} \\ &+ \underbrace{\sum_{u=1}^{\phi(k)} B_{t_u,h}(s,a)\prod_{u'=u+1}^{\phi(k)}\lambda_{u',h}(s,a)}_{=:D_3(s,a,k,h)\text{: global penalty}} \\ &+ \underbrace{\sum_{m=1}^{M}\sum_{i\in L_{k,h}^m(s,a)} \omega_{i,k,h}^m(s,a)P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m)}_{=:D_4^\pi(s,a,k,h)\text{: recursion}}, \end{aligned} \tag{23}$$

where $L_{k,h}^m(s,a) := \{1\le i\le k : (s_{i,h}^m, a_{i,h}^m) = (s,a)\}$ and $l_{k,h}^m(s,a) := \{\iota(k) < i \le k : (s_{i,h}^m, a_{i,h}^m) = (s,a)\}$. And, for simplicity, we use the shortened notations defined as

$$\lambda_{v,h}(s,a) = \begin{cases} 1 & \text{if } N_{k,h}(s,a) = 0 \\ \frac{N_{\iota(k),h}(s,a)}{N_{k,h}(s,a) + Hn_{k,h}(s,a)} & \text{otherwise} \end{cases}, \quad v = \phi(k), \tag{24a}$$

$$\omega_{0,k,h}^m(s,a) = \begin{cases} 1 & \text{if } N_{k,h}(s,a) = 0 \\ 0 & \text{otherwise} \end{cases}, \tag{24b}$$

$$\omega_{i,k,h}^m(s,a) = \frac{H+1}{N_{k,h}(s,a) + Hn_{k,h}(s,a)}\left(\prod_{x=\phi(i)}^{\phi(k)-1}\frac{N_{t_x,h}(s,a)}{N_{t_x,h}(s,a) + Hn_{t_x,h}(s,a)}\right), \quad i\in L_{k,h}^m(s,a). \tag{24c}$$

Equally favoring episodes within the same local update round. According to the decomposition (23) in Lemma B.1, for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, the Q-estimation error at episode k significantly depends on the weighted sum of transition difference for each episode where the local update occurs, namely $D_2(s,a,k,h)$. Intuitively, the weight $\omega_{i,k,h}^m(s,a)$ assigned to each episode i balances the accumulation of information from old and new updates. Our choice of learning rates, which decreases fast during local updates, as illustrated in Figure 3a, ensures that the weight $\omega_{i,k,h}^m(s,a)$ within the same local update round is always equal for all episodes and agents, as shown in (24c) and Figure 3b. The uniform weights allow the transition information of each episode to be accumulated evenly, regardless of other transitions that occur in future episodes or other agents' episodes. This is essential to keep variance arising from local updates low, especially when a synchronization period is long. Assigning equal weight to every episode allows to fully utilize transitions observed during local updates without forgetting old information, regardless of the length of the synchronization period.

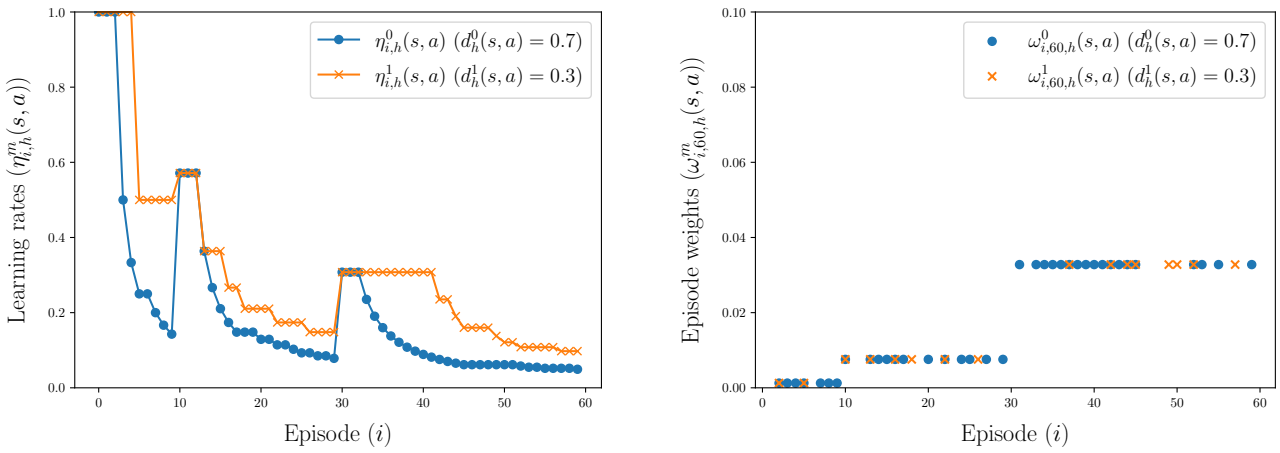
Bounded visitation counters. We introduce the following lemma regarding the visitation counters, whose proof is provided in Appendix D.2.

Lemma B.2 (Concentration bound on the visitation counters)**.** *Consider any $\delta\in(0,1)$ and some universal constant $c_1 > 0$, and let*

$$\zeta_0 := \log\left(\frac{2|\mathcal{S}||\mathcal{A}|KH}{\delta}\right) \quad \text{and} \quad K_0(s,a,h) := \frac{4\zeta_0}{c_1 M d_h^{\mathsf{avg}}(s,a)}. \tag{25}$$

Then, for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, the following holds

$$\text{when } k\ge K_0(s,a,h): \qquad \frac{1}{2}kMd_h^{\mathsf{avg}}(s,a) \le N_{k,h}(s,a) \le 2kMd_h^{\mathsf{avg}}(s,a), \tag{26a}$$



(a) Rescaled learning rates

(b) Episode weights

Figure 3: Illustration of the rescaled learning rates ($\eta^m_{i,h}(s,a)$) and the episode weights ($\omega^m_{i,60,h}(s,a)$) induced by the learning rates of two agents $m = 0, 1$ for episodes $1 \le i \le 60$, where $H = 5$, the occupancy distribution of each agent on $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [5]$ is $d^0_h(s,a) = 0.7$ and $d^1_h(s,a) = 0.3$, respectively, and the synchronization schedule is $\mathcal{T}(60) = \{10, 30, 60\}$.

$$\textit{when } k \le K_0(s,a,h): \qquad N_{k,h}(s,a) \le 8\zeta_0/c_1 \tag{26b}$$

with probability at least $1-\delta$.

We denote the event that (26) holds as $\mathcal{E}_0$.

Monotonic and pessimistic global value updates. Note that the global value estimate is always monotonically non-decreasing, i.e., for $k', k \in \mathcal{T}(K)$ it holds

$$\forall s \in \mathcal{S}: \qquad V_{k,h}(s) \ge V_{k',h}(s) \quad \text{when } k' \le k, \tag{27}$$

which follows directly from the update rule (14). Moreover, we have the following important lemma regarding the pessimistic property of the value estimate, whose proof is provided in Appendix D.3.

Lemma B.3 (Pessimistic global value)**.** *Recall $Q_{k,h}$, $V_{k,h}$, and $\pi_{k,h}$ in Algorithm 1. Let $\pi_k = \{\pi_{k,h}\}_{h\in[H]}$. Given any $\delta \in (0,1)$, for all $(k,h) \in \mathcal{T}(K) \times [H]$, it holds with probability at least $1-\delta$ that*

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad |D_2(s,a,k,h)| \le D_3(s,a,k,h) \le \sqrt{\frac{4c_B\zeta_1^2 H^4}{\max\{N_{k,h}(s,a),1\}}}, \tag{28a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_{k,h}(s,a) \le Q^{\pi_k}_h(s,a) \le Q^\star_h(s,a), \tag{28b}$$

$$\forall s \in \mathcal{S}: \qquad V_{k,h}(s) \le V^{\pi_k}_h(s) \le V^\star_h(s). \tag{28c}$$

In words, Lemma B.3 makes concrete the role of the penalty term in dominating the variability of the value estimates due to stochastic transitions, and ensures that the estimated value is a pessimistic estimate of the true optimal value function.

B.2. Proof of Theorem 3.1

Now we are ready to provide the proof of Theorem 3.1, which is divided into several key steps as follows.

Step 1: decomposition of the performance gap. The performance gap between the solution policy $\widehat{\pi}$ of Algorithm 1 after K episodes and the optimal policy $\pi^\star$ can be bounded as follows:

$$V^\star_1(\rho) - V^{\widehat{\pi}}_1(\rho) = \mathbb{E}_{s_1\sim\rho}\left[V^\star_1(s_1)\right] - \mathbb{E}_{s_1\sim\rho}\left[V^{\pi_K}_1(s_1)\right]$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \mathbb{E}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathbb{E}_{s_1\sim\rho}\left[V_{K,1}(s_1)\right] \\
&\overset{\text{(ii)}}{\leq} \frac{1}{K}\sum_{v=1}^{\phi(K)} \tau_v \big(\mathbb{E}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathbb{E}_{s_1\sim\rho}\left[V_{t_v,1}(s_1)\right]\big) \\
&= \frac{1}{K}\sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} \underbrace{d_1^{\pi^\star}(s)}_{=\rho(s)} \left(V_1^\star(s) - V_{t_v,1}(s)\right) \\
&\leq \frac{1}{K}\max_{h\in[H]} \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s) - V_{t_v,h}(s)\right),
\end{aligned} \tag{29}
$$

where (i) follows from Lemma B.3, and (ii) follows from the monotonicity property in (27) and $\sum_{v=1}^{\phi(K)} \tau_v = K$.

Since $\pi^\star = \{\pi_h^\star\}_{h\in[H]}$ is deterministic, for any $k \in \mathcal{T}(K)$ and $h \in [H]$, it follows that

$$
\begin{aligned}
\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s) - V_{k,h}(s)\right) &= \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))\left(V_h^\star(s) - V_{k,h}(s)\right) \\
&\leq \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))\big(Q_h^\star(s,\pi_h^\star(s)) - Q_{k,h}(s,\pi_h^\star(s))\big),
\end{aligned} \tag{30}
$$

where the inequality holds because $Q_{k,h}(s,\pi_h^\star(s)) \leq \max_{a\in\mathcal{A}} Q_{k,h}(s,a) \leq V_{k,h}(s)$ due to (14).

To continue, applying Lemma B.1 by setting $\pi = \pi^\star$, the Q-estimate error after k episodes is decomposed as follows:

$$
\begin{aligned}
Q_h^\star(s,a) - Q_{k,h}(s,a) &= D_1^{\pi^\star}(s,a,k,h) + D_2(s,a,k,h) + D_3(s,a,k,h) + D_4^{\pi^\star}(s,a,k,h) \\
&\leq D_1^{\pi^\star}(s,a,k,h) + D_4^{\pi^\star}(s,a,k,h) + 2D_3(s,a,k,h),
\end{aligned} \tag{31}
$$

where the second line follows from Lemma B.3. Finally, inserting the decomposition (31) and (30) back into (29), we control the performance gap with the following terms:

$$
\begin{aligned}
&V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \\
&\leq \frac{1}{K}\max_{h\in[H]} \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left[D_1^{\pi^\star}(s,\pi_h^\star(s),t_v,h) + D_4^{\pi^\star}(s,\pi_h^\star(s),t_v,h) + 2D_3(s,\pi_h^\star(s),t_v,h)\right] \\
&=: \frac{1}{K}\max_{h\in[H]}\left(D_{1,h} + D_{4,h} + D_{3,h}\right),
\end{aligned} \tag{32}
$$

for which we shall aim to bound each term individually, adopting the following short-hand notation:

$$
\begin{aligned}
D_{i,h} &:= \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) D_i^{\pi^\star}(s,\pi_h^\star(s),t_v,h) \qquad \text{for } i\in\{1,4\}, \\
D_{3,h} &:= \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) D_3(s,\pi_h^\star(s),t_v,h).
\end{aligned} \tag{33}
$$

Step 2: Bounding the decomposed terms. Here, we derive the bound of the decomposed terms separately as follows under the event $\mathcal{E}_0$, which holds with probability at least $1-\delta$.

- **Bounding $D_{1,h}$.** Using the fact that $0 \leq Q_h^\star(s,\pi_h^\star(s)) - Q_{0,h}(s,\pi_h^\star(s)) \leq H$, which follows from Lemma B.3, it follows

$$
D_{1,h} = \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))\omega_{0,t_v,h}(s,\pi_h^\star(s))(Q_h^\star(s,\pi_h^\star(s)) - Q_{0,h}(s,\pi_h^\star(s)))
$$

$$\leq \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))\omega_{0,t_v,h}(s,\pi_h^\star(s))H$$
$$= H\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s)) \sum_{v=1}^{\phi(K)} \tau_v \mathbb{I}\{N_{t_v,h}(s,\pi_h^\star(s))=0\} \tag{34}$$

where the last line follows from (24b). To continue, note that

$$\sum_{v=1}^{\phi(K)} \tau_v \mathbb{I}\{N_{t_v,h}(s,\pi_h^\star(s))=0\}$$
$$= \sum_{v\in[\phi(K)]:t_v\leq K_0(s,\pi_h^\star(s),h)} \tau_v \mathbb{I}\{N_{t_v,h}(s,\pi_h^\star(s))=0\} + \sum_{v\in[\phi(K)]:t_v> K_0(s,\pi_h^\star(s),h)} \mathbb{I}\{N_{t_v,h}(s,\pi_h^\star(s))=0\}$$
$$\leq K_0(s,\pi_h^\star(s),h),$$

where the last line follows since under the event $\mathcal{E}_0$, $N_{t_v,h}(s,\pi_h^\star(s))>0$ when $t_v > K_0(s,\pi_h^\star(s),h)$. Plugging the above inequality and the definition of $K_0(s,\pi_h^\star(s),h)$ back to (34) leads to

$$D_{1,h} \leq H\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))K_0(s,\pi_h^\star(s),h)$$
$$= H\sum_{s\in\mathcal{S}} \frac{\min\{d_h^{\pi^\star}(s,\pi_h^\star(s)),1/S\}}{d_h^{\mathsf{avg}}(s,\pi_h^\star(s))}\left(\frac{12\zeta_0}{M}\right)\frac{d_h^{\pi^\star}(s,\pi_h^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi_h^\star(s)),1/S\}}$$
$$\lesssim \frac{HC_{\mathsf{avg}}^\star S}{M}, \tag{35}$$

where the last line follows from the definition of $C_{\mathsf{avg}}^\star$ and the fact that

$$\sum_{s\in\mathcal{S}} \frac{d_h^{\pi^\star}(s,\pi_h^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi_h^\star(s)),1/S\}} \leq \sum_{s\in\mathcal{S}}\left(1+d_h^{\pi^\star}(s,\pi_h^\star(s))S\right) = \sum_{s\in\mathcal{S}}\left(1+d_h^{\pi^\star}(s)S\right) = 2S.$$

- **Bounding $D_{3,h}$.** The range of $D_3(s,a,k,h)$ is bounded as shown in the following Lemma, whose proof is provided in Appendix D.5.

Lemma B.4. *For any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ and $k\in\mathcal{T}(K)$, if $N_{k,h}(s,a)=0$, $D_3(s,a,k,h)=0$, and if, $N_{k,h}(s,a)>0$, the following holds:*

$$D_3(s,a,k,h) \in \left[\sqrt{\frac{c_B\zeta_1^2H^4}{N_{k,h}(s,a)}}, \sqrt{\frac{4c_B\zeta_1^2H^4}{N_{k,h}(s,a)}}\right]. \tag{36}$$

With the above lemma in hand, recalling (33) gives

$$D_{3,h} = \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))(2D_3(s,\pi_h^\star(s),t_v,h))$$
$$\leq \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s)) \sum_{v=1}^{\phi(K)} 2\tau_v\sqrt{\frac{4c_B\zeta_1^2H^4}{\max\{N_{t_v,h}(s,\pi_h^\star(s)),1\}}} \tag{37}$$

According to Lemma B.2, $N_{t_v,h}(s,a)\geq \frac{1}{2}t_vMd_h^{\mathsf{avg}}(s,a)$ holds if $t_v\geq K_0(s,a,h)$ under the event $\mathcal{E}_0$. Therefore,

$$\sum_{v=1}^{\phi(K)} \tau_v\sqrt{\frac{H^4}{\max\{N_{t_v,h}(s,a),1\}}} \lesssim \sum_{v:t_v\leq K_0(s,a,h)} \tau_v H^2 + \sum_{v:t_v>K_0(s,a,h)} \tau_v\sqrt{\frac{H^4}{\max\{N_{t_v,h}(s,a),1\}}}$$

$$\lesssim H^2 K_0(s,a,h) + \sum_{v:t_v > K_0(s,a,h)} \tau_v \sqrt{\frac{H^4}{\max\{N_{t_v,h}(s,a),1\}}}$$
$$\lesssim H^2 K_0(s,a,h) + \sum_{v=1}^{\phi(K)} \tau_v \sqrt{\frac{H^4}{M t_v d_h^{\mathsf{avg}}(s,a)}}. \tag{38}$$

Plugging the above inequality and the definitions of $K_0(s,\pi_h^\star(s),h)$ and $C_{\mathsf{avg}}^\star$ to (37), we obtain

$$\begin{aligned}
D_{3,h} &\lesssim \frac{H^2}{M}\sum_{s\in\mathcal{S}} \frac{d_h^{\pi^\star}(s,\pi_h^\star(s))}{d_h^{\mathsf{avg}}(s,\pi_h^\star(s))} + \sum_{v=1}^{\phi(K)}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))\tau_v \sqrt{\frac{H^4}{M t_v d_h^{\mathsf{avg}}(s,\pi_h^\star(s))}} \\
&\lesssim \frac{H^2 C_{\mathsf{avg}}^\star}{M}\sum_{s\in\mathcal{S}} \frac{d_h^{\pi^\star}(s,\pi_h^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi_h^\star(s)),1/S\}} + \sum_{v=1}^{\phi(K)} \sqrt{\frac{H^4 C_{\mathsf{avg}}^\star \tau_v^2}{M t_v}} \sum_{s\in\mathcal{S}} \sqrt{\frac{(d_h^{\pi^\star}(s,\pi_h^\star(s)))^2}{\min\{d_h^{\pi^\star}(s,\pi_h^\star(s)),1/S\}}} \\
&\overset{(i)}{\lesssim} \frac{H^2 C_{\mathsf{avg}}^\star S}{M} + \sqrt{\frac{H^4 C_{\mathsf{avg}}^\star S}{M}} \sum_{v=1}^{\phi(K)} \sqrt{\tau_v}\sqrt{\frac{\tau_v}{t_v}} \\
&\overset{(ii)}{\lesssim} \frac{H^2 C_{\mathsf{avg}}^\star S}{M} + \sqrt{\frac{H^4 S K C_{\mathsf{avg}}^\star}{M}},
\end{aligned} \tag{39}$$

where (i) holds due to Cauchy-Schwarz inequality and the fact that

$$\sum_{s\in\mathcal{S}} \frac{d_h^{\pi^\star}(s,\pi_h^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi_h^\star(s)),1/S\}} \le \sum_{s\in\mathcal{S}} \Big(1 + d_h^{\pi^\star}(s,\pi_h^\star(s))S\Big) = \sum_{s\in\mathcal{S}} \Big(1 + d_h^{\pi^\star}(s)S\Big) = 2S,$$

and the last line (ii) follows from the Cauchy-Schwarz inequality and the fact that $\sum_{v=1}^{\phi(K)} \tau_v = K$ and $\sum_{v=1}^{\phi(K)} \frac{\tau_v}{t_v} \le 1 + \log K$, with the latter following from Lemma C.2.

- **Bounding $D_{4,h}$.** In the following lemma, whose proof is provided in Section D.6, we extract the recursive formulation of $D_{4,h}$ as follows:

Lemma B.5. *Consider any $\delta \in (0,1)$. For any $h \in [H]$, the following holds with probability at least $1-\delta$:*

$$\begin{aligned}
&\sum_{v=1}^{\phi(K)} \tau_v \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{m=1}^{M} \sum_{i\in L_{t_v,h}^m(s,a)} \omega_{i,t_v,h}^m(s,a) P_{h,s,a}(V_{h+1}^\star - V_{\iota(i),h+1}) \\
&\lesssim \sigma_{\mathsf{aux}} + (1+\frac{1}{H})\sum_{u=1}^{\phi(K)} \tau_u \sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)(V_{h+1}^\star(s) - V_{t_{u-1},h+1}(s))
\end{aligned} \tag{40}$$

where

$$\sigma_{\mathsf{aux}} = \sqrt{\frac{H^2 K S C_{\mathsf{avg}}^\star}{M}} + \frac{H^2 S C_{\mathsf{avg}}^\star}{M}. \tag{41}$$

Step 3: Recursion. Combining the bounds of the decomposed errors ((35), (39), and (40)), for any $h \in [H]$, we obtain the following recursive relation:

$$\begin{aligned}
&\sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_h^\star(s) - V_{t_v,h}(s)\right) \\
&\lesssim \theta_K + (1+\frac{1}{H})\sum_{u=1}^{\phi(K)} \tau_u \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_{h+1}^\star(s) - V_{t_{u-1},h+1}(s)\right)
\end{aligned}$$

$$\overset{(i)}{\lesssim} (\theta_K + H\tau_1) + (1+\frac{1}{H}) \sum_{u=1}^{\phi(K)-1} \tau_{u+1} \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s) \left(V_{h+1}^\star(s) - V_{t_u,h+1}(s)\right)$$
$$\overset{(ii)}{\lesssim} (\theta_K + H\tau_1) + (1+\frac{2}{H})^2 \sum_{u=1}^{\phi(K)-1} \tau_{u} \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s) \left(V_{h+1}^\star(s) - V_{t_u,h+1}(s)\right) \tag{42}$$

where (i) holds because $V_{h+1}^\star(s) - V_{t_u,h+1}(s) \le H$ and (ii) holds due to the condition $\frac{\tau_{u+1}}{\tau_u} \le 1 + \frac{2}{H}$ for all $1 \le u \le \phi(K)$ and the fact that $V_{h+1}^\star(s) \ge V_{t_u,h+1}(s)$ shown in Lemma B.3, and we denote

$$\theta_K := \frac{HC_{\mathsf{avg}}^\star S}{M} + \frac{H^2 C_{\mathsf{avg}}^\star S}{M} + \sqrt{\frac{H^4 S C_{\mathsf{avg}}^\star K}{M}} + \sqrt{\frac{H^2 K S C_{\mathsf{avg}}^\star}{M}} + \frac{H^2 S C_{\mathsf{avg}}^\star}{M}. \tag{43}$$

Then, by invoking the recursion $(H-h+1)$ times, it follows that

$$\begin{aligned}
&\sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s) \left(V_h^\star(s) - V_{t_v,h}(s)\right) \\
&\lesssim (\theta_K + H\tau_1) + (1+\frac{2}{H})^2(\theta_{t_{\phi(K)-1}} + H\tau_1) + (1+\frac{2}{H})^4 \sum_{u=1}^{\phi(K)-2} \tau_u \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s) \left(V_{h+2}^\star(s) - V_{t_u,h+2}(s)\right) \\
&\lesssim (\theta_K + H\tau_1) + (1+\frac{2}{H})^2(\theta_{t_{\phi(K)-1}} + H\tau_1) + \cdots + (1+\frac{2}{H})^{2(H-h+1)}(\theta_{t_{\phi(K)-H+h-1}} + H\tau_1) \\
&\lesssim H\theta_K + H^2\tau_1
\end{aligned} \tag{44}$$

where the second line follows from the fact that $V_{H+1}^\star(s) - V_{k,H+1}(s) = 0$ for any $k \in [K]$, and the last line holds because $\theta_k \le \theta_K$ for any $k \le K$ and $(1+\frac{1}{H})^{2(H-h+1)} \le (1+\frac{2}{H})^{2H} \le e^4$.

Finally, by plugging the above bound into (29), we obtain the bound of the performance gap as follows:

$$\begin{aligned}
V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &\le \frac{1}{K} \max_{h\in[H]} \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_{t_v,h}(s)\right) \\
&\lesssim \frac{1}{K}(H\theta_K + H^2\tau_1) \\
&\lesssim \frac{H^3 S C_{\mathsf{avg}}^\star}{MK} + \sqrt{\frac{H^6 S C_{\mathsf{avg}}^\star}{MK}} + \frac{H^2\tau_1}{K} \\
&\overset{T=HK}{\lesssim} \sqrt{\frac{H^7 S C_{\mathsf{avg}}^\star}{MT}} + \frac{H^4 S C_{\mathsf{avg}}^\star}{MT},
\end{aligned} \tag{45}$$

where the last line holds if $\tau_1 \le \sqrt{\frac{HSC_{\mathsf{avg}}^\star T}{M}}$, and this completes the proof.

C. Technical lemmas

Freedman's inequality. We provide a user-friendly version of Freedman's inequality (Freedman, 1975). See Li et al. (2024a, Theorem 6) for more details.

Theorem C.1 (Li et al. (2024a, Theorem 6))**.** *Consider a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, and let $\mathbb{E}_k$ stand for the expectation conditioned on $\mathcal{F}_k$. Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \le R \qquad \textit{and} \qquad \mathbb{E}_{k-1}\left[X_k\right] = 0 \qquad \textit{for all } k \ge 1$$

for some quantity $R < \infty$. We also define

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

In addition, suppose that $W_n \le \sigma^2$ holds deterministically for some given quantity $\sigma^2 < \infty$. Then for any positive integer $m \ge 1$, with probability at least $1-\delta$ one has

$$|Y_n| \le \sqrt{8 \max\left\{W_n, \frac{\sigma^2}{2^m}\right\} \log \frac{2m}{\delta}} + \frac{4}{3} R \log \frac{2m}{\delta}. \tag{46}$$

We next present a basic analytical result that is useful in the proof.

Lemma C.2. *Consider any sequence $\{x_z\}_{z=1,\cdots,Z}$ where $x_z \ge 1$ for all z and let $X_z = \sum_{z'=1}^{z} x_{z'}$. Then, for any $Z \ge 1$, it follows that*

$$X(Z) = \sum_{z=1}^{Z} \frac{x_z}{X_z} \le 1 + \log X_Z.$$

Proof. For $Z = 1$, $X(1) = \frac{x_1}{x_1} = 1$. For $Z > 1$, suppose the claim holds for $Z-1$. Then, it holds for Z as follows:

$$\begin{aligned} X(Z) = X(Z-1) + \frac{x_Z}{X_Z} &\le 1 + \log X_{Z-1} + 1 - \frac{X_{Z-1}}{X_Z} \\ &\le 1 + \log X_{Z-1} - \log\left(\frac{X_{Z-1}}{X_Z}\right) = 1 + \log X_Z, \end{aligned} \tag{47}$$

where the first inequality follows from the induction hypothesis and $x_Z = X_Z - X_{Z-1}$, the second inequality follows from $\log y \le y - 1$ for any $y > 0$. By induction, this completes the proof. □

Last but not least, we have the following useful properties regarding the parameters introduced in (24c).

Lemma C.3. *For any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $k' \le k \in \mathcal{T}(K)$, where we denote $u = \phi(k)$, and $i \in L^m_{k,h}(s,a)$. Then, it follows that:*

$$\omega^m_{i,k,h}(s,a) \le \frac{2H}{N_{k,h}(s,a) + Hn_{k,h}(s,a)}, \tag{48a}$$

$$\sum_{m=1}^{M} \sum_{j \in L^m_{k,h}(s,a)} \omega^m_{j,k,h}(s,a) \le 1, \tag{48b}$$

$$\sum_{m=1}^{M} \sum_{j \in l^m_{k',h}(s,a)} \omega^m_{j,k,h}(s,a) \le \frac{(H+1)n_{k',h}}{N_{k,h} + Hn_{k,h}}, \tag{48c}$$

$$\sum_{m=1}^{M} \sum_{j \in L^m_{k,h}(s,a)} (\omega^m_{i,k,h}(s,a))^2 \le \frac{2H}{N_{k,h}(s,a) + Hn_{k,h}(s,a)}, \tag{48d}$$

$$\sum_{v \ge u}^{\infty} n_{t_v,h}(s,a) \sum_{m=1}^{M} \sum_{i \in l^m_{k,h}(s,a)} \omega^m_{i,t_v,h}(s,a) \le n_{k,h}(s,a)\left(1 + \frac{1}{H}\right). \tag{48e}$$

Proof. For notation simplicity, we will omit (s,a) for the following proofs. Moreover, $u = \phi(k)$ and $t_u = k$.

Proof of (48a)**.** Recalling the definition of $\omega^m_{i,k,h}$ in (24c) and using the fact that $H \ge 1$,

$$\omega^m_{i,k,h} = \frac{H+1}{N_{k,h} + Hn_{k,h}} \left(\prod_{x=\phi(i)}^{\phi(k)-1} \frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \le \frac{2H}{N_{k,h} + Hn_{k,h}}. \tag{49}$$

Proof of (48b)**.** By rearranging the terms,

$$\sum_{m=1}^{M} \sum_{j \in L^m_{k,h}(s,a)} \omega^m_{j,k,h} = \sum_{v=1}^{\phi(k)} \sum_{m=1}^{M} \sum_{j \in l^m_{t_v,h}} \frac{H+1}{N_{t_v,h} + Hn_{t_v,h}} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}}\right)$$

$$
\begin{aligned}
&= \sum_{v=1}^{\phi(k)} \frac{(H+1)n_{t_v,h}}{N_{t_v,h}+Hn_{t_v,h}} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h}+Hn_{t_x,h}} \right) \\
&= \sum_{v=1}^{\phi(k)} \left(1 - \frac{N_{t_{v-1},h}}{N_{t_v,h}+Hn_{t_v,h}} \right) \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h}+Hn_{t_x,h}} \right) \\
&= \sum_{v=1}^{\phi(k)} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h}+Hn_{t_x,h}} - \prod_{x=v}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h}+Hn_{t_x,h}} \right) \\
&= 1 - \prod_{x=1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h}+Hn_{t_x,h}} \leq 1. \qquad (50)
\end{aligned}
$$

Proof of (48c)**.** Let $v = \phi(k')$, i.e., $k' = t_v$. Similarly to the proof of (48b), by arranging some terms, we obtain the upper bound as follows:

$$
\begin{aligned}
\sum_{m=1}^{M} \sum_{j \in l^m_{k',h}(s,a)} \omega^m_{j,k,h}(s,a) &= \sum_{m=1}^{M} \sum_{j \in l^m_{t_v,h}(s,a)} \frac{H+1}{N_{t_v,h}+Hn_{t_v,h}} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h}+Hn_{t_x,h}} \right) \\
&= \frac{(H+1)n_{t_v,h}}{N_{t_v,h}+Hn_{t_v,h}} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h}+Hn_{t_x,h}} \right) \\
&= \frac{(H+1)n_{t_v,h}}{N_{k,h}+Hn_{k,h}} \left(\prod_{x=v}^{\phi(k)-1} \frac{N_{t_x,h}}{N_{t_x,h}+Hn_{t_x,h}} \right) \\
&\leq \frac{(H+1)n_{k',h}}{N_{k,h}+Hn_{k,h}}. \qquad (51)
\end{aligned}
$$

Proof of (48d)**.** Using the bound in (48a) and (48b),

$$
\sum_{m=1}^{M} \sum_{j \in L^m_{k,h}} (\omega^m_{j,k,h})^2 = \left(\max_{m \in [M], j \in L^m_{k,h}} \omega^m_{j,k,h} \right) \sum_{m=1}^{M} \sum_{j \in L^m_{k,h}} \omega^m_{j,k,h} \leq \max_{m \in [M], j \in L^m_{k,h}} \omega^m_{j,k,h} \leq \frac{2H}{N_{k,h}+Hn_{k,h}}. \qquad (52)
$$

Proof of (48e)**.** Recall that $k = t_u$. Then, reusing the intermediate result derived in (51),

$$
\begin{aligned}
\sum_{v \geq u}^{\infty} n_{t_v,h}(s,a) \sum_{m=1}^{M} \sum_{i \in l^m_{t_u,h}(s,a)} \omega^m_{i,t_v,h}(s,a) &= \sum_{v \geq u}^{\infty} n_{t_v,h} \frac{(H+1)n_{t_u,h}}{N_{t_v,h}+Hn_{t_v,h}} \left(\prod_{x=u}^{v-1} \underbrace{\frac{N_{t_x,h}}{N_{t_x,h}+Hn_{t_x,h}}}_{:=\beta_{x,h}} \right) \\
&= (H+1)n_{t_u,h} \sum_{v \geq u}^{\infty} \frac{n_{t_v,h}}{N_{t_v,h}+Hn_{t_v,h}} \left(\prod_{x=u}^{v-1} \beta_{x,h} \right) \\
&= (H+1)n_{t_u,h} \sum_{v \geq u}^{\infty} \frac{1}{H}(1-\beta_{v,h}) \left(\prod_{x=u}^{v-1} \beta_{x,h} \right) \\
&\leq n_{k,h}(1+\frac{1}{H}). \qquad (53)
\end{aligned}
$$

□

D. Proofs for main results

D.1. Proof of Lemma B.1

For any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ and $k \in \mathcal{T}(K)$, according to the pessimistic aggregation update rule in (13), the estimate error of Q function at the k-th iteration can be written as follows:

$$\begin{aligned} Q_h^\pi(s,a) - Q_{k,h}(s,a) &= Q_h^\pi(s,a) - \left(\sum_{m=1}^M \alpha_{k,h}^m(s,a) Q_{k,h}^m(s,a)\right) + B_{k,h}(s,a) \\ &= \sum_{m=1}^M \alpha_{k,h}^m(s,a)\left(Q_h^\pi(s,a) - Q_{k,h}^m(s,a)\right) + B_{k,h}(s,a), \end{aligned} \tag{54}$$

where the last equality holds by the fact $\sum_{m=1}^M \alpha_{k,h}^m(s,a) = 1$.

Then, invoking the local update rule in (12), for any i such that $(s_{i,h}^m, a_{i,h}^m) = (s,a)$, the local Q-estimate error at each agent m can be written as follows:

$$\begin{aligned} &Q_h^\pi(s,a) - Q_{i,h}^m(s,a) \\ &= (1-\eta_{i,h}^m(s,a))(Q_h^\pi(s,a) - Q_{i-1,h}^m(s,a)) + \eta_{i,h}^m(s,a)(Q_h^\pi(s,a) - r_h(s,a) - P_{i,h}^m V_{i-1,h+1}^m) \\ &= (1-\eta_{i,h}^m(s,a))(Q_h^\pi(s,a) - Q_{i-1,h}^m(s,a)) + \eta_{i,h}^m(s,a)(r_h(s,a) + P_{h,s,a}V_{h+1}^\pi - r_h(s,a) - P_{i,h}^m V_{i-1,h+1}^m) \\ &= (1-\eta_{i,h}^m(s,a))(Q_h^\pi(s,a) - Q_{i-1,h}^m(s,a)) \\ &\quad + \eta_{i,h}^m(s,a)P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) + \eta_{i,h}^m(s,a)(P_{h,s,a} - P_{i,h}^m)V_{i-1,h+1}^m, \end{aligned} \tag{55}$$

where the second line follows from the Bellman's equation. Then, by invoking the relation recursively, the local Q-estimate error at each agent m obeys the following relation:

$$\begin{aligned} Q_h^\pi(s,a) - Q_{k,h}^m(s,a) = &\prod_{i \in l_{k,h}^m(s,a)} (1-\eta_{i,h}^m(s,a))\left(Q_h^\pi(s,a) - Q_{\iota(k),h}(s,a)\right) \\ &+ \sum_{i \in l_{k,h}^m(s,a)} \eta_{i,h}^m(s,a) \prod_{\{j>i: j \in l_{k,h}^m(s,a)\}} (1-\eta_{j,h}^m(s,a)) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) \\ &+ \sum_{i \in l_{k,h}^m(s,a)} \eta_{i,h}^m(s,a) \prod_{\{j>i: j \in l_{k,h}^m(s,a)\}} (1-\eta_{j,h}^m(s,a)) (P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m, \end{aligned} \tag{56}$$

where $l_{k,h}^m(s,a)$ denotes a set of episodes where agent m has visited (s,a) at step h within $(\iota(k), k]$.

By inserting (56) to (54) and letting $v = \phi(k)$, we obtain the following recursive relation for u-th local updates:

$$\begin{aligned} &Q_h^\pi(s,a) - Q_{k,h}(s,a) \\ &= \underbrace{\left(\sum_{m=1}^M \alpha_{k,h}^m(s,a) \prod_{i \in l_{k,h}^m(s,a)} (1-\eta_{i,h}^m(s,a))\right)}_{:=\lambda_{v,h}(s,a)} \left(Q_h^\pi(s,a) - Q_{\iota(k),h}(s,a)\right) + B_{k,h}(s,a) \\ &\quad + \sum_{m=1}^M \sum_{i \in l_{k,h}^m(s,a)} \left(\alpha_{k,h}^m(s,a)\eta_{i,h}^m(s,a) \prod_{\{j>i: j \in l_{k,h}^m(s,a)\}} (1-\eta_{j,h}^m(s,a))\right) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) \\ &\quad + \sum_{m=1}^M \sum_{i \in l_{k,h}^m(s,a)} \left(\alpha_{k,h}^m(s,a)\eta_{i,h}^m(s,a) \prod_{\{j>i: j \in l_{k,h}^m(s,a)\}} (1-\eta_{j,h}^m(s,a))\right) (P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m \\ &= \lambda_{v,h}(s,a)\left(Q_h^\pi(s,a) - Q_{\iota(k),h}(s,a)\right) + B_{k,h}(s,a) \\ &\quad + \frac{(H+1)}{N_{t_v,h}(s,a) + Hn_{t_v,h}(s,a)} \sum_{m=1}^M \sum_{i \in l_{k,h}^m(s,a)} P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) \end{aligned}$$

$$+ \frac{(H+1)}{N_{t_v,h}(s,a) + Hn_{t_v,h}(s,a)} \sum_{m=1}^{M} \sum_{i \in l_{k,h}^m(s,a)} (P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m. \tag{57}$$

Here, the last line holds by invoking the definitions in (16) and (17) and observing with abuse of notation (omit (s,a) when it is clear)

$$\begin{aligned}
&\alpha_{k,h}^m(s,a)\eta_{i,h}^m(s,a) \prod_{\{j>i: j \in l_{k,h}^m(s,a)\}} (1-\eta_{j,h}^m(s,a)) \\
&= \frac{1}{M} \frac{N_{\iota(k),h} + M(H+1)n_{k,h}^m}{N_{k,h} + Hn_{k,h}} \frac{M(H+1)}{N_{\iota(i),h} + M(H+1)n_{i,h}^m} \left(\prod_{j=1}^{n_{k,h}^m - n_{i,h}^m} \left(\frac{N_{\iota(i),h} + M(H+1)(n_{i,h}^m + j - 1)}{N_{\iota(i),h} + M(H+1)(n_{i,h}^m + j)} \right) \right) \\
&= \frac{1}{M} \frac{N_{\iota(k),h} + M(H+1)n_{k,h}^m}{N_{k,h} + Hn_{k,h}} \frac{M(H+1)}{N_{\iota(i),h} + M(H+1)n_{i,h}^m} \frac{N_{\iota(i),h} + M(H+1)n_{i,h}^m}{N_{\iota(i),h} + M(H+1)n_{k,h}^m} \\
&= \frac{(H+1)}{N_{k,h} + Hn_{k,h}} = \frac{(H+1)}{N_{t_v,h} + Hn_{t_v,h}}
\end{aligned} \tag{58}$$

where the last line holds since $\iota(i) = \iota(k)$ for $i \in l_{k,h}^m(s,a)$ and $k \in \mathcal{T}(K)$ leads to $k = t_{\phi(k)} = t_v$.

Then, by invoking the above recursive relation for each aggregation, the Q-estimate error after k episodes is decomposed as follows:

$$\begin{aligned}
&Q_h^\pi(s,a) - Q_{k,h}(s,a) \\
&= \underbrace{\prod_{u=1}^{\phi(k)} \lambda_{u,h}(s,a)}_{:=\omega_{0,k,h}(s,a)} (Q_h^\pi(s,a) - Q_{0,h}(s,a)) + \sum_{u=1}^{\phi(k)} B_{t_u,h}(s,a) \prod_{x=u+1}^{\phi(k)} \lambda_{x,h}(s,a) \\
&\quad + \sum_{u=1}^{\phi(k)} \sum_{m=1}^{M} \sum_{i \in l_{t_u,h}^m(s,a)} \underbrace{\left(\frac{H+1}{N_{t_u,h} + Hn_{t_u,h}} \prod_{x=u+1}^{\phi(k)} \lambda_{x,h}(s,a) \right)}_{:=\omega_{i,k,h}(s,a)} (P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m \\
&\quad + \sum_{u=1}^{\phi(k)} \sum_{m=1}^{M} \sum_{i \in l_{t_u,h}^m(s,a)} \left(\frac{H+1}{N_{t_u,h} + Hn_{t_u,h}} \prod_{x=u+1}^{\phi(k)} \lambda_{x,h}(s,a) \right) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) \\
&= \omega_{0,k,h}(s,a)(Q_h^\pi(s,a) - Q_{0,h}(s,a)) \\
&\quad + \sum_{m=1}^{M} \sum_{i \in L_{k,h}^m(s,a)} \omega_{i,k,h}^m(s,a)(P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m \\
&\quad + \sum_{u=1}^{\phi(k)} B_{t_u,h}(s,a) \prod_{x=u+1}^{\phi(k)} \lambda_{x,h}(s,a) \\
&\quad + \sum_{m=1}^{M} \sum_{i \in L_{k,h}^m(s,a)} \omega_{i,k,h}^m(s,a) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m).
\end{aligned} \tag{59}$$

Here, $\lambda_{u,h}(s,a)$, $\omega_{0,k,h}(s,a)$, and $\omega_{i,k,h}(s,a)$ can be simply written as described in (24a), (24b), and (24c), respectively, which will be proved momentarily. For notational simplicity, we omit (s,a) in the derivations.

Proof of (24a)**.** Consider $k = t_v$. First, consider a case that $N_{\iota(k),h} = 0$. If $n_{k,h} = 0$, $\lambda_{v,h} = \sum_{m=1}^M \alpha_{k,h}^m = 1$. Otherwise, if $n_{k,h} > 0$, where there exists at least one agent $m \in [M]$ that visits the state-action at least once until k-th episode, it follows that

$$\lambda_{v,h} = \sum_{m=1}^{M} \frac{1}{M} \frac{(H+1)Mn_{k,h}^m}{(H+1)n_{k,h}} \prod_{j=1}^{n_{k,h}^m} \left(\frac{M(H+1)(j-1)}{M(H+1)j} \right)$$

$$= \sum_{m\in[M]:n^m_{k,h}=0} \underbrace{\frac{n^m_{k,h}}{n_{k,h}}}_{=0} + \sum_{m\in[M]:n^m_{k,h}>0} \frac{n^m_{k,h}}{n_{k,h}} \underbrace{\prod_{j=1}^{n^m_{k,h}} \left(\frac{(H+1)(j-1)}{(H+1)j}\right)}_{=0} = 0. \tag{60}$$

On the other hand, when $N_{\iota(k),h} > 0$,

$$\begin{aligned}\lambda_{v,h} &= \sum_{m=1}^{M} \frac{1}{M} \frac{N_{\iota(k),h} + M(H+1)n^m_{k,h}}{N_{\iota(k),h} + (H+1)n_{k,h}} \prod_{j=1}^{n^m_{k,h}} \left(\frac{N_{\iota(k),h} + M(H+1)(j-1)}{N_{\iota(k),h} + M(H+1)j}\right) \\ &= \sum_{m=1}^{M} \frac{1}{M} \frac{N_{\iota(k),h} + M(H+1)n^m_{k,h}}{N_{\iota(k),h} + (H+1)n_{k,h}} \frac{N_{\iota(k),h}}{N_{\iota(k),h} + M(H+1)n^m_{k,h}} = \frac{N_{\iota(k),h}}{N_{k,h} + Hn_{k,h}}. \end{aligned} \tag{61}$$

Proof of (24b)**.** According to (24a), if $N_{k,h}(s,a) = 0$, then $\lambda_{u,h}(s,a) = 1$ for all $1 \le u \le \phi(k)$. Thus, $\omega_{0,k,h}(s,a) = 1$. Otherwise, let the epsiode when (s,a) is visited at step h by any of the agents for the first time be j. Then, $\lambda_{\phi(j),h} = 0$ because $N_{\iota(j),h}(s,a) = 0$. Thus, if $N_{k,h}(s,a) > 0$, it always holds that $\omega_{0,k,h}(s,a) = \prod_{u=1}^{\phi(k)} \lambda_{u,h}(s,a) = 0$.

Proof of (24c)**.** For i such that $\phi(i) = u$, by rearranging terms and applying (24a),

$$\begin{aligned}\omega^m_{i,k,h} &= \frac{(H+1)}{N_{t_u,h} + Hn_{t_u,h}} \left(\prod_{x=u+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \\ &= \frac{H+1}{N_{k,h} + Hn_{k,h}} \left(\prod_{x=u}^{\phi(k)-1} \frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}}\right). \end{aligned} \tag{62}$$

D.2. Proof of Lemma B.2

Consider any given $\delta \in (0,1)$ and $(k,s,a,h) \in [K] \times \mathcal{S} \times \mathcal{A} \times [H]$. Note that $N^m_{k,h}(s,a) \sim \mathsf{Binomial}(k, d^m_h(s,a))$ for all $m \in [M]$. Then recall the definition of $N_{k,h}(s,a)$ in Section 3.2, we can view $N_{k,h}(s,a) = \sum_{m=1}^{M} N^m_{k,h}(s,a)$ as a sum of kM independent Bernoulli variables with expectation $\nu := \mathbb{E}[N_{k,h}(s,a)] = kMd^{\mathsf{avg}}_h(s,a)$. Therefore, applying Chernoff bound (see Mitzenmacher & Upfal (2005, Theorem 4.4)) yields:

$$\forall t \in [0,1] \quad : \quad \mathbb{P}(\left|N^m_{k,h}(s,a) - \nu\right| \ge \nu t) \le \exp\left(-c_1 \nu t^2\right), \tag{63a}$$

$$\forall t \ge 1 \quad : \quad \mathbb{P}(N^m_{k,h}(s,a) - \nu \ge t\nu) \le \exp\left(-c_1 \nu t\right), \tag{63b}$$

for some universal constant $c_1 > 0$.

Armed with above facts and notations, now we are ready to prove (26). First, applying (63a) with $t = \frac{1}{2}$, we arrive at:

$$\mathbb{P}(\left|N^m_{k,h}(s,a) - \nu\right| \ge \frac{\nu}{2} \le \exp(-\frac{c_1 \nu}{4}) \le \delta, \tag{64}$$

where the last line follows from the condition that $\nu = kMd^{\mathsf{avg}}_h(s,a) \ge \frac{4}{c_1} \log\left(\frac{1}{\delta}\right)$.

To continue, when $\nu = kMd^{\mathsf{avg}}_h(s,a) \le \frac{4}{c_1} \log\left(1/\delta\right)$, applying (63b) with $t = \frac{4\log(1/\delta)}{\nu c_1} \ge 1$ gives:

$$\mathbb{P}(N^m_{k,h}(s,a) - \nu \ge \frac{4\log(1/\delta)}{c_1}) \le \exp(-4\log(1/\delta)) \le \delta. \tag{65}$$

Summing up (64) and (65) and taking the union bound over $(k,s,a,h) \in [K] \times \mathcal{S} \times \mathcal{A} \times [H]$ complete the proof by showing that:

$$\text{when } k \ge \frac{4\log(\frac{|\mathcal{S}||\mathcal{A}|KH}{\delta})}{c_1 M d^{\mathsf{avg}}_h} \quad : \quad \frac{kMd^{\mathsf{avg}}_h}{2} = \frac{\nu}{2} \le N^m_{k,h}(s,a) \le \frac{3\nu}{2} \le 2kMd^{\mathsf{avg}}_h \tag{66}$$

$$\text{when } k \le \frac{4\log(\frac{|\mathcal{S}||\mathcal{A}|KH}{\delta})}{c_1 M d^{\mathsf{avg}}_h} \quad : \quad N^m_{k,h}(s,a) \le \frac{8}{c_1} \log(\frac{|\mathcal{S}||\mathcal{A}|KH}{\delta}) \tag{67}$$

holds with probability at least $1 - 2\delta$.

D.3. Proof of Lemma B.3

D.3.1. PROOF OF (28a).

Noticing that the (28a) involves two terms of interest, we start from the first one $D_2(s,a,k,h)$. For any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and any $k \in \mathcal{T}(K)$, we can rewrite $D_2(s,a,k,h)$ as

$$D_2(s,a,k,h) = \sum_{i=1}^{k}\sum_{m=1}^{M} X_{i,k,h}^m(s,a), \tag{68}$$

where $X_{i,k,h}^m(s,a) = \omega_{i,k,h}^m(s,a)(P_{h,s,a} - P_{i,h}^m)V_{i-1,h+1}^m \mathbb{I}\{(s_{i,h}^m, a_{i,h}^m) = (s,a)\}$. To further control $\sum_{i=1}^k\sum_{m=1}^M X_{i,k,h}^m(s,a)$, we first introduce the following Lemma D.1, whose proof is provided in Appendix D.4, with $N_{k,h}(s,a) = N$.

Lemma D.1. *For any $(k,s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and $N \in [1, MK]$, let*

$$\widetilde{X}_{i,k,h}^m(s,a;N) = \widetilde{\omega}_{i,k,h}^m(s,a;N)(P_{h,s,a} - P_{i,h}^m)V_{i-1,h+1}^m \mathbb{I}\{(s_{i,h}^m, a_{i,h}^m) = (s,a)\}, \tag{69}$$

where

$$\widetilde{\omega}_{i,k,h}^m(s,a;N) := \frac{H+1}{N + Hn_{k,h}(s,a)}\left(\prod_{x=\phi(i)}^{\phi(k)-1}\frac{N_{t_x,h}(s,a)}{N_{t_x,h}(s,a) + Hn_{t_x,h}(s,a)}\right) I_{i,h}^m(s,a;N), \tag{70}$$

and $I_{i,h}^m(s,a;N) := \mathbb{I}\{\sum_{m'=1}^M N_{i-1,h}^{m'}(s,a) + \sum_{m'=1}^m \mathbb{I}\{(s_{i,h}^{m'}, a_{i,h}^{m'}) = (s,a)\} \le N\}$. Then, for any $\delta \in (0,1)$, the following holds:

$$\left|\sum_{i=1}^{k}\sum_{m=1}^{M}\widetilde{X}_{i,k,h}^m(s,a;N)\right| \le \sqrt{\frac{81H^4\zeta_1^2}{N}} \tag{71}$$

at least with probability $1-\delta$, where we denote $\zeta_1 = \log\left(\frac{|\mathcal{S}||\mathcal{A}|MK^2H}{\delta}\right)$.

Armed with above lemma, for any $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]$ where $k \in \mathcal{T}(K)$, the following holds:

$$\text{when } N_{k,h}(s,a) > 0: \quad |D_2(s,a,k,h)| \le \left|\sum_{i=1}^{k}\sum_{m=1}^{M}\widetilde{X}_{i,k,h}^m(s,a;N_{k,h}(s,a))\right| \le \sqrt{\frac{81H^4\zeta_1^2}{N_{k,h}(s,a)}} \tag{72}$$

with probability at least $1-\delta$. As it is obvious that $D_2(s,a,k,h) = 0$ when $N_{k,h}(s,a) = 0$ from the definition of $D_2(s,a,k,h)$, we arrive at

$$|D_2(s,a,k,h)| \le \left|\sum_{i=1}^{k}\sum_{m=1}^{M}\widetilde{X}_{i,k,h}^m(s,a;N_{k,h}(s,a))\right| \le \sqrt{\frac{81H^4\zeta_1^2}{N_{k,h}(s,a)}}. \tag{73}$$

Finally, combining the results for $D_2(s,a,k,h)$ (cf. (73)) and $D_3(s,a,k,h)$ (cf. (36) in Lemma B.4), we conclude that for any $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]$ with $k \in \mathcal{T}(K)$, it holds with probability at least $1-\delta$ that

$$|D_2(s,a,k,h)| \le \sqrt{\frac{81H^4\zeta_1^2}{N_{k,h}(s,a)}} = \sqrt{\frac{c_B\zeta_1^2H^4}{N_{k,h}(s,a)}} \le D_3(s,a,k,h). \tag{74}$$

D.3.2. PROOF OF (28b) AND (28c).

For all $(h,s,a,k) \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{T}(K)$, it is clear that $Q_h^{\pi_k}(s,a) \le Q_h^\star(s,a)$ and $V_h^{\pi_k}(s) \le V_h^\star(s)$ due to the definition. So it suffices to show that

$$Q_{k,h}(s,a) \le Q_h^{\pi_k}(s,a) \quad \text{and} \quad V_{k,h}(s) \le V_h^{\pi_k}(s)$$

for all $(h,s,a,k) \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{T}(K)$, which we will prove by an induction argument as below.

- **Base case.** When $h = H+1$, for all $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{T}(K)$, the relation always holds since $Q_{k,H+1}(s,a) = 0 \leq Q^{\pi_k}_{H+1}(s,a)$ and $V_{k,H+1}(s) = 0 \leq V^{\pi_k}_{H+1}(s)$ according to the definition of $Q_{k,H+1}$ and $V_{k,H+1}$, respectively.
- **Induction.** When $h \in [H]$, suppose the relation holds for $h+1$, i.e., $Q_{k,h+1}(s,a) \leq Q^{\pi_k}_{h+1}(s,a)$ and $V_{k,h+1}(s) \leq V^{\pi_k}_{h+1}(s)$ for all $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{T}(K)$. First, we will verify Q-estimates at step h are pessimistic. For any $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{T}(K)$, applying Lemma B.1,

$$
\begin{aligned}
Q^{\pi_k}_h(s,a) - Q_{k,h}(s,a) = & \underbrace{\omega_{0,k,h}(s,a)(Q^{\pi_k}_h(s,a) - Q_{0,h}(s,a))}_{D^{\pi_k}_1(s,a,k,h)} \\
& + \underbrace{\sum_{m=1}^{M} \sum_{i \in L^m_{k,h}(s,a)} \omega^m_{i,k,h}(s,a)(P_{h,s,a} - P^m_{i,h}) V^m_{i-1,h+1}}_{D_2(s,a,k,h)} \\
& + \underbrace{\sum_{u=1}^{\phi(k)} B_{t_u,h}(s,a) \prod_{u'=u+1}^{\phi(k)} \lambda_{u',h}(s,a)}_{D_3(s,a,k,h)} \\
& + \underbrace{\sum_{m=1}^{M} \sum_{i \in L^m_{k,h}(s,a)} \omega^m_{i,k,h}(s,a) P_{h,s,a}(V^{\pi_k}_{h+1} - V_{\iota(i),h+1})}_{D^{\pi_k}_4(s,a,k,h)} .
\end{aligned} \tag{75}
$$

Then we control the above four terms one at a time. Here, $D^{\pi_k}_1(s,a,k,h) \geq 0$ since $Q^{\pi_k}_h(s,a) \geq Q_{0,h}(s,a) = 0$. In addition, according to the fact (28a) in Lemma B.3, $|D_2(s,a,k,h)| \leq D_3(s,a,k,h)$. And it is clear that $D_4 \geq 0$ due to

$$
V^{\pi_k}_{h+1} \geq V_{k,h+1} \geq V_{\iota(i),h+1}, \tag{76}
$$

where the first inequality holds by the induction assumption, and the last inequality arises from the monotonicity guarantee of the global update in (14). Therefore, it is clear that for any $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{T}(K)$, Q-estimates at step h are pessimistic, i.e.,

$$
Q^{\pi_k}_h(s,a) - Q_{k,h}(s,a) \geq 0. \tag{77}
$$

Next, to show that value estimates at step h are pessimistic, recalling the global update in (14),

$$
\begin{aligned}
V^{\pi_k}_h(s) - V_{k,h}(s) &= Q^{\pi_k}_h(s, \pi_{k,h}(s)) - \max\{\max_a Q_{k,h}(s,a), V_{\iota(k),h}(s)\} \\
&= Q^{\pi_k}_h(s, \pi_{k,h}(s)) - \max_a Q_{k_0,h}(s,a) \\
&= Q^{\pi_k}_h(s, \pi_{k_0,h}(s)) - Q_{k_0,h}(s, \pi_{k_0,h}(s)) \geq 0,
\end{aligned} \tag{78}
$$

where k_0 denotes the most recent episode satisfying $V_{k,h}(s) = \max_a Q_{k_0,h}(s,a)$ and $k \geq k_0 \in \mathcal{T}(K)$, and the last inequality holds because $\pi_{k,h}(s) = \pi_{k_0,h}(s)$ and $Q^{\pi_k}_h(s,a) - Q_{k_0,h}(s,a) \geq 0$ can be similarly verified using (75) and (76) for k_0. Now, we verify that $Q^{\pi_k}_h(s,a) \geq Q_{k,h}(s,a)$ and $V^{\pi_k}_h(s) \geq V_{k,h}(s)$ holds at step h for any $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{T}(K)$, and this directly completes the induction argument.

D.4. Proof of Lemma D.1

To begin with, for any time step $h \in [H]$, we denote the expectation conditioned on the trajectories $j \leq i$ of all agent as

$$
\forall (i,m) \in [k] \times [M]: \quad \mathbb{E}_{(i,m)}[\cdot] = \mathbb{E}\big[\, \cdot \mid \{s^{m'}_{j,h}, a^{m'}_{j,h}, V^m_{j,h+1}\}_{j<i, m' \in [M]}, \ \{s^{m'}_{i,h}, a^{m'}_{i,h}\}_{m' \leq m}\big]. \tag{79}
$$

Armed with this notation, fixing N, it is easily verified that $\mathbb{E}_{(i,m)}[\widetilde{X}^m_{i,k}(s,a;N)] = 0$ since then $V^m_{i-1,h+1}$ can be regarded as fixed and $(P_{h,s,a} - P^m_{i,h})$ is independent from $\widetilde{\omega}^m_{i,k,h}(s,a;N)$.

Consequently, we can apply Freedman's inequality (see the user-friendly version of Freedman's inequality provided in Theorem C.1) and control the term of interest for any $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$ and $N \in [1, MK]$ as below:

$$\sum_{i=1}^{k}\sum_{m=1}^{M} \widetilde{X}^m_{i,k,h}(s,a;N) \overset{\text{(i)}}{\leq} \sqrt{8B_1\zeta_1} + \frac{4}{3}B_2\zeta_1 \overset{\text{(ii)}}{\leq} \sqrt{\frac{32H^4\zeta_1}{N}} + \frac{3H^2\zeta_1}{N} \leq \sqrt{\frac{81H^4\zeta_1^2}{N}} \tag{80}$$

at least with probability $1-\delta$. Here, (i) and (ii) arises from the following definition and facts about B_1 and B_2:

$$B_1 \coloneqq \sum_{i=1}^{k}\sum_{m=1}^{M} \mathbb{E}_{(i,m)}\left[\left(\widetilde{X}^m_{i,k,h}(s,a;N)\right)^2\right] \leq \frac{4H^4}{N}, \tag{81}$$

$$B_2 \coloneqq \max_{(i,m)\in[k]\times[M]} \left|\widetilde{X}^m_{i,k,h}(s,a;N)\right| \leq \frac{2H^2}{N} \tag{82}$$

where the proofs of (82) and (81) are provided as below, respectively.

Proof of (81)**.** In view of that the events happen at any time step h is independent from the transitions in later time steps including $P^m_{i,h}$, we have $\widetilde{\omega}^m_{i,k,h}(s,a;N)$ is independent from $(P_{h,s,a} - P^m_{i,h})V^m_{i-1,h+1}$, which yields

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{m=1}^{M} \mathbb{E}_{(i,m)}[(\widetilde{X}^m_{i,k,h}(s,a;N))^2] &= \sum_{i=1}^{k}\sum_{m=1}^{M} \mathbb{E}_{(i,m)}[(\widetilde{\omega}^m_{i,k,h}(s,a;N))^2]\mathsf{Var}_{P_{h,s,a}}(V^m_{i-1,h+1}) \\
&\leq H^2 \sum_{i=1}^{k}\sum_{m=1}^{M} \mathbb{E}_{(i,m)}[(\widetilde{\omega}^m_{i,k,h}(s,a;N))^2] \\
&\leq H^2 N \left(\frac{2H}{N}\right)^2 \\
&= \frac{4H^4}{N}.
\end{aligned} \tag{83}$$

where the penultimate inequality holds by the fact that $|\widetilde{\omega}^m_{i,k,h}(s,a;N)| \leq \frac{2H}{N}$.

Proof of (82)**.** For any $(i,m,h) \in [k]\times[M]\times[H]$ and fixed $N \in [1, MK]$, it is observed that

$$\begin{aligned}
\left|\widetilde{X}^m_{i,k,h}(s,a;N)\right| &= \left|\widetilde{\omega}^m_{i,k,h}(s,a;N)(P_{h,s,a} - P^m_{i,h})V^m_{i-1,h+1}\mathbb{I}\{(s^m_{i,h}, a^m_{i,h}) = (s,a)\}\right| \\
&\leq |\widetilde{\omega}^m_{i,k,h}(s,a;N)|\|P_{h,s,a} - P^m_{i,h}\|_1\|V^m_{i-1,h+1}\|_\infty \leq \frac{2H^2}{N}
\end{aligned} \tag{84}$$

where the last inequality follows from the facts $\|V^m_{i-1,h+1}\|_\infty \leq H$, $\|P_{h,s,a} - P^m_{i,h}\|_1 \leq 1$, and $|\widetilde{\omega}^m_{i,k,h}(s,a;N)| \leq \frac{H+1}{N} \leq \frac{2H}{N}$.

D.5. Proof of Lemma B.4

With slightly abuse of notation, we will omit (s,a) from some notations when it is clear for simplicity throughout this section. Recall the definition of $D_3(s,a,k,h)$ in (23) and the global penalty defined in (15). When $N_{k,h}(s,a) = 0$, the global penalties are all 0, which yields $D_3(s,a,k,h) = 0$. Therefore, now it suffices to focus on the cases when $N_{k,h}(s,a) > 0$ and show that for $c_B = 81$, $c_u = 4$ and $c_l = 1$,

$$D_3(s,a,k,h) = \sum_{u=1}^{\phi(k)} B_{t_u,h}(s,a) \prod_{u'=u+1}^{\phi(k)} \lambda_{u',h}(s,a) \in \left[\sqrt{\frac{c_l c_B \zeta_1^2 H^4}{N_{k,h}(s,a)}}, \sqrt{\frac{c_u c_B \zeta_1^2 H^4}{N_{k,h}(s,a)}}\right]. \tag{85}$$

Towards this, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we consider a more general term as below: for any integer $z \geq 1$,

$$\sum_{u=1}^{z} B_{t_u,h} \prod_{u'=u+1}^{z} \lambda_{u',h} = \sum_{u=1}^{z} \frac{(H+1)n_{t_u,h}}{N_{k,h} + Hn_{t_u,h}} \sqrt{\frac{c_B\zeta_1^2 H^4}{N_{t_u,h}}} \prod_{u'=u+1}^{z} \lambda_{u',h}$$

$$= \sqrt{c_B \zeta_1^2 H^4} \sum_{u=1}^{z} \sqrt{\frac{1}{N_{t_u,h}}} (1 - \lambda_{u,h}) \prod_{u'=u+1}^{z} \lambda_{u',h}$$
$$= \sqrt{c_B \zeta_1^2 H^4} Y(z) \tag{86}$$

where the penultimate equality follows from $\frac{(H+1)n_{t_u,h}(s,a)}{N_{t_u,h}+Hn_{t_u,h}(s,a)} = (1 - \lambda_{u,h}(s,a))$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and the last equality arises by defining

$$Y(z) := \sum_{u=1}^{z} \sqrt{\frac{1}{N_{t_u,h}}} (1 - \lambda_{u,h}) \prod_{u'=u+1}^{z} \lambda_{u',h}. \tag{87}$$

As a result, to show (85), it suffices to verify that

$$Y(z) \in \left[\sqrt{\frac{c_l}{N_{t_z,h}(s,a)}}, \sqrt{\frac{c_u}{N_{t_z,h}(s,a)}} \right]. \tag{88}$$

Proof of (88) **by induction.** We will verify (88) by a induction argument. To begin with, for the basic case $z = 1$, it is easily verified that

$$Y(1) = \begin{cases} \sqrt{\frac{1}{N_{t_1,h}}} & \text{if } n_{t_1,h} > 0 \\ 0 & \text{if } n_{t_1,h} = 0, \end{cases} \tag{89}$$

since when $n_{t_1,h} > 0$ we have $\lambda_{1,h}(s,a) = 0$, and otherwise $\lambda_{1,h}(s,a) = 1$. Then suppose (88) holds for $z - 1$, namely,

$$Y(z-1) \in \left[\sqrt{\frac{c_l}{N_{t_{z-1},h}}}, \sqrt{\frac{c_u}{N_{t_{z-1},h}}} \right] \tag{90}$$

we hope to show (88) holds for z. Towards this, we first show the upper bound in (88) holds for z as follows:

$$\begin{aligned}
Y(z) &= Y(z-1)\lambda_{z,h} + \sqrt{\frac{1}{N_{t_z,h}}}(1 - \lambda_{z,h}) \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{c_u}{N_{t_{z-1},h}}} \frac{N_{t_{z-1},h}}{N_{t_z,h} + Hn_{t_z,h}} + \sqrt{\frac{1}{N_{t_z,h}}} \frac{(H+1)n_{t_z,h}}{N_{t_z,h} + Hn_{t_z,h}} \\
&\leq \sqrt{\frac{c_u}{N_{t_z,h}}} \sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h} + Hn_{t_z,h}}} + \sqrt{\frac{1}{N_{t_z,h}}} \frac{(H+1)n_{t_z,h}}{N_{t_z,h} + Hn_{t_z,h}} \\
&= \sqrt{\frac{c_u}{N_{t_z,h}}} \left(\sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h} + Hn_{t_z,h}}} + \sqrt{\frac{1}{c_u}} \frac{(H+1)n_{t_z,h}}{N_{t_z,h} + Hn_{t_z,h}} \right) \\
&= \sqrt{\frac{c_u}{N_{t_z,h}}} \left(\sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h} + Hn_{t_z,h}}} + \sqrt{\frac{1}{c_u}} \left(1 - \sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h} + Hn_{t_z,h}}} \right) \left(1 + \sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h} + Hn_{t_z,h}}} \right) \right) \\
&\leq \sqrt{\frac{c_u}{N_{t_z,h}}},
\end{aligned} \tag{91}$$

where (i) follows from the induction assumption and $\frac{(H+1)n_{t_z,h}(s,a)}{N_{t_z,h}+Hn_{t_z,h}(s,a)} = (1 - \lambda_{z,h}(s,a))$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, the penultimate equality holds by $1 - \frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}} = \frac{N_{t_z,h} - N_{t_{z-1},h} + Hn_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}} = \frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}}$, and the last inequality arises from $\sqrt{\frac{1}{c_u}} \left(1 + \sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}}} \right) \leq 1$ as long as $c_u \geq 4$.

Analogous to (91), the lower bound of $Y(z)$ is derived as below:

$$Y(z) = Y(z-1)\lambda_{z,h} + \sqrt{\frac{1}{N_{t_z,h}}}(1 - \lambda_{z,h})$$

$$
\begin{aligned}
&\geq \sqrt{\frac{c_l}{N_{t_{z-1},h}}}\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}} + \sqrt{\frac{1}{N_{t_z,h}}}\frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}} \\
&\geq \sqrt{\frac{c_l}{N_{t_z,h}}}\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}} + \sqrt{\frac{1}{N_{t_z,h}}}\frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}} \\
&\geq \sqrt{\frac{c_l}{N_{t_z,h}}}, \qquad (92)
\end{aligned}
$$

where the first inequality follows from the induction assumption and $\frac{(H+1)n_{t_z,h}(s,a)}{N_{t_z,h}+Hn_{t_z,h}(s,a)} = (1-\lambda_{z,h}(s,a))$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and the last equality holds when $1 \geq c_l$. Finally, by induction arguments, (88) holds for any $z \in \phi(K)$, and this completes the proof.

D.6. Proof of Lemma B.5

Recall the definition of $D_{4,h}$ (see (33) and (23)), $D_{4,h}$ can be rewritten as follows:

$$
\begin{aligned}
D_{4,h} &\sum_{v=1}^{\phi(K)} \tau_v \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{m=1}^{M} \sum_{i\in L^m_{t_v,h}(s,a)} \omega^m_{i,t_v,h}(s,a) P_{h,s,a}(V^\star_{h+1} - V_{\iota(i),h+1}) \\
&\overset{(i)}{=} \sum_{v=1}^{\phi(K)} \tau_v \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1}) \underbrace{\sum_{m=1}^{M}\left(\sum_{i\in l^m_{t_u,h}(s,a)} \omega^m_{i,t_v,h}(s,a)\right)}_{:=\psi_{u,v,h}(s,a)} \\
&= \sum_{v=1}^{\phi(K)} \tau_v \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) \\
&= \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m'=1}^{M} \sum_{v=1}^{\phi(K)} \sum_{t_{v-1}<j\leq t_v} d_h^{\pi^\star}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) \qquad (93)
\end{aligned}
$$

where (i) holds by rewriting the sum as $\sum_{i\in L^m_{t_v,h}(s,a)} = \sum_{u=1}^{v}\sum_{i\in l^m_{t_u,h}(s,a)}$ and the last equality holds by the definition of τ_ν.

To further control (93), we introduce the following lemma that bound the expectation form (93) by an empirical version; the proof is postponed to Appendix D.7.

Lemma D.2. *Consider any $\delta \in (0,1)$. For any $h \in [H]$, the following holds:*

$$
\begin{aligned}
&\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m'=1}^{M} \sum_{v=1}^{\phi(K)} \sum_{t_{v-1}<j\leq t_v} d_h^{\pi^\star}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) \\
&\lesssim \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} n_{t_v,h}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) + M\sigma_{\mathsf{aux},1} \qquad (94)
\end{aligned}
$$

at least with probability $1-\delta$, where

$$
\sigma_{\mathsf{aux},1} \lesssim \sqrt{\frac{H^2KSC^\star_{\mathsf{avg}}}{M}} + \frac{H^2SC^\star_{\mathsf{avg}}}{M} \qquad (95)
$$

Then, applying concentration bounds, D_4 is bounded as follows:

$$
D_4 \overset{(i)}{\lesssim} \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \sum_{u=1}^{v} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} n_{t_v,h}(s,a) P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) + \sigma_{\mathsf{aux},1}
$$

$$= \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{u=1}^{\phi(K)} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1}) \sum_{v=u}^{\phi(K)} n_{t_v,h}(s,a)\psi_{u,v,h}(s,a) + \sigma_{\mathsf{aux},1}$$

$$\overset{(ii)}{\leq} \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{u=1}^{\phi(K)} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1}) n_{t_u,h}(s,a)(1+\frac{1}{H}) + \sigma_{\mathsf{aux},1} \tag{96}$$

where (i) follows from Lemma D.2, and (ii) holds because

$$\sum_{v\geq u}^{\infty} n_{t_v,h}(s,a) \sum_{m=1}^{M} \sum_{i\in l_{t_u,h}^m(s,a)} \omega_{i,t_v,h}^m(s,a) \leq n_{t_u,h}(s,a)(1+\frac{1}{H}) \tag{97}$$

according (48e) in Lemma C.3.

To continue, we introduce the following lemma that transfer the distribution at time step h to the distribution of $h+1$; the proof is provided in Appendix D.8.

Lemma D.3. *Consider any $\delta \in (0,1)$. For any $h \in [H]$, the following holds:*

$$\sum_{u=1}^{\phi(K)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_{t_u,h}(s,a)}{M d_h^{\mathsf{avg}}(s,a)} d_h^{\pi^\star}(s,a) P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1})$$
$$\lesssim \sum_{u=1}^{\phi(K)} \tau_u \sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)(V_{h+1}^\star(s) - V_{t_{u-1},h+1}(s)) + \sigma_{\mathsf{aux},2} \tag{98}$$

at least with probability $1-\delta$, where

$$\sigma_{\mathsf{aux},2} = \sqrt{\frac{H^2 K S C_{\mathsf{avg}}^\star}{M}} + \frac{H S C_{\mathsf{avg}}^\star}{M}. \tag{99}$$

Armed with above lemma, rearranging the terms in (96) and applying Lemma D.3,

$$D_4 \lesssim (1+\frac{1}{H}) \sum_{u=1}^{\phi(K)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_{t_u,h}(s,a)}{M d_h^{\mathsf{avg}}(s,a)} d_h^{\pi^\star}(s,a) P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1}) + \sigma_{\mathsf{aux},1}$$
$$\lesssim (1+\frac{1}{H}) \sum_{u=1}^{\phi(K)} \tau_u \sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)(V_{h+1}^\star(s) - V_{t_{u-1},h+1}(s)) + \underbrace{\sigma_{\mathsf{aux},1} + \sigma_{\mathsf{aux},2}}_{=:\sigma_{\mathsf{aux}}}, \tag{100}$$

and this completes the proof.

D.7. Proof of Lemma D.2

Consider any given $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $v \in [1, \phi(K)]$. Before proceeding, we introduce some notations and auxiliary terms. Let

$$G_{v,h}(s,a) := \sum_{u=1}^{v} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a). \tag{101}$$

Then, for any $t_{v-1} < j \leq t_v$, we introduce the following auxiliary variables:

$$Y_{j,h}^m := \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^{\mathsf{avg}}(s,a) - \mathbb{I}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}\right) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} G_{v,h}(s,a) \tag{102}$$

$$\widetilde{Y}_{j,h}^m := \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^m(s,a) - \mathbb{I}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}\right) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} \widetilde{G}_{v,h}^{-j,m}(s,a), \tag{103}$$

where we define

$$\widetilde{G}_{v,h}^{-j,m}(s,a) := \begin{cases} \widetilde{\psi}_{v,v,h}^{-j,m}(s,a)P_{h,s,a}(V_{h+1}^{\star} - V_{t_{v-1},h+1}) + (1-\widetilde{\psi}_{v,v,h}^{-j,m}(s,a))G_{v-1,h}(s,a) & \text{if } v>1 \\ P_{h,s,a}(V_{h+1}^{\star} - V_{0,h+1}) & \text{if } v=1 \end{cases} \tag{104}$$

and

$$\begin{aligned} \widetilde{\psi}_{v,v,h}^{-j,m}(s,a) &:= \frac{(H+1)(n_{t_v,h}(s,a) - \mathbb{I}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\})}{N_{t_{v-1},h}(s,a) + (H+1)(n_{t_v,h}(s,a) - \mathbb{I}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\})} \\ &= \frac{(H+1)(\sum_{(m',j')\in[M]\times(t_{v-1},t_v]\setminus\{(j,m)\}} \mathbb{I}\{(s,a)=(s_{j',h}^{m'}, a_{j',h}^{m'})\})}{N_{t_{v-1},h}(s,a) + (H+1)(\sum_{(m',j')\in[M]\times(t_{v-1},t_v]\setminus\{(j,m)\}} \mathbb{I}\{(s,a)=(s_{j',h}^{m'}, a_{j',h}^{m'})\})}. \end{aligned} \tag{105}$$

We replaced $G_{v,h}(s,a)$ with an approximate $\widetilde{G}_{v,h}^{-j,m}(s,a)$, where the visits of agent m on (s,a) at the j-th episode are masked regardless of the actual visits of agent m on (s,a). The approximate is carefully designed to remove the dependency on the event $\mathbb{I}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\}$ from $G_{v,h}(s,a)$ while maintaining close distance to the original value $G_{v,h}(s,a)$.

Before continuing, we introduce some useful properties of above defined auxiliary terms whose proofs are provided in Section D.7.1:for any $v \in [\phi(K)]$,

$$G_{v,h}(s,a) = \begin{cases} \psi_{v,v,h}(s,a)P_{h,s,a}(V_{h+1}^{\star} - V_{t_{v-1},h+1}) + (1-\psi_{v,v,h}(s,a))G_{v-1,h}(s,a) & \text{if } v>1 \\ P_{h,s,a}(V_{h+1}^{\star} - V_{0,h+1}) & \text{if } v=1 \end{cases}, \tag{106a}$$

$$0 \le \widetilde{G}_{v,h}^{-j,m}(s,a),\ G_{v,h}(s,a) \le H, \tag{106b}$$

$$|\widetilde{G}_{v,h}^{-j,m}(s,a) - G_{v,h}(s,a)| \le \min\left\{H, \frac{2H^2}{N_{t_v,h}(s,a)}\right\}. \tag{106c}$$

Now, we are ready to prove (94). Towards this, we first observe that putting the first term in the right hand side of (94) to the left hand side yields

$$\begin{aligned} &\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \left(\sum_{m=1}^{M} \sum_{t_{v-1}<j\le t_v} d_h^{\pi^\star}(s,a) - \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} n_{t_v,h}(s,a) \right) \sum_{u=1}^{v} P_{h,s,a}(V_{h+1}^{\star} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) \\ &\overset{\text{(i)}}{=} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \left(\sum_{m=1}^{M} \sum_{t_{v-1}<j\le t_v} d_h^{\mathsf{avg}}(s,a) - \sum_{m=1}^{M} n_{t_v,h}^m(s,a) \right) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} G_{v,h}(s,a) \\ &\overset{\text{(ii)}}{=} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m=1}^{M} \left(\sum_{j=1}^{K} d_h^{\mathsf{avg}}(s,a) - \sum_{j=1}^{K} \mathbb{I}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\} \right) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} G_{v,h}(s,a) \\ &= \sum_{j=1}^{K} \sum_{m=1}^{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^{\mathsf{avg}}(s,a) - \mathbb{I}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\} \right) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} G_{v,h}(s,a) = \sum_{j=1}^{K} \sum_{m=1}^{M} Y_{j,h}^m \end{aligned} \tag{107}$$

where (i) holds by plugging in (101), (ii) follows from $\sum_{v=1}^{\phi(K)} \sum_{t_{v-1}<j\le t_v} 1 = K$ and $\sum_{v=1}^{\phi(K)} n_{t_v,h}^m(s,a) = \sum_{j=1}^{K} \mathbb{I}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\}$, and the last equality arise from the definition of $Y_{j,h}^m$ in (D.7).

Therefore, the above fact shows that to prove (94), it is suffices to show:

$$\left| \sum_{j=1}^{K} \sum_{m=1}^{M} Y_{j,h}^m \right| \le \left| \sum_{j=1}^{K} \sum_{m=1}^{M} \widetilde{Y}_{j,h}^m \right| + \left| \sum_{j=1}^{K} \sum_{m=1}^{M} \left(Y_{j,h}^m - \widetilde{Y}_{j,h}^m \right) \right| \lesssim M\sigma_{\mathsf{aux},1}. \tag{108}$$

We will control the two essential terms separately as below:

- **Controlling $\left|\sum_{j=1}^{K} \sum_{m=1}^{M} \widetilde{Y}_{j,h}^m\right|$.** To begin with, we observe that the approximate $\widetilde{G}_{v,h}^{-j,m}(s,a)$ (defined in (104)) is independent of agent m's visits on (s,a) at j-th episode since $V_{t_{v-1},h+1}$, $G_{v-1,h}(s,a)$ are independent of the j-th episode and $\widetilde{\psi}_{v,v,h}^{-j,m}(s,a)$ is independent from agent m's visits on (s,a) at the j-th episode (see (105)).

$\mathbb{E}_{j-1}[\widetilde{Y}_{j,h}^m] = 0$, where we denote

$$\mathbb{E}_{j-1}[\cdot] = \mathbb{E}\left[\cdot \mid \{(s_{i,h}^{m'}, a_{i,h}^{m'}), V_{i,h+1}^{m'}\}_{i<j, m' \in [M]}\right].$$

Thus, applying the Freedman's inequality for each $h \in [H]$, we can show that the following holds:

$$\begin{aligned}\left|\sum_{j=1}^{K}\sum_{m=1}^{M}\widetilde{Y}_{j,h}^m\right| &\le \sqrt{8W\log\frac{2H}{\delta}} + \frac{8}{3}B\log\frac{2H}{\delta} \\ &\lesssim \sqrt{H^2MKSC_{\mathsf{avg}}^\star} + HSC_{\mathsf{avg}}^\star \end{aligned} \tag{109}$$

at least with probability $1-\delta$, where B and W is obtained as follows:

$$\left|\widetilde{Y}_{j,h}^m\right| \le 2C_{\mathsf{avg}}^\star(1 + d_h^{\pi^\star}(s,\pi^\star(s))S)\max_{s\in\mathcal{S}}\widetilde{G}_{\phi(j),h}^{-j,m}(s,\pi^\star(s)) \le 4SC_{\mathsf{avg}}^\star H =: B \tag{110}$$

$$\begin{aligned}\sum_{j=1}^{K}\sum_{m=1}^{M}\mathbb{E}_{j-1}\left[\left(\widetilde{Y}_{j,h}^m\right)^2\right] &\le \sum_{j=1}^{K}\sum_{m=1}^{M}\mathbb{E}_{(s_{j,h}^m, a_{j,h}^m)\sim d_h^m}\left[\left(\frac{d_h^{\pi^\star}(s_{j,h}^m, a_{j,h}^m)}{d_h^{\mathsf{avg}}(s_{j,h}^m, a_{j,h}^m)}\widetilde{G}_{\phi(j),h}^{-j,m}(s_{j,h}^m, a_{j,h}^m)\right)^2\right] \\ &\le \sum_{j=1}^{K}\sum_{m=1}^{M}\sum_{s\in\mathcal{S}} d_h^m(s,\pi^\star(s))\left(\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{d_h^{\mathsf{avg}}(s,\pi^\star(s))}\widetilde{G}_{\phi(j),h}^{-j,m}(s,\pi^\star(s))\right)^2 \\ &\le H^2C_{\mathsf{avg}}^\star\sum_{j=1}^{K}\sum_{s\in\mathcal{S}}\sum_{m=1}^{M} d_h^m(s,\pi^\star(s))\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{d_h^{\mathsf{avg}}(s,\pi^\star(s))}(1 + d_h^{\pi^\star}(s,\pi^\star(s))S) \\ &\le H^2C_{\mathsf{avg}}^\star\sum_{j=1}^{K}\sum_{s\in\mathcal{S}} Md_h^{\pi^\star}(s,\pi^\star(s))(1 + d_h^{\pi^\star}(s,\pi^\star(s))S) \\ &\le 2H^2SC_{\mathsf{avg}}^\star MK =: W \end{aligned} \tag{111}$$

using the fact that $|\widetilde{G}_{\phi(j),h}^{-j,m}(s_{j,h}^m, a_{j,h}^m)| \le H$ shown in (106b) and $\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi^\star(s)), 1/S\}} \le 1 + d_h^{\pi^\star}(s,\pi^\star(s))S$.

- **Bound on the approximation gap of $\widetilde{Y}_{j,h}^m$.** The approximation gap of $\widetilde{Y}_{j,h}^m$ is bounded as follows:

$$\begin{aligned}&\left|\sum_{j=1}^{K}\sum_{m=1}^{M}\left(\widetilde{Y}_{j,h}^m - Y_{j,h}^m\right)\right| \\ &= \left|\sum_{v=1}^{\phi(K)}\sum_{m=1}^{M}\sum_{t_{v-1}<j\le t_v}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(d_h^m(s,a) - \mathbb{I}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}\right)\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)}(\widetilde{G}_{v,h}^{-j,m}(s,a) - G_{v,h}(s,a))\right| \\ &\overset{(i)}{=} \sum_{v=1}^{\phi(K)}\sum_{m=1}^{M}\sum_{t_{v-1}<j\le t_v}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{I}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}(1 - d_h^m(s,a))\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)}\left|\widetilde{G}_{v,h}^{-j,m}(s,a) - G_{v,h}(s,a)\right| \\ &\overset{(ii)}{\le} \sum_{v=1}^{\phi(K)}\sum_{m=1}^{M}\sum_{t_{v-1}<j\le t_v}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{I}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)}\min\left\{\frac{2H^2}{N_{t_v,h}(s,a)}, H\right\} \\ &\overset{(iii)}{\le} C_{\mathsf{avg}}^\star\sum_{s\in\mathcal{S}}\sum_{v=1}^{\phi(K)} n_{t_v,h}(s,\pi^\star(s))\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi^\star(s)), 1/S\}}\min\left\{\frac{2H^2}{N_{t_v,h}(s,\pi^\star(s))}, H\right\} \\ &\overset{(iv)}{\le} 2H^2C_{\mathsf{avg}}^\star\sum_{s\in\mathcal{S}}(1 + d_h^{\pi^\star}(s,\pi^\star(s))S)\sum_{v=1}^{\phi(K)}\min\left\{\frac{n_{t_v,h}(s,\pi^\star(s))}{N_{t_v,h}(s,\pi^\star(s))}, n_{t_v,h}(s,\pi^\star(s))\right\} \\ &\overset{(v)}{\lesssim} C_{\mathsf{avg}}^\star H^2S \end{aligned} \tag{112}$$

where (i) holds because $\widetilde{\psi}_{v,v,h}^{-j,m}(s,a) = \psi_{v,v,h}^{-j,m}(s,a)$ if $(s_{j,h}^m, a_{j,h}^m) \neq (s,a)$ and $\widetilde{G}_{v,h}^{-j,m}(s,a) = G_{v,h}(s,a)$ according to (106a), (ii) follows from (106c), (iii) naturally holds according to the definition of $C_{\mathsf{avg}}^\star$, (iv) holds because $\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi^\star(s)),1/S\}} \leq 1 + d_h^{\pi^\star}(s,\pi^\star(s))S$, and (v) holds because for any $z \in [\phi(K)]$,

$$\sum_{v=1}^{z} \frac{n_{t_v,h}(s,\pi^\star(s))}{N_{t_v,h}(s,\pi^\star(s))} \leq 1 + \log\left(N_{t_z,h}(s,\pi^\star(s))\right), \tag{113}$$

according to Lemma C.2.

Now, combining the bounds obtained above ((109) and (112)) into (108), we conclude that

$$\left|\sum_{j=1}^{K}\sum_{m=1}^{M} Y_{j,h}^m\right| \lesssim \sqrt{H^2MKSC_{\mathsf{avg}}^\star} + H^2SC_{\mathsf{avg}}^\star = M\left(\sqrt{\frac{H^2KSC_{\mathsf{avg}}^\star}{M}} + \frac{H^2SC_{\mathsf{avg}}^\star}{M}\right) \tag{114}$$

which completes the proof.

D.7.1. Proof of (106)

Proof of (106a). We will proof (106a) by considering different cases separately. When $v = 1$, we have

$$\begin{aligned} G_{v,h}(s,a) &= P_{h,s,a}(V_{h+1}^\star - V_{t_{v-1},h+1})\psi_{1,1,h}(s,a) \\ &= P_{h,s,a}(V_{h+1}^\star - V_{0,h+1})\sum_{m=1}^{M}\left(\sum_{i\in l_{t_1,h}^m(s,a)} \omega_{i,t_1,h}^m(s,a)\right) = P_{h,s,a}(V_{h+1}^\star - V_{0,h+1}) \end{aligned} \tag{115}$$

where the second equality follows from the definition of $\psi_{u,v,h}(s,a)$ in (93), and the last equality holds since

$$\sum_{m=1}^{M}\sum_{i\in l_{t_1,h}^m(s,a)} \omega_{i,t_1,h}^m(s,a) = \frac{(H+1)n_{t_1,h}}{N_{t_1,h} + Hn_{t_1,h}} = \frac{(H+1)n_{t_1,h}}{(H+1)n_{t_1,h}} = 1.$$

When $v > 1$, invoking the definition of $\omega_{i,t_v,h}^m$ in (24c) yields that for any $u < v$,

$$\begin{aligned} \psi_{u,v,h}(s,a) &= \sum_{m=1}^{M}\sum_{i\in l_{t_u,h}^m(s,a)} \omega_{i,t_v,h}^m(s,a) \\ &= \frac{(H+1)n_{t_u,h}}{N_{t_v,h} + Hn_{t_v,h}}\left(\prod_{x=u}^{v-1}\frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \\ &= \frac{(H+1)n_{t_u,h}}{N_{t_{v-1},h} + Hn_{t_{v-1},h}}\left(\prod_{x=u}^{v-2}\frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}}\right)\frac{N_{t_{v-1},h}}{N_{t_v,h} + Hn_{t_v,h}} \\ &= \psi_{u,v-1,h}(s,a)(1 - \psi_{v,v,h}(s,a)). \end{aligned} \tag{116}$$

where the second equality holds by $\phi(i) = u$ for all $i \in l_{t_u,h}^m(s,a)$ and the fact $\sum_{m=1}^{M}\sum_{i\in l_{t_u,h}^m(s,a)} 1 = n_{t_u,h}$, and the last equality holds by $1 - \psi_{v,v,h}(s,a) = 1 - \frac{(H+1)n_{t_v,h}}{N_{t_v,h}+Hn_{t_v,h}} = \frac{N_{t_{v-1},h}+(H+1)n_{t_v,h}-(H+1)n_{t_v,h}}{N_{t_v,h}+Hn_{t_v,h}} = \frac{N_{t_{v-1},h}}{N_{t_v,h}+Hn_{t_v,h}}$.

Consequently, inserting above fact back into (101) complete the proof by showing that

$$\begin{aligned} G_{v,h}(s,a) &= \sum_{u=1}^{v} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) \\ &= P_{h,s,a}(V_{h+1}^\star - V_{t_{v-1},h+1})\psi_{v,v,h}(s,a) + \sum_{u=1}^{v-1} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) \end{aligned}$$

$$= P_{h,s,a}(V^\star_{h+1} - V_{t_{v-1},h+1})\psi_{v,v,h}(s,a) + (1-\psi_{v,v,h}(s,a))\sum_{u=1}^{v-1} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v-1,h}(s,a)$$
$$= P_{h,s,a}(V^\star_{h+1} - V_{t_{v-1},h+1})\psi_{v,v,h}(s,a) + (1-\psi_{v,v,h}(s,a))G_{v-1,h}(s,a). \tag{117}$$

Proof of (106b)**.** First, applying (28c) in Lemma B.3 gives $G_{v,h}(s,a) \geq 0$. Then we focus on deriving the upper bound $G_{v,h}(s,a)$. Towards this, we observe that

$$\begin{aligned} G_{v,h}(s,a) &= \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) \\ &\leq P_{h,s,a}(V^\star_{h+1} - V_{0,h+1})\sum_{u=1}^{v}\psi_{u,v,h}(s,a) \\ &\leq H\sum_{u=1}^{v}\psi_{u,v,h}(s,a) \\ &= H\sum_{u=1}^{v}\sum_{m=1}^{M}\left(\sum_{i\in l^m_{t_{u,h}}(s,a)}\omega^m_{i,t_v,h}(s,a)\right) \leq H, \end{aligned} \tag{118}$$

where the first and second inequalities hold by the fact $P_{h,s,a}(V^\star_{h+1} - V_{t_x,h+1}) \leq P_{h,s,a}(V^\star_{h+1} - V_{0,h+1}) \leq H$ for any $x \in [\phi(K)]$ (see the monotonicity of the value estimates in (14) and the basic bound $\|V^\star_{h+1}\|_\infty \leq H$), the last equality arises from the definition of $\psi_{u,v,h}(s,a)$ in (93), and the last inequality follows from (48b) in Lemma C.3.

Similarly, the same facts holds for $\widetilde{G}^{-j,m}_{v,h}(s,a)$, which can be derived in the same manner. We omit it for conciseness.

Proof of (106c)**.** Consider $v = \phi(j)$. If $v = 1$, combing (106a) and (104) directly gives $\widetilde{G}^{-j,m}_{v,h}(s,a) = G_{v,h}(s,a)$. Then we turn to the case when $v > 1$ and bound the term of interest in two different cases, respectively.

- When $(s^m_{j,h}, a^m_{j,h}) \neq (s,a)$. In this case, invoking the definition in (105) gives

$$\widetilde{\psi}^{-j,m}_{v,v,h}(s,a) = \frac{(H+1)n_{t_v,h}(s,a)}{N_{t_{v-1},h}(s,a) + (H+1)n_{t_v,h}(s,a)} = \psi^{-j,m}_{v,v,h}(s,a), \tag{119}$$

  which indicates (see the definition in (104))

$$\widetilde{G}^{-j,m}_{v,h}(s,a) = G_{v,h}(s,a) \tag{120}$$

- When $(s^m_{j,h}, a^m_{j,h}) = (s,a)$. In view of (106a) and (104), it holds that:

$$\begin{aligned} &|\widetilde{G}^{-j,m}_{v,h}(s,a) - G_{v,h}(s,a)| \\ &= \left|(\widetilde{\psi}^{-j,m}_{v,v,h}(s,a) - \psi_{v,v,h}(s,a))P_{h,s,a}(V^\star_{h+1} - V_{t_{v-1},h+1}) + (\psi_{v,v,h}(s,a) - \widetilde{\psi}^{-j,m}_{v,v,h}(s,a))G_{v-1,h}(s,a)\right| \\ &= \left|(\psi_{v,v,h}(s,a) - \widetilde{\psi}^{-j,m}_{v,v,h}(s,a))(G_{v-1,h}(s,a) - P_{h,s,a}(V^\star_{h+1} - V_{t_{v-1},h+1})\right| \\ &\leq \left|\psi_{v,v,h}(s,a) - \widetilde{\psi}^{-j,m}_{v,v,h}(s,a)\right| \max\left\{G_{v-1,h}(s,a), \|P_{h,s,a}\|_1 \left\|V^\star_{h+1} - V_{t_{v-1},h+1}\right\|_\infty\right\} \\ &\overset{(i)}{\leq} H\left|\psi_{v,v,h}(s,a) - \widetilde{\psi}^{-j,m}_{v,v,h}(s,a)\right| \\ &\overset{(ii)}{\leq} \min\left\{H, \frac{2H^2}{N_{t_v,h}(s,a)}\right\}, \end{aligned} \tag{121}$$

  where (i) holds by (106b), $\|P_{h,s,a}\|_1 = 1$, and $\left\|V^\star_{h+1} - V_{t_{v-1},h+1}\right\|_\infty \leq H$. Here, (ii) can be verified by

$$0 \overset{(iii)}{\leq} \psi_{v,v,h}(s,a) - \widetilde{\psi}^{-j,m}_{v,v,h}(s,a)$$

$$
\begin{aligned}
&= \frac{(H+1)n_{t_v,h}(s,a)}{N_{t_{v-1},h}(s,a)+(H+1)n_{t_v,h}(s,a)} - \frac{(H+1)(n_{t_v,h}(s,a)-\mathbb{I}\{(s,a)=(s^m_{j,h},a^m_{j,h})\})}{N_{t_{v-1},h}(s,a)+(H+1)(n_{t_v,h}(s,a)-\mathbb{I}\{(s,a)=(s^m_{j,h},a^m_{j,h})\})} \\
&= \frac{(H+1)n_{t_v,h}(s,a)}{N_{t_{v-1},h}(s,a)+(H+1)n_{t_v,h}(s,a)} - \frac{(H+1)(n_{t_v,h}(s,a)-1)}{N_{t_{v-1},h}(s,a)+(H+1)(n_{t_v,h}(s,a)-1)} \\
&\le \frac{(H+1)}{N_{t_{v-1},h}(s,a)+(H+1)n_{t_v,h}(s,a)} \\
&\le \min\left\{1, \frac{2H}{N_{t_v,h}(s,a)}\right\}. \qquad (122)
\end{aligned}
$$

where (iii) holds by the fact that $\frac{x}{a+x}$ is monotone increasing with x when $a, x > 0$.

D.8. Proof of Lemma D.3

For each $j \in [K]$, let

$$
Z^m_{j,h} := \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(\mathbb{I}\{(s,a)=(s^m_{j,h},a^m_{j,h})\} - d^m_h(s,a)\right) \frac{d^{\pi^\star}_h(s,a)}{M d^{\mathsf{avg}}_h(s,a)} P_{h,s,a}(V^\star_{h+1} - V_{t_{\phi(j)-1},h+1}). \qquad (123)
$$

Then, to prove Lemma D.3, it suffices to show $\left|\sum_{j=1}^K \sum_{m=1}^M Z^m_{j,h}\right| \lesssim \sigma_{\mathsf{aux},2}$.

Since $V_{t_{\phi(j)-1},h+1}$ is fully determined by the events before j-th episode, $\mathbb{E}_{j-1}[Z^m_{j,h}] = 0$, where we denote

$$
\mathbb{E}_{j-1}[\cdot] = \mathbb{E}[\cdot|\{(s^{m'}_{i,h}, a^{m'}_{i,h}),\ V^{m'}_{i,h+1}\}_{i<j,m'\in[M]}].
$$

Thus, we can apply the Freedman's inequality as follows:

$$
\left|\sum_{j=1}^K \sum_{m=1}^M Z^m_{j,h}\right| \le \sqrt{8W\log\frac{2H}{\delta}} + \frac{8}{3}B\log\frac{2H}{\delta} \lesssim \sqrt{\frac{H^2KSC^\star_{\mathsf{avg}}}{M}} + \frac{HSC^\star_{\mathsf{avg}}}{M} \qquad (124)
$$

using the following properties:

$$
|Z^m_{j,h}| \le \frac{2C^\star_{\mathsf{avg}}H}{M}\left(\sum_{s\in\mathcal{S}}(1+d^{\pi^\star}_h(s,\pi^\star(s))S)\right) \le \frac{4HSC^\star_{\mathsf{avg}}}{M} =: B \qquad (125)
$$

$$
\begin{aligned}
\sum_{j=1}^K\sum_{m=1}^M \mathbb{E}_{j-1}[(Z^m_{j,h})^2] &\le \sum_{j=1}^K\sum_{m=1}^M \mathbb{E}_{(s^m_{j,h},a^m_{j,h})\sim d^m_h}\left[\left(\frac{d^{\pi^\star}_h(s^m_{j,h},a^m_{j,h})}{M d^{\mathsf{avg}}_h(s^m_{j,h},a^m_{j,h})}P_{h,s,a}(V^\star_{h+1}-V_{t_{\phi(j)-1},h+1})\right)^2\right] \\
&\le H^2\sum_{j=1}^K\sum_{m=1}^M\sum_{s\in\mathcal{S}} d^m_h(s,\pi^\star(s))\left(\frac{d^{\pi^\star}_h(s,\pi^\star(s))}{M d^{\mathsf{avg}}_h(s,\pi^\star(s))}\right)^2 \\
&\le \frac{H^2C^\star_{\mathsf{avg}}}{M}\sum_{s\in\mathcal{S}}\sum_{j=1}^K\left(\frac{d^{\pi^\star}_h(s,\pi^\star(s))}{M d^{\mathsf{avg}}_h(s,\pi^\star(s))}\right)(1+d^{\pi^\star}_h(s,\pi^\star(s))S)\sum_{m=1}^M d^m_h(s,\pi^\star(s)) \\
&= \frac{H^2C^\star_{\mathsf{avg}}}{M}\sum_{s\in\mathcal{S}}\sum_{j=1}^K d^{\pi^\star}_h(s,\pi^\star(s))(1+d^{\pi^\star}_h(s,\pi^\star(s))S) \\
&= \frac{2H^2KSC^\star_{\mathsf{avg}}}{M} =: W, \qquad (126)
\end{aligned}
$$

which follows from that fact $0 \le \|V^\star_{h+1} - V_{t_{\phi(j)-1},h+1}\|_\infty \le H$ and $\frac{d^{\pi^\star}_h(s,\pi^\star(s))}{\min\{d^{\pi^\star}_h(s,\pi^\star(s)),1/S\}} \le 1 + d^{\pi^\star}_h(s,\pi^\star(s))S$.

D.9. Proof of Corollary 3.2

Note that if $T \asymp \frac{H^7 SC^\star_{\mathsf{avg}}}{M\varepsilon^2}$, it always holds that

$$MT \gtrsim H^5 SC^\star_{\mathsf{avg}} \text{ and } H \leq \sqrt{\frac{HSC^\star_{\mathsf{avg}}T}{M}}, \tag{127}$$

as long as $\varepsilon \leq H$ and $\varepsilon \leq \frac{H^3 SC^\star_{\mathsf{avg}}}{M}$. Now, we obtain the number of communication rounds of the specified schedules, periodic and exponential synchronization.

Periodic synchronization. Consider $\tau \asymp \sqrt{\frac{HSC^\star_{\mathsf{avg}}T}{M}}$. Then, since $MT \gtrsim HSC^\star_{\mathsf{avg}}$, the value gap is bounded as

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{H^4 SC^\star_{\mathsf{avg}}}{MT} + \sqrt{\frac{H^7 SC^\star_{\mathsf{avg}}}{MT}} + \frac{H^3}{T}\sqrt{\frac{HSC^\star_{\mathsf{avg}}T}{M}} \lesssim \sqrt{\frac{H^7 SC^\star_{\mathsf{avg}}}{MT}}. \tag{128}$$

In this case, the number of synchronizations $\phi(K) = |\mathcal{T}_{\mathsf{period}}(K,\tau)|$ is

$$\phi(K) = \left\lceil \frac{K}{\tau} \right\rceil \lesssim \sqrt{\frac{MK}{H^2 SC^\star_{\mathsf{avg}}}} \asymp \sqrt{\frac{MT}{H^3 SC^\star_{\mathsf{avg}}}} \asymp \frac{H^2}{\varepsilon}.$$

Exponential synchronization. Using the fact that $MT \gtrsim HSC^\star_{\mathsf{avg}}$ and $\tau_1 = H \leq \sqrt{\frac{HSC^\star_{\mathsf{avg}}T}{M}}$ when $\varepsilon \leq \frac{H^3 SC^\star_{\mathsf{avg}}}{M}$, the value gap is bounded as

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{H^4 SC^\star_{\mathsf{avg}}}{MT} + \sqrt{\frac{H^7 SC^\star_{\mathsf{avg}}}{MT}} + \frac{H^3}{T}\sqrt{\frac{HSC^\star_{\mathsf{avg}}T}{M}} \lesssim \sqrt{\frac{H^7 SC^\star_{\mathsf{avg}}}{MT}}. \tag{129}$$

To continue, note that if $\gamma = \frac{2}{H}$ and $\tau_1 = H$, for any $u \geq 1$, τ_u is bounded as

$$(1+\frac{1}{H})^{u-1}H \leq \tau_u \leq (1+\frac{2}{H})^{u-1}H.$$

since

$$(1+\frac{1}{H})\tau_i \leq (1+\frac{2}{H})\tau_i - 1 \leq \tau_{i+1} = \lfloor (1+\frac{2}{H})\tau_i \rfloor \leq (1+\frac{2}{H})\tau_i$$

given the fact that $\tau_i \geq H$ for any $i \geq 1$. Then, computing the minimum number of synchronizations $\phi(K) = |\mathcal{T}_{\mathsf{exp}}(K,\gamma)|$ satisfying

$$\sum_{u=1}^{\phi(K)} \tau_u \geq H \sum_{u=1}^{\phi(K)} (1+\frac{1}{H})^{u-1} = H^2((1+\frac{1}{H})^{\phi(K)} - 1) \geq K,$$

we obtain

$$\phi(K) = \left\lceil \frac{\log\left(\frac{K}{H^2}+1\right)}{\log\left(1+\frac{1}{H}\right)} \right\rceil \leq 1 + (1+H)\log\left(\frac{K}{H^2}+1\right) \lesssim H \tag{130}$$

because $\frac{x}{x+1} \leq \log(1+x)$ for any $x > -1$.