Federated Reinforcement Learning for Efficient and Collaborative Decision Making

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Jiin Woo

B.S., Mathematical Sciences, KAIST
M.S., Electrical Engineering, KAIST

Carnegie Mellon University
Pittsburgh, PA
May 2026

Acknowledgments

The research presented in this thesis was supported in part by the National Science Foundation (NSF) under grants CCF-2007834, CCF-2007911, CCF-2045694, CCF-2106778, CNS-2112471, and CNS-2148212; and by the Office of Naval Research (ONR) under grants N00014-19-1-2404 and N00014-23-1-2149. This work was also supported by funds from federal agencies and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program. Furthermore, I gratefully acknowledge the support provided by the CIT Dean's Fellowship and the Hsu Chang Memorial Fellowship at Carnegie Mellon University.

First and foremost, I would like to express my deepest gratitude to my exceptional advisors, Professor Gauri Joshi and Professor Yuejie Chi. This Ph.D. journey has been the greatest exploration of my life. At first, I was apprehensive and worried about the worst-case scenarios. However, under the amazing guidance of Gauri and Yuejie, those fears completely vanished. In fact, I found myself happier here than anywhere else. This is all thanks to their endless support and patience. Because of them, this bold exploration transformed into a period of deep confidence and growth, with no regrets.

I am profoundly grateful to Gauri for her mentorship, which has been fundamental to my development. Her visionary advice consistently broadened my perspective, helping me navigate the intricate complexities of my research with greater clarity and depth. What I appreciate most is how she truly believed in me and patiently waited for my growth, providing unsparing support throughout the process. Beyond her academic guidance, her constant encouragement was a vital source of strength that empowered me to persevere through the most challenging moments without giving up. Gauri's profound expertise and steadfast belief in my potential have been instrumental in my academic journey, and I am deeply honored to have learned from her.

I am equally indebted to Yuejie, who has been an extraordinary mentor, offering me unwavering support and invaluable guidance. Whenever I felt lost, she provided the precise resources I needed and acted as a steady compass with her sharp intuition. In particular,

her rigorous mathematical insights pushed me to refine my own thinking, challenging me to never settle for suboptimal solutions but instead to constantly strive for global optimality. More importantly, she provided all the intellectual and professional nourishment I needed to flourish as a researcher in every direction. Thanks to her genuine care and mentorship, my doctoral years were so incredibly joyful and fulfilling that a part of me almost didn't want this journey to end.

I would also like to extend my sincere thanks to my thesis committee members. Professor Andrea Zanette has provided invaluable feedback and guidance that significantly improved the quality of my research. I am especially grateful to Professor Cong Ma, who graciously agreed to join my committee on short notice. He has been a great source of inspiration and support, offering insightful feedback that helped me refine my work. I would also like to thank Professor Chi Jin; although he was unable to attend my final defense, his insightful feedback during my proposal was immensely helpful in developing this thesis further. I am deeply grateful for their time, expertise, and constructive feedback, all of which have been instrumental in shaping the direction and quality of my research.

In addition, I would like to express my sincere gratitude to all the academic collaborators and colleagues who have contributed to my research journey. Prof. Laixi Shi has been an incredible mentor, collaborator, and friend. I am deeply grateful for the opportunity to have shared such an enjoyable research experience with her; our discussions were not only a joy but also highly constructive. I would also like to thank Dr. Yuchen Jiao for her collaboration and support, which have been instrumental in advancing my research. Our conversations were consistently insightful and significantly contributed to the development of my work. Also, a special thanks to Baris Askin, Anupam Nayak, and Shivam Patel for our wonderful collaborations. Working alongside you all has been a tremendous learning experience, and I have taken away so much from our time together.

I am also incredibly grateful to the collaborators and colleagues I had the privilege of working with during my internships at Amazon. A very special thank you goes to Dr. Youngsuk Park, who was not only a supportive and inspiring mentor but also a true role model. I learned a tremendous amount under his leadership, and his guidance has left a

lasting impact on my professional and academic growth. Furthermore, I want to express my sincere appreciation to my wonderful mentors, Dr. Shaowei Zhu, Dr. Allen Nie, Dr. Alireza Bagheri Garakani, Dr. Tianchen Zhou, Dr. Zhishen Huang, and Dr. Yan Gao. Our discussions and collaborations were incredibly enriching, and I am so grateful to have learned from such talented individuals. Finally, I would like to thank the entire Amazon team for their warm support and collaborative spirit, which made my internship experience truly rewarding.

Beyond the research, I have been incredibly fortunate to belong to such a warm and supportive research group at CMU. Sharing this PhD journey with my amazing lab mates has made my time here truly enjoyable and fulfilling. My deepest appreciation goes to my research family members: Harlin Lee, Vince Monardo, Tian Tong, Boyue Li, Laixi Shi, Zhize Li, Shicong Cen, Pedro Valdeira, Sudeep Salgia, Harry Dong, Lingjing Kong, Zixin Wen, He Wang, Xingyu Xu, Tong Yang, Timofey Efimov, Ankur Mallick, Jianyu Wang, Samarth Gupta, Yae Jee Cho, Tuhinangshu Choudhury, Shuli Jiang, Divyansh Jhunjhunwala, Pranay Sharma, Neharika Jali, Baris Askin, Arian Raje, Anupam Nayak, Shivam Patel, and many more. I am so grateful for the shared moments, daily encouragement, and camaraderie inside and outside the lab that made this long journey so much more rewarding.

I also want to thank my friends at CMU, Jaeyeon Pyo, Daye Nam, Yuhyun Lee, Eunji Cho, Seonghwan Hong, Sanha Park, Chunghee Kim, Soyong Shin, Jinkyoung Kim, Seonghan Cho, Jungyoon Choi, Haejoon Lee, Jungwon Yoon, Sangyoon Lee, Bokyoung Suh, Byoungjoo Ahn, Juyong Kim, Jisoo Shon, Hun Namkung, Jongik Park, Younggeun Lee, Woojun Kim, Jiyoon Kim, Seyun Kim, and many more. I am so grateful for the countless fun moments, support, and friendship that have made my time at CMU truly special. In addition, I would like to thank those who supported my life outside of the academy. Sungpil Woo welcomed me with open arms when I first arrived in Pittsburgh, providing a pillar of strength when I had no other connections. My warm and caring friends, Cecilia and Nicolas, made my time in this city filled with joy. Lastly, I send my heartfelt thanks to my long-time friends in Korea, Jiyoung Park, Kyounghee Roh, Hayeon Kim, Hyerin Park, and

Hyunho Jang. Despite the distance, your unwavering warmth and constant encouragement gave me the strength to reach the finish line.

I would also like to take a special moment to honor the memory of my Master's advisor, the late Professor Yung Yi. He was the very first person to truly recognize my potential, and it was his strong encouragement that pushed me to pursue a Ph.D. Although he is no longer with us, his belief in me laid the foundation for this entire journey and made all of these achievements possible. May he rest in peace.

Finally, I want to express my deepest gratitude to my family for their unwavering love and support. To my parents, who have always believed in me and encouraged me to pursue my dreams: thank you for your endless sacrifices and unconditional love. To my brothers, who have walked this journey alongside me and have always been reliable pillars of support, thank you for always being there for me. Lastly, a special thank you to Guzzi; your lovely presence and warmth have been a tremendous source of strength, and I am so grateful for the joy you brought into my daily life. I am truly blessed to have such a wonderful family, and I dedicate this thesis to all of you.

Abstract

Reinforcement Learning (RL) faces significant challenges in real-world applications due to the high cost of data collection and computation required to learn optimal decisions in high-dimensional state-action spaces, which is often prohibitive for individual agents. Federated Learning (FL) has emerged as a powerful paradigm for collaborative learning across distributed agents while preserving data privacy. This thesis investigates federated RL, where multiple agents collaboratively learn a global policy with the aid of a central server while keeping data local. We develop principled federated RL algorithms across synchronous, Markovian, and offline settings, and analyze how effective collaboration can maximize sample efficiency and state-action coverage while minimizing communication overhead.

First, in the synchronous setting, we show that simple averaging yields a near-optimal linear speedup in sample complexity relative to the number of agents in both discounted and average-reward settings, given appropriate parameter choices such as learning rates and discount factors. We extend these findings to regimes with Markovian sampling and offline datasets by introducing novel weighted averaging and pessimistic value aggregation schemes, which compensate for the imbalanced training progress of agents and penalize out-of-distribution actions. These methods enable the maximal utilization of heterogeneous and limited per-agent coverage, achieving near-optimal sample efficiency. We further demonstrate that collaboration substantially relaxes per-agent data requirements, requiring only collective coverage of the optimal trajectories. Building on this, we show that different environments among agents can induce such coverage in online settings; when agents strictly pursue their own interests through greedy execution, their collective disagreement fulfills the exploration requirements for one another without necessitating any individual sacrifice. Lastly, we design communication-efficient protocols, including periodic and adaptive aggregation, and prove that they preserve near-optimal sample complexity while significantly reducing communication rounds. These results establish a theoretically grounded and practical framework for collaborative decision making, highlighting the power of fed-

erated learning in addressing the data scarcity and computational bottlenecks inherent in large-scale RL while maintaining data sovereignty for individual agents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

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

Driven by the need to harness data and computation across agents, there is growing interest in implementing RL in a federated manner: multiple agents can jointly learn a global policy with the help of a central server without sharing raw data (Figure 1.1). Although federated learning (FL) has been widely studied in other areas such as supervised learning (Bonawitz et al., 2019; Kairouz et al., 2021; McMahan et al., 2017; Wang et al., 2020), RL brings unique challenges that distinguish it from general FL algorithms. Notably, long decision horizons induce delayed and highly correlated feedback, and data collection

is driven by agents' behavior policies rather than i.i.d. draws, which can significantly affect sample efficiency. Moreover, coverage of the state-action space is critical in RL: an incorrect estimate for a single state or action can propagate through the dynamics and increase risks across the domain. Consequently, federated methods must be specifically tailored for RL to realize the benefits of collaboration in both efficiency and coverage.

This thesis focuses on principled designs of federated RL algorithms that leverage collaboration across agents to achieve the following primary objectives:

- **Sample efficiency:** We characterize how federated collaboration reduces the number of samples and iterations needed to learn optimal policies, highlighting near-optimal linear speedup in the number of agents without the need for sharing local datasets at agents. Our results demonstrate that, with novel aggregation schemes, federated Q-learning can retain full linear speedup even under significant behavior and reward heterogeneity.
- **Collaborative coverage:** To ensure robust performance across diverse scenarios, it is crucial for RL agents to explore and cover a wide range of state-action pairs. In federated settings, multiple agents can collect data from different regions of the state-action space, thereby enhancing overall coverage and reducing the risk of insufficient exploration. This highlights a new perspective on how federated framework can benefit RL beyond sample efficiency, which has not been explored in FL literature.
- **Communication efficiency:** Communication between agents and a central server is often the dominant bottleneck in federated settings. To ensure practical viability, we design communication-efficient protocols that significantly reduce the number of synchronization rounds while preserving near-optimal sample efficiency. Our results show that by adaptively tuning learning rates and communication intervals, one can minimize communication overhead without sacrificing the benefits of collaboration.

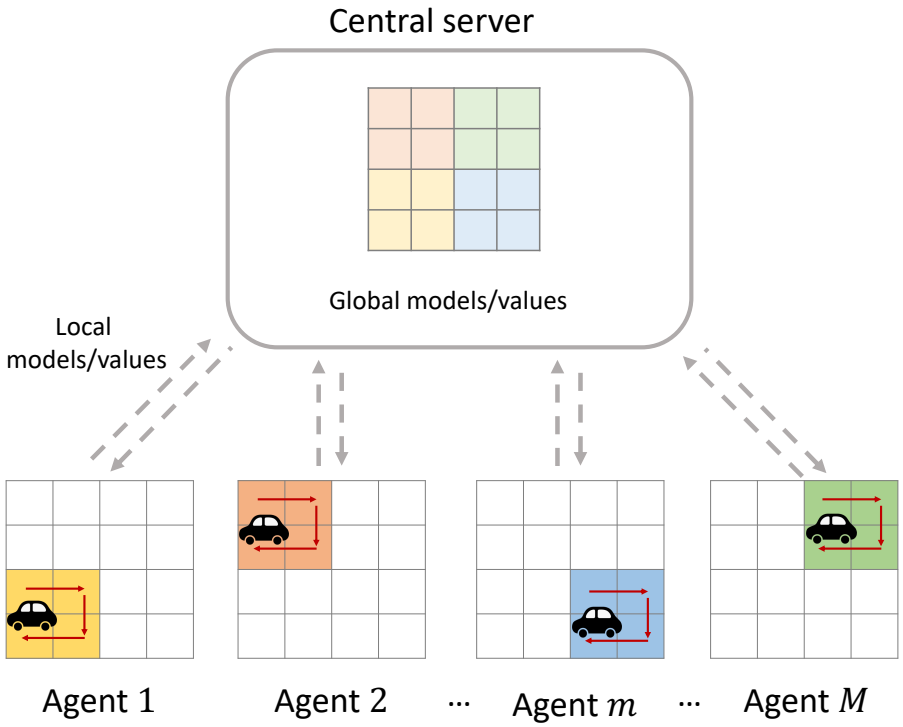


Figure 1.1: The overview of the federated RL framework, where M agents interact with their local environments or data and share their local models with a central server to learn a global policy collaboratively, without sharing local datasets.

1.1 Thesis Contribution

This thesis makes the following contributions to the design and analysis of federated RL algorithms in various RL settings, including synchronous, Markovian, offline, and online regimes:

1.1.1 Sample Efficiency and Linear Speedup

We first present federated Q-learning algorithms in the generative-model setting, where each agent has access to a simulator that returns independent samples for every state-action pair, and show that simple averaging yields a linear improvement in sample complexity (Woo et al., 2023). Within the same synchronous framework, we further develop federated Q-learning for average-reward MDPs, where the undiscounted infinite-horizon nature typically requires substantially more samples and thus makes federated collaboration es-

pecially valuable. With tailored parameter choices, we provide the first sample-complexity guarantee for federated Q-learning in average-reward settings, showing linear speedup with the number of agents (Jiao et al., 2026). We then extend the analysis to more challenging regimes with Markovian sampling (Woo et al., 2023) or offline datasets (Woo et al., 2024), where agents may have limited coverage and uneven training progress.

1.1.2 Collaborative Coverage and the Blessing of Heterogeneity

We analyze the benefits of collaboration in federated Q-learning, showing that federated settings can substantially relax per-agent coverage requirements (Woo et al., 2025). We also introduce pessimistic value aggregation to guard against overestimation when agents have very limited coverage, which is especially pronounced with offline datasets, and provide sample complexity guarantees under significantly weaker coverage conditions, requiring only that the agents' datasets collectively cover the optimal trajectories (Woo et al., 2024). Building on this, we further demonstrate that the inherent heterogeneity of agents' rewards can naturally induce such collaborative coverage, enabling efficient learning through purely greedy execution without any explicit exploration mechanism.

1.1.3 Communication-Efficient Synchronization Protocols

We develop and analyze communication-efficient protocols that substantially reduce communication while retaining near-optimal sample efficiency. In particular, we study periodic averaging (agents synchronize at fixed intervals) and adaptive aggregation schemes that progressively increase communication intervals as training proceeds (Woo et al., 2023, 2024). Our communication-complexity results, derived for offline RL, show that by adaptively tuning learning rates and communication intervals one can greatly lower the number of communication rounds without degrading sample efficiency (Woo et al., 2024). We also establish near-optimal communication complexity guarantees, making federated RL more practical for real-world deployment.

1.2 Related Work

Sample complexity of Q-learning. Q-learning (Watkins and Dayan, 1992) is one of the most widely studied model-free RL algorithms, particularly in the context of discounted MDPs. The finite-time sample complexity of Q-learning has been analyzed in various settings: Beck and Srikant (2012); Even-Dar and Mansour (2003); Li et al. (2024a); Wainwright (2019a) have investigated the synchronous case with a generative model, while Li et al. (2022b); Qu and Wierman (2020); Shi et al. (2022); Szepesvári (1998); Yan et al. (2022) have focused on asynchronous and offline Q-learning. In addition, Li et al. (2022b); Shi et al. (2022); Sidford et al. (2018); Wainwright (2019b) have proposed variance-reduced variants of Q-learning and improved sample complexity bounds.

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 and Gu, 2021; Fujimoto et al., 2019; Siegel et al., 2020) or penalizing value estimates on out-of-distribution state-action pairs (Kostrikov et al., 2022; Kumar et al., 2020; Liu et al., 2020; 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 (Jin et al., 2021; Kidambi et al., 2020; Kim and Oh, 2023; Li et al., 2022a; Rashidinejad et al., 2021; Xie et al., 2021b; Yin and Wang, 2021; Yu et al., 2020), policy-based approaches (Xie et al., 2021a; Zanette et al., 2021), and model-free approaches (Shi et al., 2022; Uehara et al., 2023; Yan et al., 2022). 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., 2022), and the distribution mismatch among the two visitation distributions governs the hardness of offline RL (Li et al., 2022a). 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.

Average-reward RL. The average-reward MDP framework was first introduced by Howard (1960), and it has been investigated in various settings. Model-based algorithms for average-reward MDPs include Jin and Sidford (2021); Tuynman et al. (2024); Wang et al. (2022, 2024b); Zurek and Chen (2024, 2025a,b), while model-free methods are studied by Chen (2025); Ganesh et al. (2024); Wan et al. (2021, 2024); Wei et al. (2020). Policy gradient approaches are explored in Bai et al. (2024); Kumar et al. (2025). Sample complexity lower bounds are established by Bravo and Contreras (2024); Jin and Sidford (2021); Wang et al. (2022). Most prior work focuses on the synchronous setting with a generative model, though some, such as Chen (2025), consider the asynchronous setting with Markovian trajectories. Many existing methods rely on prior knowledge of problem-dependent parameters, such as the mixing time or the span of the bias function, which are typically challenging to estimate in practice. Recently, Lee et al. (2025) developed a model-free algorithm that achieves near-optimal sample complexity for average-reward MDPs without requiring prior knowledge of problem parameters, leveraging recursive sampling as a variance reduction technique. While effective, this approach introduces additional algorithmic and computational complexity, which may limit its practical adoption. In contrast, Jin et al. (2024) proposed a simpler adaptation of vanilla Q-learning, relying only on dynamic updates of horizon factors and learning rates. However, their method remains suboptimal in terms of sample efficiency.

Federated RL. Federated RL (FRL) enables multiple agents to collaboratively learn optimal policies, addressing limitations in sample availability and computational resources. Recent work has studied sample complexity improvements in discounted and finite-horizon MDPs, showing that federated Q-learning achieves linear speedup with respect to the number of agents, without data sharing (Khodadadian et al., 2022; Salgia and Chi, 2024; Woo et al., 2023, 2025). In addition, Zheng et al. (2024) analyzed regret for online federated Q-learning, Woo et al. (2024) studied offline variants and the impact of collective data coverage. Also, communication efficiency is crucial; Salgia and Chi (2024) established

lower bounds on communication complexity for discounted MDPs.

While most FRL literature assumes homogeneous environments, recent work addresses the more realistic heterogeneous setting where agents differ in transition dynamics and rewards (Labbi et al., 2024; Yang et al., 2024a; Zhang et al., 2024). Many existing approaches optimize a shared global policy based on average agent values (Wang et al., 2024a; Yang et al., 2024b), often sacrificing individual utility for consensus. To mitigate this, personalized FedRL algorithms allow for individual policies. Zhang and Azizan (2026) achieved linear speedup via affinity-based variance reduction, and Xiong et al. (2025) proposed a personalized federated Q-learning algorithm with similar efficiency. However, these still rely on environmental similarity or high-coverage fixed behavior policies. Recently, Labbi et al. (2024) introduced online federated value iteration for heterogeneous environments, achieving sublinear regret and linear speedup through active exploration. Yet, its reliance on explicit UCB bonuses remains a hurdle for real-world deployment.

1.3 Notation

Throughout the thesis, we use $\Delta(\mathcal{S})$ to refer to the probability simplex over a set $\mathcal{S}$, and $[K] \coloneqq \{1, \cdots, K\}$ for any positive integer $K > 0$. In addition, $f(\cdot) = \widetilde{O}(g(\cdot))$ or $f \lesssim g$ (resp. $f(\cdot) = \widetilde{\Omega}(g(\cdot))$ or $f \gtrsim g$) 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.

Chapter 2

Background and Model

In this chapter, we establish the formal mathematical and algorithmic foundations for the federated reinforcement learning (RL) frameworks investigated in this thesis.

First, we define the core Markov Decision Process (MDP) formulations, presenting the objectives and the corresponding Bellman equations for three primary settings: infinite-horizon discounted, infinite-horizon average-reward, and finite-horizon MDPs. This comprehensive characterization of MDPs sets the stage for our subsequent algorithmic developments and theoretical analyses in federated reinforcement learning, as each setting captures different aspects of long-term decision-making and requires distinct solution techniques. Second, we formalize the data sampling models, ranging from synchronous generative models to asynchronous Markovian trajectories and offline datasets, that dictate how agents collect information from their environments. Finally, we provide a high-level taxonomy of reinforcement learning methodologies, distinguishing between model-based and model-free approaches, as well as value-based and policy-based paradigms. This background situates our work within the broader RL literature and justifies our focus on Q-learning as a primary tool for collaborative decision-making.

Together, these components serve as the structural basis for the federated designs and theoretical analyses presented in the subsequent chapters of this thesis.

2.1 Markov Decision Processes (MDPs)

We consider a Markov decision process (MDP) represented by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space and the action space, respectively. Furthermore, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ indicates the transition kernel such that $P(s' \mid s, a)$ denotes the probability that action a in state s leads to state s', and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes a deterministic reward function, where $r(s, a)$ is the immediate reward for action a in state s. Depending on the optimization objective and the decision horizon, we categorize MDPs into infinite-horizon (including discounted and average-reward) and finite-horizon settings, each governed by distinct optimality criteria.

2.1.1 Infinite-Horizon Discounted MDPs

We first consider the scenario where the agent's interaction with the environment continues indefinitely, a framework known as the *infinite-horizon* setting. When the decision-making process has no fixed terminal time, the cumulative reward involves an infinite sum, which may not naturally converge to a finite value. To ensure a well-defined and bounded objective, it is standard to introduce a discount factor $\gamma \in [0, 1)$, which assigns progressively diminishing weights to rewards received further in the future. This *discounted* formulation not only provides mathematical tractability by ensuring the convergence of total rewards but also reflects the practical preference for immediate gains over distant ones. Under this framework, the MDP is formally represented as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where the components are defined as above, with the addition of the discount factor γ.

Policy, value function, and Q-function. To interact with the MDP, the agent follows an action-selection rule called a **policy**. Formally, a policy is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\pi(a|s)$ denotes the probability of taking action a in state s. A fixed policy π, combined with the transition kernel P, induces a probability distribution over the sequence of states and actions, known as a *trajectory*. To quantify the long-term discounted total reward of following a particular policy π, we define the **value function** $V^\pi : \mathcal{S} \to \mathbb{R}$ as

the expected discounted cumulative reward starting from state s:

$$\forall s \in \mathcal{S}: \qquad V^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]. \tag{2.1}$$

In this expectation, the trajectory $\{s_t, a_t, r_t\}_{t=0}^{\infty}$ is generated by the agent's behavior $a_t \sim \pi(\cdot|s_t)$ and the environment's dynamics $s_{t+1} \sim P(\cdot|s_t, a_t)$. Similarly, the **Q-function** $Q^{\pi}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ evaluates the quality of taking a specific action a in state s and thereafter following policy π:

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

Since the rewards are bounded in $[0, 1]$, the values for any policy π are also bounded as $0 \le V^{\pi}, Q^{\pi} \le \frac{1}{1-\gamma}$.

Optimal policy and Bellman's principle of optimality. The ultimate goal of reinforcement learning is to find an **optimal policy** $\pi^{\star}$ that yields the maximum possible value simultaneously across all states, i.e., $V^{\pi^{\star}}(s) = \max_{\pi} V^{\pi}(s)$ for all $s \in \mathcal{S}$. The existence of such a policy is guaranteed (Puterman, 2014). We denote the corresponding optimal value and Q-functions as $V^{\star}$ and $Q^{\star}$, respectively. The optimal Q-function $Q^{\star}$ is the unique fixed point of the Bellman optimality operator $\mathcal{T}$, satisfying:

$$Q^{\star}(s, a) = (\mathcal{T}Q^{\star})(s, a) \coloneqq r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) \max_{a' \in \mathcal{A}} Q^{\star}(s', a'). \tag{2.2}$$

Uniform ergodicity and mixing time. We introduce a subclass of infinite-horizon MDPs that exhibit mixing behavior, formally defined as follows:

Definition 1 (Uniform ergodicity)**.** *Assume that for any policy π, there exists a stationary distribution ν^{π}, such that for any initial distribution $q \in \Delta(\mathcal{S})$, the Markov chain induced by π converges to ν^{π} with a bounded mixing time t_{mix}, i.e.,*

$$t_{\mathsf{mix}} \coloneqq \max_{\pi} \min\left\{t : \max_{q \in \Delta(\mathcal{S})} d_{\mathsf{TV}}((P^{\pi})^t(q), \nu^{\pi}) \le \frac{1}{4}\right\},$$

where $d_{\mathsf{TV}}(\cdot,\cdot)$ *denotes the total variation distance between two distributions, and* $P^\pi : \mathbb{R}^S \to \mathbb{R}^S$ *is the transition probability induced by policy* π*, defined as*

$$(P^\pi(q))(s) = \sum_{s' \in \mathcal{S}} P(s|s', \pi(s'))q(s').$$

We assume that $t_{\mathsf{mix}} < \infty$*.*

The uniform ergodicity condition ensures that the Markov chain induced by any policy converges to a unique stationary distribution at a geometric rate, which is crucial for analyzing the convergence properties of learning algorithms in MDPs. The mixing time t_{mix} quantifies how quickly this convergence occurs, and it plays a significant role in determining the sample complexity of algorithms in infinite-horizon MDPs.

2.1.2 Infinite-Horizon Average-Reward MDPs

While the discounted formulation is mathematically convenient for ensuring convergence, many long-term decision-making tasks, such as continuous control or steady-state optimization, are more naturally captured by the average-reward framework (Howard, 1960). In this setting, the agent aims to maximize the expected reward per time step over an infinite horizon without diminishing the importance of future gains.

An infinite-horizon Average-reward Markov Decision Process (AMDP) is represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S} = \{1, \ldots, S\}$ and $\mathcal{A} = \{1, \ldots, A\}$ denote the finite state and action spaces, respectively. The transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ specifies the probability distribution over the next state given a state-action pair, i.e., $P(\cdot|s, a) \in \Delta(\mathcal{S})$ denotes the transition probability when action a is taken in state s. The reward function $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ assigns a deterministic and bounded immediate reward $r(s, a)$ to each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Average reward and bias function. To evaluate the performance of a policy π in this non-discounted regime, we consider the long-term average reward. Under policy π, the

average reward starting from an initial state $s \in \mathcal{S}$ is defined as:

$$J^\pi(s) \coloneqq \liminf_{T\to\infty} \mathbb{E}_\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) | s_0 = s \right], \quad \forall s \in \mathcal{S},$$

where the trajectory $\{s_t, a_t\}_{t=0}^{\infty}$ is generated by following policy π and evolving according to the transition kernel P. Unlike the discounted setting where the value function remains bounded by $(1-\gamma)^{-1}$, the total reward here is typically infinite. Therefore, we focus on the relative difference in cumulative rewards, known as the **bias function** $h^\pi(s)$:

$$h^\pi(s) \coloneqq \text{C-}\lim_{T\to\infty} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} (r(s_t, a_t) - J^\pi(s_t)) \right],$$

where C-lim denotes the Cesaro limit. If the Markov chain induced by π is aperiodic, then C-lim coincides with the standard limit.

The optimal policy $\pi^\star$ is defined as the policy that maximizes the average reward:

$$\pi^\star = \arg\max_\pi J^\pi.$$

For convenience, we denote the optimal value by $J^\star = J^{\pi^\star}$ and the corresponding optimal bias function as $h^\star$.

Weakly communicating MDPs and Bellman equation. A **weakly communicating MDP** is one in which there exists a set of states such that, under some policy, each state in the set is reachable from every other state in that set, and states outside this set are transient under all policies (Puterman, 2014). In the weakly communicating setting, there always exists a unichain optimal policy, such that the optimal average reward $J^\star(s)$ is identical across all states $s \in \mathcal{S}$ (Puterman, 2014; Zurek and Chen, 2024). In this case, it is well-known that the optimal value $J^\star$ and the corresponding optimal bias function $h^\star$

satisfy the average-reward Bellman equation:

$$J^\star + h^\star(s) = \max_{a\in\mathcal{A}} \left[r(s,a) + \sum_{s'\in\mathcal{S}} P(s'|s,a)h^\star(s') \right], \quad \forall s \in \mathcal{S}. \tag{2.3}$$

The span norm $h^\star$ is often used to characterize the sample complexity in the average-reward setting, defined as:

$$\|h^\star\|_{\mathsf{sp}} := \max_s h^\star(s) - \min_s h^\star(s).$$

Uniform ergodicity and mixing time. We consider a subclass of infinite-horizon MDPs that exhibit mixing behavior, formally defined as follows:

Definition 2 (Uniform ergodicity)**.** *Assume that for any policy π, there exists a stationary distribution ν^π, such that for any initial distribution $q \in \Delta(\mathcal{S})$, the Markov chain induced by π converges to ν^π with a bounded mixing time t_{mix}, i.e.,*

$$t_{\mathsf{mix}} := \max_\pi \min \left\{ t : \max_{q\in\Delta_\mathcal{S}} d_{\mathsf{TV}}((P^\pi)^t(q), \nu^\pi) \le \frac{1}{4} \right\},$$

where $d_{\mathsf{TV}}(\cdot,\cdot)$ denotes the total variation distance between two distributions, and $P^\pi : \mathbb{R}^S \to \mathbb{R}^S$ is the transition probability induced by policy π, defined as

$$(P^\pi(q))(s) = \sum_{s'\in\mathcal{S}} P(s|s',\pi(s'))q(s').$$

We assume that $t_{\mathsf{mix}} < \infty$.

This assumption implies that the Markov chain induced by any policy is uniformly ergodic, i.e., it has a unique stationary distribution and converges to it at a uniformly exponential rate. In particular, any MDP satisfying this assumption must be unichain, meaning that under every policy, the induced Markov chain consists of a single recurrent class, along with a possibly empty set of transient states.

Since unichain MDPs are a subset of weakly communicating MDPs, any MDP satisfying this assumption is also weakly communicating. In particular, it has been shown that (Wang

et al., 2022)

$$\|h^\star\|_{\mathsf{sp}} \lesssim t_{\mathsf{mix}}.$$

Relationship with Discounted MDPs. Many prior works (Jin et al., 2024; Jin and Sidford, 2021) build upon well-established techniques developed for Discounted Markov Decision Processes (DMDPs), defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\gamma \in [0, 1)$ denotes the discount factor. To avoid confusion between the two settings when both are considered in the same context, for a discount factor γ and a policy π, we denote the value function and Q-function for DMDPs with a discount factor γ as V_γ^π and Q_γ^π, respectively, defined as follows.

$$V_\gamma^\pi(s) \coloneqq (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right], \tag{2.4}$$

$$Q_\gamma^\pi(s, a) \coloneqq (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a\right]. \tag{2.5}$$

When the rewards lie within $[0, 1]$, it follows that for any policy π,

$$0 \le V_\gamma^\pi \le 1, \qquad 0 \le Q_\gamma^\pi \le 1.$$

We note that the above definition differs from their typical forms presented in Section 2.1.1 by a factor of $1/(1-\gamma)$.

2.1.3 Finite-Horizon MDPs

For episodic tasks with a fixed duration, we consider an episodic finite-horizon MDP represented by

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

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. Unlike the

infinite-horizon settings, here the transition kernel and reward may depend on the time step h.

Value Functions and Optimality. 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.,

$$V_h^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=h}^{H} r_t\big(s_t, a_t\big) \,\Big|\, s_h = s\right], \tag{2.6}$$

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

$$Q_h^\pi(s,a) \coloneqq r_h(s,a) + \mathbb{E}\left[\sum_{t=h+1}^{H} r_t(s_t, a_t) \,\Big|\, s_h = s, a_h = a\right], \tag{2.7}$$

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

$$V_h^\star(s) \coloneqq V_h^{\pi^\star}(s) = \max_\pi V_h^\pi(s), \qquad Q_h^\star(s,a) \coloneqq Q_h^{\pi^\star}(s,a) = \max_\pi Q_h^\pi(s,a)$$

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

$$V_1^{\pi}(\rho) \coloneqq \mathbb{E}_{s_1\sim\rho}\big[V_1^{\pi}(s_1)\big] \qquad \text{and} \qquad V_1^{\star}(\rho) \coloneqq \mathbb{E}_{s_1\sim\rho}\big[V_1^{\star}(s_1)\big]. \tag{2.8}$$

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

$$Q_h^{\pi}(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_{h,s,a}}\big[V_{h+1}^{\pi}(s')\big] \tag{2.9}$$

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

Optimal Value Functions and Bellman Optimality. The primary objective in a finite-horizon MDP is to find a policy that maximizes the expected cumulative reward. To this end, we define the optimal Q-function $Q_h^{\star}(s,a)$ as the maximum value achievable across all possible policies for each state-action pair at every time step h:

$$Q_h^{\star}(s,a) \coloneqq \max_{\pi} Q_h^{\pi}(s,a), \quad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]. \tag{2.10}$$

It is well-established that this optimal Q-function satisfies the **Bellman optimality equation**, which characterizes the optimal behavior through a recursive relationship. Specifically, for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, $Q_h^{\star}$ must satisfy:

$$\begin{aligned} Q_h^{\star}(s,a) &= r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\big[\max_{a'\in\mathcal{A}} Q_{h+1}^{\star}(s',a')\big] \\ &= r_h(s,a) + \sum_{s'\in\mathcal{S}} P_h(s'|s,a)\max_{a'\in\mathcal{A}} Q_{h+1}^{\star}(s',a'), \end{aligned} \tag{2.11}$$

subject to the boundary condition $Q_{H+1}^{\star}(s,a) = 0$. By solving this equation, typically via backward induction, one can derive an optimal policy $\pi^{\star}$ that satisfies $\pi_h^{\star}(s) \in \arg\max_{a\in\mathcal{A}} Q_h^{\star}(s,a)$.

2.2 Data Sampling

In this thesis, we investigate federated RL algorithms under different paradigms of agent-environment interaction. Each paradigm dictates how samples are collected from the underlying MDP, which in turn determines the fundamental limits of learning efficiency. We categorize these into three main data sampling models.

2.2.1 Generative Model (Synchronous Sampling)

A generative model, or simulator, provides the agent with the most flexible access to the environment (Kearns and Singh, 1999). In this setting under infinite-horizon MDPs, an agent can query any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and receive an independent transition sample:

$$s' \sim P(\cdot|s, a), \quad \text{and} \quad r = r(s, a). \tag{2.12}$$

This allows for a synchronous coverage of the state-action space, making it a powerful tool for establishing baseline sample complexity results, as explored in Chapter 3.

2.2.2 Markovian Trajectories (Asynchronous Sampling)

In many real-world scenarios, agents do not have access to a simulator and must learn from a single, continuous stream of experience. Under infinite-horizon MDPs, an agent follows a behavior policy π_{b} to generate a Markovian trajectory $\{s_t, a_t, r_t\}_{t=0}^{\infty}$ such that:

$$a_t \sim \pi_{\mathsf{b}}(\cdot \,|\, s_t), \quad r_t = r(s_t, a_t), \quad s_{t+1} \sim P(\cdot \,|\, s_t, a_t). \tag{2.13}$$

for all $t \geq 0$, where the initial state is s_0. The behavior policy π_{b} may be different from the optimal policy $\pi^\star$. Contrary to the generative model considered in the synchronous setting, the samples collected under the asynchronous setting are Markovian (dependent across time). The sample trajectory can be viewed as a time-homogeneous Markov chain over the set of state-action pairs. In Chapter 4, we analyze the sample complexity of federated Q-learning under this asynchronous sampling paradigm.

2.2.3 Offline Datasets

In offline RL, agents are prohibited from interacting with the environment during training. Instead, an agent has access to an offline dataset containing pre-collected episodes by following some behavior policy. Consider the finite-horizon Setting. The offline dataset $\mathcal{D}$ at an agent is composed of K episodes, each generated independently according to a behavior policy $\mu = \{\mu_h\}_{h=1}^H$, resulting in

$$\mathcal{D} \coloneqq \left\{ \left(s_{k,1},\, a_{k,1},\, r_{k,1},\, \ldots,\, s_{k,H},\, a_{k,H},\, r_{k,H} \right) \right\}_{k=1}^K,$$

where the initial state $s_{k,1} \sim \rho$ is drawn from some initial state distribution $\rho \in \Delta(\mathcal{S})$, $s_{k,h}, a_{k,h}, r_{k,h}$ are the state, action and reward at step h in the k-th episode, $a_{k,h} \sim \mu_h(\cdot \,|\, s_{k,h})$ and $r_{k,h} = r_h(s_{k,h}, a_{k,h})$. In Chapter 5, we analyze the sample complexity of federated offline RL under this offline dataset paradigm.

2.2.4 Online Sampling

In online RL, agents aim to learn the optimal policy by interacting with the environment using a policy that is updated iteratively based on the data collected so far. In an episodic finite-horizon MDP, at episode k, an agent selects a policy π_k based on the data collected from the previous episodes, and then executes π_k to generate a trajectory $\{s_{k,h}, a_{k,h}, r_{k,h}\}_{h=1}^H$, where the initial state $s_{k,1}$ is drawn from some initial state distribution $\rho \in \Delta(\mathcal{S})$. This setting requires balancing exploration and exploitation to balance the trade-off between gathering informative data and maximizing rewards. In Chapter 6, we analyze the sample complexity of federated online RL under this online sampling paradigm.

2.3 Reinforcement Learning Algorithms

Building upon the theoretical foundations of MDPs, RL provides a suite of algorithmic paradigms to find optimal policies from interactions with the environment. In this section, we introduce the main classes of RL algorithms that are relevant to this thesis, using the

infinite-horizon setting as the primary context.

2.3.1 Model-Based RL

Model-based RL is a paradigm where the agent explicitly learns an estimate of the MDP's transition dynamics and reward function. An agent utilizes collected samples to construct an empirical estimate of the MDP's underlying components, namely the transition kernel $\widehat{P}$ and the reward function $\widehat{r}$. Once these estimates are obtained, the agent solves the Bellman equations as if the empirical model were the true environment. While model-based methods are often more sample-efficient as they fully exploit the learned dynamics, they require access to batch samples to accurately estimate the model, which can be memory and computationally intensive.

One of the most widely studied model-based algorithms is *value iteration* (VI), which computes the optimal value function by treating the Bellman optimality equation as a fixed-point problem. Specifically, given the empirical estimates $\widehat{P}$ and $\widehat{r}$, VI iteratively applies the empirical Bellman optimality operator $\widehat{\mathcal{T}}$ to an initial estimate $\widehat{Q}_0$ until convergence:

$$\widehat{Q} \leftarrow \widehat{\mathcal{T}}(\widehat{Q}) = \widehat{r} + \gamma \widehat{P} \max_{a' \in \mathcal{A}} \widehat{Q}(\cdot, a'). \tag{2.14}$$

Due to the contraction property of the empirical Bellman operator, this process is guaranteed to converge to the unique optimal Q-function $\widehat{Q}^\star$ of the empirical MDP, which serves as an approximation to the true optimal Q-function $Q^\star$ of the underlying MDP.

2.3.2 Model-Free RL

In contrast, model-free RL algorithms bypass the explicit estimation of the transition dynamics. Instead, they learn the value functions or policies directly from interaction samples, which is typically more computationally efficient and requires significantly less memory as the agent does not need to store large transition matrices.

The most prominent class of model-free methods is value-based learning, with *Q-learning* (Watkins and Dayan, 1992) being the most celebrated example. Fundamentally,

Q-learning can be viewed as a stochastic approximation of the value iteration algorithm described above. Rather than performing a full expectation over the transition kernel P, it uses individual transition samples (s, a, r, s') to iteratively update the Q-function toward the Bellman target:

$$Q(s,a) \leftarrow (1-\eta)Q(s,a) + \eta\left(r + \gamma \max_{a' \in \mathcal{A}} Q(s',a')\right), \tag{2.15}$$

where $\eta \in (0,1]$ is the learning rate. This update rule allows the agent to learn optimal policies without ever constructing an explicit model of the environment, making it particularly suitable for large or complex MDPs where model estimation is infeasible.

Rather than indirectly deriving a policy from a value function, *policy optimization* (or policy-based) methods (Williams, 1992) directly parameterize a policy π_θ and optimize the expected return $J(\pi_\theta)$ via gradient ascent, where the gradient is given by:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)\right]. \tag{2.16}$$

These methods are particularly effective in high-dimensional or continuous action spaces where finding the maximum of a Q-function at every step is computationally prohibitive.

In this thesis, we primarily focus on Q-learning and value iteration methods, analyzing their sample complexity and convergence properties under various federated learning settings.

Chapter 3

Sample-Efficient Federated RL with Generative Models

3.1 Federated Synchronous Q-Learning for Discounted MDPs

In this section, we study federated synchronous Q-learning, where all the state-action pairs are updated simultaneously assuming access to a generative model or simulator for infinite-horizon discounted MDPs at all the agents.

Synchronous Q-learning with generative models. In the synchronous sampling, all state-action pairs are sampled uniformly assuming access to a generative model or a simulator (Kearns and Singh, 1999). In every iteration t, an agent generates a transition sample

$$s_t(s,a) \sim P(\cdot|s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{3.1}$$

for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ *independently* at every iteration t.

In the synchronous Q-learning for infinite-horizon MDPs, starting with certain initial-

sampling	reference	number of agents	coverage	sample complexity
synchronous	Chen et al. (2020); Wainwright (2019a)	1	full	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
	(Li et al., 2023)	1	full	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
	FedSynQ (Theorem 1)	M	full	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{M(1-\gamma)^5\varepsilon^2}$

Table 3.1: Comparison of sample complexity upper bounds of single-agent and federated Q-learning algorithms under synchronous sampling protocols to learn an ε-optimal Q-function in the ℓ_∞ sense, where logarithmic factors and burn-in costs are hidden. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, γ is the discount factor, and M is the total number of agents.

ization Q_0, at every iteration $t \geq 1$, the Q-function is updated according to

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q_t(s,a) = (1-\eta) Q_{t-1}(s,a) + \eta \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t(s,a), a') \right), \tag{3.2}$$

where $s_t(s,a) \sim P(\cdot|s,a)$ is drawn independently for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, and η denotes the constant learning rate.

3.1.1 Problem setting

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

$$s_t^m(s,a) \sim P(\cdot|s,a) \tag{3.3}$$

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

3.1.2 Algorithm: Federated Synchronous Q-learning (FedSynQ)

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

$$\forall s \in \mathcal{S}: \qquad V_t^m(s) := \max_{a \in \mathcal{A}} Q_t^m(s,a). \tag{3.4}$$

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

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

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_{t-\frac{1}{2}}^m(s,a) = (1-\eta) Q_{t-1}^m(s,a) + \eta \left(r(s,a) + \gamma V_{t-1}^m(s_t^m(s,a)) \right), \tag{3.5}$$

   where $s_t^m(s,a)$ is drawn according to (3.3), and $\eta \geq 0$ is the learning rate.

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

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_t^m(s,a) = \begin{cases} \frac{1}{M} \sum_{m=1}^M Q_{t-\frac{1}{2}}^m(s,a) & \text{if } t \equiv 0 \pmod{\tau} \\ Q_{t-\frac{1}{2}}^m(s,a) & \text{otherwise} \end{cases}. \tag{3.6}$$

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

Algorithm 1: Federated Synchronous Q-learning (FedSynQ)

inputs: learning rate η, discount factor γ, number of agents M, synchronization period τ, number of iterations T.
initialization: $Q_0^m = Q_0$ for all m.
for $t = 1, \cdots, T$ **do**
  for $m \in [M]$ **do**
    Draw $s_t^m(s,a) \sim P(\cdot \,|\, s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
    Compute $Q_{t-\frac{1}{2}}^m$ according to (3.5).
    Compute Q_t^m according to (3.6).
  end for
end for
return: $Q_T = \frac{1}{M} \sum_{m=1}^M Q_T^m$.

3.1.3 Performance guarantee

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

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

$$\tau \le 1 + \frac{1}{\eta} \min\left\{ \frac{1-\gamma}{8\gamma}, \frac{1}{M} \right\}. \tag{3.7a}$$

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

$$T \ge \frac{c_T}{M(1-\gamma)^5\varepsilon^2} (\log((1-\gamma)^2\varepsilon))^2 \log \frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}, \tag{3.7b}$$

$$\eta = c_\eta M (1-\gamma)^4 \varepsilon^2 \frac{1}{\log \frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}}. \tag{3.7c}$$

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

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

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

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

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

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

Communication efficiency. One key feature of our federated setting is the use of periodic averaging with the hope to improve communication efficiency. According to (3.7a), our theory requires that the synchronization period τ be inversely proportional to the learning rate η, which suggests that more frequent communication is needed to compensate the discrepancy of local updates when the learning rate is large. To provide insights, consider the parameter regime when $M \gtrsim \frac{1}{1-\gamma}$ and $\varepsilon \lesssim \frac{1}{M(1-\gamma)^2}$. Plugging the choice of the learning rate (3.7c) into the upper bound of τ in (3.7a), we can choose the synchronization period as $\tau \asymp \frac{1}{M^2(1-\gamma)^4\varepsilon^2}$ up to logarithmic factors, leading to a communication complexity no larger than $C_{\mathsf{round}} = \frac{T}{\tau} \lesssim \frac{M}{1-\gamma}$, which is almost independent of the final accuracy ε.

3.2 Federated Synchronous Q-Learning for Average-Reward MDPs

Average-reward RL presents unique computational challenges because the optimal value function and policy depend on long-term average performance, which requires substantially more samples and training iterations to achieve optimality and places a significant

Previous works	Sample complexity	No. of agents	MDP class	Target error
Jin and Sidford (2021) (lower bound)	$\widetilde{\Omega}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert t_{\text{mix}}/\varepsilon^2\right)$	1	U	$\lVert J^{\widehat{\pi}} - J^\star\rVert_\infty$
Wang et al. (2022) (lower bound)	$\widetilde{\Omega}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\lVert h^\star\rVert_{\mathsf{sp}}/\varepsilon^2\right)$	1	W	$\lVert J^{\widehat{\pi}} - J^\star\rVert_\infty$
Bravo and Contreras (2024)	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\lVert h^\star\rVert_{\mathsf{sp}}{}^7/\varepsilon^7\right)$	1	W	$\lVert J^{\widehat{\pi}} - J^\star\rVert_\infty$
Lee et al. (2025)	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\lVert h^\star\rVert_{\mathsf{sp}}{}^2/\varepsilon^2\right)$	1	W	$\lVert J^{\widehat{\pi}} - J^\star\rVert_\infty$
Jin et al. (2024)	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert t_{\text{mix}}^8/\varepsilon^8\right)$	1	U	$\lVert J^{\widehat{\pi}} - J^\star\rVert_\infty$
Ours (Theorem 4)	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\lVert h^\star\rVert_{\mathsf{sp}}{}^5/M\varepsilon^5\right)$	M	W	$\lVert J^{\widehat{\pi}} - J^\star\rVert_\infty$
Jin et al. (2024)	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert t_{\text{mix}}^5/\varepsilon^5\right)$	1	U	$\lVert \widehat{Q} - J^\star\rVert_\infty$
Ours (Theorem 2)	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\lVert h^\star\rVert_{\mathsf{sp}}{}^3/\varepsilon^3\right)$	1	W	$\lVert \widehat{Q} - J^\star\rVert_\infty$
Ours (Theorem 3)	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\lVert h^\star\rVert_{\mathsf{sp}}{}^3/M\varepsilon^3\right)$	M	W	$\lVert \widehat{Q} - J^\star\rVert_\infty$

Table 3.2: The table summarizes the leading-order sample complexity of model-free algorithms for obtaining an ε-optimal value or policy with probability at least $1-\delta$ in the synchronous setting without prior knowledge, where $\widetilde{O}(\cdot)$ omits logarithmic factors. The 'MDP class' column indicates the applicable MDP type: (U) uniformly mixing or ergodic MDPs with finite mixing time t_{mix}; (W) weakly communicating MDPs with bias vector $h^\star$.

burden on a single agent. Federated learning addresses these challenges by distributing the computational and sampling load across multiple agents, enabling collaborative learning of a shared model without directly sharing local datasets at agents.

3.2.1 Single-Agent Settings

Before introducing the federated algorithm, we first consider Q-learning for average-reward MDPs in the single-agent setting and identify effective choices of the learning rate and discount factor. We begin with a brief overview of the algorithm, and then present its sample-complexity analysis. We focus on the synchronous sampling, where all state-action pairs are sampled uniformly assuming access to a generative model or a simulator (Kearns and Singh, 1999).

Synchronous Q-learning for average-reward RL. Q-learning (Watkins and Dayan, 1992) stands out as one of the most widely used model-free approaches, which directly estimates the Q-function without needing model estimation or prior knowledge. In this paper, we focus on a Q-learning approach to directly estimate the optimal value $J^\star$ for

average-reward RL. In the synchronous setting, Q-learning operates as follows: beginning with an initial Q-function Q_0, the algorithm updates the Q-function at each iteration $t \geq 1$ using the following rule for all $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$Q_t(s,a) = (1-\eta_t)Q_{t-1}(s,a) + \eta_t\bigg((1-\gamma_t)r(s,a) + \gamma_t \max_{a' \in \mathcal{A}} Q_{t-1}(s_t(s,a),a')\bigg), \tag{3.8}$$

where $s_t(s,a) \sim P(\cdot|s,a)$ represents an independent sample drawn for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, while η_t and γ_t correspond to the learning rate and discount factor at iteration t, respectively. A key distinction from traditional Q-learning in discounted MDPs—which applies a constant discount factor to future values—is that Q-learning for average-reward RL necessitates dynamic adjustment of discount factors when evaluating future values (Jin et al., 2024).

Algorithm. We consider a stage-wise Q-learning algorithm in which the discount factor varies dynamically across epochs. To be specific, in the t-th iteration of the k-th epoch, we maintain an estimate $Q_{k,t} \in \mathbb{R}^{S \times A}$ and update it using a variant of the standard Q-learning update rule for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with an epoch-dependent discount factor γ_k, sample size N_k, and the step-size $\eta_{k,t}$:

$$Q_{k,t}(s,a) = (1-\eta_{k,t})Q_{k,t-1}(s,a) + \eta_{k,t}\bigg((1-\gamma_k)r(s,a) + \gamma_k V_{k,\iota(k,t)}(s_{k,t}(s,a))\bigg), \tag{3.9}$$

where $s_{k,t}(s,a) \sim P(\cdot|s,a)$. The value function $V_{\iota(k,t)}(s_{k,t}(s,a)) \coloneqq \max_{a' \in \mathcal{A}} Q_{\iota(k,t)}(s_{k,t}(s,a),a')$ used in the update is calculated using a historical estimate $Q_{\iota(k,t)}$, where $\iota(k,t)$ denotes a historical index.

It is obvious that the choice of the parameters— including γ_k, N_k, $\eta_{k,t}$ and $\iota(k,t)$— has a significant influence on the algorithm's performance. In the following, we shall provide two groups of parameters that do not require prior knowledge of the mixing time or $\|h^\star\|_{\mathsf{sp}}$. Assume $N_0 = 0$. For some sufficiently large constant $c_N > 0$, two groups of parameters are as follows.

- The first group decays the discount factor $1-\gamma_k$ at an exponential rate, such that the

Algorithm 2: Average-reward Q-learning

Input: number of epochs K, discount factor $\{\gamma_k\}_{k=1}^K$, sample size $\{N_k\}_{k=0}^K$, step-size $\{\eta_{k,t}\}_{(k,t)=(1,1)}^{(K,N_k)}$, and historical index $\iota(k,t)$.

Output: average reward estimation Q_K.

Initialize $Q_{0,0} = 0$, $V_{0,0} = 0$.
for $k = 1, \cdots, K$ **do**
  Initialize $Q_{k,0} = Q_{k-1,N_{k-1}}$.
  for $t = 1, \cdots, N_k$ **do**
    Draw $s_{k,t}(s,a) \sim P(\cdot|s,a)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    Compute $Q_{k,t}(s,a)$ according to (3.9).
    Compute $V_{k,t}(s) = \max_{a' \in \mathcal{A}} Q_{k,t}(s,a')$ for all $s \in \mathcal{S}$.
  end for
end for

initial estimate $Q_{k,0}$ converges to the average reward $J^\star$ exponential fast. In each epoch, the step-size $\eta_{k,t}$ and the historical index $\iota(k,t)$ are chosen in a manner similar to those in the standard Q-learning algorithm, such that $Q_{k,t}$ converges to the γ_k-discounted optimal Q-function $Q^\star_{\gamma_k}$at a comparable rate. The sample size N_k is selected to ensure that Q_{k,N_k} sufficiently approximates $Q^\star_{\gamma_k}$.

$$N_k = c_N 2^k, \quad \gamma_k = 1 - \frac{2\log(4N_k)}{N_k^{1/3}},$$
$$\eta_{k,t} = \frac{1}{1 + \frac{t^{2/3}}{8\log(4t)}}, \quad \iota(k,t) = (k, t-1). \tag{3.10a}$$

- The second group decays the discount factor $1 - \gamma_k$ at a rate of $1/k$, such that the difference between $Q_{k,0}$ and $J^\star$ decreases at the same rate. In each epoch, the step-size $\eta_{k,t}$ and the historical index $\iota(k,t)$ are fixed, resulting in a convergence rate of $1/\sqrt{N_k}$ for $Q_{k,t}$. Thus the sample size N_k is chosen to be in the order of k^2:

$$N_k = c_N k^2 \log^5(k+1) \log^3\left(\frac{SA}{\delta}\right), \quad \gamma_k = \frac{k}{k+1},$$
$$\eta_{k,t} = \frac{1}{1 + \frac{N_k}{4\log(\sum_{i=1}^k N_i)}}, \quad \iota(k,t) = (k-1, N_{k-1}). \tag{3.10b}$$

The two parameter groups can offer distinct advantages in practice. For example,

constant learning rates are generally more adaptive and respond faster to environmental changes, whereas rescaled linear rates yield more stable estimates. Although our theoretical analysis assumes a fixed MDP, practical systems often exhibit slow temporal drift; in such cases, providing guarantees for both schedules gives practitioners greater flexibility.

Analysis: optimal value estimation The following theorem presents the sample complexity of Algorithm 2, under the mild assumption of a weakly communicating MDP. The proof is provided in Appendix C.2.

Theorem 2. *Assume* $\|h^\star\|_{\mathsf{sp}} < \infty$. *For both of the two groups of parameters in* (3.10), *with probability at least* $1-\delta$, *the output of Algorithm 2 satisfies*

$$\|Q_{K,N_K} - J^\star\|_\infty \le \frac{C\|h^\star\|_{\mathsf{sp}}}{(T_K)^{1/3}} \log^4 \frac{4SAT_k}{\delta},$$

where C *is a positive constant,* $T_K := \sum_{k=1}^K N_k$ *denotes the total number of iterations. For* $\varepsilon \in (0,1]$, *to achieve* $\|Q_{K,N_K} - J^\star\|_\infty \le \varepsilon$, *the total number of samples is bounded as*

$$SAT_K = \widetilde{O}\left(SA\frac{\|h^\star\|_{\mathsf{sp}}^3}{\varepsilon^3}\right).$$

Theorem 2 shows that the proposed Q-learning algorithm attains a sample complexity of $\widetilde{O}(SA\|h^\star\|_{\mathsf{sp}}^3/\varepsilon^3)$ for both parameter groups. This is reminiscent to the discounted setting (Li et al., 2024a), where different learning rate schedules, such as constant or rescaled linear rates, achieve the same optimal rate. Moreover, this result provides the best known sample complexity for Q-learning algorithms in average-reward RL, improving upon the bound in Jin et al. (2024) by a factor of $\widetilde{O}(\|h^\star\|_{\mathsf{sp}}^2/\varepsilon^2)$.

Previous studies have established stronger results by leveraging more sophisticated algorithmic techniques. For example, Jin and Sidford (2021) extended learning algorithms with optimal sample complexity in discounted MDPs to the average-reward setting, thereby achieving better bounds. Zhang and Xie (2023) refined Q-learning via the use of the upper confidence bound (UCB) technique and obtained an improved dependency on $\|h^\star\|_{\mathsf{sp}}$ and ε. Policy optimization methods adopted by Li et al. (2024b); Wang (2017) achieved a $\widetilde{O}(1/\varepsilon^2)$

rate. All of them required prior knowledge of the mixing time or $\|h^\star\|_{\mathsf{sp}}$. Moreover, Lee et al. (2025) established the state-of-the-art by incorporating the Halpern iteration with variance reduction, implemented via differential techniques with sample batching.

In contrast, our result focuses on the vanilla Q-learning algorithm, without variance reduction or UCB enhancements. The performance is constrained by the behavior of Q-learning in discounted MDPs, which is known to be suboptimal (Li et al., 2024a), explaining why our bound falls short of the optimal rate. Nevertheless, Q-learning remains a classical and fundamental algorithm, valued for its simplicity of implementation and practical applicability. A rigorous theoretical analysis of its behavior is therefore essential, and it provides foundation for subsequent analysis with variance reduction or UCB techniques.

3.2.2 Federated Settings

In this section, we develop federated variants of average-reward Q-learning involving M agents and analyze the sample complexity of the variants. We study a federated framework where M agents independently update their local Q-function estimates using their respective datasets or generative models, while periodically communicating with a central server to aggregate these local estimates into a global model. Through this collaborative approach, our objective is to efficiently learn the optimal average reward $J^\star$ or optimal policy $\pi^\star$ by leveraging the combined learning experience of all participating agents. We first begin with the federated variants of Q-learning in the synchronous setting with generative models, where all the state-action pairs are updated simultaneously at all agents. In the synchronous setting, at every epoch $k \in [K]$ and iteration $t \in [N_k]$, each agent $m \in [M]$ has access to a generative model, and generates a new sample

$$s_{k,t}^m(s,a) \sim P(\cdot|s,a) \tag{3.11}$$

for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ independently. We assume agents communicate their estimates to the central server only at iterations when communication is scheduled, and we denote the communication schedule at epoch k as $\mathcal{C}(k)$. Then, synchronous feder-

ated Q-learning proceeds according to the following steps.

1. *Local updates:* For each epoch k, local Q-functions at agents are initialized as $Q^m_{k,0} = Q_{k-1}$, where Q_{k-1} is a global Q-estimate computed in the previous epoch. Then, for each iteration $t \in [N_k]$, each agent independently updates *all* entries of its Q-estimate $Q^m_{k,t-1}$ to reach some *intermediate* estimate following the update rule:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^m_{k,t}(s,a) = (1-\eta_{k,t}) Q^m_{k,t-1}(s,a) + \eta_{k,t}\Big((1-\gamma_k) r(s,a) + \gamma_k V_{k,\iota(k,t)}(s^m_{k,t}(s,a))\Big), \tag{3.12}$$

where $s^m_t(s,a)$ is drawn according to (3.11), and $\eta_{k,t}$ and γ_k are the learning rate and discount rate scheduled for epoch k and iteration t at all agents $m \in [M]$. Here, $\iota(k,t)$ is the most recent iteration before t when agents communicate with the central server, where $\iota(k,t) = \max\{j \in \mathcal{C}(k) : j < t\}$ if such j exists, and $\iota(k,t) = 0$ otherwise.

2. *Global aggregation:* At each stage k, whenever $t \in \mathcal{C}(k)$ and communication is scheduled, all agents send their local Q-function estimates to the central server. The server then aggregates these by averaging the local Q-estimates $Q^m_{k,t}$ from all agents to form the updated global estimate $Q_{k,t}$:

$$\forall (s,a): \quad Q_{k,t}(s,a) = \frac{1}{M} \sum_{m=1}^{M} Q^m_{k,t}(s,a). \tag{3.13}$$

Accordingly, after aggregation, the global value function is updated as $V_{k,t}(s) = \max_a Q_{k,t}(s,a)$, and all agents synchronize their local Q-functions to this updated global Q-function. The complete procedure is summarized in Algorithm 3.

Assume $N_0 = 0$. In the federated setting with M agents, for some sufficiently large constant $c_N > 0$, two groups of parameters are given as below.

- The first group of parameters accelerates the decay of the discount factor $(1-\gamma_k)$ by a factor of $M^{1/3}$, thereby increasing the effective planning horizon and expediting the convergence of $Q^\star_{\gamma_k}$ to $J^\star$. Simultaneously, the learning rate $\eta^m_{k,t}$ decays more gradually, also by a factor of $M^{1/3}$, reflecting the variance reduction achieved through collaborative

Algorithm 3: Federated average-reward Q-learning

inputs: number of agents M, number of epochs K, discount factor $\{\gamma_k\}_{k=1}^K$, sample size $\{N_k\}_{k=0}^K$, step-size $\{\eta_{k,t}\}_{(k,t)=(1,1)}^{(K,N_k)}$.
Initialize $Q_{0,0}^m = 0$, $V_{0,0}^m = 0$ for all $m \in [M]$.
for $k = 1, \cdots, K$ **do**
  Initialize $Q_{k,0}^m = Q_{k-1,N_{k-1}}$, $V_{k,0}^m = V_{k-1,N_{k-1}}$ for all $m \in [M]$
  for $t = 1, \cdots, N_k$ **do**
    for $m \in [M]$ **do**
      Draw $s_t^m(s,a) \sim P(\cdot \,|\, s,a)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
      Compute $Q_{k,t}^m$ according to (3.12)
      Compute $V_{k,t}^m(s) = \max_{a' \in \mathcal{A}} Q_{k,t}^m(s,a')$ for all $s \in \mathcal{S}$
      if $t \in \mathcal{C}(k)$ **then**
        Compute global Q-estimates $Q_{k,t}$ via aggregation according to (3.13).
        Update global value estimates $V_{k,t}(s) = \max_a Q_{k,t}(s,a)$ for all $s \in \mathcal{S}$.
        Synchronize local Q-estimates $Q_{k,t}^m = Q_{k,t}$ for all $m \in [M]$.
      end if
    end for
  end for
end for
return: Q_{K,N_K}.

learning among agents. These adjustments jointly reduce both the bias between $Q_{\gamma_k}^\star$ and $J^\star$ and the error in the convergence of the global Q-estimates $Q_{k,t}$ to $Q_{\gamma_k}^\star$, and crucially, balance these two sources of error so that they decrease at the same rate as M increases, thereby improving overall convergence:

$$N_k = c_N 2^k, \quad \gamma_k = 1 - \frac{2\log(4MN_k)}{(MN_k)^{1/3}},$$
$$\eta_{k,t} = \frac{1}{1 + \frac{t^{2/3}}{8M^{1/3}\log(4Mt)}}, \quad \mathcal{C}(k) = \{i \le N_k | i \equiv 0 \pmod{g_k}\} \cup \{N_k\}, \tag{3.14a}$$

where $g_k \coloneqq \lceil \frac{-\log(1-\gamma_k)^2}{\eta_{k,N_k}} \rceil$.

- The second group retains the same rates for the discount factor $(1-\gamma_k)$ and learning rate $\eta_{k,t}^m$ as in the single-agent setting, but reduces the sample size N_k by a factor of M. This reduction is due to the improved convergence rate of $Q_{k,t}$, as the variance reduction from multiple agents allows the algorithm to achieve the same error level with M times

fewer samples:

$$N_k = c_N \max\left\{\frac{k^2}{M}\log^5(k+1)\log^3\left(\frac{SA}{\delta}\right), \log(Mk\log\frac{SA}{\delta})\right\}, \qquad \gamma_k = \frac{k}{k+1},$$
$$\eta_{k,t} = \frac{1}{1+\frac{N_k}{4\log(M\sum_{i=1}^{k}N_i)}}, \quad \mathcal{C}_k = \{N_k\}. \tag{3.14b}$$

Analysis: optimal value estimation The next theorem provides a theoretical analysis of Algorithm 3, characterizing both the sample complexity and the communication round requirements for estimating the optimal reward $J^\star$.

Theorem 3. *Assume $\|h^\star\|_{\mathsf{sp}} < \infty$. For both of the two groups of parameters in (3.14), with probability at least $1-\delta$, the output of Algorithm 3 satisfies*

$$\|Q_{K,N_K} - J^\star\|_\infty \le \frac{C\|h^\star\|_{\mathsf{sp}}}{(MT_K)^{1/3}}\log^4\frac{4MSAT_K}{\delta},$$

where C is a positive constant, $T_K \coloneqq \sum_{k=1}^{K} N_k$ denotes the total number of iterations. For $\varepsilon \in (0,1]$, to achieve $\|Q_{K,N_K} - J^\star\|_\infty \le \varepsilon$, the total number of samples required per agent is bounded as

$$SAT_K = \widetilde{O}\left(\frac{SA\|h^\star\|_{\mathsf{sp}}^3}{M\varepsilon^3}\right),$$

and the number of communication rounds can be bounded as

$$\sum_{k=1}^{K}|\mathcal{C}(k)| = \widetilde{O}\left(\frac{\|h^\star\|_{\mathsf{sp}}}{\varepsilon}\right).$$

The proof of Theorem 3 is provided in Appendix C.3. We provide a few remarks on the implications as follows

Sample complexity with linear speedup. Theorem 3 shows that federated collaboration among M agents accelerates the convergence of Q-function estimates to the optimal average reward $J^\star$. Specifically, compared to the single-agent case shown in Theorem 2, the sample complexity per agent is reduced by a factor of M, demonstrating a linear speedup in convergence as the number of agents increases. Remarkably, this means that the federated

algorithm achieves the same sample efficiency as if all data were sampled centrally, despite agents not sharing their individual datasets.

Communication efficiency. In federated settings, communication between agents and the central server can be costly, making it crucial to minimize the communication frequency. Theorem 3 indicates that the total number of communication rounds required to achieve an ε-optimal value is $\widetilde{O}(\|h^\star\|_{\mathsf{sp}}/\varepsilon)$, which is independent of the number of agents M. This means that adding more agents improves sample efficiency without increasing the frequency of communication rounds, allowing scalability in the number of agents without incurring additional communication costs.

Furthermore, the communication complexity bound holds for both parameter groups in (3.14), enabling flexible choices of communication schedules based on resource constraints without compromising communication efficiency. For the first group, communication is scheduled periodically within each epoch at fixed intervals $g_k = \widetilde{O}((\frac{N_k^2}{M})^{1/3})$, with N_k increasing exponentially. For the second group, communication occurs only at the end of each epoch, with the interval determined by the polynomially growing epoch size N_k. Depending on the available communication resources, communication rounds can be scheduled either at constant intervals with exponential jumps or at polynomially growing intervals.

Analysis: optimal policy learning We now analyze the problem of learning an ε-optimal policy, providing sample complexity and communication complexity for learning a policy π such that $J^\star - J^\pi \le \varepsilon$.

Theorem 4. *Assume $\|h^\star\|_{\mathsf{sp}} < \infty$. Take $N_k = c_N 2^k$ for sufficiently large c_N, $\gamma_k = 1 - \frac{1}{(N_k M)^{1/5}}$, $\eta_{k,t} = (1+\frac{t}{8(N_k M)^{1/5}\log(MN_k)})^{-1}$, and $\mathcal{C}(k) = \left\{ \left\lceil N_k \left(\frac{1+\gamma_k}{2}\right)^{i-1} \right\rceil \middle| i \in \left[\left\lceil \frac{4\log(1-\gamma_k)}{\log((1+\gamma_k)/2)} \right\rceil \right] \right\}$. Take the output policy of Algorithm 3 as $\widehat{\pi}(s) := \arg\max_a Q_{K,N_K}(s,a)$. Then with probability as least $1-\delta$, the average reward $J^{\widehat{\pi}}$ satisfies $J^\star - J^{\widehat{\pi}} \le \varepsilon$ as long as*

$$T_K = \sum_{k=1}^{K} N_k \gtrsim \frac{\|h^\star\|_{\mathsf{sp}}^5}{M\varepsilon^5} \log^5(N_K M) \log^{5/2} \frac{SAMT_K}{\delta}.$$

The number of communication rounds is bounded as $\sum_{k=1}^{K} |\mathcal{C}_k| = \widetilde{O}\left(\|h^\star\|_{\mathsf{sp}}/\varepsilon\right)$.

Theorem 4 shows that the algorithm can learn an ε-optimal policy with a sample complexity of $\widetilde{O}\big(\frac{SA\|h^\star\|_{\mathsf{sp}}^5}{M\varepsilon^5}\big)$ per agent, also achieving linear speedup in terms of the number of agents M. Notably, this improves the result of Jin et al. (2024) by a factor of $\widetilde{O}\big(\|h^\star\|_{\mathsf{sp}}^3/\varepsilon^3\big)$ in the single-agent case, where $M = 1$. The communication complexity remains $\widetilde{O}\big(\frac{\|h^\star\|_{\mathsf{sp}}}{\varepsilon}\big)$, independent of the number of agents M. The proof of Theorem 4 is provided in Appendix C.4.

Chapter 4

Collaborative Federated RL with Markovian Sampling

In this section, we study a federated asynchronous Q-learning, where M agents sample local trajectories using different behavior policies under infinite-horizon discounted MDPs.

4.1 Problem Setting

4.1.1 Background

Single-agent asynchronous Q-learning with Markovian trajectories In the asynchronous setting, an agent collects a sample trajectory $\{s_t, a_t, r_t\}_{t=0}^{\infty}$ from the underlying MDP $\mathcal{M}$ by following a stationary behavior policy π_{b} such that

$$a_t \sim \pi_{\mathsf{b}}(\cdot \,|\, s_t), \quad r_t = r(s_t, a_t), \quad s_{t+1} \sim P(\cdot \,|\, s_t, a_t) \tag{4.1}$$

for all $t \geq 0$, where the initial state is s_0. The behavior policy π_{b} may be different from the optimal policy $\pi^\star$. Contrary to the generative model considered in the synchronous setting, the samples collected under the asynchronous setting are Markovian (dependent across time). The sample trajectory can be viewed as a time-homogeneous Markov chain over the set of state-action pairs.

sampling	reference	number of agents	coverage	sample complexity
asynchronous	Qu and Wierman (2020)	1	full	$\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$
	Li et al. (2021)	1	full	$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$
	Li et al. (2023)	1	full	$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2}$
	FedAsynQ-EqAvg (Khodadadian et al., 2022)	M	full	$\frac{\lvert\mathcal{S}\rvert^2}{M\mu_{\mathsf{min}}^5(1-\gamma)^9\varepsilon^2}$
	FedAsynQ-EqAvg (Theorem 5)	M	partial	$\frac{C_{\mathsf{het}}}{M\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}$
	FedAsynQ-ImAvg (Theorem 6)	M	partial	$\frac{1}{M\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2}$

Table 4.1: Comparison of sample complexity upper bounds of single-agent and federated Q-learning algorithms under asynchronous sampling protocols to learn an ε-optimal Q-function in the ℓ_∞ sense, where logarithmic factors and burn-in costs are hidden. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, γ is the discount factor, M is the total number of agents, and t_{mix} is the mixing time of the behavior policy. In addition, $\mu_{\mathsf{min}} = \min_{m,s,a} \mu_{\mathsf{b}}^m(s,a)$ denotes the minimum entry of the stationary state-action occupancy distributions μ_{b}^m of all agents, $\mu_{\mathsf{avg}} := \min_{s,a} \frac{1}{M}\sum_{k=1}^M \mu_{\mathsf{b}}^m(s,a)$ denotes the minimum entry of the average stationary state-action occupancy distribution of all agents, and $C_{\mathsf{het}} := \max_{m,s,a} M\mu_{\mathsf{b}}^m(s,a)/\big(\sum_{m=1}^M \mu_{\mathsf{b}}^m(s,a)\big)$ captures the heterogeneity across the agents.

In asynchronous Q-learning for infinite-horizon MDPs, at each iteration $t \geq 1$, upon receiving a transition (s_{t-1}, a_{t-1}, s_t), the Q-estimate is updated via

$$Q_t(s,a) = \begin{cases} (1-\eta)Q_{t-1}(s,a) + \eta\left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_{t-1}(s_t,a')\right), & \text{if } (s,a) = (s_{t-1}, a_{t-1}), \\ Q_t(s,a) & \text{otherwise,} \end{cases} \tag{4.2}$$

where η denotes the learning rate.

4.1.2 Federated RL with Markovian Sampling

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

policy π_{b}^m such that

$$a_t^m \sim \pi_{\mathsf{b}}^m(\cdot|s_t^m), \quad r_t^m = r(s_t^m, a_t^m), \quad s_{t+1}^m \sim P(\cdot|s_t^m, a_t^m) \tag{4.3}$$

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

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

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

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

$$t_{\mathsf{mix}}^m := \min\left\{t \;\middle|\; \max_{(s_0,a_0)\in\mathcal{S}\times\mathcal{A}} d_{\mathsf{TV}}\big(P_t^m(\cdot \,|\, s_0, a_0),\, \mu_{\mathsf{b}}^m\big) \leq \frac{1}{4}\right\}, \tag{4.4}$$

where $P_t^m(\cdot \,|\, s_0, a_0)$ denote the distribution of (s_t, a_t) conditioned on (s_0, a_0) for agent m, and $d_{\mathsf{TV}}(\cdot, \cdot)$ is the total variation distance. Further, let the largest mixing time of all the

Markov chains induced by local behavior policies be

$$t_{\mathsf{mix}}^{\mathsf{max}} \coloneqq \max_{m\in[M]} t_{\mathsf{mix}}^{m}. \tag{4.5}$$

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

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

$$\mu_{\mathsf{min}} \coloneqq \min_{m\in[M]} \mu_{\mathsf{min}}^{m}, \qquad \text{where} \qquad \mu_{\mathsf{min}}^{m} \coloneqq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_{\mathsf{b}}^{m}(s,a). \tag{4.6}$$

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

$$\mu_{\mathsf{avg}} \coloneqq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{1}{M}\sum_{m=1}^{M} \mu_{\mathsf{b}}^{m}(s,a). \tag{4.7}$$

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

$$\mu_{\mathsf{avg}} = \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{1}{M}\sum_{m=1}^{M} \mu_{\mathsf{b}}^{m}(s,a) \geq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}, m\in[M]} \mu_{\mathsf{b}}^{m}(s,a) = \mu_{\mathsf{min}}. \tag{4.8}$$

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

$$C_{\mathsf{het}} \coloneqq \max_{m\in[M]} \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mu_{\mathsf{b}}^{m}(s,a)}{\frac{1}{M}\sum_{m=1}^{M} \mu_{\mathsf{b}}^{m}(s,a)}, \tag{4.9}$$

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

4.2 Algorithm: Federated Asynchronous Q-learning (FedAsynQ)

Similar to the synchronous setting, we describe a federated asynchronous Q-learning algorithm, called FedAsynQ (see Algorithm 4), that learns the optimal Q-function by periodically averaging the local Q-estimates with the aid of a central server. Inheriting the notation of Q_t^m and V_t^m from the synchronous setting (cf. (3.4)), FedAsynQ proceeds as follows in the rest of the t-th iteration.

1. *Local updates:* Each agent m samples a transition $(s_{t-1}^m, a_{t-1}^m, r_{t-1}^m, s_t^m)$ from its Markovian trajectory generated by the behavior policy π_{b}^m according to (4.3) and updates a *single* entry of its local Q-estimate Q_{t-1}^m:

$$Q_{t-\frac{1}{2}}^m(s,a) = \begin{cases} (1-\eta) Q_{t-1}^m(s,a) + \eta\big(r_{t-1}^m + \gamma V_{t-1}^m(s_t^m)\big) & \text{if } (s,a) = (s_{t-1}^m, a_{t-1}^m) \\ Q_{t-1}^m(s,a), & \text{otherwise} \end{cases}, \tag{4.10}$$

   where η denotes the learning rate.

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

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_t^m(s,a) = \begin{cases} \sum_{m=1}^M \alpha_t^m(s,a) Q_{t-\frac{1}{2}}^m(s,a) & \text{if } t \equiv 0 \pmod{\tau} \\ Q_{t-\frac{1}{2}}^m(s,a) & \text{otherwise} \end{cases}, \tag{4.11}$$

Algorithm 4: Federated Asynchronous Q-learning (FedAsynQ)

inputs: learning rate $\{\eta\}$, discount factor γ, number of agents M, synchronization period τ, total number of iterations T.
initialization: $Q_0^m = Q_0$ for all $m \in [M]$.
for $t = 1, \cdots, T$ **do**
  for $m \in [M]$ **do**
    Draw action $a_{t-1}^m \sim \pi_{\mathsf{b}}^m(s_{t-1}^m)$, observe reward $r_{t-1}^m = r(s_{t-1}^m, a_{t-1}^m)$, and draw next state $s_t^m \sim P(\cdot \,|\, s_{t-1}^m, a_{t-1}^m)$.
    Compute $Q_{t-\frac{1}{2}}^m$ according to (4.10).
    Compute Q_t^m according to (4.11).
  end for
end for
return: $Q_T(s,a) = \sum_{m=1}^M \alpha_T^m(s,a) Q_T^m(s,a)$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

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

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

After a total of T iterations, FedAsynQ outputs a global Q-estimate

$$Q_T(s,a) = \sum_{m=1}^M \alpha_T^m(s,a) Q_T^m(s,a)$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. In the subsections below, we provide two possible ways (equal and importance weighting) to choose α_t^m and their corresponding sample complexity analyses.

4.3 Performance Guarantees with Equal Averaging

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

$$\alpha_t^m(s,a) = \frac{1}{M}. \tag{4.12}$$

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

Theorem 5 (Finite-time convergence of FedAsynQ-EqAvg)**.** *Consider any given* $\delta \in (0,1)$ *and* $\varepsilon \in (0, \frac{1}{1-\gamma}]$. *Suppose that the initialization of* FedAsynQ-EqAvg *satisfies* $0 \le Q_0 \le \frac{1}{1-\gamma}$. *There exist some sufficiently large constant* $c_T > 0$ *and sufficiently small constant* $c_\eta > 0$, *such that with probability at least* $1-\delta$, *the output of* FedAsynQ-EqAvg *satisfies* $\|Q_T - Q^\star\|_\infty \le \varepsilon$, *provided that the synchronization period* τ, *the sample size per agent* T, *and the learning rate* η *satisfy*

$$\tau_0 \le \tau \le \frac{1}{4\eta}\min\left\{\frac{1-\gamma}{4}, \frac{1}{M}\right\}, \tag{4.13a}$$

$$T \ge c_T\left(\frac{C_{\mathsf{het}}}{M\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2} + T_0\right)(\log((1-\gamma)^2\varepsilon))^2 \log(TM)\log\frac{|\mathcal{S}||\mathcal{A}|T^2M}{\delta}, \tag{4.13b}$$

$$\eta = c_\eta \min\left\{\frac{M(1-\gamma)^4\varepsilon^2}{C_{\mathsf{het}}}, \eta_0\right\}\frac{1}{\log(TM)\log\frac{|\mathcal{S}||\mathcal{A}|T^2M}{\delta}}, \tag{4.13c}$$

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

Theorem 5 implies that to achieve an ε-accurate estimate (in the ℓ_∞ sense), the sample complexity per agent of FedAsynQ-EqAvg is no more than

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

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

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

$\widetilde{O}\Big(\frac{|\mathcal{S}|^2}{M\mu_{\mathsf{min}}^5(1-\gamma)^9\varepsilon^2}\Big)$ obtained in Khodadadian et al. (2022) by at least a factor of

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

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

4.4 Performance Guarantees with Importance Averaging

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

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

FedAsynQ-ImAvg with weights given by

$$\alpha_t^m(s,a) = \frac{(1-\eta)^{-N_{t-\tau,t}^m(s,a)}}{\sum_{m'=1}^M (1-\eta)^{-N_{t-\tau,t}^{m'}(s,a)}} \tag{4.14}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $m \in [M]$, where $N_{t-\tau,t}^m(s,a)$ represents the number of iterations between $[t-\tau, t)$ when the agent m visits (s,a). The weights in (4.14) can be calculated at the server based on the number of visits to each state-action pair by the agents in one synchronization period. Therefore, each agent needs to send its $N_{t-\tau,t}^m(s,a)$ for each (s,a) along with its local Q-estimate, and FedAsynQ-ImAvg incurs twice the communication cost of FedAsynQ-EqAvg per iteration.

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

Theorem 6 (Finite-time convergence of FedAsynQ-ImAvg)**.** *Consider any given $\delta \in (0,1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$. Suppose that the initialization of* FedAsynQ-ImAvg *satisfies $0 \le Q_0 \le \frac{1}{1-\gamma}$, and the synchronization period τ obeys*

$$\tau \le \frac{1}{4\eta} \min\left\{\frac{1-\gamma}{4}, \frac{1}{M}\right\}. \tag{4.15a}$$

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

$$T \ge c_T \left(\frac{1}{M\mu_{\mathsf{avg}}(1-\gamma)^5\varepsilon^2} + \widetilde{T}_0\right) (\log((1-\gamma)^2\varepsilon))^2 \log(TM) \log\frac{|\mathcal{S}||\mathcal{A}|T^2M}{\delta}, \tag{4.15b}$$

$$\eta = c_\eta \min\left\{M(1-\gamma)^4\varepsilon^2, \widetilde{\eta}_0\right\} \frac{1}{\log(TM)\log\frac{|\mathcal{S}||\mathcal{A}|T^2M}{\delta}}, \tag{4.15c}$$

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

Theorem 6 implies that to achieve an ε-accurate estimate (in the ℓ_∞ sense), the sample complexity per agent of FedAsynQ-ImAvg is no more than

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

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

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

Communication efficiency. To provide further insights on the communication complexity of FedAsynQ-ImAvg, consider again the regime when ε is sufficiently small and $M \gtrsim \frac{1}{1-\gamma}$. To minimize the communication frequency while preserving the sample efficiency, we again plug the choice of the learning rate (4.15c) into (4.15a) and select the synchronization period as large as $\tau \asymp \frac{1}{M^2(1-\gamma)^4\varepsilon^2}$ up to logarithmic factors. Then, this ensures the communication complexity $C_{\mathsf{round}} = T/\tau$ is no more than $\widetilde{O}\Big(\frac{M}{\mu_{\mathsf{avg}}(1-\gamma)}\Big)$.

4.5 Numerical Experiments

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

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

of transitioning to the other state as $P(1-s|s,a) = 1 - P(s|s,a)$ for each $s \in \mathcal{S}$.

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

Faster convergence of FedAsynQ-ImAvg**.** Figure 4.1 shows the normalized Q-estimate error $(1-\gamma)\|Q_T - Q^\star\|_\infty$ with respect to the sample size T, with $M = 20$ and $\tau = 50$. Given the trajectories of agents collectively cover the entire state-action space, the global Q-estimates of both FedAsynQ-EqAvg and FedAsynQ-ImAvg converge to the optimal Q-function, yet at different speeds. Although FedAsynQ-EqAvg converges in the end, we can see that it converges much slower compared to FedAsynQ-ImAvg, because each entry of the Q-function is trained by only one agent while the other $M-1$ agents never contribute useful information. However, the vacuous values of the $M-1$ agents significantly slow down the global convergence under equal averaging.

Convergence speedup. Figure 4.2 demonstrates the impact of the number of agents on the convergence speed of FedAsynQ-EqAvg and FedAsynQ-ImAvg. It can be observed that there is indeed a speedup in terms of the number of agents M with respect to the squared ℓ_∞ error $\|Q_T - Q^\star\|_\infty^{-2}$, which is poised to scale linearly with respect to the number of agents. In particular, the speedup is more rapid with FedAsynQ-ImAvg as M increases, while it increases much slower with FedAsynQ-EqAvg. This shows that FedAsynQ-ImAvg achieves much better convergence speedup in terms of the number of agents.

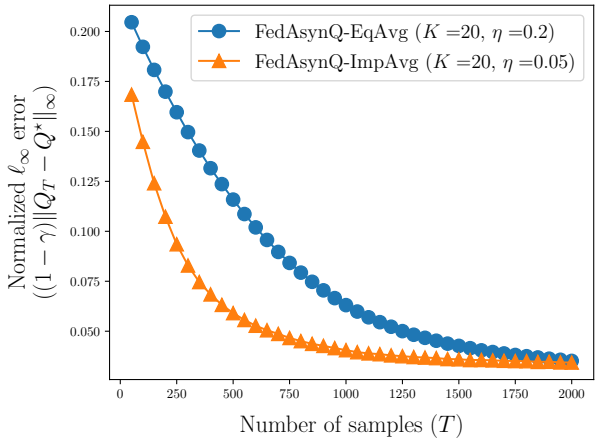


Figure 4.1: The normalized ℓ_∞ error of the Q-estimates $(1-\gamma)\|Q_T - Q^\star\|_\infty$ with respect to the number of samples T for both FedAsynQ-EqAvg and FedAsynQ-ImAvg, with $M = 20$ and $\tau = 50$. Here, the learning rates of FedAsynQ-ImAvg and FedAsynQ-EqAvg are set as $\eta = 0.05$ and $\eta = 0.2$, where each algorithm converges to the same error floor at the fastest speed, respectively.

Communication efficiency. Figure 4.3 demonstrates the impact of the synchronization period τ on the convergence of FedAsynQ-ImAvg and FedAsynQ-EqAvg. With frequent averaging ($\tau = 1$), FedAsynQ-ImAvg slightly outperforms FedAsynQ-EqAvg, but there is no significant difference because the heterogeneity between local Q-functions after just one local update is very small. The performance of FedAsynQ-EqAvg degrades as we increase τ since FedAsynQ-EqAvg cannot cope with the increased heterogeneity between local Q-estimates as we increase the number of local steps. On the other end, the performance of FedAsynQ-ImAvg improves first (i.e., $\tau = 10,\ 25,\ 50$) as it balances the heterogeneity much better than FedAsynQ-EqAvg, but drops later if τ is too large (i.e., $\tau = 75,\ 100$) due to the high variance of the averaged Q-estimates.

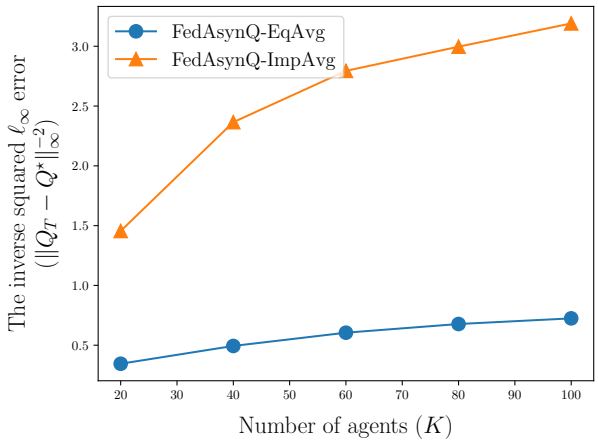


Figure 4.2: The inverse squared ℓ_∞ error $\|Q_T - Q^\star\|_\infty^{-2}$ with respect to the number of agents $M = 20, 40, 60, 80, 100$ for both FedAsynQ-EqAvg and FedAsynQ-ImAvg, with $T = 300$ and $\tau = 50$.

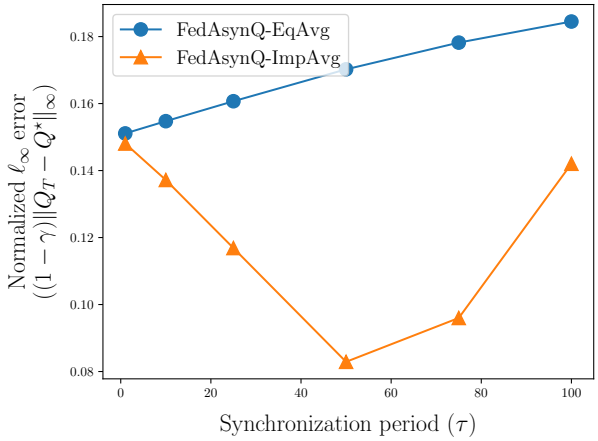


Figure 4.3: The normalized ℓ_∞ error of the Q-estimates $(1-\gamma)\|Q_T - Q^\star\|_\infty$ with respect to the synchronization period $\tau = 1, 10, 25, 50, 75, 100$ for both FedAsynQ-EqAvg and FedAsynQ-ImAvg, with $M = 20$ and $T = 300$.

Chapter 5

Collaborative Federated RL with Offline Data

We study a federated offline Q-learning that effectively elicits the collaborative benefit of agents to overcome limited coverage of local offline datasets at each agent.

5.1 Problem Setting

5.1.1 Background

Single-agent offline Data In offline RL, an agent has access to a offline dataset containing pre-collected episodes by following some behavior policy. The offline dataset $\mathcal{D}$ at an agent is composed of K episodes, each generated independently according to a behavior policy $\mu = \{\mu_h\}_{h=1}^H$, resulting in

$$\mathcal{D} \coloneqq \left\{ \left(s_{k,1},\, a_{k,1},\, r_{k,1},\, \ldots,\, s_{k,H},\, a_{k,H},\, r_{k,H} \right) \right\}_{k=1}^K,$$

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

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

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., 2022a)	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 7)	M	collaborative	$\frac{H^7SC^\star_{\mathsf{avg}}}{M\varepsilon^2}$	H

Table 5.1: 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. (5.3) and (5.4)), 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.

suboptimality gap

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

using the history dataset $\mathcal{D}$.

Pessimistic Q-learning (Shi et al., 2022). To this end, there exist offline Q-learning algorithm that update the Q-function estimates based on the offline dataset $\mathcal{D}$ by introducing pessimism to handle distributional shifts between the learned policy and the behavior policy μ. For $(s, a, r, s') = (s^m_{k,h}, a^m_{k,h})$ sampled from the offline dataset $\mathcal{D}$ for episode $k \in [K]$

and step $h \in [H]$, the pessimistic Q-learning update is given by

$$Q_{k,h}(s,a) \leftarrow (1-\eta_{k,h}(s,a))Q^m_{k-1,h}(s,a) + \eta_{k,h}(s,a)(r + V_{k-1,h+1}(s') - b_{k,h}(s,a)) \qquad (5.1)$$

where $\eta^m_{k,h}(s,a)$ is the learning rate, $V_{k,h}(s) \leftarrow \max\Big\{V_{k-1,h}(s),\, \max_a Q_h(s,a)\Big\}$, and the penalty term $b_{k,h}(s,a) > 0$ reflects the uncertainty of the corresponding Q-estimate and implements pessimism in the face of uncertainty.

5.1.2 Federated Offline RL

In federated 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 $\mu^m = \{\mu^m_h\}_{h=1}^H$, resulting in

$$\mathcal{D}^m \coloneqq \Big\{\big(s^m_{k,1},\, a^m_{k,1},\, r^m_{k,1},\, \ldots, s^m_{k,H},\, a^m_{k,H},\, r^m_{k,H}\big)\Big\}_{k=1}^K,$$

where the initial state $s^m_{k,1} \sim \rho$ is drawn from some initial state distribution $\rho \in \Delta(\mathcal{S})$, and $s^m_{k,h}, a^m_{k,h}, r^m_{k,h}$ are the state, action and reward at step h in the k-th episode, $a^m_{k,h} \sim \mu^m_h(\cdot \,|\, s^m_{k,h})$ and $r^m_{k,h} = r_h(s^m_{k,h}\, a^m_{k,h})$.

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

$$V^\star_1(\rho) - V^{\widehat{\pi}}_1(\rho) \le \varepsilon$$

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

[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.

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

$$d_h^\pi(s) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi), \qquad d_h^\pi(s,a) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi)\, \pi_h(a \,|\, s). \tag{5.2}$$

Recent works (Rashidinejad et al., 2021; Shi et al., 2022; Xie et al., 2021b) 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. (2022a) offered a more refined notion called *single-policy clipped concentrability*, defined as follows.

Definition 3 (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*

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a),\, 1/S\}}{d_h^\mu(s,a)} \le C^\star, \tag{5.3}$$

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

$$d_h^{\mathsf{avg}}(s) = \frac{1}{M}\sum_{m=1}^{M} d_h^m(s) \quad \text{and} \quad d_h^{\mathsf{avg}}(s,a) = \frac{1}{M}\sum_{m=1}^{M} d_h^m(s,a). \tag{5.4}$$

Definition 4 (average single-policy clipped concentrability)**.** *The average single-policy concentrability coefficient $C^\star_{\mathsf{avg}} \in [1/S, \infty)$ of multiple behavior policies $\{\mu^m\}_{m\in[M]}$ is defined to be the smallest quantity that satisfies*

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a), 1/S\}}{d_h^{\mathsf{avg}}(s,a)} \le C^\star_{\mathsf{avg}}, \tag{5.5}$$

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 $C^\star_{\mathsf{avg}} < \infty$ is finite. Therefore, this is much weaker than the coverage requirement in the single-agent case.

5.2 Algorithm: Federated Lower Confidence Bound Q-learning (FedLCB-Q)

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. 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 $\mathcal{C}(K)$, 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 $m \in [M]$

Algorithm 5: Federated pessimistic Q-learning (FedLCB-Q)

Parameters: horizon length H, number of agents M, total number of episodes per agent K, synchronization schedule $\mathcal{C}(K)$, 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 $k \in \mathcal{C}(K)$ **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 broadcasts $Q_{k,h}$, $V_{k,h}$ 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**

$Q^m_{k,h}(s,a) = Q_{k,h}(s,a)$, $V^m_{k,h}(s) = V_{k,h}(s)$

return: $\widehat{Q} = \{Q_{K,h}\}_{h\in[H]}$ and $\widehat{\pi} = \{\pi_{K,h}\}_{h\in[H]}$.

and the global estimates ($Q_{0,h}$ and $V_{0,h}$) at the server as follows:

$$Q^m_{0,h}(s,a) = 0, \qquad V^m_{0,h}(s,a) = 0, \qquad \text{for all } (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H+1], \tag{5.6a}$$

$$Q_{0,h}(s,a) = 0, \qquad V_{0,h}(s,a) = 0, \qquad \text{for all } (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H+1]. \tag{5.6b}$$

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 $Q^m_{k,h}$ as follows:

$$Q^m_{k,h}(s,a) = \begin{cases} (1-\eta^m_{k,h}(s,a))Q^m_{k-1,h}(s,a) & \\ \qquad + \eta^m_{k,h}(s,a)(r^m_{k,h} + V^m_{k-1,h+1}(s^m_{k,h+1})) & \text{if } (s,a) = (s^m_{k,h}, a^m_{k,h}) \\ Q^m_{k-1,h}(s,a) & \text{otherwise} \end{cases}, \tag{5.7}$$

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

$$V^m_{k-1,h}(s) = V^m_{\iota(k),h}(s) = V_{\iota(k),h}(s), \quad \text{for all } (m,s,h,k) \in [M] \times \mathcal{S} \times [H+1] \times [K], \tag{5.8}$$

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

$$\iota(k) \coloneqq \max_{k'} \{1 \le k' < k : k' \in \mathcal{C}(K)\}.$$

2. **Pessimistic aggregation:** If synchronization is scheduled at episode k, i.e., $k \in \mathcal{C}(K)$, 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 as follows:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_{k,h}(s,a) = \left(\sum_{m=1}^{M} \alpha^m_{k,h}(s,a) Q^m_{k,h}(s,a) \right) - B_{k,h}(s,a), \tag{5.9}$$

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, the global value estimate is updated as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad V_{k,h}(s) = \max \left\{ V_{\iota(k),h}(s), \max_{a\in\mathcal{A}} Q_{k,h}(s,a) \right\}. \tag{5.10}$$

Algorithm 6: `Local-Q-learning` (agents)

1: **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**
      $Q^m_{k,h}(s,a) = Q^m_{k-1,h}(s,a)$, $V^m_{k,h}(s) = V^m_{k-1,h}(s)$
    `// Update the local counters and learning rates`
1   $n^m_{k,h}(s^m_{k,h}, a^m_{k,h}) = n^m_{k-1,h}(s^m_{k,h}, a^m_{k,h}) + 1$
2   $\eta^m_{k,h}(s^m_{k,h}, a^m_{k,h}) = \frac{M(H+1)}{N_{\iota(k),h}(s^m_{k,h}, a^m_{k,h}) + M(H+1) n^m_{k,h}(s^m_{k,h}, a^m_{k,h})}$
    `// Update local Q-estimates`
3   $Q^m_{k,h}(s^m_{k,h}, a^m_{k,h}) =$
      $\big(1 - \eta^m_{k,h}(s^m_{k,h}, a^m_{k,h})\big) Q^m_{k-1,h}(s^m_{k,h}, a^m_{k,h}) + \eta^m_{k,h}(s,a)(r^m_{k,h} + V^m_{k-1,h+1}(s^m_{k,h+1}))$

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

$$\forall (k,m) \in \mathcal{C}(K) \times [M]: \qquad Q^m_{k,h} = Q_{k,h}, \quad V^m_{k,h} = V_{k,h}. \tag{5.11}$$

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., $K \in \mathcal{C}(K)$.

5.3 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., $n^m_{k,h}(s,a) \coloneqq |\{\iota(k) < i \leq k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.

- $N^m_{k,h}(s,a)$: the number of episodes in the interval $[1,k]$ during which agent m visits (s,a) at step h, i.e., $N^m_{k,h}(s,a) \coloneqq |\{1 \leq i \leq k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.
- $n_{k,h}(s,a)$: the total number of episodes in the interval $(\iota(k),k]$ during which all agents visit (s,a) at step h, i.e., $n_{k,h}(s,a) \coloneqq \sum_{m=1}^M n^m_{k,h}(s,a) = |\{\iota(k) < i \leq k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.
- $N_{k,h}(s,a)$: the total number of episodes in the interval $[1,k]$ during which all agents visit (s,a) at step h, i.e., $N_{k,h}(s,a) \coloneqq \sum_{m=1}^M N^m_{k,h}(s,a) = |\{1 \leq i \leq k : (s^m_{i,h}, a^m_{i,h}) = (s,a)\}|$.

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 $k \in \mathcal{C}(K)$:

$$B_{k,h}(s,a) \coloneqq \begin{cases} \frac{(H+1)n_{k,h}(s,a)}{N_{k,h}(s,a)+Hn_{k,h}(s,a)} \sqrt{\frac{c_B \zeta_1^2 H^4}{N_{k,h}(s,a)}} & \text{if } \; N_{k,h}(s,a) > 0 \\ 0 & \text{if } \; N_{k,h}(s,a) = 0 \end{cases}, \tag{5.12}$$

where $\zeta_1 = \log\left(\frac{SAMK^2H}{\delta}\right)$ 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:

$$\alpha_{k,h}^m(s,a) := \begin{cases} \frac{1}{M}\frac{N_{\iota(k),h}(s,a)+(H+1)Mn_{k,h}^m(s,a)}{N_{k,h}(s,a)+Hn_{k,h}(s,a)} & \text{if } \; n_{k,h}(s,a) > 0 \\ \frac{1}{M} & \text{if } \; n_{k,h}(s,a) = 0 \end{cases}. \tag{5.13}$$

  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 $n^m_{k,h}$:

$$\eta^m_{k,h}(s,a) \coloneqq \frac{M(H+1)}{N_{\iota(k),h}(s,a) + M(H+1)n^m_{k,h}(s,a)}. \tag{5.14}$$

  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.

The computation of the global penalty (5.12) and importance averaging (5.13) at a server requires local counters $n^m_{k,h}(s,a)$ from every agent, and determining the learning rates (5.14) 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.

5.4 Theoretical Guarantees

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

Theorem 7. *Consider $\delta \in (0,1)$ and let $\widehat{\pi}$ be the solution policy of* FedLCB-Q*. If a syn-*

chronization schedule $\mathcal{C}(K)$ is independent of trajectories in datasets $\mathcal{D}$ and satisfies

$$\tau_1 \leq \sqrt{\frac{H^2 S C^\star_{\mathsf{avg}} K}{M}} \quad \textit{and} \quad \frac{\tau_{u+1}}{\tau_u} \leq 1 + \frac{2}{H} \tag{5.15}$$

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:

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \leq c\left(\sqrt{\frac{H^7 S C^\star_{\mathsf{avg}} \zeta_1^2}{MT}} + \frac{H^4 S C^\star_{\mathsf{avg}} \zeta_1}{MT}\right) \tag{5.16}$$

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

Theorem 7 implies that as long as the initial synchronization occurs early and the synchronization intervals do not increase too rapidly (cf. (5.15)), 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

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

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 $C^\star_{\mathsf{avg}}$ (cf. (5.5)) 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. (5.3)), 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 $\widetilde{O}\Big(\frac{H^6SC^\star}{\varepsilon^2}\Big)$ 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., 2022a), as if we collect all the agents' datasets at a central location. Note that the effective single-policy concentrability coefficient (cf. (5.3)) for the combined datasets $\mathcal{D}_{\mathsf{all}} = \cup_{m=1}^{M} \mathcal{D}^m$ becomes

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a),\, 1/S\}}{\sum_{m=1}^{M} d_h^m(s,a)} = \max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{\min\{d_h^{\pi^\star}(s,a),\, 1/S\}}{M d_h^{\mathsf{avg}}(s,a)} = \frac{C^\star_{\mathsf{avg}}}{M}, \quad (5.17)$$

leading to the minimax lower bound (Li et al., 2022a)

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

Comparing with the sample complexity bound of FedLCB-Q, obtained as $\widetilde{O}\Big(\frac{H^7SC^\star_{\mathsf{avg}}}{M\varepsilon^2}\Big)$, 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.

5.5 Near-Optimal Communication Efficiency

In federated settings, communication with a central server is often the dominant bottleneck. We therefore explore protocols that reduce the number of communication rounds while preserving near-optimal sample performance. We consider two schemes: periodic aggregation with a fixed interval, and adaptive schedules that progressively lengthen synchronization intervals as global estimates stabilize.

Theorem 7 suggests initiating the first synchronization early and avoiding rapid increases in synchronization intervals (cf. (5.15)) to ensure fast convergence. This is at-

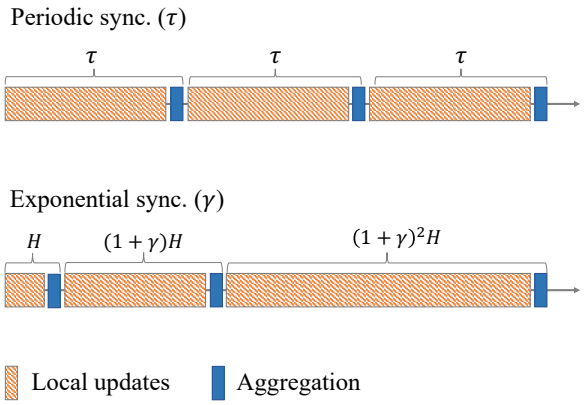


Figure 5.1: Illustration of the periodic synchronization with constant period τ and the exponential synchronization with a rate γ.

tributed 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 is available for every τ episodes, i.e., $\tau_i = \tau$ for all $i \geq 1$, and we denote the synchronization schedule as $\mathcal{C}_{\mathsf{period}}(K, \tau)$.

(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 $\mathcal{C}_{\mathsf{exp}}(K, \gamma)$.

Now, we analyze the number of communication rounds required to achieve a target accuracy, for each scheduling scheme.

Corollary 1. *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.,* $\mathcal{C}(K) = \mathcal{C}_{\mathsf{period}}(K, \tau)$, *with* $\tau \asymp \sqrt{\frac{HSC_{\mathsf{avg}}^{\star} T}{M}}$, *or the exponential synchronization scheduling, i.e.,* $\mathcal{C}(K) = \mathcal{C}_{\mathsf{exp}}(K, \gamma)$, *with* $\gamma = \frac{2}{H}$. *Then, each schedule requires the number of synchronizations at most*

$$\text{(Periodic)} \quad |\mathcal{C}_{\mathsf{period}}(K, \tau)| \lesssim \sqrt{\frac{MK}{H^2 SC_{\mathsf{avg}}^{\star}}}, \tag{5.18a}$$

$$\text{(Exponential)} \quad |\mathcal{C}_{\mathsf{exp}}(K, \gamma)| \lesssim H, \tag{5.18b}$$

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

Corollary 1 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, this matches the lower bound on communication rounds established by Salgia and Chi (2024), demonstrating the communication-efficiency optimality of our communication protocols in federated Q-learning. 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 (5.15) due to which it necessitates more communication rounds, which increase with K and M.

Chapter 6

Personalized Federated RL with Heterogeneous Environments

In real-world federated reinforcement learning (RL), agents often operate with highly diverse goals and reward structures. Traditionally, such heterogeneity has been viewed as a significant hurdle, a source of *client drift* that hinders consensus or a conflict of interest that necessitates performance compromises. Consequently, existing literature primarily focuses on either mitigating this divergence to learn a shared global policy (Wang et al., 2024a; Yang et al., 2024b) or developing personalization techniques that treat heterogeneity as a challenge to be tackled.

In this chapter, we propose a fundamental shift in this perspective: reward heterogeneity is not a hurdle, but a structural catalyst for exploration. To preserve individual utility and avoid negative transfer, we develop a personalized federated reinforcement learning framework that explicitly decouples the learning of shared transition dynamics from personalized reward functions. As a foundational step, we introduce Personalized Federated Upper Confidence Bound Value Iteration (PF-UCBVI). By globally aggregating transition counts while strictly isolating local reward estimates and personalized exploration bonuses, PF-UCBVI allows agents to collaborate efficiently even when pursuing arbitrarily different goals. This approach achieves a provably optimal linear speedup in sample efficiency, demonstrating that collaborative exploration is highly effective in the presence of reward

heterogeneity.

Despite the theoretical guarantees of UCB-based methods, explicit exploration is often risky, unethical, or prohibitively expensive in practical domains such as robotics or clinical trials. If a population of agents possesses sufficiently diverse objectives, their collective, purely greedy executions will naturally span the state-action space. Leveraging this insight, we propose Personalized Federated Exploration-Free Value Iteration (PF-EFVI). We demonstrate that under a novel reward-diversity condition, asymptotically optimal collective learning can be achieved relying entirely on pure greedy exploitation, completely eliminating the need for explicit exploratory actions.

6.1 Preliminaries

To situate our work within the broader context of reinforcement learning and federated optimization, we review two critical research thrusts: heterogeneous federated RL and exploration-free RL.

Federated RL with Heterogeneous Environments. While the majority of federated RL literature assumes homogeneous environments (Salgia and Chi, 2024; Woo et al., 2025; Zheng et al., 2024), recent works have begun to explore the more realistic setting where agents possess different transition dynamics and reward functions. Existing literature in this area primarily focuses on two directions:

1. Consensus-based methods: These algorithms (Labbi et al., 2024; Wang et al., 2024a; Yang et al., 2024b; Zhang et al., 2024) aim to learn a shared global policy that optimizes the average values across all environments. However, this approach often sacrifices individual agent utility to reach a global consensus.
2. Personalized methods: These algorithms allow agents to learn individualized policies tailored to their local environments. For instance, Zhang and Azizan (2026) proved linear speedup via affinity-based variance reduction under similarity conditions, and Xiong et al. (2025) proposed a personalized federated TD-learning algorithm that

achieves linear speedup with shared representation.

Despite these advances, most personalized algorithms assume a degree of similarity between agents or rely on a fixed behavior policy that ensures sufficient coverage of the state-action space.

Exploration-free Learning and Covariate Diversity. In practice, many RL applications avoid explicit exploration in favor of a greedy approach, particularly in domains such as robotics or clinical trials where explorative actions may be risky, unethical, or prohibitively expensive. Beyond safety concerns, explorative solutions are often harder to implement, and their outcomes are less predictable or interpretable. Consequently, common sense often suggests a greedy strategy in real-world scenarios, even when standard theory dictates the necessity of exploration.

Foundational studies by Bastani et al. (2021) and Kannan et al. (2018) have investigated this phenomenon in linear contextual bandits. They established that under a sufficient diversity of contexts, known as the covariate-diversity assumption, a purely greedy selection can achieve sublinear regret without any explicit exploration mechanisms. Recently, Civitavecchia and Papini (2025) further characterized the covariate-diversity condition for reinforcement learning. However, their results still require strong environment stochasticity to ensure state-action coverage, which remains a stringent requirement for a single agent.

6.2 Problem Setting

We consider M agents operating in M finite-horizon MDPs $\{\mathcal{M}^m\}_{m=1}^M$. Each agent m is characterized by a local MDP $\mathcal{M}^m = \big(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h^m\}_{h=1}^H\big)$, where agents share a common transition kernel P_h but pursue distinct, personalized reward functions r_h^m. In online RL, for K episodes, agents interact with their local environments using their behavior policies updated based on the data collected previously. At episode k, each agent m selects a policy π_k^m based on the data collected from previous episodes and executes π_k^m to generate a trajectory $\{s_{k,h}^m, a_{k,h}^m, r_{k,h}^m\}_{h=1}^H$, where the initial state $s_{k,1}^m$ is drawn from some

initial state distribution $\rho \in \Delta(\mathcal{S})$.

Unlike prior works that sacrifice individual utility for global consensus (Labbi et al., 2024; Zhang et al., 2024), the goal of each agent is to learn their personalized optimal policy $\pi^{m,\star}$ maximizing its local value function $V_1^{m,\star}(s_1)$, while the overall system aims to minimize the total regret across all agents over K episodes:

$$\mathsf{Regret}(K) = \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbb{E}_{s_1 \sim \rho} \left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1) \right] \tag{6.1}$$

where π_k^m is the policy of agent m at episode k based on its local reward and the current global estimate of the transition kernel.

6.3 Algorithm: Personalized Federated Upper Confidence Bound Value Iteration (PF-UCBVI)

To solve the personalized regret minimization problem defined in the previous section, we must leverage the shared structure of the environments without forcing a single global objective. Since the transition kernel P_h is common across all agents, sharing state-action visitation data allows the ensemble to estimate the dynamics significantly faster than any single agent could alone.

However, because each agent pursues a distinct reward function r_h^m, their respective value functions, and consequently, their uncertainty about future returns, differ significantly. A standard federated approach with a single global exploration bonus would be heavily suboptimal here; it might force an agent to explore regions of the state space that are entirely irrelevant to its personalized goals.

To bridge this gap, we propose Personalized Federated Upper Confidence Bound Value Iteration (PF-UCBVI). The core principle of PF-UCBVI is to globally aggregate empirical transition counts to build a highly accurate shared dynamics model, while strictly maintaining *personalized* UCB bonuses and local Bellman backups tailored to each agent's individual objectives. The algorithm proceeds as follows:

1. **Initialization:** The central server initializes global counts $N_h(s,a) = 0$ and $N_h(s,a,s') = 0$ for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}, h \in [H]$. Each agent $m \in [M]$ initializes $V_h^m(s) = H-h+1$ and $Q_h^m(s,a) = H-h+1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}, h \in [H]$ and $V_{H+1}^m = 0$. Ties in $\arg\max$ are broken uniformly at random.

2. **Federated Learning Loop:** For each episode $k = 1, \ldots, K$:

   - **Optimistic Execution:** Agent m follows $\pi_{k,h}^m(s) = \arg\max_a Q_h^m(s,a)$ to collect transitions.

   - **Global Aggregation:** Server updates global counts $N_h(s,a) \leftarrow N_h(s,a) + \sum_m \mathbb{I}\{(s_{k,h}^m, a_{k,h}^m) = (s,a)\}$ and $N_h(s,a,s')$ similarly, then computes and broadcasts:

$$\widehat{P}_h(s' \mid s,a) = \frac{N_h(s,a,s')}{\max\{1, N_h(s,a)\}}.$$

   - **Local Bellman Backup:** For $h = H, \ldots, 1$, each agent m updates its personalized local bonus:

$$B_h^m(s,a) = \sqrt{\frac{4\mathsf{Var}_{\widehat{P}_{h,s,a}}(V_{h+1}^m)\iota}{\max\{N_h(s,a),1\}}} + \frac{11SH^2\iota}{\max\{N_h(s,a),1\}} \tag{6.2}$$

   where $\iota = \log(4MSAKH/\delta)$. Then performs value iteration:

$$Q_h^m(s,a) = \min\left\{H-h+1, r_h^m(s,a) + B_h^m(s,a) + \sum_{s'\in\mathcal{S}} \widehat{P}_h(s' \mid s,a) V_{h+1}^m(s')\right\}$$
$$V_h^m(s) = \max_{a\in\mathcal{A}} Q_h^m(s,a)$$

6.4 Regret Analysis of PF-UCBVI

To characterize the benefits of the personalized exploration strategy in PF-UCBVI, we establish the following regret bound:

Theorem 8 (Regret of Personalized Federated UCB-VI)**.** *Consider a federated MDP setting with M agents, S states, A actions, and a planning horizon H. Let the agents perform*

the PF-UCBVI *algorithm with a confidence parameter* $\delta \in (0,1)$*. Then, with probability at least* $1-\delta$*, the total regret satisfies:*

$$\mathsf{Regret}(K) \leq C \cdot \left(\sqrt{SAMKH^3 \log^2\left(\frac{4SAMKH}{\delta}\right)} + S^2AH^3 \log^2\left(\frac{4SAMKH}{\delta}\right) \right) \quad (6.3)$$

where $C > 0$ *is a universal constant.*

The proof of Theorem 8 is provided in Appendix E.2. The key implications of this result are as follows:

Optimal Linear Speedup and Accelerated Burn-in. The regret bound in Theorem 8 demonstrates an optimal linear speedup with respect to the number of participating agents M. Specifically, since the total regret across all agents scales as $\widetilde{\mathcal{O}}(\sqrt{SAMKH^3})$, the average regret per agent is tightly bounded by $\widetilde{\mathcal{O}}(\sqrt{SAKH^3/M})$. This yields a strict $1/\sqrt{M}$ reduction in sample complexity compared to learning in isolation. Furthermore, this collaborative efficiency significantly accelerates the initial exploration phase; the "burn-in" period required for the dominant sub-linear term to overtake the lower-order term is reduced to $K \geq \frac{S^3AH^3}{M}$, strictly dividing the required local exploration time by M.

Improved Regret over Existing Baselines. Furthermore, our algorithm achieves an improved regret bound compared to the best-known federated UCBVI result for online settings (Labbi et al., 2024), with a gap of $\widetilde{\mathcal{O}}(H)$. This shows that PF-UCBVI effectively exploits shared structure across agents to accelerate personalized learning while maintaining tight individual performance limits.

Collaboration with Diverse Goals. Crucially, the benefits of collaboration in PF-UCBVI are fully realized as a seamless synergy, where reward heterogeneity acts as a facilitator rather than a barrier. By strictly isolating personalized reward estimators and local value functions while globally aggregating shared transition dynamics, our framework ensures that diverse objectives never interfere with the collaborative learning process. Agents can pursue arbitrarily different goals without their value updates conflicting, allowing each

participant to harvest the full benefits of collective information while maintaining absolute policy autonomy. This confirms that in our federated setting, reward diversity is a structural feature that enables unhindered, high-efficiency learning.

Beyond Linear Speedup: Turning Diversity into a Structural Catalyst. While PF-UCBVI effectively navigates reward heterogeneity using explicit UCB-based exploration, the necessity of such exploratory actions remains a significant practical burden in safety-critical or resource-constrained domains. However, the collaboration benefits despite the agents' diverse goals suggests a deeper opportunity. If agents pursuing different objectives naturally visit different regions of the environment, their collective self-interest can eliminate the need for explicit exploration altogether. This insight shifts our perspective on heterogeneity: it is not merely a condition to be tolerated, but a potent resource that can provide exploration "for free." Motivated by this, we now introduce a paradigm where reward diversity itself ensures sufficient state-action coverage, enabling optimal learning through pure exploitation.

6.5 Algorithm: Personalized Federated Exploration-Free Value Iteration (PF-EFVI)

To enable exploration-free learning in this heterogeneous setting, we introduce a key structural assumption. While single-agent RL typically requires strong environment stochasticity (e.g., covariate diversity) (Civitavecchia and Papini, 2025), we show that reward heterogeneity can itself induce sufficient coverage. Let $d_h^\pi(s,a)$ denote the marginal visitation probability of observing state-action pair (s,a) at step h under policy π.

Assumption 2 (Federated State-Action Coverage via Reward Diversity)**.** *Let $\mathcal{P}$ be the set of all valid transition kernels. For any assumed transition kernel $\widetilde{P} \in \mathcal{P}$ and each agent $m \in [M]$, let $\widetilde{Q}_h^m$ denote the action-value function computed using its true private reward r^m and the transition $\widetilde{P}$. Let $\widetilde{\pi}^m$ be the corresponding greedy policy induced by $\widetilde{Q}_h^m$, such that $\widetilde{\pi}_h^m(s) \in \arg\max_{a\in\mathcal{A}} \widetilde{Q}_h^m(s,a)$. We assume there exists a strictly positive constant*

$\lambda_0 > 0$ *such that, for any assumed transition kernel* $\widetilde{P} \in \mathcal{P}$ *and each step* $h \in [H]$*:*

$$\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{1}{M} \sum_{m=1}^{M} d_h^{\widetilde{\pi}^m}(s,a) \geq \lambda_0, \tag{6.4}$$

where $d_h^{\widetilde{\pi}^m}(s,a)$ *is the marginal visitation probability of state-action pair* (s,a) *at step* h *when executing policy* $\widetilde{\pi}^m$ *in the true environment* P*.*

This structural assumption posits that the diversity of true goals is potent enough to prevent the ensemble's policies from collapsing into a narrow subspace, regardless of early estimation errors in the dynamics.

Reward-diversity-induced coverage relaxation. Assumption 2 significantly relaxes the stringent requirements of single-agent RL by relying on the macroscopic diversity of the agents' goals rather than environment stochasticity. Intuitively, if $\{r^m\}_{m=1}^M$ are sufficiently diverse, their collective greedy executions will naturally span the state-action space, achieving exploration *for free.* The parameter λ_0 captures the degree of this system-wide heterogeneity; a larger λ_0 indicates broader collective coverage, facilitating faster convergence, whereas a smaller λ_0 implies more concentrated trajectories over limited state-action space.

Based on the shared transition dynamics and the global coverage assumption, we propose Personalized Federated Exploration-Free Value Iteration (PF-EFVI). Unlike standard federated RL algorithms that require meticulous tuning of exploration bonuses, PF-EFVI relies solely on the empirical average of the global transitions and pure greedy exploitations at the local level.

1. **Initialization:** The central server initializes global counts $N_h(s,a) = 0$ and $N_h(s,a,s') = 0$ for all (s,a,s',h). Each agent m initializes $V_h^m(s) = 0$ and $Q_h^m(s,a) = 0$ for all (s,a,h), and $V_{H+1}^m(s) = 0$. Ties in arg max are broken uniformly at random.
2. **Federated Learning Loop:** For each episode $k = 1, \ldots, K$:
   - **Local execution (pure exploitation):** Each agent $m \in [M]$ executes its current greedy policy $\pi_{k,h}^m(s) = \arg\max_a Q_h^m(s,a)$ to collect a local trajectory $\{s_{k,h}^m, a_{k,h}^m, r_h^m(s_{k,h}^m, a_{k,h}^m)\}_{h=1}^H$.

- **Global aggregation:** Server updates global counts $N_h(s,a) \leftarrow N_h(s,a) + \sum_m \mathbb{I}\{(s_{k,h}^m, a_{k,h}^m) = (s,a)\}$ and $N_h(s,a,s')$ similarly, then computes and broadcasts:

$$\widehat{P}_h(s' \mid s,a) = \frac{N_h(s,a,s')}{\max\{1, N_h(s,a)\}}.$$

- **Local Bellman backup:** Each agent m performs backward induction. For $h = H, H-1, \ldots, 1$:

$$Q_h^m(s,a) \leftarrow r_h^m(s,a) + \sum_{s'} \widehat{P}_h(s' \mid s,a) V_{h+1}^m(s'),$$
$$V_h^m(s) \leftarrow \max_a Q_h^m(s,a).$$

6.6 Regret Analysis of PF-EFVI

In this section, we present the theoretical guarantees of the PF-EFVI algorithm. We show that under the heterogeneous coverage condition, a purely greedy approach without any explicit exploration or optimistic initialization achieves constant cumulative regret.

Theorem 9 (Instance-Dependent Regret Bound for PF-EFVI)**.** *Suppose Assumption 2 holds with a global coverage parameter $\lambda_0 > 0$. When M agents follow the* PF-EFVI *algorithm, then with probability at least $1-\delta$, the total system regret over K episodes is bounded by:*

$$\mathsf{Regret}(K) \le \mathcal{O}\left(\frac{H^5 \log(4SAKMH/\delta)}{M\lambda_0} \sum_{m=1}^{M} \frac{1}{\Delta_{\min}^m} + \frac{H \log(4SAKH/\delta)}{\lambda_0^2}\right) \tag{6.5}$$

where the instance-dependent gap $\Delta_{\min}^m$ for agent m is defined as the minimum suboptimality gap across all state-action pairs, i.e., $\Delta_{\min}^m = \min_{(s,a,h):\Delta_h^m(s,a)>0} (V_h^{m,\star}(s) - Q_h^{m,\star}(s,a))$.

The proof of Theorem 9 is provided in Appendix E.3. The key implications of this result are as follows:

Asymptotic convergence through heterogeneous exploitation. Theorem 9 demonstrates that PF-EFVI achieves a nearly constant cumulative regret bound that remains independent of the number of episodes K. This result signifies that when a sufficiently large population of agents (M) possesses diverse enough objectives to ensure global coverage (λ_0), exploitation alone becomes sufficient for identifying optimal policies without any explicit exploration mechanism.

Crucially, this implies that heterogeneous agents can achieve optimal collective learning while remaining *fully greedy* toward their local rewards, without ever compromising their individual interests. Contrary to the conventional wisdom that heterogeneity is a hurdle to consensus or a source of drift in federated learning, our findings suggest that heterogeneity acts as a structural catalyst for exploration. In this federated regime, one agent's self-interested exploitation inadvertently serves as another's information source, fulfilling the system's exploration requirements without necessitating any personal sacrifice or forced, sub-optimal actions. This fundamentally reshapes the exploration-exploitation trade-off, highlighting that in a federated setting with sufficiently diverse objectives, the *disagreement* between greedy agents is not a conflict to be managed, but a vital resource that enables highly efficient, exploration-free learning.

The Scale of λ_0 and the shift of exploration complexity While Assumption 2 appears to completely decouple the learning process from the action space size A, the constant λ_0 inherently encapsulates this dependency. Since the sum of visitation probabilities across all state-action pairs is bounded, even the most ideal scenario of perfectly uniform coverage yields $\lambda_0 \leq \mathcal{O}(1/SA)$. Consequently, λ_0 mathematically absorbs the action-space penalty.

This reveals a fundamental paradigm shift rather than a mathematical illusion: the *exploration tax* is shifted from the algorithm's temporal complexity to the environment's structural diversity. In standard UCB-based single-agent RL, the agent must pay for A temporally, spending numerous episodes to explicitly sample every suboptimal action ($\sum_{a \neq a^\star}$). In contrast, PF-EFVI resolves this spatially. The temporal penalty is fully absorbed by λ_0

and subsequently neutralized by the sheer spatial scale of the federated system (the M agents). Thus, as long as the multi-agent population M is sufficiently large, the system bypasses the necessity of active exploration over time.

Synergistic convergence. While the summation $\sum 1/\Delta_{\min}^{m}$ suggests that agents with smaller gaps contribute more to the system regret, it is important to note that our federated framework enables cross-agent acceleration. An agent with a small $\Delta_{\min}^{m}$ (a hard task) does not need to explore its own environment for an extended period; instead, it benefits from the rapid reduction of the global estimation error driven by the entire ensemble. Thus, heterogeneity does not penalize the system but rather ensures that *easy* tasks provide immediate high-quality data to *hard* tasks, stabilizing the overall learning process.

Chapter 7

Conclusion

In this dissertation, we have investigated the principled design and theoretical foundations of federated reinforcement learning (FedRL) algorithms. While the potential of RL in sequential decision-making is vast, its practical application has long been constrained by high sample complexity and the risks associated with limited state-action space coverage. By bridging federated learning with reinforcement learning, this research demonstrates how decentralized agents can collaboratively learn a global policy that is both communication-efficient and robust to data heterogeneity.

The primary contributions of this work are centered around three critical pillars that address the fundamental bottlenecks of modern RL:

- **Sample Efficiency and Linear Speedup:** We established that federated collaboration can achieve a near-optimal linear speedup in sample complexity. We first proved that in generative-model settings, simple averaging among M agents reduces the required samples by a factor of M (Woo et al., 2023). This analysis was further extended to the more challenging average-reward MDPs, providing the first sample-complexity guarantees in this regime with tailored parameter choices (Jiao et al., 2026). Furthermore, we demonstrated that this linear speedup remains attainable even under Markovian sampling and behavior heterogeneity in offline datasets (Woo et al., 2024).
- **Collaborative Coverage:** This thesis highlighted the *blessing of heterogeneity* in

federated settings. We showed that collaboration significantly relaxes the coverage requirements for individual agents, allowing for global convergence even when no single agent covers the entire state-action space (Woo et al., 2025). By employing pessimistic value aggregation, we ensured that FedRL remains robust in offline settings, requiring only that the union of agents' datasets covers the optimal trajectories (Woo et al., 2024). We also discovered that reward heterogeneity can naturally induce such coverage, enabling efficient learning through greedy execution without excessive exploration.

- **Communication-Efficient Design:** To ensure practical viability, we developed and analyzed communication-efficient protocols. We proved that *periodic averaging* and *adaptive aggregation schemes* can drastically reduce the number of synchronization rounds between agents and the central server. Specifically, in the context of offline RL, we established that by adaptively tuning learning rates and communication intervals, one can minimize communication overhead without sacrificing the near-optimal sample efficiency achieved by the collaborative fleet (Woo et al., 2024).

While this thesis establishes a rigorous theoretical foundation for federated reinforcement learning, the transition from principled design to large-scale, resilient deployments opens several promising avenues for future investigation. Building upon our findings in sample efficiency, collaborative coverage, and communication-efficient design, the following areas warrant further exploration:

1. **Scaling to Complex, High-Dimensional Models:** While this dissertation establishes fundamental guarantees within tractable settings, a natural progression is to extend these federated RL principles to complex, high-dimensional function approximators, such as deep neural networks and Large Language Models (LLMs). Reinforcement learning has recently emerged as a powerful paradigm for training LLMs, particularly through alignment techniques. However, these models still suffer from biased datasets and prohibitive computational demands. Federated RL can effectively mitigate these bottlenecks by leveraging diverse, decentralized data sources without compromising user privacy. Developing federated RL algorithms that can

efficiently fine-tune LLMs across distributed environments, while maintaining the privacy and communication efficiency analyzed in this work, represents a critical and timely next step.

2. **Resilience Against Malicious and Byzantine Agents:** The works in this thesis mainly assume that participating agents are fundamentally honest. However, federated systems are inherently vulnerable to Byzantine agents-malicious or faulty participants who provide poisoned updates. In an RL context, these adversarial behaviors can lead to reward hacking, which is particularly detrimental to the stability of the global policy. Future research should focus on developing robust aggregation rules that can detect and neutralize misaligned updates while preserving the linear speedup guarantees established in this dissertation. Ensuring safety and integrity in such adversarial environments is paramount for safety-critical deployments.
3. **Personalization for Heterogeneous Transition Dynamics:** While this thesis addressed heterogeneity in behavior policies and reward functions, real-world applications often involve agents operating in environments with diverse transition dynamics (e.g., robots with different hardware specs). A purely global policy may fail to capture these local nuances. Building on our work, a significant challenge remains in formulating a framework that balances global knowledge sharing with local policy personalization for different transition dynamics. Developing algorithms that can distinguish between universally useful information and "environment-specific dynamics" will be essential.

In conclusion, the findings of this dissertation underscore that collaboration is not merely an auxiliary feature but a fundamental necessity for scalable and robust reinforcement learning. The algorithms and theoretical bounds presented herein provide a path toward more efficient, private, and collaborative intelligent systems in the real world.

Appendix A

Preliminaries

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

Theorem 10 (Theorem 6 in Li et al. (2023)). *Suppose that* $Y_n = \sum_{k=1}^{n} X_k \in \mathbb{R}$, *where* $\{X_k\}$ *is a real-valued scalar sequence obeying*

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

Define

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

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

$$\mathbb{P}\left\{|Y_n| \ge \tau \text{ and } W_n \le \sigma^2\right\} \le 2\exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right). \tag{A.1}$$

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

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

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

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

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

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

In addition, we provide the concentration bound of the total number of visits of multiple agents with independent uniformly ergodic Markov chains, whose proof is provided in Appendix B.6.1. Denote

$$t_{\mathsf{th}}(s,a) \coloneqq \frac{2176 t_{\mathsf{mix}}^{\mathsf{max}} \log 8M \log \frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}{\mu_{\mathsf{avg}}(s,a)} \quad \text{and} \quad t_{\mathsf{th}} \coloneqq \frac{2176 t_{\mathsf{mix}}^{\mathsf{max}} \log 8M \log \frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}{\mu_{\mathsf{avg}}}. \tag{A.3}$$

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

Lemma 2. *Consider any* $\delta \in (0,1)$*. Under the asynchronous sampling, for any* $(s,a) \in \mathcal{S} \times \mathcal{A}$ *and* $0 \leq u < v \leq T$ *such that* $v - u \geq t_{\mathsf{th}}(s,a)$*, the following holds :*

$$\frac{1}{4}(v-u) M \mu_{\mathsf{avg}}(s,a) \leq \sum_{m=1}^{M} N_{u,v}^m(s,a) \leq 2(v-u) M \mu_{\mathsf{avg}}(s,a) \tag{A.4}$$

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

Appendix B

Analysis of Federated Q-Learning for Discounted MDPs

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

$$\mathsf{Var}_{s,a}(V) \coloneqq \mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V(s') - P(s,a)V\big]^2 = P(s,a)(V\circ V) - [P(s,a)V]\circ[P(s,a)V]. \quad \text{(B.1)}$$

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

B.1 Basic facts

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

and $t \geq 0$ that

$$0 \leq Q_t^m \leq \frac{1}{1-\gamma}, \qquad 0 \leq V_t^m \leq \frac{1}{1-\gamma}, \tag{B.2}$$

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

$$\|V_t^m - V^\star\|_\infty \leq \|Q_t^m - Q^\star\|_\infty \tag{B.3}$$

since

$$\|V_t^m - V^\star\|_\infty = \max_{s\in\mathcal{S}} \left| \max_{a\in\mathcal{A}} Q_t^m(s,a) - \max_{a\in\mathcal{A}} Q^\star(s,a) \right| \leq \max_{s\in\mathcal{S}, a\in\mathcal{A}} \left| Q_t^m(s,a) - Q^\star(s,a) \right| \leq \|Q_t^m - Q^\star\|_\infty.$$

Letting Q_t be the average of the local Q-estimates at the end of the t-th iteration, i.e., $Q_t = \frac{1}{M}\sum_{m=1}^M Q_t^m$, it follows from (3.6) and (4.11) that for all $t \geq 0$ that

$$Q_t = \frac{1}{M}\sum_{m=1}^M Q_t^m = \frac{1}{M}\sum_{m=1}^M Q_{t-\frac{1}{2}}^m. \tag{B.4}$$

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

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

which is the quantity we aim to control. From (B.2), it holds immediately that for all $t \geq 0$,

$$\|\Delta_t\|_\infty \leq \frac{1}{1-\gamma}. \tag{B.5}$$

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

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

B.2 Proof outline of Theorem 1

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

$$P_t^m((s,a),s') := \begin{cases} 1, & \text{if } s' = s_t^m(s,a) \\ 0, & \text{otherwise} \end{cases}, \tag{B.6}$$

then the local update rule (3.5) can be rewritten as

$$Q_{t-\frac{1}{2}}^m = (1-\eta)Q_{t-1}^m + \eta\left(r + \gamma P_t^m V_{t-1}^m\right). \tag{B.7}$$

The proof of Theorem 1 consists of the following steps.

Step 1: error decomposition. To analyze the error Δ_t, we first decompose the error into three terms, each of which can be bounded in a simple form. From (B.4), it follows that

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

where (i) follows from (B.7), and (ii) follows from Bellman's optimality equation $Q^\star = r + \gamma P V^\star$. By recursion over the above relation, we obtain

$$\Delta_t = \underbrace{(1-\eta)^t\Delta_0}_{=:E_t^1} + \underbrace{\eta\frac{\gamma}{M}\sum_{i=1}^t (1-\eta)^{t-i}\sum_{m=1}^M (P - P_i^m)V_{i-1}^m}_{=:E_t^2} + \underbrace{\eta\frac{\gamma}{M}\sum_{i=1}^t (1-\eta)^{t-i}\sum_{m=1}^M P(V^\star - V_{i-1}^m)}_{=:E_t^3}. \tag{B.8}$$

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

Step 2: bounding the error terms. Now, we obtain a bound of each of the error terms in (B.8) separately.

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

$$\|E_t^1\|_\infty = (1-\eta)^t \|\Delta_0\|_\infty \leq \frac{(1-\eta)^t}{1-\gamma}, \tag{B.9}$$

  where the last inequality follows from (B.5).

- **Bounding $\|E_t^2\|_\infty$.** Exploiting conditional independence across transitions in different iterations and applying Freedman's inequality (Freedman, 1975), the second error term is bounded using Lemma 3 below, whose proof is provided in Appendix B.5.1.

  Lemma 3. *For any given $\delta \in (0,1)$, the following holds*

$$\left\|E_t^2\right\|_\infty \leq \frac{8\gamma}{1-\gamma}\sqrt{\frac{\eta}{M}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} \tag{B.10}$$

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

- **Bounding $\|E_t^3\|_\infty$.** For E_t^3, we obtain the following recursive relation using Lemma 4 below, whose proof is provided in Appendix B.5.2.

  Lemma 4. *Let β be any integer that satisfies $0 \leq \beta \leq \phi(T)$. For any given $\delta \in (0,1)$,*

the following holds

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

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

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

$$\|\Delta_t\|_\infty \le \zeta + \gamma(1+4\eta(\tau-1))\max_{\iota(t)-\beta\tau\le i<t}\|\Delta_i\|_\infty \le \zeta + \left(\frac{1+\gamma}{2}\right)\max_{\iota(t)-\beta\tau\le i<t}\|\Delta_i\|_\infty, \quad \text{(B.11)}$$

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

$$\zeta \coloneqq \frac{4(1-\eta)^{\beta\tau}}{1-\gamma} + \frac{8\gamma}{1-\gamma}\sqrt{\frac{\eta}{M}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|MT}{\delta}}, \quad \text{(B.12)}$$

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

By invoking the recursive relation in (B.11) L times, where the choices of β and L will be made momentarily, it follows that for any $L\beta\tau \le t \le T$,

$$\begin{aligned}\|\Delta_t\|_\infty &\le \sum_{i=0}^{L-1}\left(\frac{1+\gamma}{2}\right)^i\zeta + \left(\frac{1+\gamma}{2}\right)^L\max_{\iota(t)-L\beta\tau\le i<t}\|\Delta_i\|_\infty \\ &\le \frac{2}{1-\gamma}\zeta + \left(\frac{1+\gamma}{2}\right)^L\left(\frac{1}{1-\gamma}\right), \quad \text{(B.13)}\end{aligned}$$

where the second line uses the crude bound in (B.5).

Setting $\beta = \left\lfloor \frac{1}{\tau}\sqrt{\frac{(1-\gamma)T}{2\eta}} \right\rfloor$ and $L = \left\lceil \sqrt{\frac{\eta T}{1-\gamma}} \right\rceil$, which ensures $L\beta\tau \le T$, and plugging their choices into (B.12) and (B.13) at $t = T$, we obtain that

$\|\Delta_T\|_\infty$

$$
\begin{aligned}
&\leq \frac{8(1-\eta)^{\beta\tau}}{(1-\gamma)^2} + \frac{16\gamma}{(1-\gamma)^2}\sqrt{\frac{\eta}{M}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{32\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|MT}{\delta}} \\
&\quad + \left(\frac{1+\gamma}{2}\right)^L \left(\frac{1}{1-\gamma}\right) \\
&\leq \frac{32}{(1-\gamma)^2}\left(\exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right) + \gamma\sqrt{\frac{\eta}{M}\log\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \gamma\eta\sqrt{\tau-1}\sqrt{\log\frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}}\right) \\
&\leq \frac{64}{(1-\gamma)^2}\left(\exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right) + \gamma\sqrt{\frac{\eta}{M}\log\frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}}\right), \qquad \text{(B.14)}
\end{aligned}
$$

where the second line follows from

$$
\begin{aligned}
&(1-\eta)^{\beta\tau} \leq \exp(-\eta\beta\tau) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right), \\
&\left(\frac{1+\gamma}{2}\right)^L = \left(1-\frac{1-\gamma}{2}\right)^L \leq \exp\left(-\frac{(1-\gamma)}{2}L\right) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\eta T}}{2}\right),
\end{aligned}
$$

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

$$
\tau - 1 \leq \frac{1}{\eta}\min\left\{\frac{1-\gamma}{8\gamma}, \frac{1}{M}\right\}. \qquad \text{(B.15)}
$$

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

$$
\begin{aligned}
T &\geq c_T \frac{1}{M(1-\gamma)^5\varepsilon^2}(\log((1-\gamma)^2\varepsilon))^2 \log\frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}, \\
\eta &= c_\eta M(1-\gamma)^4\varepsilon^2 \frac{1}{\log\frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}} \qquad \text{(B.16)}
\end{aligned}
$$

for some sufficiently large c_T and sufficiently small c_η.

B.3 Proof outline of Theorem 5

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

$$
\mathcal{U}^m_{v_1,v_2}(s,a) \coloneqq \{u \in [v_1, v_2) : \ (s^m_u, a^m_u) = (s,a)\},
$$

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

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

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

$$P^m_t((s,a),s') \coloneqq \begin{cases} 1 & \text{if } (s,a,s') = (s^m_{t-1}, a^m_{t-1}, s^m_t) \\ 0 & \text{otherwise} \end{cases}. \tag{B.17}$$

Then the local update rule (4.10) can be rewritten as

$$Q^m_{t-\frac{1}{2}}(s,a) = \begin{cases} (1-\eta)Q^m_{t-1}(s,a) + \eta(r^m_{t-1} + \gamma P^m_t(s,a)V^m_{t-1}) & \text{if } (s,a) = (s^m_{t-1}, a^m_{t-1}) \\ Q^m_{t-1}(s,a), & \text{otherwise} \end{cases}. \tag{B.18}$$

The proof of Theorem 5 consists of the following steps.

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

$$\begin{aligned} \Delta_t(s,a) &= \frac{1}{M}\sum_{m=1}^M (Q^\star(s,a) - Q^m_{t-\frac{1}{2}}(s,a)) \\ &= \left(\frac{1}{M}\sum_{m=1}^M (1-\eta)^{N^m_{t-\tau,t}(s,a)}\right)\Delta_{t-\tau}(s,a) \\ &\quad + \frac{\gamma}{M}\sum_{m=1}^M \sum_{u\in\mathcal{U}^m_{t-\tau,t}(s,a)} \eta(1-\eta)^{N^m_{u+1,t}(s,a)}(P(s,a) - P^m_{u+1}(s,a))V^m_u \\ &\quad + \frac{\gamma}{M}\sum_{m=1}^M \sum_{u\in\mathcal{U}^m_{t-\tau,t}(s,a)} \eta(1-\eta)^{N^m_{u+1,t}(s,a)}P(s,a)(V^\star - V^m_u), \end{aligned} \tag{B.19}$$

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

$$\begin{aligned}
&Q^\star(s,a) - Q^m_{u-\frac{1}{2}}(s,a) \\
&= (1-\eta)(Q^\star(s,a) - Q^m_{u-1}(s,a)) + \eta(Q^\star(s,a) - r^m_{u-1} - \gamma P^m_u(s,a)V^m_{u-1}) \\
&= (1-\eta)(Q^\star(s,a) - Q^m_{u-1}(s,a)) + \eta(\gamma P(s,a)V^\star - \gamma P^m_u(s,a)V^m_{u-1}) \\
&= (1-\eta)(Q^\star(s,a) - Q^m_{u-1}(s,a)) + \gamma\eta(P(s,a) - P^m_u(s,a))V^m_{u-1} + \gamma P(s,a)(V^\star - V^m_{u-1})
\end{aligned} \tag{B.20}$$

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

$$\lambda_{v_1,v_2}(s,a) \coloneqq \frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{N^m_{v_1,v_2}(s,a)} \tag{B.21}$$

for any integer $0 \le v_1 \le v_2 \le T$, we apply recursion to the relation (B.19) over the synchronization periods, and obtain

$$\begin{aligned}
&\Delta_t(s,a) \\
&= \left(\prod_{h=0}^{\phi(t)-1}\lambda_{h\tau,(h+1)\tau}(s,a)\right)\Delta_0(s,a) \\
&\quad + \sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right) \\
&\qquad \times \frac{\gamma}{M}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\eta(1-\eta)^{N^m_{u+1,(h+1)\tau}(s,a)}(P(s,a) - P^m_{u+1}(s,a))V^m_u \\
&\quad + \sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right) \\
&\qquad \times \frac{\gamma}{M}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\eta(1-\eta)^{N^m_{u+1,(h+1)\tau}(s,a)}P(s,a)(V^\star - V^m_u)
\end{aligned}$$

$$= \underbrace{\omega_{0,t}(s,a)\Delta_0(s,a)}_{=:E_t^1(s,a)} + \underbrace{\gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,t}^m(s,a)} \omega_{u,t}^m(s,a)(P(s,a) - P_{u+1}^m(s,a))V_u^m}_{=:E_t^2(s,a)}$$
$$+ \underbrace{\gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,t}^m(s,a)} \omega_{u,t}^m(s,a)P(s,a)(V^\star - V_u^m)}_{=:E_t^3(s,a)}, \tag{B.22}$$

which is decomposed in a similar manner as (B.8). Here, we define

$$\omega_{0,t}(s,a) \coloneqq \prod_{h=0}^{\phi(t)-1} \lambda_{h\tau,(h+1)\tau}(s,a), \tag{B.23a}$$

$$\omega_{u,t}^m(s,a) \coloneqq \frac{1}{M}\eta(1-\eta)^{N_{u+1,(\phi(u)+1)\tau}^m(s,a)} \prod_{l=\phi(u)+1}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a). \tag{B.23b}$$

We record the following useful lemma whose proof is provided in Appendix B.6.2.

Lemma 5. *Consider integers v_1 and v_2 such that $0 \le v_1 \le v_2 \le t \le T$, where $t \equiv 0 \ (mod\ \tau)$, and a state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Suppose that $\eta\tau \le 1$. The parameters defined in* (B.23) *satisfy*

$$\lambda_{v_1,v_2}(s,a) \le \exp\left(-\frac{\eta}{2M}\sum_{m=1}^{M} N_{v_1,v_2}^m(s,a)\right), \tag{B.24a}$$

$$\omega_{0,t}(s,a) + \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,t}^m(s,a)} \omega_{u,t}^m(s,a) = 1, \tag{B.24b}$$

$$\sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,h'\tau}^m(s,a)} \omega_{u,t}^m(s,a) \le \exp\left(-\frac{\eta}{2M}\sum_{m=1}^{M} N_{h'\tau,t}^m(s,a)\right), \quad \forall 0 \le h' \le \phi(t), \tag{B.24c}$$

$$\sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,t}^m(s,a)} (\omega_{u,t}^m(s,a))^2 \le \frac{2\eta}{M}. \tag{B.24d}$$

Step 2: bounding the error terms. Here, we derive the bound of the error terms in (B.22) separately for all the state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$.

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

$$|E_t^1(s,a)| \le \omega_{0,t}(s,a)(\|Q_0\|_\infty + \|Q^\star\|_\infty) \overset{\text{(i)}}{\le} \frac{2\omega_{0,t}(s,a)}{1-\gamma} \overset{\text{(ii)}}{\le} \frac{2}{1-\gamma}\exp\left(-\frac{\eta\mu_{\mathsf{avg}}t}{8}\right), \tag{B.25}$$

where (i) holds because $\|Q_0\|_\infty,\ \|Q^\star\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)) and (ii) follows from the fact that

$$\omega_{0,t}(s,a) \le \exp\left(-\frac{\eta}{2M}\sum_{m=1}^{M} N_{0,t}^m(s,a)\right) \le \exp\left(-\frac{\eta\mu_{\mathsf{avg}}t}{8}\right), \tag{B.26}$$

where the first inequality holds according to (B.24a) of Lemma 5, and the last inequality follows from the fact that $\sum_{m=1}^{M} N_{0,t}^m(s,a) \ge \frac{M\mu_{\mathsf{avg}}t}{4}$ for all $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[M]\times[T]$ at least with probability $1-\delta$ according to Lemma 2 and the union bound, as long as $t \ge t_{\mathsf{th}}$.

- **Bounding $|E_t^2(s,a)|$.** By carefully treating the statistical dependency via a decoupling argument and applying Freedman's inequality, we can obtain the following bound, whose proof is provided in Appendix B.6.3.

  Lemma 6. *For any given $\delta \in (0,1)$, the following holds for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $1 \le t \le T$:*

$$\left|E_t^2(s,a)\right| \le \frac{7241\gamma}{(1-\gamma)}\sqrt{\frac{C_{\mathsf{het}}\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \tag{B.27}$$

with probability at least $1-4\delta$, as long as $\tau \ge t_{\mathsf{th}}$ and

$$\frac{3}{T} \le \eta \le \min\left\{\frac{1}{16\tau}, \frac{1}{4\tau M}, \frac{1}{128MC_{\mathsf{het}}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}}\right\}.$$

- **Bounding $|E_t^3(s,a)|$.** For E_t^3, we can obtain the following recursive relation, whose proof is provided in Appendix B.6.4.

  Lemma 7. *Let β be any integer that satisfies $0 < \beta \le \phi(T)$. For any given $\delta \in (0,1)$,*

the following holds

$$|E_t^3(s,a)| \le \frac{2\gamma}{1-\gamma}\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right) + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}} + \frac{1+\gamma}{2}\max_{\phi(t)-\beta\le h\le\phi(t)-1}\|\Delta_{h\tau}\|_\infty, \tag{B.28}$$

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

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

$$\|\Delta_t\|_\infty \le \theta + \frac{1+\gamma}{2}\max_{\phi(t)-\beta\le h\le\phi(t)-1}\|\Delta_{h\tau}\|_\infty, \tag{B.29}$$

where we define

$$\theta \coloneqq \frac{4}{1-\gamma}\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right) + \frac{7241\gamma}{(1-\gamma)}\sqrt{\frac{C_{\mathsf{het}}\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}}. \tag{B.30}$$

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

$$\|\Delta_t\|_\infty \le \sum_{l=0}^{L_1-1}\left(\frac{1+\gamma}{2}\right)^l\theta + \left(\frac{1+\gamma}{2}\right)^{L_1}\max_{\phi(t)-\beta L\le i\le\phi(t)-1}\|\Delta_{i\tau}\|_\infty \le \frac{2}{1-\gamma}\left(\theta + \left(\frac{1+\gamma}{2}\right)^{L_1}\right), \tag{B.31}$$

where the last inequality follows from (B.5).

Setting $\beta = \left\lfloor\frac{1}{\tau}\sqrt{\frac{2(1-\gamma)T}{\mu_{\mathsf{avg}}\eta}}\right\rfloor$ and $L_1 = \left\lceil\frac{1}{2}\sqrt{\frac{\mu_{\mathsf{avg}}\eta T}{(1-\gamma)}}\right\rceil$, which ensures $L_1\beta\tau \le T$, and

plugging the choices into (B.30) and (B.31) at $t = T$, we obtain

$$\begin{aligned}
\|\Delta_T\|_\infty &\leq \frac{8\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right)}{(1-\gamma)^2} + \frac{14481\gamma}{(1-\gamma)^2}\sqrt{\frac{C_{\mathsf{het}}\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}} + \frac{2}{1-\gamma}\left(\frac{1+\gamma}{2}\right)^L \\
&\leq \frac{16}{(1-\gamma)^2}\exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{8}\right) + \frac{14481\gamma}{(1-\gamma)^2}\sqrt{\frac{C_{\mathsf{het}}\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}} \\
&\leq \frac{14497}{(1-\gamma)^2}\left(\exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{8}\right) + \gamma\sqrt{\frac{C_{\mathsf{het}}\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}}\right),
\end{aligned} \tag{B.32}$$

where the second line follows from

$$\exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{8}\right) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{8}\right),$$

$$\left(\frac{1+\gamma}{2}\right)^{L_1} = \left(1-\frac{1-\gamma}{2}\right)^{L_1} \leq \exp\left(-\frac{1-\gamma}{2}L_1\right) \leq \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right),$$

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

$$t_{\mathsf{th}} \leq \tau \leq \frac{1}{4\eta}\min\left\{\frac{1-\gamma}{4}, \frac{1}{M}\right\}. \tag{B.33}$$

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

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

for some sufficiently large c_T and sufficiently small c_η.

B.4 Proof outline of Theorem 6

The proof of Theorem 6 consists of the following steps.

Step 1: error decomposition. Consider any $0 \le t \le T$ such that $t \equiv 0 \pmod{\tau}$, i.e., t is a synchronization step. To analyze Δ_t, invoking the recursive relation of the local error (cf. (B.20)), we first decompose the error for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows:

$$\begin{aligned}
\Delta_t(s,a) &= \sum_{m=1}^{M} \alpha_t^m(s,a)(Q^\star(s,a) - Q^m_{t-\frac{1}{2}}(s,a)) \\
&= \left(\sum_{m=1}^{M} \alpha_t^m(s,a)(1-\eta)^{N^m_{t-\tau,t}(s,a)}\right)\Delta_{t-\tau}(s,a) \\
&\quad + \gamma \sum_{m=1}^{M} \alpha_t^m(s,a) \sum_{u \in \mathcal{U}^m_{t-\tau,t}(s,a)} \eta(1-\eta)^{N^m_{u+1,t}(s,a)}(P(s,a) - P^m_{u+1}(s,a))V^m_u \\
&\quad + \gamma \sum_{m=1}^{M} \alpha_t^m(s,a) \sum_{u \in \mathcal{U}^m_{t-\tau,t}(s,a)} \eta(1-\eta)^{N^m_{u+1,t}(s,a)}P(s,a)(V^\star - V^m_u) \\
&= \left(\frac{M}{\sum_{m=1}^{M}(1-\eta)^{-N^m_{t-\tau,t}(s,a)}}\right)\Delta_{t-\tau}(s,a) \\
&\quad + \gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}^m_{t-\tau,t}(s,a)} \frac{\eta(1-\eta)^{-N^m_{t-\tau,u+1}(s,a)}}{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{t-\tau,t}(s,a)}}(P(s,a) - P^m_{u+1}(s,a))V^m_u \\
&\quad + \gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}^m_{t-\tau,t}(s,a)} \frac{\eta(1-\eta)^{-N^m_{t-\tau,u+1}(s,a)}}{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{t-\tau,t}(s,a)}}P(s,a)(V^\star - V^m_u), \qquad \text{(B.34)}
\end{aligned}$$

where the last line uses the definition of $\alpha_t^m(s,a)$ in (4.14). Denoting

$$\widetilde{\lambda}_{v_1,v_2}(s,a) \coloneqq \frac{M}{\sum_{m=1}^{M}(1-\eta)^{N^m_{v_1,v_2}(s,a)}} \qquad \text{(B.35)}$$

for any integer $0 \le v_1 \le v_2 \le T$, we apply recursion to the relation (B.34) over the synchronization period, and obtain

$\Delta_t(s,a)$

$$
\begin{aligned}
&= \left(\prod_{h=0}^{\phi(t)-1} \widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a)\right)\Delta_0(s,a) \\
&\quad + \sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \\
&\qquad\qquad \times \gamma \sum_{m=1}^{M} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \frac{\eta(1-\eta)^{-N^m_{h\tau,u+1}(s,a)}}{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{h\tau,(h+1)\tau}(s,a)}}(P(s,a)-P^m_{u+1}(s,a))V^m_u \\
&\quad + \sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \\
&\qquad\qquad \times \gamma \sum_{m=1}^{M} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \frac{\eta(1-\eta)^{-N^m_{h\tau,u+1}(s,a)}}{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{h\tau,(h+1)\tau}(s,a)}}P(s,a)(V^\star-V^m_u) \\
&= \underbrace{\widetilde{\omega}_{0,t}(s,a)\Delta_0(s,a)}_{=:E^1_t(s,a)} + \underbrace{\gamma \sum_{m=1}^{M} \sum_{u\in\mathcal{U}^m_{0,t}(s,a)} \widetilde{\omega}^m_{u,t}(s,a)(P(s,a)-P^m_{u+1}(s,a))V^m_u}_{=:E^2_t(s,a)} \\
&\quad \underbrace{+\gamma \sum_{m=1}^{M} \sum_{u\in\mathcal{U}^m_{0,t}(s,a)} \widetilde{\omega}^m_{u,t}(s,a)P(s,a)(V^\star-V^m_u)}_{=:E^3_t(s,a)},
\end{aligned} \tag{B.36}
$$

which is again decomposed similarly as (B.8). Here, we define

$$
\widetilde{\omega}_{0,t}(s,a) \coloneqq \prod_{h=0}^{\phi(t)-1} \widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a), \tag{B.37a}
$$

$$
\widetilde{\omega}^m_{u,t}(s,a) \coloneqq \frac{\eta(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}}{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}\left(\prod_{l=\phi(u)+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right). \tag{B.37b}
$$

We record the following useful lemma whose proof is provided in Appendix B.6.5.

Lemma 8. *Consider any integers $0 \le v_1 \le v_2 \le t \le T$ where $t \equiv 0 \ (mod\ \tau)$ and any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. Suppose that $\eta\tau \le 1$, then the parameters defined in* (B.37) *satisfy*

$$
\frac{1}{3M} \le \alpha^m_t(s,a) \le \frac{3}{M}, \tag{B.38a}
$$

$$\widetilde{\omega}_{0,t}(s,a) \le (1-\eta)^{\frac{1}{M}\sum_{m=1}^{M} N_{0,t}^m(s,a)}, \tag{B.38b}$$

$$\widetilde{\omega}_{0,t}(s,a) + \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,t}^m(s,a)} \widetilde{\omega}_{u,t}^m(s,a) = 1, \tag{B.38c}$$

$$\sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,h'\tau}^m(s,a)} \widetilde{\omega}_{u,t}^m(s,a) \le (1-\eta)^{\frac{1}{M}\sum_{m=1}^{M} N_{h'\tau,t}^m(s,a)}, \quad \forall 0 \le h' \le \phi(t), \tag{B.38d}$$

$$\sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,t}^m(s,a)} (\widetilde{\omega}_{u,t}^m(s,a))^2 \le \frac{6\eta}{M}. \tag{B.38e}$$

Step 2: bounding the error terms. Here, we derive the bound of each error term in (B.36) separately for all the state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$.

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

$$\begin{aligned} |E_t^1(s,a)| &\le \widetilde{\omega}_{0,t}(\|Q_0\|_\infty + \|Q^\star\|_\infty) \\ &\overset{\text{(i)}}{\le} \frac{2\widetilde{\omega}_{0,t}}{1-\gamma} \\ &\overset{\text{(ii)}}{\le} \frac{2}{1-\gamma}(1-\eta)^{\frac{1}{M}\sum_{m=1}^{M} N_{0,t}^m(s,a)} \\ &\overset{\text{(iii)}}{\le} \frac{2}{1-\gamma}(1-\eta)^{\frac{1}{4}\mu_{\mathsf{avg}} t}, \end{aligned} \tag{B.39}$$

  where (i) holds because $\|Q_0\|_\infty,\ \|Q^\star\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)), (ii) follows from (B.38b) of Lemma 8, and (iii) holds for all $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times [T]$ with probability at least $1-\delta$ according to Lemma 2, as long as $t \ge t_{\mathsf{th}}$.

- **Bounding $|E_t^2(s,a)|$.** By carefully treating the statistical dependency via a decoupling argument and applying Freedman's inequality, we can obtain the following bound, whose proof is provided in Appendix B.6.6.

  Lemma 9. *For any given $\delta \in (0,1)$, the following holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $1 \le t \le T$:*

$$\left|E_t^2(s,a)\right| \le \frac{2064\gamma}{(1-\gamma)}\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \tag{B.40}$$

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

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

- **Bounding** $|E_t^3(s,a)|$**.** For E_t^3, similarly to Lemma 7, we can obtain the following recursive relation, whose proof is provided in Appendix B.6.7.

  Lemma 10. *Let* β *be any integer that satisfies* $\frac{t_{\mathsf{th}}}{\tau} \leq \beta \leq \phi(T)$*. For any given* $\delta \in (0,1)$*, the following holds*

$$|E_t^3(s,a)| \leq \frac{2(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}}}{1-\gamma} + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}} + \frac{1+\gamma}{2}\max_{\phi(t)-\beta\leq h\leq\phi(t)-1}\|\Delta_{h\tau}\|_\infty, \tag{B.41}$$

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

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

$$\|\Delta_t\|_\infty \leq \theta + \frac{1+\gamma}{2}\max_{\phi(t)-\beta\leq h\leq\phi(t)-1}\|\Delta_{h\tau}\|_\infty, \tag{B.42}$$

where we define

$$\widetilde{\theta} \coloneqq \frac{4}{1-\gamma}(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}} + \frac{2064\gamma}{(1-\gamma)}\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}}. \tag{B.43}$$

Then, by invoking the recursive relation for L_2 times, where the choices of β and L_2

will be made momentarily, it follows that for any $L_2\beta\tau \le t \le T$,

$$\|\Delta_t\|_\infty \le \sum_{l=0}^{L_2-1}\left(\frac{1+\gamma}{2}\right)^l \widetilde{\theta} + \left(\frac{1+\gamma}{2}\right)^{L_2} \max_{\phi(t)-\beta L \le i \le \phi(t)-1} \|\Delta_{i\tau}\|_\infty \le \frac{2}{1-\gamma}\left(\theta + \left(\frac{1+\gamma}{2}\right)^{L_2}\right), \tag{B.44}$$

where the last inequality follows from (B.5).

Setting $L_2 = \left\lceil \frac{1}{2}\sqrt{\frac{\mu_{\mathsf{avg}}\eta T}{(1-\gamma)}} \right\rceil$ and $\beta = \left\lfloor \frac{1}{\tau}\sqrt{\frac{2(1-\gamma)T}{\mu_{\mathsf{avg}}\eta}} \right\rfloor$, which ensures $L_2\beta\tau \le T$, and plugging the choices into (B.43) and (B.44) at $t = T$, we obtain

$$\begin{aligned}
\|\Delta_T\|_\infty &\le \frac{8(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}}}{(1-\gamma)^2} + \frac{4128\gamma}{(1-\gamma)^2}\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}} + \frac{2}{1-\gamma}\left(\frac{1+\gamma}{2}\right)^{L_2} \\
&\le \frac{16}{(1-\gamma)^2}\exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right) + \frac{4128\gamma}{(1-\gamma)^2}\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \\
&\quad + \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)^2}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}} \\
&\le \frac{4144}{(1-\gamma)^2}\left(\exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right) + \gamma\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}}\right),
\end{aligned} \tag{B.45}$$

where the second line follows from

$$\begin{aligned}
(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}} &\le \exp\left(-\frac{\eta\mu_{\mathsf{avg}}\beta\tau}{4}\right) \le \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right), \\
\left(\frac{1+\gamma}{2}\right)^{L_2} &= \left(1-\frac{1-\gamma}{2}\right)^{L_2} \le \exp\left(-\frac{1-\gamma}{2}L_2\right) \le \exp\left(-\frac{\sqrt{(1-\gamma)\mu_{\mathsf{avg}}\eta T}}{4}\right),
\end{aligned}$$

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

$$\tau \le \frac{1}{4\eta}\min\left\{\frac{1-\gamma}{4}, \frac{1}{M}\right\}. \tag{B.46}$$

Thus, for any given $\varepsilon \in (0, \frac{1}{1-\gamma})$, optimizing η and T to make (B.45) bounded by ε and

recalling $\beta\tau \geq t_{\mathsf{th}}$, we can guarantee that $\|\Delta_T\|_\infty \leq \varepsilon$ if

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

for some sufficiently large c_T and sufficiently small c_η.

B.5 Proofs for federated synchronous Q-learning (Theorem 1)

Define the following actions

$$a^\star(s) = \arg\max_{a\in\mathcal{A}} Q^\star(s,a), \quad a_i^m(s) = \arg\max_{a\in\mathcal{A}} Q_i^m(s,a), \quad a_i(s) = \arg\max_{a\in\mathcal{A}} \frac{1}{M}\sum_{m=1}^M Q_i^m(s,a) \tag{B.47}$$

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

B.5.1 Proof of Lemma 3

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

$$E_t^2(s,a) = \eta\frac{\gamma}{M}\sum_{i=1}^t (1-\eta)^{t-i}\sum_{m=1}^M (P(s,a) - P_i^m(s,a))V_{i-1}^m = \frac{\gamma}{M}\sum_{i=1}^t\sum_{m=1}^M z_i^m(s,a), \tag{B.48}$$

which we plan to bound by invoking Freedman's inequality (cf. Theorem 10) using the fact $z_i^m(s,a)$ is independent of the transition events of other agents $m' \neq m$ at i and has zero

mean conditioned on the events before iteration i, i.e.,

$$\mathbb{E}[z_i^m(s,a)|V_{i-1}^M,\ldots,V_{i-1}^1,\ldots,V_0^M,\ldots,V_0^1]=0,\qquad \forall m\in[M],\ 1\le i\le t. \tag{B.49}$$

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

- First, we can bound

$$\begin{aligned} B_t(s,a) &\coloneqq \max_{m\in[M],1\le i\le t}|z_i^m(s,a)| \le \max_{m\in[M],1\le i\le t}\eta\big(\|P(s,a)\|_1+\|P_i^m(s,a)\|_1\big)\|V_{i-1}^m\|_\infty \\ &\le \frac{2\eta}{1-\gamma}, \end{aligned} \tag{B.50}$$

  where the first inequality uses $(1-\eta)^{t-i}\le 1$, and the last inequality follows from $\|P(s,a)\|_1\le 1$, $\|P_i^m(s,a)\|_1\le 1$, and $\|V_{i-1}^m\|_\infty\le\frac{1}{1-\gamma}$ (cf. (B.2)).

- Next, we have

$$\begin{aligned} W_t(s,a) &\coloneqq \sum_{i=1}^t\sum_{m=1}^M \mathbb{E}\big[(z_i^m(s,a))^2|V_{i-1}^M,\ldots,V_{i-1}^1,\ldots,V_0^M,\ldots,V_0^1\big] \\ &= \sum_{i=1}^t\sum_{m=1}^M \mathsf{Var}\big(z_i^m(s,a)|V_{i-1}^M,\ldots,V_{i-1}^1,\ldots,V_0^M,\ldots,V_0^1\big) \\ &= \sum_{i=1}^t\sum_{m=1}^M \eta^2(1-\eta)^{2(t-i)}\mathsf{Var}_{s,a}(V_{i-1}^m) \\ &\le \frac{2M}{(1-\gamma)^2}\sum_{i=1}^t\eta^2(1-\eta)^{2(t-i)} \le \frac{2\eta M}{(1-\gamma)^2}\coloneqq\sigma^2, \end{aligned} \tag{B.51}$$

  where we recall the definition of $\mathsf{Var}_{s,a}$ in (B.1). Here, the first inequality holds since

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

and the last inequality follows from

$$\sum_{i=1}^{t} \eta^2 (1-\eta)^{2(t-i)} \leq \frac{\eta^2 (1-(1-\eta)^{2t})}{1-(1-\eta)^2} \leq \eta. \tag{B.52}$$

By substituting the above bounds (cf. (B.50) and (B.51)) and $m = 1$ into Freedman's inequality (see Theorem 10), it follows that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in [T]$,

$$\begin{aligned}
\left| \sum_{i=1}^{t} \sum_{m=1}^{M} z_i^m(s,a) \right| &\leq \sqrt{8 \max\{W_t(s,a), \frac{\sigma^2}{2^m}\} \log \frac{2m|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{4}{3} B_t(s,a) \log \frac{2m|\mathcal{S}||\mathcal{A}|T}{\delta} \\
&\leq \sqrt{\frac{32\eta M}{(1-\gamma)^2} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \frac{6\eta}{1-\gamma} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta} \\
&\leq \frac{8\gamma}{1-\gamma} \sqrt{\frac{\eta}{M} \log \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}
\end{aligned} \tag{B.53}$$

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

B.5.2 Proof of Lemma 4

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

$$\begin{aligned}
|E_t^3(s,a)| &= \left| \frac{\eta\gamma}{M} \sum_{i=0}^{t-1} \sum_{m=1}^{M} (1-\eta)^{t-i-1} P(s,a)(V^\star - V_i^m) \right| \\
&\leq \underbrace{\left| \frac{\eta\gamma}{M} \sum_{i=0}^{\iota(t)-\beta\tau-1} \sum_{m=1}^{M} (1-\eta)^{t-i-1} P(s,a)(V^\star - V_i^m) \right|}_{=: E_t^{3a}(s,a)} \\
&\quad + \underbrace{\left| \frac{\eta\gamma}{M} \sum_{i=\iota(t)-\beta\tau}^{t-1} \sum_{m=1}^{M} (1-\eta)^{t-i-1} P(s,a)(V^\star - V_i^m) \right|}_{=: E_t^{3b}(s,a)}.
\end{aligned} \tag{B.54}$$

We shall bound these two terms separately.

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

$$\begin{aligned} E_t^{3a}(s,a) &\le \eta\frac{\gamma}{M}\sum_{m=1}^{M}\sum_{i=0}^{\iota(t)-\beta\tau-1}(1-\eta)^{t-i}\|P(s,a)\|_1(\|V^\star\|_\infty+\|V_i^m\|_\infty) \\ &\le \frac{2\eta\gamma}{1-\gamma}\sum_{i=0}^{\iota(t)-\beta\tau-1}(1-\eta)^{t-i-1} \le \frac{2\gamma}{1-\gamma}(1-\eta)^{\beta\tau}, \end{aligned} \tag{B.55}$$

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

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

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

$$\begin{aligned} E_t^{3b}(s,a) &= \left|\frac{\eta\gamma}{M}\sum_{i=\iota(t)-\beta\tau}^{t-1}\sum_{m=1}^{M}(1-\eta)^{t-i-1}P(s,a)(V^\star-V_i^m)\right| \\ &\le \gamma\sum_{i=\iota(t)-\beta\tau}^{t-1}\eta(1-\eta)^{t-i-1}\left|\frac{1}{M}\sum_{m=1}^{M}P(s,a)(V^\star-V_i^m)\right| \\ &\le \gamma\sum_{i=\iota(t)-\beta\tau}^{t-1}\eta(1-\eta)^{t-i-1}\left\|\frac{1}{M}\sum_{m=1}^{M}(V^\star-V_i^m)\right\|_\infty, \end{aligned} \tag{B.56}$$

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

$$d_{v,w}^m(s,a) \coloneqq Q_w^m(s,a)-Q_v^m(s,a), \tag{B.57}$$

we claim the following bound for any $0 \le i < T$, which will be shown in Appendix B.5.2:

$$\left\|\frac{1}{M}\sum_{m=1}^{M}(V^\star-V_i^m)\right\|_\infty \le \|\Delta_i\|_\infty + 2\max_{m\in[M]}\left\|d_{\iota(i),i}^m\right\|_\infty. \tag{B.58}$$

In view of (B.58), it boils down to control $\max_{m\in[M]} \left\|d^m_{\iota(i),i}\right\|_\infty$. For any $(s,a)\in\mathcal{S}\times\mathcal{A}$, $m\in[M]$, and $0\le i<T$, by the definition (B.57), it follows that

$$\left|d^m_{\iota(i),i}(s,a)\right| = \left|\sum_{j=\iota(i)}^{i-1} d^m_{j,j+1}(s,a)\right| \le \underbrace{2\eta \sum_{j=\iota(i)}^{i-1} \|\Delta^m_j\|_\infty}_{:=B_1} + \underbrace{\gamma\eta \left|\sum_{j=\iota(i)}^{i-1} (P^m_{j+1}(s,a) - P(s,a))V^\star\right|}_{:=B_2}, \tag{B.59}$$

where

$$\Delta^m_j = Q^\star - Q^m_j. \tag{B.60}$$

The inequality (B.59) holds by the local update rule:

$$\begin{aligned}
d^m_{j,j+1}(s,a) &= Q^m_{j+1}(s,a) - Q^m_j(s,a) \\
&= \eta(r(s,a) + \gamma P^m_{j+1}(s,a)V^m_j - Q^m_j(s,a)) \\
&\overset{\text{(i)}}{=} \eta(r(s,a) + \gamma P^m_{j+1}(s,a)V^m_j - r(s,a) - \gamma P(s,a)V^\star + Q^\star(s,a) - Q^m_j(s,a)) \\
&= \eta(\gamma P^m_{j+1}(s,a)V^m_j - \gamma P(s,a)V^\star + Q^\star(s,a) - Q^m_j(s,a)) \\
&= \gamma\eta P^m_{j+1}(s,a)(V^m_j - V^\star) + \gamma\eta(P^m_{j+1}(s,a) - P(s,a))V^\star + \eta\Delta^m_j(s,a) \\
&\le 2\eta\|\Delta^m_j\|_\infty + \gamma\eta(P^m_{j+1}(s,a) - P(s,a))V^\star,
\end{aligned} \tag{B.61}$$

where (i) follows from Bellman's optimality equation, and the last inequality follows from $\|P^m_{j+1}(s,a)\|_1 \le 1$ and $\|V^m_j - V^\star\|_\infty \le \|\Delta^m_j\|_\infty$ (cf. (B.3)).

Next, we bound each term in (B.59) separately.

- **Bounding B_1.** The local error $\|\Delta^m_j\|_\infty$ is bounded as stated in the following lemma, whose proof is provided in Appendix B.5.2.

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

$$\|\Delta^m_i\|_\infty \le \|\Delta_{\iota(i)}\|_\infty + \frac{2}{1-\gamma}\sqrt{\eta\log\frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}} \tag{B.62}$$

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

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

$$2\eta \sum_{j=\iota(i)}^{i-1} \|\Delta_j^m\|_\infty \leq 2\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty + \frac{4\eta(\tau-1)}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}}. \tag{B.63}$$

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

$$\begin{aligned}\gamma\eta \left| \sum_{j=\iota(i)}^{i-1} (P_{j+1}^m(s,a) - P(s,a))V^\star \right| &\leq \frac{\gamma\eta}{1-\gamma}\sqrt{\frac{1}{2}\sum_{j=\iota(i)}^{i-1} \log \frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}} \\ &\leq \frac{\gamma\eta}{1-\gamma}\sqrt{(\tau-1)\log \frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}} \end{aligned} \tag{B.64}$$

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

By substituting the bound of B_1 and B_2 into (B.59), and applying the union bound, we obtain that: for any given $\delta \in (0,1)$, the following holds for any $0 \leq i \leq T$ and $m \in [M]$:

$$\begin{aligned}\|d_{\iota(i),i}^m\|_\infty &\leq 2\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty + \frac{4\eta((\tau-1)\sqrt{\eta}+\sqrt{\tau-1})}{(1-\gamma)}\sqrt{\log \frac{2|\mathcal{S}||\mathcal{A}|MT}{\delta}} \\ &\leq 2\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty + \frac{8\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log \frac{2|\mathcal{S}||\mathcal{A}|MT}{\delta}} \end{aligned} \tag{B.65}$$

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

By combining (B.65) and (B.58) and substituting it into (B.56) and using the fact that

$\sum_{i=\iota(t)-\beta\tau}^{t-1} \eta(1-\eta)^{t-i-1} \leq 1$, we can obtain the bound $E_t^{3b}(s,a)$ as follows:

$$|E_t^{3b}(s,a)| \leq \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|MT}{\delta}}$$
$$+\gamma \sum_{i=\iota(t)-\beta\tau}^{t-1} \eta(1-\eta)^{t-i-1}\left(\|\Delta_i\|_\infty + 4\eta(\tau-1)\|\Delta_{\iota(i)}\|_\infty\right)$$
$$\leq \frac{16\gamma\eta\sqrt{\tau-1}}{(1-\gamma)}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|MT}{\delta}}$$
$$+\gamma(1+4\eta(\tau-1))\max_{\iota(t)-\beta\tau\leq i<t}\|\Delta_i\|_\infty. \quad \text{(B.66)}$$

Step 3: putting all together. Now, we have the bounds of E_t^{3a} and E_t^{3b} separately derived above. By combining the bounds in (B.54), we can finally claim the advertised bound and this completes the proof.

Proof of (B.58)

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

$$\frac{1}{M}\sum_{m=1}^{M}(V^\star(s) - V_i^m(s)) = Q^\star(s,a^\star(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a_i^m(s))$$
$$\leq Q^\star(s,a^\star(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a^\star(s)) = \Delta_i(s,a^\star(s)), \quad \text{(B.67)}$$

where we use the definitions in (B.47). On the other end, it follows that

$$\frac{1}{M}\sum_{m=1}^{M}(V^\star(s) - V_i^m(s))$$
$$= Q^\star(s,a^\star(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a_{\iota(i)}(s)) + \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a_{\iota(i)}(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a_i^m(s))$$
$$\geq Q^\star(s,a_{\iota(i)}(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a_{\iota(i)}(s)) + \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a_{\iota(i)}(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s,a_i^m(s))$$

$$= \Delta_i(s, a_{\iota(i)}(s)) + \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s, a_{\iota(i)}(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s, a_i^m(s)), \tag{B.68}$$

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

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M} Q_i^m(s, a_{\iota(i)}(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s, a_i^m(s)) \\
&= \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s, a_{\iota(i)}(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_{\iota(i)}^m(s, a_{\iota(i)}(s)) + \frac{1}{M}\sum_{m=1}^{M} Q_{\iota(i)}^m(s, a_{\iota(i)}(s)) \\
&\quad - \frac{1}{M}\sum_{m=1}^{M} Q_{\iota(i)}^m(s, a_i^m(s)) + \frac{1}{M}\sum_{m=1}^{M} Q_{\iota(i)}^m(s, a_i^m(s)) - \frac{1}{M}\sum_{m=1}^{M} Q_i^m(s, a_i^m(s)) \\
&\geq \frac{1}{M}\sum_{m=1}^{M} \big(d_{\iota(i),i}^m(s, a_{\iota(i)}(s)) - d_{\iota(i),i}^m(s, a_i^m(s))\big),
\end{aligned} \tag{B.69}$$

where the inequality follows from the definition (B.57) and the fact that

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

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

Combining (B.67), (B.68) and (B.69), we obtain

$$\begin{aligned}
\Delta_i(s, a_{\iota(i)}(s)) + \frac{1}{M}\sum_{m=1}^{M} \big(d_{\iota(i),i}^m(s, a_{\iota(i)}(s)) - d_{\iota(i),i}^m(s, a_i^m(s))\big) &\leq \frac{1}{M}\sum_{m=1}^{M} (V^\star(s) - V_i^m(s)) \\
&\leq \Delta_i(s, a^\star(s)),
\end{aligned}$$

which immediately implies (B.58).

Proof of Lemma 11

By applying the decomposition in (B.8) to the local error for agent m, we decompose Δ_i^m as follows:

$$\Delta_i^m(s,a) = \underbrace{(1-\eta)^{i-\iota(i)}\Delta_{\iota(i)}^m(s,a)}_{:=D_1} + \underbrace{\gamma \sum_{j=\iota(i)+1}^{i} \eta(1-\eta)^{i-j}(P(s,a)-P_j^m(s,a))V^\star}_{:=D_2} + \underbrace{\gamma \sum_{j=\iota(i)+1}^{i} \eta(1-\eta)^{i-j}P_j^m(s,a)(V^\star - V_{j-1}^m)}_{:=D_3}. \tag{B.70}$$

We shall bound each term separately.

- **Bounding D_1.** Since $\Delta_{\iota(i)}^m = \Delta_{\iota(i)}$ for every agent m at the synchronization step $\iota(i)$,

$$|D_1| \le (1-\eta)^{i-\iota(i)}\|\Delta_{\iota(i)}\|_\infty. \tag{B.71}$$

- **Bounding D_2.** In a similar manner to (B.64), by invoking Hoeffding inequality and using the fact that $\sum_{j=\iota(i)+1}^{i}(\eta(1-\eta)^{i-j})^2 \le \eta$ (cf. (B.52)), we can claim that the following holds for all $(s,a,m,t) \in \mathcal{S}\times\mathcal{A}\times[M]\times[T]$,

$$|D_2| \le \gamma\sqrt{\sum_{j=\iota(i)+1}^{i}(\eta(1-\eta)^{i-j})^2\|V^\star\|_\infty^2 \log\frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}} \le \frac{\gamma}{1-\gamma}\sqrt{\eta\log\frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}} \tag{B.72}$$

  with probability at least $1-\delta$ for any given $\delta \in (0,1)$.

- **Bounding D_3.** By bounding $\|V^\star - V_{j-1}^m\|_\infty$ with the local error $\|\Delta_{j-1}^m\|_\infty$ (cf. (B.3))

and using $\|P_j^m(s,a)\|_1 \leq 1$, we have

$$|D_3| \leq \gamma \sum_{j=\iota(i)+1}^{i} \eta(1-\eta)^{i-j} \|P_j^m(s,a)\|_1 \|V^\star - V_{j-1}^m\|_\infty \leq \gamma \sum_{j=\iota(i)+1}^{i} \eta(1-\eta)^{i-j} \|\Delta_{j-1}^m\|_\infty. \tag{B.73}$$

By combining the bounds obtained above in (B.70), we obtain the following recursive relation

$$\|\Delta_i^m\|_\infty \leq (1-\eta)^{i-\iota(i)} \|\Delta_{\iota(i)}\|_\infty + \underbrace{\frac{\gamma}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|MT}{\delta}}}_{:=\rho} + \gamma \sum_{j=\iota(i)+1}^{i} \eta(1-\eta)^{i-j} \|\Delta_{j-1}^m\|_\infty. \tag{B.74}$$

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

$$\begin{aligned}
&\|\Delta_i^m\|_\infty \\
&\leq (1-\eta)^{i-\iota(i)} \|\Delta_{\iota(i)}\|_\infty + \rho \\
&\quad + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-j_1} \left((1-\eta)^{j_1-1-\iota(i)} \|\Delta_{\iota(i)}\|_\infty + \rho + \gamma \sum_{j_2=\iota(i)+1}^{j_1-1} \eta(1-\eta)^{j_1-1-j_2} \|\Delta_{j_2-1}^m\|_\infty \right) \\
&= \left((1-\eta)^{i-\iota(i)} + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-1-\iota(i)} \right) \|\Delta_{\iota(i)}\|_\infty + \left(1 + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-j_1} \right) \rho \\
&\quad + \gamma^2 \sum_{j_1=\iota(i)+1}^{i} \sum_{j_2=\iota(i)+1}^{j_1-1} \eta^2 (1-\eta)^{i-1-j_2} \|\Delta_{j_2-1}^m\|_\infty \\
&\leq \left((1-\eta)^{i-\iota(i)} + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-1-\iota(i)} \right) \|\Delta_{\iota(i)}\|_\infty + \left(1 + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-j_1} \right) \rho \\
&\quad + \gamma^2 \sum_{j_1=\iota(i)+1}^{i} \sum_{j_2=\iota(i)+1}^{j_1-1} \eta^2 (1-\eta)^{i-1-j_2} \left((1-\eta)^{j_2-1-\iota(i)} \|\Delta_{\iota(i)}\|_\infty + \rho + \cdots \right) \\
&\leq \left((1-\eta)^{i-\iota(i)} + \gamma \sum_{j_1=\iota(i)+1}^{i} \eta(1-\eta)^{i-1-\iota(i)} + \cdots + \gamma^l \sum_{j_1=\iota(i)+1}^{i} \cdots \sum_{j_l=\iota(i)+1}^{j_{l-1}-1} \eta^l (1-\eta)^{i-l-\iota(i)} \right) \|\Delta_{\iota(i)}\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&+\left(1+\gamma\sum_{j_1=\iota(i)+1}^{i}\eta(1-\eta)^{i-j_1}+\cdots+\gamma^l\sum_{j_1=\iota(i)+1}^{i}\cdots\sum_{j_l=\iota(i)+1}^{j_{l-1}-1}\eta^l(1-\eta)^{i-l+1-j_l}\right)\rho\\
&+\gamma^{l+1}\sum_{j_1=\iota(i)+1}^{i}\cdots\sum_{j_{l+1}=\iota(i)+1}^{j_l-1}\eta^{l+1}(1-\eta)^{i-l-j_{l+1}}\left(\|\Delta^m_{j_{l+1}-1}\|\right)\\
&\overset{\text{(i)}}{\leq}\sum_{l=0}^{i-\iota(i)}\gamma^l\binom{i-\iota(i)}{l}\eta^l(1-\eta)^{i-\iota(i)-l}\|\Delta^m_{\iota(i)}\|_\infty+\sum_{l=0}^{i-\iota(i)-1}\gamma^l\binom{i-\iota(i)}{l}\eta^l\rho\\
&\leq((1-\eta)+\gamma\eta)^{i-\iota(i)}\|\Delta^m_{\iota(i)}\|_\infty+(1+\gamma\eta)^{i-\iota(i)}\rho\\
&\overset{\text{(ii)}}{\leq}\|\Delta^m_{\iota(i)}\|_\infty+2\rho,
\end{aligned}
\tag{B.75}
$$

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

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

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

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

B.6 Proofs for federated asynchronous Q-learning (Theorem 5 and Theorem 6)

B.6.1 Proof of Lemma 2

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

by

$$P \coloneqq \begin{pmatrix} P^1 & & & \\ & P^2 & & \\ & & \ddots & \\ & & & P^M \end{pmatrix}, \tag{B.76}$$

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

$$\mu(x) \coloneqq \prod_{m=1}^{M} \mu_{\mathsf{b}}^m(x^m), \qquad \forall x = (x^1, x^2, \cdots, x^M) \in (\mathcal{S} \times \mathcal{A})^M, \tag{B.77}$$

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

$$t_{\mathsf{mix}}(\epsilon) \coloneqq \min \left\{ t \;\middle|\; \sup_{x_0 \in (\mathcal{S}\times\mathcal{A})^M} d_{\mathsf{TV}}(P_t(\cdot|x_0),\, \mu) \le \epsilon \right\} \;\text{ and }\; t_{\mathsf{mix}} \coloneqq t_{\mathsf{mix}}\left(\frac{1}{4}\right), \tag{B.78}$$

where

$$P_t(\cdot|x_0) = \prod_{m=1}^{M} P_t^m(\cdot|x_0^m) \tag{B.79}$$

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

$$t_{\mathsf{mix}}(\epsilon) \le \max_m t_{\mathsf{mix}}^m(\epsilon/M), \qquad t_{\mathsf{mix}} \le 4 \log 8M \max_{m \in [M]} t_{\mathsf{mix}}^m, \tag{B.80}$$

which will be proven at the end of the proof.

We now turn to the proof of Lemma 2. Define the event

$$\mathcal{B}_{u,v}(s,a) \coloneqq \left\{ \left| \sum_{m=1}^{M} N_{u,v}^m(s,a) - (v-u) \sum_{m=1}^{M} \mu_{\mathsf{b}}^m(s,a) \right| \ge \frac{1}{2}(v-u) \sum_{m=1}^{M} \mu_{\mathsf{b}}^m(s,a) \right\}. \tag{B.81}$$

We first establish that

$$\max_{x_0 \in (\mathcal{S}\times\mathcal{A})^M} \mathbb{P}\Big\{\mathcal{B}_{u,v}(s,a)\Big|\{(s_0^m, a_0^m)\}_{m=1}^M = x_0\Big\} \le \frac{\delta}{|\mathcal{S}||\mathcal{A}|T^2} \tag{B.82}$$

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

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

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

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

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

where

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

and

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

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

$$G_1 = \mathbb{P}_{\{(s_0^m, a_0^m)\}_{m=1}^M \sim \mu} \left[\left| \sum_{i=u+1}^{v} \bar{Z}_i(s,a) - \nu_{u,v}(s,a) \right| \geq \frac{1}{2} \nu_{u,v}(s,a) \right]$$
$$\leq 2 \exp\left(- \frac{(\nu_{u,v}(s,a)/2)^2 \gamma_{\mathrm{ps}}}{8((v-u) + 1/\gamma_{\mathrm{ps}}) V_f + 20 C (\nu_{u,v}(s,a)/2)} \right). \tag{B.83}$$

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

$$\gamma_{\mathrm{ps}} \geq \frac{1}{2 t_{\mathsf{mix}}} \tag{B.84a}$$

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

$$C \coloneqq \max_{u < i \leq v} \left| \bar{Z}_i(s,a) - \mathbb{E}[\bar{Z}_i(s,a)] \right| \leq M, \tag{B.84b}$$

$$V_f \coloneqq \mathsf{Var}(\bar{Z}_i(s,a)) = \sum_{m=1}^{M} (1 - \mu_{\mathsf{b}}^m(s,a)) \mu_{\mathsf{b}}^m(s,a) \leq \sum_{m=1}^{M} \mu_{\mathsf{b}}^m(s,a). \tag{B.84c}$$

Plugging (B.84) into (B.83), we have

$$G_1 \leq 2 \exp\left(- \frac{(\nu_{u,v}(s,a))^2}{8 t_{\mathsf{mix}} (24(v-u)(\sum_{m=1}^M \mu_{\mathsf{b}}^m(s,a)) + 10 M \nu_{u,v}(s,a))} \right)$$
$$\leq 2 \exp\left(- \frac{(v-u)(\sum_{m=1}^M \mu_{\mathsf{b}}^m(s,a))}{8 t_{\mathsf{mix}}(24 + 10M)} \right) \leq \frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2}, \tag{B.85}$$

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

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

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

long time, we have

$$G_2 := \mathbb{P}\left\{\mathcal{B}_{u,v}(s,a)\middle|\{(s_0^m,a_0^m)\}_{m=1}^M = x_0\right\} - \mathbb{P}\left\{\mathcal{B}_{u,v}(s,a)\middle|\{(s_0^m,a_0^m)\}_{m=1}^M \sim \mu\right\}$$
$$\leq d_{\mathsf{TV}}(P_u(\cdot|x_0),\mu) \leq \frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2}, \tag{B.86}$$

where the last inequality holds due to

$$u \geq \frac{t_{\mathsf{th}}(s,a)}{2} \geq 4\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}\max_{m\in[M]} t_{\mathsf{mix}}^m \geq \max_{m\in[M]} t_{\mathsf{mix}}^m\left(\frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2M}\right) \geq t_{\mathsf{mix}}\left(\frac{\delta}{2|\mathcal{S}||\mathcal{A}|T^2}\right). \tag{B.87}$$

Here, the second inequality follows from the fact that $t_{\mathsf{mix}}^m(\epsilon) \leq 2t_{\mathsf{mix}}^m \log_2 \frac{2}{\epsilon}$ (Paulin, 2015), and the last inequality follows from (B.80).

Step 3: summing things up. By combining the above bound, we complete the proof of (B.82), provided that $u \geq t_{\mathsf{th}}(s,a)/2$ and $v - u \geq t_{\mathsf{th}}(s,a)$. Then, we can obtain the following bound for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $0 \leq u < v \leq T$:

$$\mathbb{P}\left\{\frac{1}{4}(v-u)\sum_{m=1}^M \mu_{\mathsf{b}}^m(s,a) \leq \sum_{m=1}^M N_{u,v}^m(s,a) \leq 2(v-u)\sum_{m=1}^M \mu_{\mathsf{b}}^m(s,a)\right\}$$
$$\leq \mathbb{P}\left\{\left|\sum_{m=1}^M N^m_{u+\frac{t_{\mathsf{th}}(s,a)}{2},v}(s,a) - \left(v-u-\frac{t_{\mathsf{th}}(s,a)}{2}\right)\sum_{m=1}^M \mu_{\mathsf{b}}^m(s,a)\right| \geq \frac{1}{2}\left(v-u-\frac{t_{\mathsf{th}}(s,a)}{2}\right)\sum_{m=1}^M \mu_{\mathsf{b}}^m(s,a)\right\}$$
$$= \max_{x_0\in(\mathcal{S}\times\mathcal{A})^M} \mathbb{P}\left\{\mathcal{B}_{u+\frac{t_{\mathsf{th}}(s,a)}{2},v}(s,a)\middle|\{(s_0^m,a_0^m)\}_{m=1}^M = x_0\right\} \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|T^2}. \tag{B.88}$$

Proof of (B.80)**.** Notice that by the definition of d_{TV} and (B.79), we have

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

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

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

which immediately implies

$$t_{\mathsf{mix}}(\epsilon) \leq \max_m t^m_{\mathsf{mix}}(\epsilon/M).$$

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

$$t_{\mathsf{mix}} \leq \max_{m\in[M]} t^m_{\mathsf{mix}}\left(\frac{1}{4M}\right) \leq 4\log 8M \max_{m\in[M]} t^m_{\mathsf{mix}}.$$

B.6.2 Proof of Lemma 5

First, (B.24a) is derived as follows:

$$\begin{aligned}\lambda_{v_1,v_2}(s,a) = \frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{N^m_{v_1,v_2}(s,a)} &\leq \frac{1}{M}\sum_{m=1}^{M}\exp(-\eta N^m_{v_1,v_2}(s,a)) \\ &\leq 1 - \frac{1}{2}\frac{1}{M}\sum_{m=1}^{M}\eta N^m_{v_1,v_2}(s,a) \\ &\leq \exp\left(-\frac{\eta}{2M}\sum_{m=1}^{M} N^m_{v_1,v_2}(s,a)\right) \qquad \text{(B.89)}\end{aligned}$$

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

Next, we obtain (B.24b) through the following derivation:

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

$$\overset{\text{(iii)}}{=} 1 - \lambda_{0,\tau}\lambda_{\tau,2\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} = 1 - \omega_{0,t}(s,a), \tag{B.90}$$

where (i) follows from the geometric sum

$$\begin{aligned}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \eta(1-\eta)^{N^m_{u+1,(h+1)\tau}(s,a)} &= \eta + \eta(1-\eta) + \cdots + \eta(1-\eta)^{N^m_{h\tau,(h+1)\tau}(s,a)-1} \\ &= 1 - (1-\eta)^{N^m_{h\tau,(h+1)\tau}(s,a)}, \end{aligned} \tag{B.91}$$

(ii) follows from the definition (B.21), and (iii) follows by cancellation.

Similarly, (B.24c) can be obtained with some algebraic calculations as follows:

$$\begin{aligned}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}^m_{0,h'\tau}(s,a)} \omega^m_{u,t}(s,a) &= \sum_{m=1}^{M}\sum_{h=0}^{h'-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \omega^m_{u,t}(s,a) \\ &\overset{\text{(i)}}{=} \sum_{h=0}^{h'-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right)(1-\lambda_{h\tau,(h+1)\tau}(s,a)) \\ &\overset{\text{(ii)}}{\leq} \lambda_{h'\tau,(h'+1)\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} - \lambda_{0,\tau}\lambda_{\tau,2\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} \\ &\leq \lambda_{h'\tau,(h'+1)\tau}\cdots\lambda_{(\phi(t)-1)\tau,t} \qquad \text{(B.92)} \\ &\overset{\text{(iii)}}{\leq} \prod_{h=h'}^{\phi(t)-1}\exp\left(-\frac{\eta}{2M}\sum_{m=1}^{M}N^m_{h\tau,(h+1)\tau}(s,a)\right), \qquad \text{(B.93)}\end{aligned}$$

where (i) follows from similar derivations as above, (ii) follows by cancellation, and (iii) follows from (B.24a).

Finally, (B.24d) is derived as follows:

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{u\in\mathcal{U}^m_{0,t}(s,a)}(\omega^m_{u,t}(s,a))^2 \\ &= \sum_{m=1}^{M}\sum_{h=0}^{\phi(t)-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}(\omega^m_{u,t}(s,a))^2 \\ &= \frac{1}{M}\sum_{h=0}^{\phi(t)-1}\left(\prod_{l=(h+1)}^{\phi(t)-1}\lambda_{l\tau,(l+1)\tau}(s,a)\right)^2\sum_{m=1}^{M}\frac{1}{M}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\left(\eta(1-\eta)^{N^m_{u+1,(h+1)\tau}(s,a)}\right)^2\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \frac{2\eta}{M} \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a) \right) \sum_{m=1}^{M} \frac{1}{M} \left(1 - (1-\eta)^{(N^m_{h\tau,(h+1)\tau}(s,a))} \right) \\
&= \frac{2\eta}{M} \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=(h+1)}^{\phi(t)-1} \lambda_{l\tau,(l+1)\tau}(s,a) \right) \left(1 - \lambda_{h\tau,(h+1)\tau}(s,a) \right) \\
&\overset{\text{(ii)}}{\leq} \frac{2\eta}{M},
\end{aligned}
$$

where (i) holds since

$$
\begin{aligned}
\sum_{u \in \mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \left(\eta(1-\eta)^{N^m_{u+1,(h+1)\tau}(s,a)} \right)^2 &= \eta^2 + \eta^2(1-\eta)^2 + \cdots + \eta(1-\eta)^{2(N^m_{u+1,(h+1)\tau}(s,a)-1)} \\
&\leq \eta \left(1 - (1-\eta)^{2N^m_{u+1,(h+1)\tau}(s,a)} \right) \\
&\leq 2\eta \left(1 - (1-\eta)^{N^m_{u+1,(h+1)\tau}(s,a)} \right) \qquad \text{(B.94)}
\end{aligned}
$$

and (ii) can be similarly derived to the proof of (B.24c) (cf. (B.93)).

B.6.3 Proof of Lemma 6

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

$$
y^m_{u,t}(s,a) = \begin{cases} \omega^m_{u,t}(s,a)(P(s,a) - P^m_{u+1}(s,a))V^m_u & \text{if } (s^m_u, a^m_u) = (s,a) \\ 0 & \text{otherwise} \end{cases}, \qquad \text{(B.95)}
$$

where

$$
\omega^m_{u,t}(s,a) = \frac{\eta}{M}(1-\eta)^{N^m_{u+1,(\phi(u)+1)\tau}(s,a)} \prod_{h=\phi(u)+1}^{\phi(t)-1} \left(\frac{1}{M} \sum_{m'=1}^{M} (1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)} \right), \qquad \text{(B.96)}
$$

then $E^2_t(s,a) = \gamma \sum_{m=1}^{M} \sum_{u=0}^{t-1} y^m_{u,t}(s,a)$. However, due to the dependency between $P^m_{u+1}(s,a)$ and $\omega^m_{u,t}(s,a)$ arising from the Markovian sampling, it is difficult to track the sum of $y := \{y^m_{u,t}(s,a)\}$ directly. To address this issue, we will first analyze the sum using a collec-

tion of approximate random variables $\widehat{y} = \{\widehat{y}^m_{u,t}(s,a)\}$ drawn from a carefully constructed set $\widehat{\mathcal{Y}}$, which is closely coupled with the target $\{y^m_{u,t}(s,a)\}_{0\le u<t}$, i.e.,

$$D(y,\widehat{y}) \coloneqq \left|\sum_{m=1}^{M}\sum_{u=0}^{t-1}\left(y^m_{u,t}(s,a) - \widehat{y}^m_{u,t}(s,a)\right)\right| \tag{B.97}$$

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

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

$$\mathcal{B}_I(s,a) \coloneqq \bigcap_{u=0}^{t-I\tau}\left\{\frac{1}{4}\mu_{\mathsf{avg}}(s,a)MI\tau \le \sum_{m=1}^{M} N^k_{u,u+I\tau}(s,a) \le 2\mu_{\mathsf{avg}}(s,a)MI\tau\right\}, \tag{B.98}$$

where $I = I(s,a) \coloneqq \lfloor \frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau}\rfloor$. Note that $I\tau \ge \tau \ge t_{\mathsf{th}}$ (see (A.3) for the definition of $t_{\mathsf{th}}(s,a)$), and $1 \le 1/(16\eta\mu_{\mathsf{avg}}(s,a)\tau) \le I(s,a) \le 1/(8\eta\mu_{\mathsf{avg}}(s,a)\tau)$ if $\eta\tau \le 1/16$. Then, $\mathcal{B}_I(s,a)$ holds with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$ according to Lemma 2. The rest of the proof shall be carried out under the event $\mathcal{B}_I(s,a)$.

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

$$\begin{aligned}
&\omega^m_{u,t}(s,a)\\
&= \frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\frac{(1-\eta)^{N^m_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}\prod_{h=\phi(u)}^{\phi(t)-1}\left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)\\
&= \underbrace{\frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\prod_{h=\phi(u)}^{\phi(t)-1}\left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)}_{\coloneqq\bar{\omega}^m_{u,t}(s,a)}
\end{aligned}$$

$$+\underbrace{\frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\left(\frac{(1-\eta)^{N^m_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{m'=1}^M(1-\eta)^{N^{m'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}-1\right)\prod_{h=\phi(u)}^{\phi(t)-1}\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)}_{:=\chi^m_{u,t}(s,a)}.$$

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

$$\begin{aligned}
\bar{\omega}^m_{u,t}(s,a) &= \prod_{h=h_0(u,t)}^{\phi(u)-1}\left(\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)^{-1}\right)\\
&\quad\times\frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\prod_{h=\phi(u)}^{\phi(t)-1}\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)\\
&=\underbrace{\frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\prod_{h=h_0(u,t)}^{\phi(u)-1}\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)^{-1}}_{\text{dependent on history until } u}\\
&\quad\times\underbrace{\prod_{h=h_0(u,t)}^{\phi(t)-1}\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)}_{\text{dependent on history and future until } t}\\
&=\underbrace{\frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\prod_{h=h_0(u,t)}^{\phi(u)-1}\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)^{-1}}_{:=x^m_u(s,a)}\\
&\quad\times\prod_{l=1}^{l(u,t)}\underbrace{\prod_{h=\max\{0,\phi(t)-lI\}}^{\phi(t)-(l-1)I-1}\left(\frac{1}{M}\sum_{m'=1}^M(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)}_{:=z_l(s,a)},
\end{aligned}\tag{B.99}$$

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

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

$$L\coloneqq\min\left\{\left\lceil\frac{t}{I\tau}\right\rceil,\lceil 128\log(M/\eta)\rceil\right\}.\tag{B.100}$$

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

$$\begin{aligned}
\exp\left(-\frac{2\eta}{M}\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)\right) &\leq 1 - \frac{\eta}{M}\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a) \\
&\leq \frac{1}{M}\sum_{m'=1}^{M} (1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)} \\
&\leq \frac{1}{M}\sum_{m'=1}^{M} \exp(-\eta N^{m'}_{h\tau,(h+1)\tau}(s,a)) \\
&\leq 1 - \frac{1}{2}\frac{1}{M}\sum_{m'=1}^{M} \eta N^{m'}_{h\tau,(h+1)\tau}(s,a) \\
&\leq \exp\left(-\frac{\eta}{2M}\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)\right). \qquad \text{(B.101)}
\end{aligned}$$

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

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

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

$$\mathcal{Z} \coloneqq \left\{\exp\left(-\frac{1}{8}\eta\mu_{\mathsf{avg}}(s,a)I\tau - \frac{i\eta}{M}\right) \ \middle|\ i \in \mathbb{Z}:\ \ 0 \leq i < 4M\mu_{\mathsf{avg}}(s,a)I\tau\right\}. \qquad \text{(B.102)}$$

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

$$\mathcal{Z}_0 \coloneqq \left\{\exp\left(-\frac{i\eta}{M}\right) \ \middle|\ i \in \mathbb{Z}:\ \ 0 \leq i < 4M\mu_{\mathsf{avg}}(s,a)I\tau\right\}. \qquad \text{(B.103)}$$

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

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

$$\widehat{x}_u^m(s,a) = \begin{cases} x_u^m(s,a) & \text{if } \sum_{m=1}^M N_{h_0(u,t)\tau,\phi(u)\tau}^m(s,a) \le 2M\mu_{\mathsf{avg}}(s,a)I\tau \\ 0 & \text{otherwise} \end{cases}. \tag{B.104}$$

Note that the clipping never occurs and $\widehat{x}_u^m(s,a) = x_u^m(s,a)$ for all u as long as $\mathcal{B}_I(s,a)$ holds. To provide useful properties of $\widehat{x}_u^m(s,a)$ that will be useful later, we record the following lemma whose proof is provided in Appendix B.6.3.

Lemma 12. *For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, consider any integers $1 \le t \le T$ and $1 \le l \le \lceil \frac{t}{I\tau} \rceil$, where $I = \lfloor \frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau} \rfloor$. Suppose that $4\eta\tau \le 1$, then $\widehat{x}_u^m(s,a)$ defined in* (B.104) *satisfy*

$$\forall u \in [h_0, \phi(t) - (l-1)I) \quad : \quad \widehat{x}_u^m(s,a) \le \frac{9\eta}{M}, \tag{B.105a}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)} \sum_{m=1}^M \widehat{x}_u^m(s,a) \le 16\eta\mu_{\mathsf{avg}}(s,a)I\tau, \tag{B.105b}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1} \sum_{u \in \mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)} \sum_{m=1}^M (\widehat{x}_u^m(s,a))^2 \le \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M}, \tag{B.105c}$$

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

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

$$\widehat{\omega}_{u,t}^m(s,a;\boldsymbol{z}) = \widehat{x}_u^m(s,a) \prod_{l=1}^{l(u,t)} \widehat{z}_l, \tag{B.106}$$

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

$$\widehat{y}_{u,t}^m(s,a;\boldsymbol{z}) = \begin{cases} \widehat{\omega}_{u,t}^m(s,a;\boldsymbol{z})(P(s,a) - P_{u+1}^m(s,a))V_u^m & \text{if } (s_u^m, a_u^m) = (s,a) \text{ and } l(u,t) \le L \\ 0 & \text{otherwise} \end{cases}. \tag{B.107}$$

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

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

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

$$D(y, \widehat{y}_{\boldsymbol{z}}) < \frac{525}{1-\gamma}\sqrt{\frac{C_{\mathsf{het}}\eta L}{M}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}} \tag{B.108}$$

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

$$\begin{aligned}
&\min_{\widehat{y}_{\boldsymbol{z}} \in \widehat{\mathcal{Y}}} D(y, \widehat{y}_{\boldsymbol{z}}) \\
&= \min_{\boldsymbol{z} \in \mathcal{Z}^{L-1}\times\mathcal{Z}_0} \left| \sum_{m=1}^{M} \sum_{u=0}^{t-1} \left(y^m_{u,t}(s,a) - \widehat{y}^m_{u,t}(s,a;\boldsymbol{z}) \right) \right| \\
&\le \max_{\boldsymbol{z} \in \mathcal{Z}^{L-1}\times\mathcal{Z}_0} \left| \sum_{m=1}^{M} \sum_{u=0}^{t-LI\tau-1} y^m_{u,t}(s,a) - \widehat{y}^m_{u,t}(s,a;\boldsymbol{z}) \right| \\
&\quad + \min_{\boldsymbol{z} \in \mathcal{Z}^{L-1}\times\mathcal{Z}_0} \left| \sum_{m=1}^{M} \sum_{u=t-LI\tau}^{t-1} y^m_{u,t}(s,a) - \widehat{y}^m_{u,t}(s,a;\boldsymbol{z}) \right| \\
&\le \underbrace{\max_{\boldsymbol{z} \in \mathcal{Z}^{L-1}\times\mathcal{Z}_0} \left| \sum_{m=1}^{M} \sum_{u=0}^{t-LI\tau-1} y^m_{u,t}(s,a) - \widehat{y}^m_{u,t}(s,a;\boldsymbol{z}) \right|}_{=:D_1} \\
&\quad + \underbrace{\min_{\boldsymbol{z} \in \mathcal{Z}^{L-1}\times\mathcal{Z}_0} \left| \sum_{m=1}^{M} \sum_{u=t-LI\tau}^{t-1} (\bar{\omega}^m_{u,t}(s,a) - \widehat{\omega}^m_{u,t}(s,a;\boldsymbol{z}))(P(s,a) - P^m_{u+1}(s,a))V^m_u \right|}_{=:D_2} \\
&\quad + \underbrace{\left| \sum_{m=1}^{M} \sum_{u=t-LI\tau}^{t-1} \chi^m_{u,t}(s,a)(P(s,a) - P^m_{u+1}(s,a))V^m_u \right|}_{=:D_3},
\end{aligned}$$

and will bound each term separately.

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

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

  where (i) holds since $\|P(s,a)\|_1,\ \|P_u^m(s,a)\|_1 \le 1$ and $\|V_{u-1}^m\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)), (ii) follows from (B.101), (iii) holds due to $\mathcal{B}_I(s,a)$, and (iv) holds because $L \ge 128\log\frac{M}{\eta} \ge \frac{8}{\eta\mu_{\mathsf{avg}}(s,a)I\tau}\log\frac{M}{\eta}$ given that $\eta\mu_{\mathsf{avg}}(s,a)I\tau \ge 1/16$.

- **Bounding D_2.** Since $\widehat{x}_u^m(s,a) = x_u^m(s,a)$ when $\mathcal{B}_I(s,a)$ holds, in view of (B.107), we have

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

where the last inequality holds since $\|P(s,a)\|_1,\ \|P_u^m(s,a)\|_1 \le 1$ and $\|V_{u-1}^m\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)) and the definition of $\widehat{\omega}_{u,t}^m(s,a;\boldsymbol{z})$ defined in (B.106).

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

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

Then, applying the above bound and (B.105b) in Lemma 12,

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

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

- **Bounding D_3.** Applying Freedman's inequality, we can obtain the following bound, whose proof is provided in Appendix B.6.3.

  Lemma 13. *Consider any $\delta \in (0,1)$ and L defined in* (B.100). *For any $(s,a) \in \mathcal{S}\times\mathcal{A}$*

and $1 \le t \le T$*, the following holds:*

$$D_3 \le \frac{9}{1-\gamma}\sqrt{\frac{C_{\mathsf{het}}\eta L}{M}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}} \tag{B.109}$$

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

By combining the bounds obtained above,

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

since $\eta \le \frac{M}{128\log(TM)} \le M/L$ due to $L \le 128\log(TM)$.

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

$$\left|\sum_{m=1}^{M}\sum_{u=0}^{t-1}\widehat{y}_{u,t}^m(s,a;\boldsymbol{z})\right| < \frac{115}{(1-\gamma)}\sqrt{\frac{\eta L}{M}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \tag{B.110}$$

with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$. It suffices to establish (B.110) for a fixed $\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0$ with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\mathcal{Y}|}$, where

$$|\widehat{\mathcal{Y}}| = |\mathcal{Z}^{L-1}\times\mathcal{Z}_0| \le (4M\mu_{\mathsf{avg}}(s,a)I\tau)^L \le (M/\eta)^L \le (TM)^L \tag{B.111}$$

because $\eta\mu_{\mathsf{avg}}(s,a)I\tau \le 1/4$ and $\eta \ge 1/T$.

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

$$\mathbb{E}[\widehat{y}_{u,t}^m(s,a;\boldsymbol{z})|\mathcal{Y}_u] = 0, \tag{B.112}$$

where $\mathcal{Y}_u$ denotes the history of visited state-action pairs and updated values of all agents

until u, i.e., $\mathcal{Y}_u = \{(s_v^m, a_v^m), V_v^k\}_{m\in[M], v\le u}$. Before applying Freedman's inequality, we need to calculate the following quantities. First,

$$B_t(s,a) \coloneqq \max_{m\in[M],0\le u<t} |\widehat{y}_{u,t}^m(s,a;\boldsymbol{z})| \le \widehat{x}_u^m(s,a)\prod_{l=1}^{l(u,t)}\widehat{z}_l \|P(s,a)-P_{u+1}^m(s,a)\|_1\|V_u^m\|_\infty \le \frac{18\eta}{(1-\gamma)M}, \tag{B.113}$$

where the last inequality follows from $\|P(s,a)\|_1,\ \|P_u^m(s,a)\|_1 \le 1,\ \|V_{u-1}^m\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)), $\widehat{z}_l \le 1$, and (B.105a) in Lemma 12. Next, we can bound the variance as

$$\begin{aligned}
W_t(s,a) &\coloneqq \sum_{u=0}^{t}\sum_{m=1}^{M}\mathbb{E}[(\widehat{y}_{u,t}^m(s,a;\boldsymbol{z}))^2|\mathcal{Y}_u] \\
&= \sum_{l=1}^{L}\sum_{h=\max\{0,\phi(t)-lI\}}^{\phi(t)-(l-1)I-1}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)}(\widehat{x}_u^m(s,a)\prod_{l'=1}^{l}\widehat{z}_{l'})^2\mathsf{Var}_{P(s,a)}(V_u^m) \\
&\overset{\text{(i)}}{\le} \frac{2}{(1-\gamma)^2}\sum_{l=1}^{L}\left(\prod_{l'=1}^{l}\widehat{z}_{l'}^2\right)\sum_{h=max\{0,\phi(t)-lI\}}^{\phi(t)-(l-1)I-1}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)}(\widehat{x}_u^m(s,a))^2 \\
&\overset{\text{(ii)}}{\le} \frac{2}{(1-\gamma)^2}\sum_{l=1}^{L}\left(\prod_{l'=1}^{l}\widehat{z}_{l'}^2\right)\frac{64\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M} \\
&\overset{\text{(iii)}}{\le} \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M(1-\gamma)^2}\sum_{l=1}^{L}\exp\left(-1/4(l-1)\eta\mu_{\mathsf{avg}}(s,a)I\tau\right) \\
&\le \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M(1-\gamma)^2}\frac{1}{1-\exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)I\tau)} \\
&\overset{\text{(iv)}}{\le} \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M(1-\gamma)^2}\frac{8}{\eta\mu_{\mathsf{avg}}(s,a)I\tau} = \frac{1024\eta}{M(1-\gamma)^2} =: \sigma^2,
\end{aligned} \tag{B.114}$$

where (i) holds due to the fact that $\|\mathsf{Var}_P(V)\|_\infty \le \|P\|_1(\|V\|_\infty)^2 + (\|P\|_1\|V\|_\infty)^2 \le \frac{2}{(1-\gamma)^2}$ because $\|V\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)) and $\|P\|_1 \le 1$, (ii) follows from (B.105c) in Lemma 12, (iii) holds due to the range of $\mathcal{Z}$ and $\mathcal{Z}_0$ is bounded by $\exp(-1/8\eta\mu_{\mathsf{avg}}(s,a)I\tau)$ and 1, respectively, and (iv) holds since $e^{-x} \le 1-\frac{1}{2}x$ for any $0 \le x \le 1$ and $\eta\mu_{\mathsf{avg}}(s,a)I\tau/4 \le 1$

.

Now, by substituting the above bounds of W_t and B_t into Freedman's inequality (see Theorem 10) and setting $m = 1$, it follows that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $t \in [T]$ and $\widehat{y}_{\boldsymbol{z}} \in \widehat{\mathcal{Y}}$,

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

with at least probability $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}$, where (i) holds because $|\widehat{\mathcal{Y}}| \leq (TM)^L$ (cf. (B.111)), and $\frac{\eta L}{M}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta} \leq 1$ when $L \leq 128\log(TM)$ and $\eta \leq \frac{M}{128\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}}$. Therefore, it follows that (B.110) holds.

Step 4: putting things together. We now putting all the results obtained in the previous steps together to achieve the claimed bound. Under $\mathcal{B}_I(s,a)$, there exists $\widehat{y}_{\boldsymbol{z}} := \widehat{y}_{\boldsymbol{z}(y)} \in \widehat{\mathcal{Y}}$ such that (B.108) holds. Hence,

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

where the second line holds due to (B.110) and (B.108), and the last line holds because $L \leq 128\log(TM)$. By taking a union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $t \in [T]$, we complete the proof.

Proof of Lemma 12

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

$$\sum_{m=1}^{M} N^k_{h_0\tau,(h-1)\tau}(s,a) \leq 2M\mu_{\mathsf{avg}}(s,a)I\tau. \tag{B.116}$$

Then, the following holds:

$$\begin{aligned}\sum_{m=1}^{M} N^k_{h_0\tau,\overline{h}\tau}(s,a) &= \sum_{m=1}^{M} N^k_{(\overline{h}-1)\tau,\overline{h}\tau}(s,a) + \sum_{m=1}^{M} N^k_{h_0\tau,(\overline{h}-1)\tau}(s,a) \\ &\leq M\tau + 2M\mu_{\mathsf{avg}}(s,a)I\tau. \end{aligned} \tag{B.117}$$

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

$$\begin{aligned}\sum_{m'=1}^{M} \frac{(1-\eta)^{-N^{m'}_{h\tau,(h+1)\tau}(s,a)}}{M} \leq \frac{\sum_{m'=1}^{M} e^{\eta N^{m'}_{h\tau,(h+1)\tau}(s,a)}}{M} &\leq 1 + 2\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)}{M} \\ &\leq \exp\left(2\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)}{M}\right), \end{aligned} \tag{B.118}$$

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

According to (B.104), for any integer $u \in [\overline{h}\tau, t - (l-1)I\tau)$, $\widehat{x}^m_u(s,a)$ is clipped to zero. Now, we prove the bounds in Lemma 12 respectively.

Proof of (B.105a)**.** For $u \in [h_0\tau, \overline{h}\tau)$,

$$\begin{aligned}\widehat{x}^m_u(s,a) &= \frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)} \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)^{-1} \\ &\overset{\text{(i)}}{\leq} \frac{3\eta}{M} \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)^{-1} \end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \frac{3\eta}{M} \exp\left(\frac{2\eta}{M} \sum_{m'=1}^{M} N^{m'}_{h_0\tau,(\overline{h}-1)\tau}(s,a)\right)$$
$$\overset{\text{(iii)}}{\leq} \frac{3\eta}{M} \exp\left(4\eta\mu_{\mathsf{avg}}(s,a) I\tau\right) \overset{\text{(iv)}}{\leq} \frac{9\eta}{M},$$

where (i) holds since $(1+\eta)^x \leq e^{\eta x}$ and $\eta N^m_{\phi(u)\tau,u+1}(s,a) \leq \eta\tau \leq 1$, (ii) holds due to (B.101) and the fact that $\phi(u) \leq \overline{h}-1$, (iii) follows from the condition of $\overline{h}$ in (B.116), and (iv) holds because $4\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq 1$.

Proof of (B.105b)**.** By the definition of $\overline{h}$, it follows that

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} \widehat{x}^m_u(s,a) = \sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} x^m_u(s,a).$$

Using the following relation for each h:

$$\begin{aligned}
&\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} x^m_u(s,a) \\
&= \frac{1}{M}\left(\sum_{m=1}^{M} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \eta(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}\right)^{-1} \\
&= \left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-N^m_{h\tau,(h+1)\tau}(s,a)} - 1\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}\right)^{-1} \\
&\leq \left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-N^m_{h\tau,(h+1)\tau}(s,a)} - 1\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-N^m_{h'\tau,(h'+1)\tau}(s,a)}\right),
\end{aligned}$$

where the last inequality follows from Jensen's inequality, and applying (B.118), we can complete the proof as follows:

$$\begin{aligned}
\sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} x^m_u(s,a) &\leq \prod_{h'=h_0}^{\overline{h}-1} \left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-N^m_{h'\tau,(h'+1)\tau}(s,a)}\right) - 1 \\
&\leq \exp\left(\frac{2\eta\sum_{m'=1}^{M} N^{m'}_{h_0\tau,\overline{h}\tau}(s,a)}{M}\right) - 1
\end{aligned}$$

$$\overset{\text{(i)}}{\leq} \exp\left(4\eta\mu_{\mathsf{avg}}(s,a)I\tau + 2\eta\tau\right) - 1$$
$$\overset{\text{(ii)}}{\leq} 16\eta\mu_{\mathsf{avg}}(s,a)I\tau,$$

where (i) follows from (B.117), and (ii) holds because $e^x \leq 1 + 2x$ for any $x \in [0,1]$, $2\eta\tau \leq 1/2$, and $4\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq 1/2$.

Proof of (B.105c)**.** Similarly,

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (\widehat{x}^m_u(s,a))^2 = \sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (x^m_u(s,a))^2.$$

Using the following relation for each h:

$$\begin{aligned}
&\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (x^m_u(s,a))^2 \\
&= \frac{1}{M^2}\left(\sum_{m=1}^{M} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \eta^2(1-\eta)^{-2N^m_{\phi(u)\tau,u+1}(s,a)}\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}\right)^{-2} \\
&\leq \frac{\eta}{M}\left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-2N^m_{h\tau,(h+1)\tau}(s,a)} - 1\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}\right)^{-2} \\
&\leq \frac{\eta}{M}\left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-2N^m_{h\tau,(h+1)\tau}(s,a)} - 1\right) \prod_{h'=h_0}^{h-1} \left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-2N^m_{h'\tau,(h'+1)\tau}(s,a)}\right),
\end{aligned}$$

where the last inequality follows from Jensen's inequality, and applying (B.118) under the condition $2\eta\tau \leq 1$, we can complete the proof as follows:

$$\begin{aligned}
\sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (x^m_u(s,a))^2 &\leq \frac{\eta}{M} \prod_{h'=h_0}^{\overline{h}-1} \left(\frac{1}{M}\sum_{m=1}^{M}(1-\eta)^{-2N^m_{h'\tau,(h'+1)\tau}(s,a)}\right) - 1 \\
&\leq \frac{\eta}{M}\left(\exp\left(4\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h_0\tau,\overline{h}\tau}(s,a)}{M}\right) - 1\right) \\
&\overset{\text{(i)}}{\leq} \frac{\eta}{M}\left(\exp\left(8\eta\mu_{\mathsf{avg}}(s,a)I\tau + 4\eta\tau\right) - 1\right)
\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M},$$

where (i) follows from (B.117), and (ii) holds because $e^x \leq 1+4x$ for any $x \in [0,2]$, $4\eta\tau \leq 1$, and $8\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq 1$.

Proof of Lemma 13

Recall that

$$\begin{aligned}\chi^m_{u,t}(s,a) &= \frac{\eta}{M}(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}\left(\frac{(1-\eta)^{N^m_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{m'=1}^M (1-\eta)^{N^{m'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}} - 1\right)\\ &\quad\times \prod_{h=\phi(u)}^{\phi(t)-1}\left(\frac{1}{M}\sum_{m'=1}^M (1-\eta)^{N^{m'}_{h\tau,(h+1)\tau}(s,a)}\right)\\ &= \left(\frac{(1-\eta)^{N^m_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{\sum_{m'=1}^M (1-\eta)^{N^{m'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}} - 1\right)\omega^m_{u,t}(s,a).\end{aligned}$$

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

$$\begin{aligned}&\mathbb{P}\left[\left|\sum_{m=1}^M\sum_{u=t-LI\tau}^{t-1}\chi^m_{u,t}(s,a)(P(s,a)-P^m_{u+1}(s,a))V^m_u\right|\geq X\right]\\ &= \sum_{\boldsymbol{N}\in[0,\tau]^{MLI}}\mathbb{P}\left[\left|\sum_{m=1}^M\sum_{u=t-LI\tau}^{t-1}\chi^m_{u,t}(s,a)(P(s,a)-P^m_{u+1}(s,a))V^m_u\right|\geq X, \chi^m_{u,t}(s,a)=\chi^m_{u,t}(s,a;\boldsymbol{N})\right]\\ &\leq \sum_{\boldsymbol{N}\in[0,\tau]^{MLI}}\mathbb{P}\left[\left|\sum_{m=1}^M\sum_{u=t-LI\tau}^{t-1}\chi^m_{u,t}(s,a;\boldsymbol{N})(P(s,a)-P^m_{u+1}(s,a))V^m_u\right|\geq X\right],\end{aligned}$$

and it suffices to show that

$$\mathbb{P}\left[\left|\sum_{m=1}^{M}\sum_{u=t-LI\tau}^{t-1}\chi_{u,t}^{m}(s,a;\boldsymbol{N})(P(s,a)-P_{u+1}^{m}(s,a))V_{u}^{m}\right|\geq X\right]\leq\frac{\delta}{|\mathcal{S}||\mathcal{A}|T(1+\tau)^{MLI}}.$$

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

$$\mathbb{E}[\chi_{u,t}^{m}(s,a;\boldsymbol{N})(P(s,a)-P_{u+1}^{m}(s,a))V_{u}^{m}|\mathcal{Y}_{u}]=0, \tag{B.119}$$

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

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

$$\begin{aligned}
B_t(s,a) &\coloneqq \max_{m\in[M],t-LI\tau\leq u<t}|\chi_{u,t}^{m}(s,a;\boldsymbol{N})(P(s,a)-P_{u+1}^{m}(s,a))V_{u}^{m}| \\
&\leq \max_{m\in[M],t-LI\tau\leq u<t}\left|1-\frac{\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{m'}(s,a)}}{(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{m}(s,a)}}\right| \\
&\qquad\qquad \times\omega_{u,t}^{m}(s,a;\boldsymbol{N})\|P(s,a)-P_{u+1}^{m}(s,a)\|_1\|V_{u}^{m}\|_\infty \\
&\overset{\text{(i)}}{\leq} \frac{2}{1-\gamma}\max_{m\in[M],t-LI\tau\leq u<t}\left|1-\frac{\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{m'}(s,a)}}{(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{m}(s,a)}}\right|\omega_{u,t}^{m}(s,a;\boldsymbol{N}) \\
&\overset{\text{(ii)}}{\leq} \frac{8\eta\mu_{\mathsf{max}}(s,a)\tau}{1-\gamma}\max_{m\in[M],t-LI\tau\leq u<t}\omega_{u,t}^{m}(s,a;\boldsymbol{N}) \overset{\text{(iii)}}{\leq} \frac{8\eta^2\mu_{\mathsf{max}}(s,a)\tau}{(1-\gamma)M},
\end{aligned}$$

where (i) holds because $\|P(s,a)\|_1,\ \|P_u^m(s,a)\|_1\leq 1$, $\|V_{u-1}^m\|_\infty\leq\frac{1}{1-\gamma}$ (cf. (B.2)), (ii) follows from the fact that (which will be shown at the end of the proof)

$$\left|1-\frac{\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{m'}(s,a)}}{(1-\eta)^{N_{\phi(u)\tau,(\phi(u)+1)\tau}^{m}(s,a)}}\right|\leq 4\eta\mu_{\mathsf{max}}(s,a)\tau, \tag{B.120}$$

with $\mu_{\mathsf{max}}(s,a)\coloneqq\max_m\mu_{\mathsf{b}}^m(s,a)$, and (iii) holds due to the fact that $\omega_{u,t}^m(s,a;\boldsymbol{N})\leq\frac{\eta}{M}$.

Next, we can bound the variance as

$$
\begin{aligned}
W_t(s,a) &:= \sum_{u=\max\{0,t-LI\tau\}}^{t-1} \sum_{m=1}^{M} \mathbb{E}\Big[\Big(\chi_{u,t}^m(s,a;\boldsymbol{N})(P(s,a)-P_{u+1}^m(s,a))V_u^m\Big)^2 |\mathcal{Y}_u\Big] \\
&\overset{\text{(i)}}{\leq} (4\eta\mu_{\max}(s,a)\tau)^2 \sum_{h=\max\{0,\phi(t)-LI\}}^{\phi(t)-1} \sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)} \sum_{m=1}^{M} \Big(\omega_{u,t}^m(s,a;\boldsymbol{N})\Big)^2 \mathsf{Var}_{P(s,a)}(V_u^m) \\
&\overset{\text{(ii)}}{\leq} \frac{2(4\eta\mu_{\max}(s,a)\tau)^2}{(1-\gamma)^2} \sum_{h=\max\{0,\phi(t)-LI\}}^{\phi(t)-1} \sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)} \sum_{m=1}^{M} \Big(\omega_{u,t}^m(s,a;\boldsymbol{N})\Big)^2 \\
&\overset{\text{(iii)}}{\leq} \frac{2(4\eta\mu_{\max}(s,a)\tau)^2}{(1-\gamma)^2}\frac{6\eta}{M} =: \sigma^2,
\end{aligned}
$$

where (i) follows from (B.120), (ii) holds due to the fact that $\|\mathsf{Var}_P(V)\|_\infty \leq \|P\|_1(\|V\|_\infty)^2 + (\|P\|_1\|V\|_\infty)^2 \leq \frac{2}{(1-\gamma)^2}$ because $\|V\|_\infty \leq \frac{1}{1-\gamma}$ (cf. (B.2)) and $\|P\|_1 \leq 1$, (iii) follows from (B.24d) in Lemma 5.

Now, by substituting the above bounds of W_t and B_t into Freedman's inequality (see Theorem 10) and setting $c = 1$, it follows that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $t \in [T]$ and $\boldsymbol{N} = (N_{h\tau,(h+1)\tau}^m(s,a))_{m\in[M],h\in[\phi(t)-LI,\phi(t)-1]} \in [0,\tau]^{MLI}$,

$$
\begin{aligned}
&\left|\sum_{m=1}^{M}\sum_{u=t-LI\tau}^{t-1} \chi_{u,t}^m(s,a;\boldsymbol{N})(P(s,a)-P_{u+1}^m(s,a))V_u^m\right| \\
&\leq \sqrt{8\max\{W_t(s,a),\frac{\sigma^2}{2^m}\}\log\frac{4c|\mathcal{S}||\mathcal{A}|T(1+\tau)^{MLI}}{\delta}} + \frac{4}{3}B_t(s,a)\log\frac{4c|\mathcal{S}||\mathcal{A}|T(1+\tau)^{MLI}}{\delta} \\
&\leq \sqrt{96\frac{(4\eta\mu_{\max}(s,a)\tau)^2\eta}{M(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)^{MLI}}{\delta}} + \frac{12\eta^2\mu_{\max}(s,a)\tau}{M(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)^{MLI}}{\delta} \\
&\leq \sqrt{384\frac{(4\eta\tau M)(\mu_{\max}(s,a)^2\eta I\tau)L\eta}{M(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta}} \\
&\quad + \frac{12L\eta(\mu_{\max}(s,a)\eta I\tau)}{(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta} \\
&\overset{\text{(i)}}{\leq} \sqrt{48\frac{C_{\mathsf{het}}L\eta}{M(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta}} + \frac{2C_{\mathsf{het}}L\eta}{(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T(1+\tau)}{\delta} \\
&\overset{\text{(ii)}}{\leq} 9\sqrt{\frac{C_{\mathsf{het}}\eta L}{M(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}
\end{aligned} \tag{B.121}
$$

with at least probability $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T(1+\tau)^{MLT}}$, where we invoke the definition of C_{het} (cf. (4.9)). Here, (i) holds because $\eta\tau M \le 1/4$ and $\mu_{\mathsf{max}}(s,a)\eta I\tau \le C_{\mathsf{het}}\mu_{\mathsf{avg}}(s,a)\eta I\tau \le \frac{C_{\mathsf{het}}}{8}$, and (ii) follows from the fact that $\eta \le \frac{1}{128MC_{\mathsf{het}}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}} \le \frac{1}{MC_{\mathsf{het}}L\log\frac{4|\mathcal{S}||\mathcal{A}|T^2}{\delta}}$.

Proof of (B.120)**.** Using the fact that for $0<\eta<1$,

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

we can obtain the bounds as follows:

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

Thus, recalling $\mu_{\mathsf{max}}(s,a) \coloneqq \max_m \mu^m_{\mathsf{b}}(s,a)$, and using the fact that for any $(s,a,m,u) \in \mathcal{S}\times\mathcal{A}\times[M]\times[T]$:

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

at least with probability $1-\delta$, as long as $\tau \ge 443\left(\frac{t^m_{\mathsf{mix}}}{\mu_{\mathsf{max}}(s,a)}\right)\log\frac{4|\mathcal{S}||\mathcal{A}|TM}{\delta}$, which naturally holds if $\tau \ge t_{\mathsf{th}}$ (see (A.3) for the definition of t_{th}), according to Lemma 1,

$$\begin{aligned}
&\left|1-\frac{\frac{1}{M}\sum_{m'=1}^{M}(1-\eta)^{N^{m'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}{(1-\eta)^{N^{m}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)}}\right| \\
&\le 2\eta\max\left\{N^{m}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a), \frac{1}{M}\sum_{m'=1}^{M} N^{m'}_{\phi(u)\tau,(\phi(u)+1)\tau}(s,a)\right\} \\
&\le 4\eta\mu_{\mathsf{max}}(s,a)\tau.
\end{aligned}$$

B.6.4 Proof of Lemma 7

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

$$\begin{aligned} E_t^3(s,a) &= \gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,t}^m(s,a)} \omega_{u,t}^m(s,a) P(s,a)(V^\star - V_u^m) \\ &= \underbrace{\gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,(\phi(t)-\beta)\tau}^m(s,a)} \omega_{u,t}^m(s,a) P(s,a)(V^\star - V_u^m)}_{=:E_t^{3a}(s,a)} \\ &\qquad + \underbrace{\gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{(\phi(t)-\beta)\tau,t}^m(s,a)} \omega_{u,t}^m(s,a) P(s,a)(V^\star - V_u^m)}_{=:E_t^{3b}(s,a)} . \end{aligned} \tag{B.122}$$

We shall these two terms separately.

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

$$\begin{aligned} |E_t^{3a}(s,a)| &\leq \gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,(\phi(t)-\beta)\tau}^m(s,a)} \omega_{u,t}^m(s,a) \|P(s,a)\|_1 \|(V^\star - V_u^m)\|_\infty \\ &\overset{\text{(i)}}{\leq} \frac{2\gamma}{1-\gamma} \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{0,(\phi(t)-\beta)\tau}^m(s,a)} \omega_{u,t}^m(s,a) \\ &\overset{\text{(ii)}}{\leq} \frac{2\gamma}{1-\gamma} \exp\left(-\frac{\eta}{2M} \sum_{m=1}^{M} N_{(\phi(t)-\beta)\tau,t}^m(s,a) \right) \\ &\overset{\text{(iii)}}{\leq} \frac{2\gamma}{1-\gamma} \exp\left(-\frac{\eta \mu_{\mathsf{avg}} \beta\tau}{8} \right), \end{aligned} \tag{B.123}$$

  where (i) holds because $\|V_u^m\|_\infty, \|V^\star\|_\infty \leq \frac{1}{1-\gamma}$ (cf. (B.2)) and $\|P(s,a)\|_1 \leq 1$, (ii) holds due to (B.24c) in Lemma 5, and (iii) follows from the fact that $\sum_{m=1}^{M} N_{(\phi(t)-\beta)\tau,t}^m(s,a) \geq \frac{M\mu_{\mathsf{avg}}\beta\tau}{4}$ according to Lemma 2 as long as $\beta\tau \geq t_{\mathsf{th}}$.

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

$$|E_t^{3b}(s,a)| \leq \gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}_{(\phi(t)-\beta)\tau,t}^m(s,a)} \omega_{u,t}^m(s,a) \, \|V^\star - V_u^m\|_\infty$$

$$\overset{\text{(i)}}{\leq} \gamma \sum_{m=1}^{M} \sum_{h=\phi(t)-\beta}^{\phi(t)-1} \sum_{u \in \mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \omega^m_{u,t}(s,a)(\|\Delta_{h\tau}\|_\infty + \|Q^m_u - Q^m_{h\tau}\|_\infty)$$

$$\overset{\text{(ii)}}{\leq} \gamma \sum_{m=1}^{M} \sum_{h=\phi(t)-\beta}^{\phi(t)-1} \sum_{u \in \mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \omega^m_{u,t}(s,a)((1+2\eta\tau)\|\Delta_{h\tau}\|_\infty + \sigma_{\mathsf{local}}) \tag{B.124}$$

where (i) follows from the following bound, which will be shown in Appendix B.6.4,

$$\|V^\star - V^m_u\|_\infty \leq \|\Delta^m_{\iota(u)}\|_\infty + \|Q^m_u - Q^m_{\iota(u)}\|_\infty, \tag{B.125}$$

and (ii) holds due to the following lemma.

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

$$\|Q^m_u - Q^m_{\iota(u)}\|_\infty \leq 2\eta\tau\|\Delta^m_{\iota(u)}\|_\infty + \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}} \tag{B.126}$$

with probability at least $1-\delta$.

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

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

$$\begin{aligned}
|E^{3b}_t(s,a)| &\overset{\text{(i)}}{\leq} \sigma_{\mathsf{local}} + \gamma \sum_{h=\phi(t)-\beta}^{\phi(t)-1} (1+2\eta\tau)\|\Delta_{h\tau}\|_\infty \sum_{m=1}^{M} \sum_{u \in \mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \omega^m_{u,t}(s,a) \\
&\overset{\text{(ii)}}{\leq} \sigma_{\mathsf{local}} + \frac{1+\gamma}{2} \max_{\phi(t)-\beta \leq h < \phi(t)} \|\Delta_{h\tau}\|_\infty \sum_{m=1}^{M} \sum_{h=\phi(t)-\beta}^{\phi(t)-1} \sum_{u \in \mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \omega^m_{u,t}(s,a) \\
&\overset{\text{(iii)}}{\leq} \sigma_{\mathsf{local}} + \frac{1+\gamma}{2} \max_{\phi(t)-\beta \leq h < \phi(t)} \|\Delta_{h\tau}\|_\infty,
\end{aligned} \tag{B.127}$$

where (i) holds according to (B.24b) of Lemma 5, (ii) holds when η is small enough that $\eta \leq \frac{1-\gamma}{4\gamma\tau}$, and (iii) follows from (B.24b) of Lemma 5.

Now we have the bounds of $E^{3a}_t(s,a)$ and $E^{3b}_t(s,a)$ separately obtained above. By

combining the bounds in (B.122), we can claim the advertised bound, which completes the proof.

Proof of (B.125)

We prove the claim by showing

$$\Delta^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) - d^m_{\iota(u),u}(s, a^\star(s)) \leq V^\star(s) - V^m_u(s) \leq \Delta^m_{\iota(u)}(s, a^\star(s)) - d^m_{\iota(u),u}(s, a^\star(s)) \tag{B.128}$$

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

$$\begin{aligned}
&V^\star(s) - V^m_u(s) \\
&= Q^\star(s, a^\star(s)) - Q^m_u(s, a^m_u(s)) \\
&\leq Q^\star(s, a^\star(s)) - Q^m_u(s, a^\star(s)) \\
&= Q^\star(s, a^\star(s)) - Q^m_{\iota(u)}(s, a^\star(s)) - \underbrace{(Q^m_u(s, a^\star(s)) - Q^m_{\iota(u)}(s, a^\star(s)))}_{d^m_{\iota(u),u}(s, a^\star(s))}
\end{aligned} \tag{B.129}$$

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

$$\begin{aligned}
&V^\star(s) - V^m_u(s) \\
&= Q^\star(s, a^\star(s)) - Q^m_u(s, a^m_u(s)) \\
&= Q^\star(s, a^\star(s)) - Q^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) + Q^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) - Q^m_u(s, a^m_u(s)) \\
&= Q^\star(s, a^\star(s)) - Q^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) + Q^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) - Q^m_{\iota(u)}(s, a^m_u(s)) - d^m_{\iota(u),u}(s, a^m_u(s)) \\
&\geq Q^\star(s, a^m_{\iota(u)}(s)) - Q^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) + Q^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) - Q^m_{\iota(u)}(s, a^m_u(s)) - d^m_{\iota(u),u}(s, a^m_u(s)) \\
&\geq Q^\star(s, a^m_{\iota(u)}(s)) - Q^m_{\iota(u)}(s, a^m_{\iota(u)}(s)) - d^m_{\iota(u),u}(s, a^m_u(s))
\end{aligned} \tag{B.130}$$

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

Proof of Lemma 14

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

$$|Q_u^m(s,a) - Q_{\iota(u)}^m(s,a)| \le \underbrace{2\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} \|\Delta_v^m\|_\infty}_{:=B_1} + \underbrace{\left| \gamma\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} (P_{v+1}^m(s,a) - P(s,a))V^\star \right|}_{:=B_2}. \tag{B.131}$$

The inequality holds by the local update rule:

$$\begin{aligned} Q_{v+1}^m(s,a) - Q_v^m(s,a) &= (1-\eta)Q_v^m(s,a) + \eta(r(s,a) + \gamma P_{v+1}^m(s,a)V_v^m) - Q_v^m(s,a) \\ &= \eta(r(s,a) + \gamma P_{v+1}^m(s,a)V_v^m - Q_v^m(s,a)) \\ &= \eta(\gamma P_{v+1}^m(s,a)V_v^m - \gamma P(s,a)V^\star + Q^\star(s,a) - Q_v^m(s,a)) \\ &= \gamma\eta P_{v+1}^m(s,a)(V_v^m - V^\star) + \gamma\eta(P_{v+1}^m(s,a) - P(s,a))V^\star + \eta\Delta_v^m(s,a), \end{aligned} \tag{B.132}$$

and

$$\begin{aligned} |Q_u^m(s,a) - Q_{\iota(u)}^m(s,a)| &\le \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} |Q_{v+1}^m(s,a) - Q_v^m(s,a)| \\ &\le \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} \big(\eta|\Delta_v^m(s,a)| + \gamma\eta|P_{v+1}^m(s,a)(V_v^m - V^\star)|\big) \\ &\quad + \left| \gamma\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} (P_{v+1}^m(s,a) - P(s,a))V^\star \right| \\ &\le \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} 2\eta\|\Delta_v^m\|_\infty + \left| \gamma\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} (P_{v+1}^m(s,a) - P(s,a))V^\star \right|, \end{aligned} \tag{B.133}$$

where the last inequality holds since $\|P_{v+1}^m(s,a)\|_1 \le 1$ and $\|V_v^m - V^\star\|_\infty \le \|Q_v^m - Q^\star\|_\infty$ (cf. (B.3)).

Now, we shall bound each term separately.

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

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

$$\|\Delta_u^m\|_\infty \leq \|\Delta_{\iota(u)}^m\|_\infty + \frac{2\gamma}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|TM}{\delta}} \tag{B.134}$$

  with probability at least $1-\delta$.

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

$$\begin{aligned} 2\eta \sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} \|\Delta_v^m\|_\infty &\leq 2\eta |\mathcal{U}_{\iota(u),u}^m(s,a)| \max_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} \|\Delta_v^m\|_\infty \\ &\leq 2\eta(\tau-1)\left(\|\Delta_{\iota(u)}^m\|_\infty + \frac{2}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|TM}{\delta}}\right) \end{aligned} \tag{B.135}$$

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

$$\begin{aligned} B_2 &\leq \gamma\eta \sqrt{\sum_{v \in \mathcal{U}_{\iota(u),u}^m(s,a)} |(P_{v+1}^m(s,a) - P(s,a))V^\star| \log \frac{|\mathcal{S}||\mathcal{A}|TM}{\delta}} \\ &\leq \frac{2\gamma\eta}{1-\gamma}\sqrt{(\tau-1)\log \frac{|\mathcal{S}||\mathcal{A}|TM}{\delta}} \end{aligned} \tag{B.136}$$

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

By substituting the bound on B_1 and B_2 into (B.131) and using the condition that $\eta\tau < 1$, we can claim the stated bound holds and this completes the proof.

Proof of Lemma 15

For each state-action $(s,a) \in \mathcal{S} \times \mathcal{A}$ and agent m, by invoking the recursive relation (B.20) derived from the local Q-learning update in (4.10), Δ_u^m is decomposed as follows:

$$\Delta_u^m(s,a) = \underbrace{(1-\eta)^{N^m_{\iota(u),u}(s,a)}\Delta^m_{\iota(u)}(s,a)}_{=:D_1} + \underbrace{\gamma \sum_{v \in \mathcal{U}^m_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^m_{v+1,u}(s,a)}(P(s,a) - P^m_{v+1}(s,a))V^\star}_{=:D_2}$$
$$\underbrace{+\gamma \sum_{v \in \mathcal{U}^m_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^m_{v+1,u}(s,a)} P^m_{v+1}(s,a)(V^\star - V^m_v)}_{=:D_3}. \tag{B.137}$$

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

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

$$|D_1| \le (1-\eta)^{N^m_{\iota(u),u}(s,a)} \|\Delta^m_{\iota(u)}\|_\infty. \tag{B.138}$$

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

$$\begin{aligned} |D_2| &\le \gamma \sqrt{\sum_{v \in \mathcal{U}^m_{\iota(u),u}(s,a)} (\eta(1-\eta)^{N^m_{v+1,u}(s,a)})^2 (\|V^\star\|_\infty)^2 \log \frac{|\mathcal{S}||\mathcal{A}|TM}{\delta}} \\ &\le \frac{\gamma}{1-\gamma}\sqrt{\eta \log \frac{|\mathcal{S}||\mathcal{A}|TM}{\delta}} := \rho \end{aligned} \tag{B.139}$$

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

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

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

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

$$\begin{aligned}|D_3| &\leq \gamma \sum_{v \in \mathcal{U}^m_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^m_{v+1,u}(s,a)} \|P^m_{v+1}(s,a)\|_1 \|V^\star - V^m_v\|_\infty \\ &\leq \gamma \sum_{v \in \mathcal{U}^m_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^m_{v+1,u}(s,a)} \|\Delta^m_v\|_\infty, \end{aligned} \tag{B.140}$$

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

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

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

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

$$|\Delta^m_u(s,a)| \leq (1-\eta)^{N^m_{\iota(u),u}(s,a)} \|\Delta^m_{\iota(u)}\|_\infty + \rho + \gamma \sum_{v \in \mathcal{U}^m_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^m_{v+1,u}(s,a)} \|\Delta^m_v\|_\infty. \tag{B.141}$$

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

$$\|\Delta^m_{\iota(u)+i}\|_\infty \leq \|\Delta^m_{\iota(u)}\|_\infty + 2\rho, \tag{B.142}$$

which completes the proof of Lemma 15. First, if $i = 0$, the claim is obviously true. Suppose the claim holds for $\iota(u), \iota(u)+1, \cdots, \iota(u)+i-1$. Then, for $u = \iota(u)+i$, by invoking the recursive relation (B.141), we can show that the claim (B.142) holds for i as follows:

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

$$
\begin{aligned}
&= ((1-\eta)^{N^m_{\iota(u),u}(s,a)} + \gamma \sum_{v\in\mathcal{U}^m_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^m_{v+1,u}(s,a)})\|\Delta^m_{\iota(u)}\|_\infty \\
&\quad + (1+2\gamma \sum_{v\in\mathcal{U}^m_{\iota(u),u}(s,a)} \eta(1-\eta)^{N^m_{v+1,u}(s,a)})\rho \\
&= ((1-\eta)^{N^m_{\iota(u),u}(s,a)} + \gamma(1-(1-\eta)^{N^m_{\iota(u),u}(s,a)})\|\Delta^m_{\iota(u)}\|_\infty + (1+2\gamma(1-(1-\eta)^{N^m_{\iota(u),u}(s,a)}))\rho \\
&\leq \|\Delta^m_{\iota(u)}\|_\infty + 2\rho, \qquad \text{(B.143)}
\end{aligned}
$$

where the last inequality holds since

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

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

B.6.5 Proof of Lemma 8

First, using the fact that

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

given that $\eta\tau \leq 1$, by the definition of α^m_t (cf. (4.14)), we derive (B.38a) as follows:

$$
\begin{aligned}
\frac{1}{3M} \leq \frac{1}{M\max_{m'\in[M]}(1-\eta)^{-N^{m'}_{t-\tau,t}(s,a)}} \leq \alpha^m_t(s,a) &= \frac{(1-\eta)^{-N^m_{t-\tau,t}(s,a)}}{\sum_{m'=1}^M (1-\eta)^{-N^{m'}_{t-\tau,t}(s,a)}} \\
&\leq \frac{(1-\eta)^{-N^m_{t-\tau,t}(s,a)}}{M} \leq \frac{3}{M}.
\end{aligned}
$$

Moving onto (B.38b), it follows that

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

$$\overset{\text{(i)}}{=} \prod_{h=0}^{\phi(t)-1} \frac{M}{\sum_{m=1}^{M}(1-\eta)^{-N^m_{h\tau,(h+1)\tau}(s,a)}}$$
$$\overset{\text{(ii)}}{\leq} \prod_{h=0}^{\phi(t)-1} \frac{1}{(1-\eta)^{-\frac{1}{M}\sum_{m=1}^{M} N^m_{h\tau,(h+1)\tau}(s,a)}}$$
$$= (1-\eta)^{\sum_{h=0}^{\phi(t)-1}\frac{1}{M}\sum_{m=1}^{M} N^m_{h\tau,(h+1)\tau}(s,a)} = (1-\eta)^{\frac{1}{M}\sum_{m=1}^{M} N^m_{0,t}(s,a)},$$

where (i) follows from the definition of α_t^m (cf. (4.14)), (ii) follows from Jensen's inequality.

Next, we obtain (B.38c) through the following derivation:

$$\sum_{m=1}^{M}\sum_{u\in\mathcal{U}^m_{0,t}(s,a)} \widetilde{\omega}^m_{u,t}(s,a) = \sum_{m=1}^{M}\sum_{h=0}^{\phi(t)-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \widetilde{\omega}^m_{u,t}(s,a)$$
$$= \sum_{m=1}^{M}\sum_{h=0}^{\phi(t)-1} \alpha^m_{(h+1)\tau}(s,a) \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \eta(1-\eta)^{N^m_{u+1,(h+1)\tau}(s,a)} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right)$$
$$= \sum_{m=1}^{M}\sum_{h=0}^{\phi(t)-1} \alpha^m_{(h+1)\tau}(s,a) \left(1-(1-\eta)^{N^m_{h\tau,(h+1)\tau}(s,a)}\right) \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right)$$
$$\overset{\text{(i)}}{=} \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \sum_{m=1}^{M} \alpha^m_{(h+1)\tau}(s,a) \left(1-(1-\eta)^{N^m_{h\tau,(h+1)\tau}(s,a)}\right)$$
$$\overset{\text{(ii)}}{=} \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \left(1-\sum_{m=1}^{M} \alpha^m_{(h+1)\tau}(s,a)(1-\eta)^{N^m_{h\tau,(h+1)\tau}(s,a)}\right)$$
$$= \sum_{h=0}^{\phi(t)-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a)\right) \left(1-\widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a)\right)$$
$$\overset{\text{(iii)}}{=} 1-\widetilde{\lambda}_{0,\tau}(s,a)\widetilde{\lambda}_{\tau,2\tau}(s,a)\cdots\widetilde{\lambda}_{(\phi(t)-1)\tau,t}(s,a) = 1-\widetilde{\omega}_{0,t}(s,a), \tag{B.144}$$

where (i) follows by reordering the summation, (ii) follows by $\sum_{m=1}^{M}\alpha_t^m(s,a)=1$, and (iii) holds by cancellation.

In a similar manner, (B.38d) is derived as follows:

$$\sum_{m=1}^{M}\sum_{u\in\mathcal{U}^m_{0,h'\tau}(s,a)} \widetilde{\omega}^m_{u,t}(s,a) = \sum_{m=1}^{M}\sum_{h=0}^{h'-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \widetilde{\omega}^m_{u,t}(s,a)$$

$$
\begin{aligned}
&= \sum_{h=0}^{h'-1} \left(\prod_{l=h+1}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a) \right) \left(1 - \widetilde{\lambda}_{h\tau,(h+1)\tau}(s,a) \right) \\
&\leq \prod_{l=h'}^{\phi(t)-1} \widetilde{\lambda}_{l\tau,(l+1)\tau}(s,a) \\
&\leq (1-\eta)^{\frac{1}{M}\sum_{m=1}^{M} N^m_{h'\tau,t}(s,a)},
\end{aligned}
$$

where the last inequality follows from

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

due to Jensen's inequality.

Finally, with basic algebraic calculations, (B.38e) is derived as follows:

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

where (i) holds because

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

$$\leq 2\eta(1-(1-\eta)^{\left(N^m_{h\tau,(h+1)\tau}(s,a)\right)}) \tag{B.145}$$

given that $2x - x^2 \geq x$ for $x \leq 1$ and $(1-x^2) \leq 2(1-x)$, (ii) follows from (B.38a), and (iii) follows from the same reasoning of (B.144).

B.6.6 Proof of Lemma 9

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

$$\widetilde{y}^m_{u,t}(s,a) = \begin{cases} \widetilde{\omega}^m_{u,t}(s,a)(P(s,a) - P^m_{u+1}(s,a))V^m_u & \text{if } (s^m_u, a^m_u) = (s,a) \\ 0 & \text{otherwise} \end{cases}, \tag{B.146}$$

where

$$\widetilde{\omega}^m_{u,t}(s,a) = \frac{\eta(1-\eta)^{-N^m_{\phi(u)\tau,u+1}(s,a)}}{M} \prod_{h=\phi(u)}^{\phi(t)-1} \frac{M}{\sum_{m'=1}^M (1-\eta)^{-N^{m'}_{h\tau,(h+1)\tau}(s,a)}}, \tag{B.147}$$

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

$$D(\widetilde{y}, \widehat{y}) \coloneqq \left| \sum_{m=1}^M \sum_{u=0}^{t-1} \left(\widetilde{y}^m_{u,t}(s,a) - \widehat{y}^m_{u,t}(s,a) \right) \right| \tag{B.148}$$

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

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

$$\mathcal{B}_I \coloneqq \bigcap_{u=0}^{t-I\tau} \left\{ \frac{1}{4}\mu_{\mathsf{avg}}(s,a)MI\tau \le \sum_{m=1}^{M} N_{u,u+I\tau}^{m}(s,a) \le 2\mu_{\mathsf{avg}}(s,a)MI\tau \right\}, \tag{B.149}$$

where $I = I(s,a) := \lfloor \frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau} \rfloor$. Note that $I \ge \frac{1}{16\eta\mu_{\mathsf{avg}}(s,a)\tau}$ since $\eta\tau \le 1/16$. Combining this with the assumption $\eta \le \frac{1}{16 t_{\mathsf{th}}(s,a)\mu_{\mathsf{avg}}(s,a)}$ (see (A.3) for the definition of $t_{\mathsf{th}}(s,a)$), it follows that $I\tau \ge t_{\mathsf{th}}(s,a)$ always holds. Then, $\mathcal{B}_I$ holds with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$ according to Lemma 2. The rest of the proof shall be carried out under the event $\mathcal{B}_I$.

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

$$\begin{aligned}
\widetilde{\omega}_{u,t}^{m}(s,a) &= \prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{M}{\sum_{m'=1}^{M}(1-\eta)^{-N_{h\tau,(h+1)\tau}^{m'}(s,a)}} \frac{\sum_{m'=1}^{M}(1-\eta)^{-N_{h\tau,(h+1)\tau}^{m'}(s,a)}}{M} \right) \\
&\quad \times \frac{\eta(1-\eta)^{-N_{\phi(u)\tau,u+1}^{m}(s,a)}}{M} \prod_{h=\phi(u)}^{\phi(t)-1} \frac{M}{\sum_{m'=1}^{M}(1-\eta)^{-N_{h\tau,(h+1)\tau}^{m'}(s,a)}} \\
&= \underbrace{\left(\prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{\sum_{m'=1}^{M}(1-\eta)^{-N_{h\tau,(h+1)\tau}^{m'}(s,a)}}{M} \right) \frac{\eta(1-\eta)^{-N_{\phi(u)\tau,u+1}^{m}(s,a)}}{M} \right)}_{\text{dependent on history until } u} \\
&\quad \times \underbrace{\left(\prod_{h=h_0(u,t)}^{\phi(t)-1} \frac{M}{\sum_{m'=1}^{M}(1-\eta)^{-N_{h\tau,(h+1)\tau}^{m'}(s,a)}} \right)}_{\text{dependent on history and future until } t} \\
&= \underbrace{\left(\prod_{h=h_0(u,t)}^{\phi(u)-1} \left(\frac{\sum_{m'=1}^{M}(1-\eta)^{-N_{h\tau,(h+1)\tau}^{m'}(s,a)}}{M} \right) \frac{\eta(1-\eta)^{-N_{\phi(u)\tau,u+1}^{m}(s,a)}}{M} \right)}_{\coloneqq x_u^m(s,a)} \\
&\quad \times \prod_{l=1}^{l(u,t)} \underbrace{\left(\prod_{h=\max\{0,\phi(t)-lI\}}^{\phi(t)-(l-1)I-1} \frac{M}{\sum_{m'=1}^{M}(1-\eta)^{-N_{h\tau,(h+1)\tau}^{m'}(s,a)}} \right)}_{\coloneqq z_l(s,a)}.
\end{aligned} \tag{B.150}$$

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

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

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

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

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

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

$$\mathcal{Z} \coloneqq \left\{ \exp\left(-\frac{1}{4}\eta\mu_{\mathsf{avg}}(s,a)I\tau - \frac{i\eta}{M}\right) \ \middle|\ i \in \mathbb{Z} :\ \ 0 \le i < 4M\mu_{\mathsf{avg}}(s,a)I\tau \right\}. \tag{B.151}$$

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

$$\mathcal{Z}_0 \coloneqq \left\{ \exp\left(-\frac{i\eta}{M}\right) \ \middle|\ i \in \mathbb{Z} :\ \ 0 \le i < 4M\mu_{\mathsf{avg}}(s,a)I\tau \right\}. \tag{B.152}$$

In sum, we can always find $(\widehat{z}_1, \cdots, \widehat{z}_l, \cdots, \widehat{z}_L) \in \mathcal{Z}^{L-1} \times \mathcal{Z}_0$ where its entry-wise distance

to $(z_l(s,a))_{l\in[L-1]}$ (resp. $z_L(s,a)$) is at most $\eta/Me^{-1/4\eta\mu_{\mathsf{avg}}(s,a)I\tau}$ (resp. η/M).

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

$$\widehat{x}_u^m(s,a) = \begin{cases} x_u^m(s,a) & \text{if } \sum_{m=1}^M N^m_{h_0(u,t)\tau,\phi(u)\tau}(s,a) \le 2M\mu_{\mathsf{avg}}(s,a)I\tau \\ 0 & \text{otherwise} \end{cases}. \tag{B.153}$$

Note that the clipping never occurs and $\widehat{x}_u^m(s,a) = x_u^m(s,a)$ for all u as long as $\mathcal{B}_I$ holds. To provide useful properties of $\widehat{x}_u^m(s,a)$ that will be useful later, we record the following lemma whose proof is provided in Appendix B.6.6.

Lemma 16. *For any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, consider any integers $1 \le t \le T$ and $1 \le l \le \lceil \frac{t}{I\tau} \rceil$, where $I = \lfloor \frac{1}{8\eta\mu_{\mathsf{avg}}(s,a)\tau} \rfloor$. Suppose that $4\eta\tau \le 1$, then $\widehat{x}_u^m(s,a)$ defined in* (B.153) *satisfy*

$$\forall u \in [h_0, \phi(t) - (l-1)I) \quad : \quad \widehat{x}_u^m(s,a) \le \frac{9\eta}{M}, \tag{B.154a}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^M \widehat{x}_u^m(s,a) \le 16\eta\mu_{\mathsf{avg}}(s,a)I\tau, \tag{B.154b}$$

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^M (\widehat{x}_u^m(s,a))^2 \le \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M}, \tag{B.154c}$$

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

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

$$\widehat{y}_{u,t}^m(s,a;\boldsymbol{z}) = \begin{cases} \widehat{\omega}_{u,t}^m(s,a;\boldsymbol{z})(P(s,a) - P_{u+1}^m(s,a))V_u^m & \text{if } (s_u^m, a_u^m) = (s,a) \text{ and } l(u,t) \le L \\ 0 & \text{otherwise} \end{cases}. \tag{B.155}$$

If $t > LI\tau$, for any $u < t - LI\tau$, i.e., $l(u,t) > L$, we set $\widehat{y}_{u,t}^m(s,a;\boldsymbol{z}) = 0$ since the magnitude of $\widetilde{\omega}_{u,t}^m(s,a)$ becomes negligible when the time difference between u and t is

large enough, and the fine-grained approximation using $\mathcal{Z}$ is no longer needed, as shall be seen momentarily. Finally, denote a collection of the approximates induced by $\mathcal{Z}^{L-1} \times \mathcal{Z}_0$ as

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

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

$$D(\widetilde{y}, \widehat{y}_{\boldsymbol{z}}) < \frac{129}{1-\gamma}\sqrt{\frac{L\eta}{M}}. \tag{B.156}$$

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

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

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

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

where (i) holds since $\|P(s,a)\|_1,\ \|P_u^m(s,a)\|_1 \le 1$ and $\|V_{u-1}^m\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)), (ii) follows from (B.38d) in Lemma 8, (iii) holds due to $\mathcal{B}_I$, and (iv) holds because $L \ge 64\log\frac{M}{\eta} \ge \frac{4}{\eta\mu_{\mathsf{avg}}(s,a)I\tau}\log\frac{M}{\eta}$ given that $\eta\mu_{\mathsf{avg}}(s,a)I\tau \ge 1/16$.

- **Bounding D_2.** Since $\widehat{x}_u^m(s,a) = x_u^m(s,a)$ when $\mathcal{B}_I$ holds, in view of (B.155), we have

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

where the last inequality holds since $\|P(s,a)\|_1,\ \|P_u^m(s,a)\|_1 \le 1$ and $\|V_{u-1}^m\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)).

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

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

Then, applying the above bound and (B.154b) in Lemma 16,

$$\min_{\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0}\left|\sum_{m=1}^{M}\sum_{u=t-LI\tau}^{t-1}\widetilde{y}_{u,t}^m(s,a) - \widehat{y}_{u,t}^m(s,a;\boldsymbol{z})\right|$$

$$
\begin{aligned}
&\leq \frac{2}{1-\gamma}\sum_{l=1}^{L}\sum_{h=\phi(t)-lI}^{\phi(t)-(l-1)I-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\sum_{m=1}^{M}\widehat{x}^m_u(s,a)\left|\prod_{l'=1}^{l}z_{l'}(s,a)-\prod_{l'=1}^{l}\widehat{z}^\star_{l'}\right| \\
&\leq \frac{2}{1-\gamma}\frac{L\eta}{M}\sum_{l=1}^{L}\exp\Big(-\frac{1}{4}(l-1)\eta\mu_{\mathsf{avg}}(s,a)I\tau\Big)\sum_{h=\phi(t)-lI}^{\phi(t)-(l-1)I-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\sum_{m=1}^{M}\widehat{x}^m_u(s,a) \\
&\leq \frac{2}{1-\gamma}\frac{L\eta}{M}\frac{1}{1-\exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)I\tau)}(16\eta\mu_{\mathsf{avg}}(s,a)I\tau) \\
&\overset{\text{(i)}}{\leq} \frac{2}{1-\gamma}\frac{L\eta}{M}\frac{8}{\eta\mu_{\mathsf{avg}}(s,a)I\tau}16\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq \frac{256L\eta}{(1-\gamma)M},
\end{aligned}
$$

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

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

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

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

$$
\left|\sum_{m=1}^{M}\sum_{u=0}^{t-1}\widehat{y}^m_{u,t}(s,a;\boldsymbol{z})\right| < \frac{624}{(1-\gamma)}\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} \tag{B.157}
$$

with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T}$. It suffices to establish (B.157) for a fixed $\boldsymbol{z}\in\mathcal{Z}^{L-1}\times\mathcal{Z}_0$ with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\mathcal{Y}|}$, where

$$
|\widehat{\mathcal{Y}}| = |\mathcal{Z}^{L-1}\times\mathcal{Z}_0| \leq (4M\mu_{\mathsf{avg}}(s,a)I\tau)^L \leq (M/\eta)^L \leq (TM)^L. \tag{B.158}
$$

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

$$
\mathbb{E}[\widehat{y}^m_{u,t}(s,a;\boldsymbol{z})|\mathcal{Y}_u] = 0, \tag{B.159}
$$

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

$$B_t(s,a) \coloneqq \max_{m\in[M],0\le u<t} |\widehat{y}_{u,t}^m(s,a;\boldsymbol{z})| \le \widehat{x}_u^m(s,a)\prod_{l=1}^{l(u,t)} \widehat{z}_l \|P(s,a) - P_{u+1}^m(s,a)\|_1 \|V_u^m\|_\infty \le \frac{18\eta}{(1-\gamma)M}, \tag{B.160}$$

where the last inequality follows from $\|P(s,a)\|_1$, $\|P_u^m(s,a)\|_1 \le 1$, $\|V_{u-1}^m\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)), $\widehat{z}_l \le 1$, and (B.154a) in Lemma 16. Next, we can bound the variance as

$$\begin{aligned}
W_t(s,a) &\coloneqq \sum_{u=t-LI\tau}^{t-1}\sum_{m=1}^{M} \mathbb{E}[(\widehat{y}_{u,t}^m(s,a;\boldsymbol{z}))^2|\mathcal{Y}_u] \\
&= \sum_{l=1}^{L}\sum_{h=\max\{0,\phi(t)-lI\}}^{\phi(t)-(l-1)I-1}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)} (\widehat{x}_u^m(s,a)\prod_{l'=1}^{l}\widehat{z}_{l'})^2 \mathsf{Var}_{P(s,a)}(V_u^m) \\
&\overset{\text{(i)}}{\le} \frac{2}{(1-\gamma)^2}\sum_{l=1}^{L}\left(\prod_{l'=1}^{l}\widehat{z}_{l'}^2\right)\sum_{h=max\{0,\phi(t)-lI\}}^{\phi(t)-(l-1)I-1}\sum_{m=1}^{M}\sum_{u\in\mathcal{U}_{h\tau,(h+1)\tau}^m(s,a)} (\widehat{x}_u^m(s,a))^2 \\
&\overset{\text{(ii)}}{\le} \frac{2}{(1-\gamma)^2}\sum_{l=1}^{L}\left(\prod_{l'=1}^{l}\widehat{z}_{l'}^2\right)\frac{64\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M} \\
&\overset{\text{(iii)}}{\le} \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M(1-\gamma)^2}\sum_{l=1}^{L}\exp\left(-1/2(l-1)\eta\mu_{\mathsf{avg}}(s,a)I\tau\right) \\
&\le \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M(1-\gamma)^2}\frac{1}{1-\exp(-1/2\eta\mu_{\mathsf{avg}}(s,a)I\tau)} \\
&\overset{\text{(iv)}}{\le} \frac{128\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M(1-\gamma)^2}\frac{4}{\eta\mu_{\mathsf{avg}}(s,a)I\tau} = \frac{512\eta}{M(1-\gamma)^2} \coloneqq \sigma^2,
\end{aligned} \tag{B.161}$$

where (i) holds due to the fact that $\|\mathsf{Var}_P(V)\|_\infty \le \|P\|_1(\|V\|_\infty)^2 + (\|P\|_1\|V\|_\infty)^2 \le \frac{2}{(1-\gamma)^2}$ because $\|V\|_\infty \le \frac{1}{1-\gamma}$ (cf. (B.2)) and $\|P\|_1 \le 1$, (ii) follows from (B.154c) in Lemma 16, (iii) holds due to the range of $\mathcal{Z}$ and $\mathcal{Z}_0$ is bounded by $\exp(-1/4\eta\mu_{\mathsf{avg}}(s,a)I\tau)$ and 1, respectively, and (iv) holds since $e^{-x} \le 1 - \frac{1}{2}x$ for any $0 \le x \le 1$ and $1/2\eta\mu_{\mathsf{avg}}(s,a)I\tau \le 1$

.

Now, by substituting the above bounds of W_t and B_t into Freedman's inequality (see

Theorem 10) and setting $c = 1$, it follows that for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $t \in [T]$ and $\widehat{y}_{\boldsymbol{z}} \in \widehat{\mathcal{Y}}$,

$$\begin{aligned}\left|\sum_{m=1}^{M}\sum_{u=0}^{t-1}\widehat{y}_{u,t}^{m}(s,a;\boldsymbol{z})\right| &\leq \sqrt{8\max\{W_t(s,a),\frac{\sigma^2}{2^c}\}\log\frac{4c|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}} + \frac{4}{3}B_t(s,a)\log\frac{4c|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta} \\ &\leq \sqrt{4096\frac{\eta}{M(1-\gamma)^2}\log\frac{4|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta}} + \frac{24\eta}{M(1-\gamma)}\log\frac{4|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}{\delta} \\ &\overset{\text{(i)}}{\leq} \frac{78}{(1-\gamma)}\sqrt{\frac{\eta L}{M}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}},\end{aligned} \tag{B.162}$$

with at least probability $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|T|\widehat{\mathcal{Y}}|}$, where (i) holds because $|\widehat{\mathcal{Y}}| \leq (TM)^L$ given that $\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq 1/4$, and $\frac{4\eta L}{M}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta} \leq 1$. Therefore, it follows that (B.157) holds.

Step 4: putting things together. We now putting all the results obtained in the previous steps together to achieve the claimed bound. Under $\mathcal{B}_I$, there always exists $\widehat{y}_{\boldsymbol{z}} := \widehat{y}_{\boldsymbol{z}(\widetilde{y})} \in \widehat{\mathcal{Y}}$ such that (B.156) holds. Hence, setting $q = \frac{2064}{(1-\gamma)}\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}}$,

$$\begin{aligned}\sum_{m=1}^{M}\sum_{u=0}^{t-1}\widetilde{y}_{u,t}^{m}(s,a) &\leq \left|\sum_{m=1}^{M}\sum_{u=0}^{t-1}\widehat{y}_{u,t}^{m}(s,a;\boldsymbol{z})\right| + D(\widetilde{y},\widehat{y}_{\boldsymbol{z}}) \\ &\leq \frac{78}{(1-\gamma)}\sqrt{\frac{\eta L}{M}\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}} + \frac{129}{1-\gamma}\sqrt{\frac{L\eta}{M}} \\ &\leq \frac{2064}{(1-\gamma)}\sqrt{\frac{\eta}{M}\log(TM)\log\frac{4|\mathcal{S}||\mathcal{A}|T^2M}{\delta}},\end{aligned} \tag{B.163}$$

where the second line holds due to (B.157) and (B.156), and the last line holds due to $L \leq 64\log(TM)$. By taking a union bound over all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $t \in [T]$, we complete the proof.

Proof of Lemma 16

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

$$\sum_{m=1}^{M} N^m_{h_0\tau,(h-1)\tau}(s,a) \leq 2M\mu_{\mathsf{avg}}(s,a)I\tau. \tag{B.164}$$

Then, the following holds:

$$\begin{aligned}\sum_{m=1}^{M} N^m_{h_0\tau,\overline{h}\tau}(s,a) &= \sum_{m=1}^{M} N^m_{(\overline{h}-1)\tau,\overline{h}\tau}(s,a) + \sum_{m=1}^{M} N^m_{h_0\tau,(\overline{h}-1)\tau}(s,a) \\ &\leq M\tau + 2M\mu_{\mathsf{avg}}(s,a)I\tau. \end{aligned} \tag{B.165}$$

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

$$\begin{aligned}\sum_{m'=1}^{M} \frac{(1-\eta)^{-N^{m'}_{h\tau,(h+1)\tau}(s,a)}}{M} \leq \frac{\sum_{m'=1}^{M} e^{\eta N^{m'}_{h\tau,(h+1)\tau}(s,a)}}{M} &\leq 1 + 2\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)}{M} \\ &\leq \exp\left(2\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)}{M}\right), \end{aligned} \tag{B.166}$$

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

According to (B.153), for any integer $u \in [\overline{h}\tau, t - (l-1)I\tau)$, $\widehat{x}^m_u(s,a)$ is clipped to zero. Now, we prove the bounds in Lemma 16 respectively.

Proof of (B.154a)**.** For $u \in [h_0\tau, \overline{h}\tau)$,

$$\begin{aligned}\widehat{x}^m_u(s,a) &= \prod_{h=h_0}^{\phi(u)-1} \left(\frac{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{h\tau,(h+1)\tau}(s,a)}}{M}\right) \frac{\eta(1-\eta)^{-N^{m}_{\phi(u)\tau,u+1}(s,a)}}{M} \\ &\overset{\text{(i)}}{\leq} \prod_{h=h_0}^{\phi(u)-1} \left(\frac{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{h\tau,(h+1)\tau}(s,a)}}{M}\right) \frac{3\eta}{M} \end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \exp\left(\frac{2\eta}{M}\sum_{m'=1}^{M} N^{m'}_{h_0\tau,(\overline{h}-1)\tau}(s,a)\right)\frac{3\eta}{M}$$
$$\overset{\text{(iii)}}{\leq} \exp(4\eta\mu_{\mathsf{avg}}(s,a)I\tau)\frac{3\eta}{M} \overset{\text{(iv)}}{\leq} \frac{9\eta}{M}, \tag{B.167}$$

where (i) holds since $(1+\eta)^x \leq e^{\eta x}$ and $\eta N^m_{\phi(u)\tau,u+1}(s,a) \leq \eta\tau \leq 1$, (ii) holds due to (B.166) and the fact that $\phi(u) \leq \overline{h}-1$, (iii) follows from the definition of $\overline{h}$ in (B.164), and (iv) holds because $4\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq 1$.

Proof of (B.154b)**.** By the definition of $\overline{h}$, it follows that

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\sum_{m=1}^{M}\widehat{x}^m_u(s,a) = \sum_{h=h_0}^{\overline{h}-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\sum_{m=1}^{M}x^m_u(s,a).$$

Using the following relation for each h:

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

and applying (B.166), we can complete the proof as follows:

$$\begin{aligned}
\sum_{h=h_0}^{\overline{h}-1}\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)}\sum_{m=1}^{M}x^m_u(s,a) &\leq \prod_{h'=h_0}^{\overline{h}-1}\exp\left(\frac{2\eta\sum_{m'=1}^{M}N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}{M}\right)-1\\
&\leq \exp\left(\frac{2\eta\sum_{m'=1}^{M}N^{m'}_{h_0\tau,\overline{h}\tau}(s,a)}{M}\right)-1
\end{aligned}$$

$$\overset{\text{(i)}}{\leq} \exp\left(4\eta\mu_{\mathsf{avg}}(s,a)I\tau + 2\eta\tau\right) - 1$$
$$\overset{\text{(ii)}}{\leq} 16\eta\mu_{\mathsf{avg}}(s,a)I\tau,$$

where (i) follows from (B.165), and (ii) holds because $e^x \leq 1 + 2x$ for any $x \in [0,1]$ and $2\eta\tau \leq 4\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq 1/2$.

Proof of (B.154c)**.** Similarly,

$$\sum_{h=h_0}^{\phi(t)-(l-1)I-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (\widehat{x}^m_u(s,a))^2 = \sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (x^m_u(s,a))^2.$$

Using the following relation for each h:

$$\begin{aligned}
&\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (x^m_u(s,a))^2 \\
&= \left(\prod_{h'=h_0}^{h-1} \frac{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}}{M}\right)^2 \sum_{m=1}^{M} \frac{\sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \eta^2(1-\eta)^{-2N^m_{h\tau,u+1}(s,a)}}{M^2} \\
&\leq \left(\prod_{h'=h_0}^{h-1} \frac{\sum_{m'=1}^{M}(1-\eta)^{-N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}}{M}\right)^2 \sum_{m=1}^{M} \frac{\eta((1-\eta)^{-2N^m_{h\tau,(h+1)\tau}(s,a)} - 1)}{M^2} \\
&\leq \frac{\eta}{M}\left(\prod_{h'=h_0}^{h-1} \exp\left(2\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h'\tau,(h'+1)\tau}(s,a)}{M}\right)\right)^2 \left(\exp\left(4\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)}{M}\right) - 1\right) \\
&= \frac{\eta}{M}\exp\left(4\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h_0\tau,h\tau}(s,a)}{M}\right)\left(\exp\left(4\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h\tau,(h+1)\tau}(s,a)}{M}\right) - 1\right) \\
&= \frac{\eta}{M}\left(\exp\left(4\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h_0\tau,(h+1)\tau}(s,a)}{M}\right) - \exp\left(4\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h_0\tau,h\tau}(s,a)}{M}\right)\right), \qquad \text{(B.168)}
\end{aligned}$$

where the inequality is derived similarly to (B.166) under the condition $2\eta\tau \leq 1$, we can complete the proof as follows:

$$\sum_{h=h_0}^{\overline{h}-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \sum_{m=1}^{M} (x^m_u(s,a))^2 \leq \frac{\eta}{M}\left(\exp\left(4\eta\frac{\sum_{m'=1}^{M} N^{m'}_{h_0\tau,\overline{h}\tau}(s,a)}{M}\right) - 1\right)$$

$$\overset{\text{(i)}}{\leq} \frac{\eta}{M}\left(\exp\left(8\eta\mu_{\mathsf{avg}}(s,a)I\tau + 4\eta\tau\right) - 1\right)$$
$$\overset{\text{(ii)}}{\leq} \frac{64\eta^2\mu_{\mathsf{avg}}(s,a)I\tau}{M}, \tag{B.169}$$

where (i) follows from (B.165), and (ii) holds because $e^x \leq 1 + 4x$ for any $x \in [0, 2]$ and $4\eta\tau \leq 8\eta\mu_{\mathsf{avg}}(s,a)I\tau \leq 1$.

B.6.7 Proof of Lemma 10

The proof follows a similar structure to that of Lemma 7. We omit common parts of the proofs and refer to Appendix B.6.4 to check the detailed derivations. First, we decompose the error term as follows:

$$E_t^3(s,a) = \underbrace{\gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}^m_{0,(\phi(t)-\beta)\tau}(s,a)} \widetilde{\omega}^m_{u,t}(s,a) P(s,a)(V^\star - V^m_u)}_{=:E_t^{3a}(s,a)}$$
$$+ \underbrace{\gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}^m_{(\phi(t)-\beta)\tau,t}(s,a)} \widetilde{\omega}^m_{u,t}(s,a) P(s,a)(V^\star - V^m_u)}_{=:E_t^{3b}(s,a)}. \tag{B.170}$$

We shall bound these two terms separately.

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

$$|E_t^{3a}(s,a)| \leq \gamma \sum_{m=1}^{M} \sum_{u \in \mathcal{U}^m_{0,(\phi(t)-\beta)\tau}(s,a)} \widetilde{\omega}^m_{u,t}(s,a) \|P(s,a)\|_1 \|V^\star - V^m_u\|_\infty$$
$$\overset{\text{(i)}}{\leq} \frac{2}{1-\gamma}(1-\eta)^{\frac{1}{M}\sum_{m=1}^{M} N^m_{(\phi(t)-\beta)\tau,t}(s,a)}$$
$$\overset{\text{(ii)}}{\leq} \frac{2}{1-\gamma}(1-\eta)^{\frac{\mu_{\mathsf{avg}}\beta\tau}{4}}, \tag{B.171}$$

where (i) holds due to Lemma 8 (cf. (B.38d)), and (ii) follows fromapplying Lemma 2

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

$$\sum_{m=1}^{M} N^m_{(\phi(t)-\beta)\tau,t}(s,a) \geq \frac{M\beta\tau\mu_{\mathsf{avg}}}{4}$$

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

- **Bounding $E^{3b}_t(s,a)$.** Combining (B.125) and Lemma 14 to bound $\|V^\star - V^m_u\|_\infty$, we bound $E^{3b}_t(s,a)$ as follows:

$$\begin{aligned}
|E^{3b}_t(s,a)| &\leq \gamma \sum_{m=1}^{M} \sum_{u\in\mathcal{U}^m_{(\phi(t)-\beta)\tau,t}(s,a)} \widetilde{\omega}^m_{u,t}(s,a)\,\|V^\star - V^m_u\|_\infty \\
&\leq \gamma \sum_{m=1}^{M} \sum_{h=\phi(t)-\beta}^{\phi(t)-1} \sum_{u\in\mathcal{U}^m_{h\tau,(h+1)\tau}(s,a)} \widetilde{\omega}^m_{u,t}(s,a)((1+2\eta\tau)\|\Delta_{h\tau}\|_\infty + \sigma_{\mathsf{local}}) \\
&\leq \sigma_{\mathsf{local}} + \frac{1+\gamma}{2} \max_{\phi(t)-\beta\leq h<\phi(t)} \|\Delta_{h\tau}\|_\infty \qquad \text{(B.172)}
\end{aligned}$$

  where we denote $\sigma_{\mathsf{local}} \coloneqq \frac{8\gamma\eta\sqrt{\tau-1}}{1-\gamma}\sqrt{\log\frac{2|\mathcal{S}||\mathcal{A}|TM}{\delta}}$ for notational simplicity, and the last inequality follows from Lemma 8 (cf. (B.38c)) and the assumption that $\eta \leq \frac{1-\gamma}{4\gamma\tau}$.

Now we have the bounds of $E^{3a}_t(s,a)$ and $E^{3b}_t(s,a)$ separately obtained above. By combining the bounds in (B.170), we can claim the advertised bound, which completes the proof.

Appendix C

Analysis Federated Q-Learning in Average-Reward MDPs

C.1 Preliminaries

Solutions to Bellman equation. By (2.3), the optimal bias function $h^\star$ satisfies the Bellman optimality equation. For convenience, we restate it here:

$$J^\star + h^\star(s) = \max_{a\in\mathcal{A}} \left[r(s,a) + \sum_{s'\in\mathcal{S}} P(s'|s,a)h^\star(s') \right] = \max_{a\in\mathcal{A}} \left[r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[h^\star(s')] \right]. \tag{C.1}$$

By introducing the bias Q-function $h_q^\star(s,a)$ defined as

$$h_q^\star(s,a) \coloneqq r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[h^\star(s')] - J^\star,$$

we can rewrite (C.1) as

$$J^\star + h_q^\star(s,a) = r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[h^\star(s')], \quad \text{and} \tag{C.2}$$

$$\max_{a\in\mathcal{A}} h_q^\star(s,a) = \max_{a\in\mathcal{A}} \left[r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[h^\star(s')] \right] - J^\star = h^\star(s). \tag{C.3}$$

Motivated by the observation that (C.2) is shift-invariant, we define normalized value functions $V^\star$ and $Q^\star$ by shifting $h^\star(s)$ and $h_q^\star(s,a)$ by a constant c:

$$V^\star(s) \coloneqq h^\star(s) - c, \qquad Q^\star(s,a) \coloneqq h_q^\star(s,a) - c, \quad \text{where} \quad c \coloneqq \frac{\max_{s'} h^\star(s') + \min_{s'} h^\star(s')}{2}. \tag{C.4}$$

This normalization ensures that $\|V^\star\|_\infty = \|h^\star\|_{\mathsf{span}}$. One can then verify that $(V^\star, Q^\star)$ satisfy the same Bellman relations, namely

$$J^\star + Q^\star(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\star(s')], \quad \text{and} \quad \max_{a \in \mathcal{A}} Q^\star(s,a) = V^\star(s). \tag{C.5}$$

By virtue of (C.5), one can straightforwardly verify that provided $\|h^\star\|_{\mathsf{span}} \gtrsim 1$, the ℓ_∞ norm of $V^\star$ and $Q^\star$ are controlled by the span semi-norm of the bias function:

$$\|V^\star\|_\infty = \|h^\star\|_{\mathsf{span}}, \qquad \|Q^\star\|_\infty \le \|V^\star\|_\infty + \|r\|_\infty + J^\star \le \|h^\star\|_{\mathsf{span}} + 2 \lesssim \|h^\star\|_{\mathsf{span}}. \tag{C.6}$$

Two auxiliary sequences. Next, we introduce two auxiliary sequences. The first is based on the discounted value and Q functions $V_{\gamma_k}^\star(s)$ and $Q_{\gamma_k}^\star(s,a)$, $k = 1, \ldots, K$, whose general definitions are given in (2.4). Under the assumption that the AMDP is weakly communicating, Lemmas 6-8 of Wang et al. (2022) imply the following bounds:

Lemma 17. *If an AMDP is a weakly communicating MDP, for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$|V_\gamma^\star(s) - J^\star| \le 4(1-\gamma)\|h^\star\|_{\mathsf{sp}} \quad \textit{and} \quad |Q_\gamma^\star(s,a) - J^\star| \le 4(1-\gamma)\|h^\star\|_{\mathsf{sp}}. \tag{C.7}$$

Also, note that for any epoch k and iteration t, and state $s \in \mathcal{S}$,

$$|V_{k,t}(s) - V_\gamma^\star(s)| \le \max_{a \in \mathcal{A}} |Q_{k,t}(s,a) - Q_\gamma^\star(s,a)| \le 1. \tag{C.8}$$

The second auxiliary sequence builds on the normalized functions $V^\star$ and $Q^\star$ from

(C.4). Define the sequence of value functions $\{V_k^\star\}_{k\geq 1} \subset \mathbb{R}^S$ by

$$V_k^\star(s) \coloneqq J^\star + \frac{1}{k} V^\star(s), \tag{C.9}$$

which converges to $J^\star \mathbf{1}$ as $k \to \infty$. Correspondingly, we introduce the Q-functions $Q_{k+1}^\star \in \mathbb{R}^{SA}$ via a Bellman-like update:

$$Q_{k+1}^\star(s,a) \coloneqq \frac{1}{k+1} r(s,a) + \frac{k}{k+1} \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_k^\star(s')], \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{C.10}$$

with initialization $Q_1^\star(s, \pi^\star(s)) = V_1^\star(s)$ and $Q_1^\star(s,a) = 0$ for $a \neq \pi^\star(s)$. By virtue of (C.6), we obtain

$$\|V_k^\star - J^\star\|_\infty = \frac{1}{k} \|V^\star\|_\infty \leq \frac{\|h^\star\|_{\mathsf{sp}}}{k}. \tag{C.11}$$

Furthermore, the sequence $Q_k^\star$ satisfies the following identities for all $k \geq 1$:

$$\begin{aligned} Q_{k+1}^\star(s,a) &= J^\star + \frac{1}{k+1} Q^\star(s,a), \quad V_k^\star(s) \\ &= \max_{a \in \mathcal{A}} Q_k^\star(s,a), \quad \|Q_{k+1}^\star\|_\infty \leq 1 + \frac{2 + \|h^\star\|_{\mathsf{span}}}{k+1}. \end{aligned} \tag{C.12}$$

Proof of (C.12)**.** Substituting (C.9) into (C.10), we obtain

$$\begin{aligned} Q_{k+1}^\star(s,a) &= \frac{1}{k+1} \left(r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\star(s')] \right) + \frac{k}{k+1} J^\star \\ &\overset{\text{(C.5)}}{=} \frac{1}{k+1} \left(J^\star + Q^\star(s,a) \right) + \frac{k}{k+1} J^\star = J^\star + \frac{1}{k+1} Q^\star(s,a), \qquad k \geq 1. \end{aligned} \tag{C.13}$$

Moreover, we have

$$\begin{aligned} \max_{a \in \mathcal{A}} Q_k^\star(s,a) &\overset{\text{(C.13)}}{=} \max_{a \in \mathcal{A}} \left[J^\star + \frac{1}{k} Q^\star(s,a) \right] \\ &= J^\star + \frac{1}{k} \max_{a \in \mathcal{A}} Q^\star(s,a) \\ &\overset{\text{(C.5)}}{=} J^\star + \frac{1}{k} V^\star(s) \end{aligned}$$

$$\overset{(C.9)}{=} V_k^\star(s), \qquad \forall k \geq 1.$$

Furthermore, by virtue of (C.6), and (C.13), we have

$$\|Q_{k+1}^\star\|_\infty \leq \|J^\star\|_\infty + \frac{\|Q^\star\|_\infty}{k+1} \leq 1 + \frac{2 + \|h^\star\|_{\mathsf{span}}}{k+1}.$$

Useful properties of the learning rates. The learning rates specified in (3.10a) and (3.14a) have the following properties:

$$\forall i \leq t \ : \quad \eta_{k,i} \prod_{j=i+1}^{t} (1 - \eta_{k,j}) \leq \eta_{k,t} \tag{C.14a}$$

$$\forall t \geq 1 \ : \quad \sum_{i=1}^{t} \eta_{k,i} \prod_{j=i+1}^{t} (1 - \eta_{k,j}) + \prod_{j=1}^{t} (1 - \eta_{k,j}) = 1. \tag{C.14b}$$

Proof of (C.14)**.** For the learning rates defined in (3.10a), (C.14a) can be proved as follows:

$$\begin{aligned}
\eta_{k,i} \prod_{j=i+1}^{t} (1 - \eta_{k,j}) &= \frac{1}{1 + \frac{i^{2/3}}{8\log(4i)}} \frac{\frac{(i+1)^{2/3}}{8\log(4(i+1))}}{1 + \frac{(i+1)^{2/3}}{8\log(4(i+1))}} \cdots \frac{\frac{t^{2/3}}{8\log(4t)}}{1 + \frac{t^{2/3}}{8\log(4t)}} \\
&= \frac{\frac{(i+1)^{2/3}}{8\log(4(i+1))}}{1 + \frac{i^{2/3}}{8\log(4i)}} \frac{\frac{(i+2)^{2/3}}{8\log(4(i+2))}}{1 + \frac{(i+1)^{2/3}}{8\log(4(i+1))}} \cdots \frac{1}{1 + \frac{t^{2/3}}{8\log(4t)}} \\
&\leq \frac{1}{1 + \frac{t^{2/3}}{8\log(4t)}} = \eta_{k,t},
\end{aligned} \tag{C.15}$$

where the last inequality follows from the fact that $(1+i)^{2/3} \leq 1 + i^{2/3}$ for any $i \geq 1$ due to subadditivity. Similarly, the property (C.14a) can be proved for the learning rates defined in (3.14a). Next, we prove (C.14b) by induction. For $t = 1$, it holds that

$$\sum_{i=1}^{1} \eta_{k,i} \prod_{j=i+1}^{1} (1 - \eta_{k,j}) + \prod_{j=1}^{1} (1 - \eta_{k,j}) = \eta_{k,1} + (1 - \eta_{k,1}) = 1. \tag{C.16}$$

Then, suppose the statement holds for $t-1$, and we show it also holds for t:

$$\begin{aligned}
&\sum_{i=1}^{t} \eta_{k,i} \prod_{j=i+1}^{t} (1-\eta_{k,j}) + \prod_{j=1}^{t}(1-\eta_{k,j}) \\
&= \sum_{i=1}^{t-1} \eta_{k,i} \prod_{j=i+1}^{t-1} (1-\eta_{k,j})(1-\eta_{k,t}) + \eta_{k,t} + \prod_{j=1}^{t-1}(1-\eta_{k,j})(1-\eta_{k,t}) \\
&= (1-\eta_{k,t}) \left(\sum_{i=1}^{t-1} \eta_{k,i} \prod_{j=i+1}^{t-1} (1-\eta_{k,j}) + \prod_{j=1}^{t-1}(1-\eta_{k,j}) \right) + \eta_{k,t} \\
&= (1-\eta_{k,t})(1) + \eta_{k,t} = 1. \qquad \text{(C.17)}
\end{aligned}$$

Notation. In the remainder of the proof, we use vectors $V^\star, V_k^\star \in \mathbb{R}^S$ and $Q^\star, Q_k^\star, r \in \mathbb{R}^{SA}$ to denote the respective mappings from $\mathcal{S}$ or $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$. Let $P \in \mathbb{R}^{SA \times S}$ denote the transition probability matrix, where $P(s,a) = P(\cdot|s,a)$ is the probability vector corresponding to the state transition at the state-action pair (s,a). In addition, we define the local empirical transition matrix at the t-th iteration in epoch k in the single-agent setting as $P_{k,t} \in \{0,1\}^{SA \times S}$ as

$$P_{k,t}((s,a),s') \coloneqq \begin{cases} 1, & \text{if } s' = s_{k,t}(s,a) \\ 0, & \text{otherwise} \end{cases}. \qquad \text{(C.18)}$$

In the federated setting, we denote the local empirical transition matrix for the m-th agent as $P_{k,t}^m \in \{0,1\}^{SA \times S}$ in a similar way. Let $P^\pi \in \mathbb{R}^{S \times S}$ be the transition matrix under policy π, with the s-th row equal to $P(\cdot|s,\pi(s))$. Similarly, define $r^\pi \in \mathbb{R}^S$ as the reward vector under policy π, with the s-th entry given by $r(s,\pi(s))$. For convenience, we write $P^\star = P^{\pi^\star}$ and $r^\star = r^{\pi^\star}$. Notation $J^\star$ may refer to a scalar or to a vector with dimension implied by the context. Finally, define the quantity $T_k \coloneqq \sum_{j=1}^{k} N_j$. The inequalities $\leq$ and $\geq$ between vectors are understood entry-wise.

Armed with these notation, update rules (3.9) in single-agent setting and (3.5) in fed-

erated setting are expressed by

$$Q_{k,t} = (1-\eta_{k,t})Q_{k,t-1} + \eta_{k,t}\bigg((1-\gamma_k)r + \gamma_k P_{k,t}V_{\iota(k,t)}\bigg), \tag{C.19a}$$

$$Q^m_{k,t} = (1-\eta_{k,t})Q^m_{k,t-1} + \eta_{k,t}\bigg((1-\gamma_k)r + \gamma_k P^m_{k,t}V_{k,\iota(k,t)}\bigg), \qquad \forall 1 \le m \le M. \tag{C.19b}$$

C.2 Analysis in the single-agent setting (Theorem 2)

C.2.1 Analysis for the first group of parameters

For each epoch $k \in [K]$, we decompose the error as

$$\|Q_{k,t} - J^\star\|_\infty \le \|Q^\star_{\gamma_k} - J^\star\|_\infty + \|Q_{k,t} - Q^\star_{\gamma_k}\|_\infty. \tag{C.20}$$

The first error $\|Q^\star_{\gamma_k} - J^\star\|_\infty$, which arises from horizon mismatch, can be directly bounded according to Lemma 17 as follows:

$$\|Q^\star_{\gamma_k} - J^\star\|_\infty \le 3(1-\gamma_k)\|h^\star\|_{\mathsf{sp}}. \tag{C.21}$$

Now, it suffices to show that the second error is bounded as

$$\|Q_{k,t} - Q^\star_{\gamma_k}\|_\infty \lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}}. \tag{C.22}$$

Denote $\Delta_{k,t} = Q_{k,t} - Q^\star_{\gamma_k}$. Since γ_k remains fixed within stage k, the analysis follows a similar structure to the discounted MDP case in Li et al. (2024a). The key difference is that we derive error bounds in terms of both the discount factor and the span norm of the bias, ensuring that the second error converges at the same rate as the first error (C.21), thereby balancing the convergence bottleneck.

Step 1: error decomposition. Using the fact that $Q^\star_{\gamma_k} = (1-\gamma_k)r + \gamma_k P V^\star_{\gamma_k}$, we can write the error as

$$\Delta_{k,t} = (1-\eta_{k,t})\Delta_{k,t-1} + \eta_{k,t}\gamma_k \left(P_{k,t}V_{k,\iota(k,t)} - PV^\star_{\gamma_k}\right).$$

Then, it follows that

$$\begin{aligned}
\Delta_{k,t} &= \prod_{i=1}^{t}(1-\eta_{k,i})\Delta_{k,0} + \gamma_k \sum_{i=1}^{t}\eta_{k,i}\left(\prod_{j=i+1}^{t}(1-\eta_{k,j})\right)\left(P_{k,i}V_{k,\iota(k,i)} - PV^\star_{\gamma_k}\right) \\
&= \underbrace{\prod_{i=1}^{t}(1-\eta_{k,i})\Delta_{k,0} + \gamma_k \sum_{i=1}^{t-g_k}\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})P_{k,i}\left(V_{k,\iota(k,i)} - V^\star_{\gamma_k}\right)}_{=:E^1_{k,t}} \\
&\quad + \underbrace{\gamma_k \sum_{i=1}^{t}\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})\left(P_{k,i}-P\right)V^\star_{\gamma_k}}_{=:E^2_{k,t}} \\
&\quad + \underbrace{\gamma_k \sum_{i=t-g_k+1}^{t}\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})P_{k,i}\left(V_{k,\iota(k,i)} - V^\star_{\gamma_k}\right)}_{=:E^3_{k,t}},
\end{aligned} \tag{C.23}$$

where $g_k \coloneqq \left\lceil \frac{-\log(1-\gamma_k)^2}{\eta_{k,N_k}} \right\rceil$.

Step 2: bounding the decomposed errors. We can bound the error terms separately as follows:

- Bounding the initialization error $E^1_{k,t}$. For sufficiently large $t \geq g_k = \left\lceil \frac{-\log(1-\gamma_k)^2}{\eta_{k,N_k}} \right\rceil$, with the proposed learning rates (C.14), the initialization error decays at least in the rate of

$$\prod_{i=\lfloor t-g_k \rfloor+1}^{t}(1-\eta_{k,i}) \leq (1-\eta_{k,t})^{g_k} \leq \exp(-\eta_{k,t}g_k) \leq (1-\gamma_k)^2. \tag{C.24}$$

  Since $\|V_{k,\iota(k,i)} - V^\star_{\gamma_k}\|_\infty,\ \|\Delta_{k,0}\|_\infty \leq 1$ according to (C.8), it follows that $\|E_1\|_\infty \leq 2(1-\gamma_k)^2$.

- Bounding the transition variance $E_{k,t}^2$. By applying Bernstein inequality, we obtain

Lemma 18. *For any $((k,t),s,a) \in [T_k] \times \mathcal{S} \times \mathcal{A}$ and $\delta \in (0,1)$, the following holds*

$$\|E_{k,t}^2\|_\infty \lesssim (1-\gamma_k)^2 \|h^\star\|_{\mathsf{sp}} \log \frac{2|\mathcal{S}||\mathcal{A}|T_K}{\delta} \tag{C.25}$$

at least with probability $1-\delta$, where $T_K = \sum_{k=1}^K N_k$.

Proof of Lemma 18. Let $z_{k,i}$ be a random vector defined as $z_{k,i} = \eta_{k,i} \prod_{j=i+1}^t (1-\eta_{k,j}) (P_{k,i} - P) V_{\gamma_k}^\star$, then

$$E_{k,t}^2 = \sum_{i=1}^t z_{k,i}. \tag{C.26}$$

Since each $z_{k,i}$ is an independent random variable with zero mean, we aim to apply the Hoeffding inequality to bound $E_{k,t}^2$. Due to the properties of the learning rates (C.14), we first derive the following bounds:

$$\sum_{i=1}^t \left(\eta_{k,i} \prod_{j=i+1}^t (1-\eta_{k,j}) \right)^2 \|V_{\gamma_k}^\star\|_{\mathsf{sp}}^2 \le \eta_{k,t} \|V_{\gamma_k}^\star\|_{\mathsf{sp}}^2 \le 16(1-\gamma_k)^2 (6(1-\gamma_k)\|h^\star\|_{\mathsf{sp}})^2, \tag{C.27}$$

where $\eta_{k,t} \le 16(1-\gamma_k)^2$ for $t \ge \frac{N_k}{2}$ and

$$\|V_{\gamma_k}^\star\|_{\mathsf{sp}} = \left| \max_s V_{\gamma_k}^\star(s) - \min_s V_{\gamma_k}^\star(s) \right| \le 6(1-\gamma_k)\|h^\star\|_{\mathsf{sp}}$$

according to Lemma 17. Applying the Hoeffding inequality yields

$$\begin{aligned}
\left\| \sum_{i=1}^t z_{k,i} \right\|_\infty &\le \sqrt{\sum_{i=1}^t \left(\eta_{k,i} \prod_{j=i+1}^t (1-\eta_{k,j}) \right)^2 \|V_{\gamma_k}^\star\|_{\mathsf{sp}}^2 \log \frac{2|\mathcal{S}||\mathcal{A}|N}{\delta}} \\
&\le \sqrt{576(1-\gamma_k)^4 \|h^\star\|_{\mathsf{sp}}^2 \log \frac{2|\mathcal{S}||\mathcal{A}|T_K}{\delta}} \\
&\le 24(1-\gamma_k)^2 \|h^\star\|_{\mathsf{sp}} \log \frac{2|\mathcal{S}||\mathcal{A}|T_K}{\delta},
\end{aligned} \tag{C.28}$$

where $T_K = \sum_{k=1}^{K} N_k$, and this completes the proof.

- Bounding the recursive optimality gap $E_{k,t}^3$. Using the fact that $\|V_{k,i} - V_{\gamma_k}^\star\|_\infty \leq \|\Delta_{k,i}\|_\infty$, following from (C.8),

$$\|E_{k,t}^3\|_\infty \leq \gamma_k \max_{\iota(k,t-g_k+1)\leq i<t} \|\Delta_{k,i}\|_\infty. \tag{C.29}$$

Then, by combining the error bounds altogether, we obtain

$$\|\Delta_{k,t}\|_\infty \lesssim (1-\gamma_k)^2 \|h^\star\|_{\mathsf{sp}} + \gamma_k \max_{t-2g_k\leq i<t} \|\Delta_{k,i}\|_\infty \tag{C.30}$$

for sufficiently large $t \geq 2g_k = 2\left\lceil \frac{-\log(1-\gamma_k)^2}{\eta_{k,N_k}} \right\rceil$. Note that given $N_k \geq 1000$ for any $k \in [K]$, it follows $N_k \geq 2g_k$ since $(\eta_k)^{-1} \leq 2(N_k)^{2/3}$ and $(N_k)^{-1/3} \leq (1-\gamma_k) \leq 1$.

Step 3: solving recursion. Now, we solve the recursive relation we obtained from the previous step as follows:

$$\begin{aligned}
\|\Delta_{k,N_k}\|_\infty &\lesssim (1-\gamma_k)^2 \|h^\star\|_{\mathsf{sp}} + \gamma_k \max_{N_k-2g_k\leq i<N_k} \|\Delta_{k,i}\|_\infty \\
&\lesssim (1-\gamma_k)^2 \|h^\star\|_{\mathsf{sp}} + \gamma_k(1-\gamma_k)^2 \|h^\star\|_{\mathsf{sp}} + (\gamma_k)^2 \max_{N_k-4g_k\leq i<N_k} \|\Delta_{k,i}\|_\infty \\
&\lesssim \frac{1}{1-\gamma_k}(1-\gamma_k)^2 \|h^\star\|_{\mathsf{sp}} + (\gamma_k)^L \max_{N_k-2Lg_k\leq i<N_k} \|\Delta_{k,i}\|_\infty \\
&\lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}} + e^{-(1-\gamma_k)L} \\
&\lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}}
\end{aligned} \tag{C.31}$$

for $L = \left\lceil \frac{-\log(1-\gamma_k)}{(1-\gamma_k)} \right\rceil$ and $\frac{N_k}{2} \geq 2\left(\frac{\log(4/(1-\gamma_k))}{(1-\gamma_k)}\right)\frac{\log(1/(1-\gamma_k)^2)}{\eta_{k,N_k}} \geq 2Lg_k$ because the learning rates and discount factors in (3.10a) satisfy $(1-\gamma_k)^{-1} \leq N_k$ and

$$\frac{2\log(4N_k)}{(N_k)^{1/3}} \leq (1-\gamma_k) \quad \text{and} \quad \frac{4\log(N_k)}{(N_k)^{2/3}} \leq \eta_{k,N_k}. \tag{C.32}$$

Finally, by plugging (C.31) and (C.21) into (C.20), we conclude that after K stages, the error of Q-estimates is bounded as

$$\|Q_{K,N_K} - J^\star\|_\infty \lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}} \lesssim \frac{\|h^\star\|_{\mathsf{sp}}}{(N_K)^{1/3}}. \tag{C.33}$$

To achieve $\|Q_{K,N_K} - J^\star\|_\infty \le \varepsilon$, K should be large enough to satisfy

$$N_K = \max(1000, 2^K) \gtrsim \left(\frac{\|h^\star\|_{\mathsf{sp}}}{\varepsilon}\right)^3 \tag{C.34}$$

and the total number of samples is bounded as

$$SA\sum_{k=1}^{K} N_k \le SA\left(10 + \sum_{k=1}^{K} 2^k\right) = \widetilde{O}\left(SA\left(\frac{\|h^\star\|_{\mathsf{sp}}}{\varepsilon}\right)^3\right). \tag{C.35}$$

Here, note that we need $N_k \ge 1000$ to make sure a large enough number of iterations is given, $N_k \ge 2g_k$ for small k, which is required for the error bounds in Step 2 to hold for all $k \in [K]$.

C.2.2 Analysis for the second group of parameters

Step 1: error decomposition. We begin from the k-th epoch, which aims to approximate $Q^\star_{k+1}$. Combining the update rule (C.19a) and the definition of $Q^\star_{k+1}$ (cf. (C.10)), we express the estimation error at the t-th step as:

$$Q_{k,t+1} - Q^\star_{k+1} = (1-\eta_k)(Q_{k,t} - Q^\star_{k+1}) + \eta_k\left((1-\gamma_k)r + \gamma_k P_{k,t}V_{\iota(k,t)} - \frac{r}{k+1} - \frac{k}{k+1}PV^\star_k\right)$$

Note that $\iota(k,t) = (k-1, N_{k-1})$, and $\gamma_k = k/(k+1)$. For ease of notation, we denote

$$V_k \coloneqq V_{k,0} = V_{k-1,N_{k-1}}.$$

The estimation error can be further simplified as:

$$Q_{k,t+1} - Q^\star_{k+1} = (1-\eta_k)(Q_{k,t} - Q^\star_{k+1}) + \frac{k\eta_k}{k+1}(P_{k,t}V_k - PV^\star_k)$$

By recursion, the estimation error at the end of the k-th stage becomes:

$$\begin{aligned}
&Q_{k+1,0} - Q^\star_{k+1} = Q_{k,N_k} - Q^\star_{k+1} \\
&= (1-\eta_k)^{N_k}(Q_{k,0} - Q^\star_{k+1}) + \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P_{k,i}V_k - PV^\star_k) \\
&= (1-\eta_k)^{N_k}(Q_{k,0} - Q^\star_k) + (1-\eta_k)^{N_k}(Q^\star_k - Q^\star_{k+1}) + \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P_{k,i}V_k - PV^\star_k).
\end{aligned} \tag{C.36}$$

Define the error term as

$$\Delta_{k+1} \coloneqq Q_{k+1,0} - Q^\star_{k+1} \in \mathbb{R}^{SA}.$$

Identity (C.36) then tells us that

$$\Delta_{k+1} = (1-\eta_k)^{N_k}\Delta_k + (1-\eta_k)^{N_k}(Q^\star_k - Q^\star_{k+1}) + \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P_{k,i}V_k - PV^\star_k)$$

By recursion, Δ_{k+1} can be decomposed as:

$$\begin{aligned}
\Delta_{k+1} = &\underbrace{\prod_{j=1}^{k}(1-\eta_j)^{N_j}\Delta_1}_{=:E^1_k} + \underbrace{\sum_{j=1}^{k}\prod_{l=j}^{k}(1-\eta_l)^{N_l}(Q^\star_j - Q^\star_{j+1})}_{=:E^2_k} \\
&+ \sum_{j=1}^{k}\underbrace{\prod_{l=j+1}^{k}(1-\eta_l)^{N_l}\frac{j\eta_j}{j+1}\sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i}(P_{j,i}V_j - PV^\star_j)}_{=:E^3_{k,j}}.
\end{aligned} \tag{C.37}$$

Step 2: bounding the decomposed errors. Now we intend to bound the error terms in the right-hand-side of (C.37) separately as follows:

- Bounding the initialization error E_k^1 and errors from historical epochs E_k^2 and $E_{k,j}^3$ for $j < k$. Notice that as long as $N_k \geq 4 \log T_k$, which is satisfied by $c_N \geq 4 \log c_N + 3 \log 2$, then the following decay factor at the k-th stage satisfies

$$(1-\eta_k)^{N_k} \leq \exp(-N_k \eta_k) \overset{\text{(i)}}{=} \exp\left(-\frac{N_k}{1+\frac{N_k}{4\log T_k}}\right) \leq \exp\left(-\frac{N_k}{\frac{N_k}{2\log T_k}}\right) = \frac{1}{T_k^2}, \quad \text{(C.38)}$$

where (i) arises from the value of $\eta_{k,t}$ in (3.10b). Consequently, the first error term E_k^1 in the right-hand-side of (C.37) is controlled by:

$$\left\|E_k^1\right\|_\infty \leq \frac{\|\Delta_1\|_\infty}{T_k^2} = \frac{\|Q_1^\star - Q_{1,0}\|_\infty}{T_k^2} \overset{\text{(i)}}{=} \frac{\|Q_1^\star\|_\infty}{T_k^2} \overset{\text{(ii)}}{\leq} \frac{1+\|h^\star\|_{\mathsf{span}}}{T_k^2}, \quad \text{(C.39)}$$

where (i) holds since $Q_{1,0} = Q_{0,0} = 0$, and (ii) uses property (C.12). Similarly, using (C.38) yields the following bound for the error term E_k^2:

$$\begin{aligned}\|E_k^2\|_\infty &\leq \sum_{j=1}^k \frac{\|Q_j^\star - Q_{j+1}^\star\|_\infty}{T_k^2} \overset{\text{(i)}}{\leq} \frac{1}{T_k^2}\sum_{j=1}^k \left\|J^\star + \frac{Q^\star}{j} - J^\star - \frac{Q^\star}{j+1}\right\|_\infty \\ &= \frac{\|Q^\star\|_\infty}{T_k^2}\sum_{j=1}^k \frac{1}{j(j+1)} \overset{\text{(ii)}}{\leq} \frac{\|Q^\star\|_\infty}{T_k^2} \overset{\text{(iii)}}{\leq} \frac{2+\|h^\star\|_{\mathsf{span}}}{T_k^2}, \end{aligned} \quad \text{(C.40)}$$

where (i) holds since (C.12), (ii) arises from $\sum_{j=1}^k 1/(j(j+1)) = 1 - 1/(k+1)$, and (iii) uses the bound of $\|Q^\star\|_\infty$ given in (C.6).

Moreover, the error terms $E_{k,j}^3$ for $j < k$ in the right-hand-side of (C.37) satisfy:

$$\begin{aligned}\left\|\sum_{j=1}^{k-1} E_{k,j}^3\right\|_\infty &= \left\|\sum_{j=1}^{k-1}\prod_{l=j+1}^{k}(1-\eta_l)^{N_l}\frac{j\eta_j}{j+1}\sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i}(P_{j,i}V_j - PV_j^\star)\right\|_\infty \\ &\leq \sum_{j=1}^{k-1}\prod_{l=j+1}^{k}(1-\eta_l)^{N_l}\eta_j\sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i}(\|V_j\|_\infty + \|V_j^\star\|_\infty) \\ &\leq \frac{1}{T_k^2}\sum_{j=1}^{k-1}(\|V_j\|_\infty + \|V_j^\star\|_\infty), \end{aligned} \quad \text{(C.41)}$$

where the last inequality arises from (C.38) and $\eta_j \sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i} \leq \eta_j/\eta_j = 1$.

By virtue of (C.11), we have $\|V_j^\star\|_\infty \le 1 + \|h^\star\|_{\mathsf{span}}/j$ and $\|V_j\|_\infty \le 1$, which further yields

$$\left\|\sum_{j=1}^{k-1} E_{k,j}^3\right\|_\infty \le \frac{k(2+\|h^\star\|_{\mathsf{sp}})}{T_k^2} \lesssim \frac{k\|h^\star\|_{\mathsf{sp}}}{T_k^2}, \tag{C.42}$$

provided that $\|h^\star\|_\infty \gtrsim 1$.

- Bounding the error term from the last epoch $E_{k,k}^3$. Note that the error term $E_{k,k}^3$ does not contain the factor $(1-\eta_k)^{N_k}$, and thus its ℓ_∞ norm cannot be bounded in the same manner as (C.42) To control $\|E_{k,k}^3\|_\infty$, we need to establish a new recursive expression. Inserting (C.39), (C.40), and (C.42) into (C.37), we bound Δ_{k+1} as:

$$\begin{aligned}
\Delta_{k+1} &= \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P_{k,i}V_k - PV_k^\star) + O\left(\frac{k\|h^\star\|_{\mathsf{sp}}}{T_k^2}\right) \\
&= \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}P(V_k - V_k^\star) \\
&\qquad + \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P_{k,i}-P)V_k + O\left(\frac{k\|h^\star\|_{\mathsf{sp}}}{T_k^2}\right),
\end{aligned} \tag{C.43}$$

where the first term reflects the bias and the second term comes from the randomness of sampling.

Let $\pi_k(s) \coloneqq \arg\max_{a\in\mathcal{A}} Q_k(s,a)$. Recalling that $P(V_k - V_k^\star) = P^{\pi_k}Q_k - P^{\pi_\star}Q_k^\star \le P^{\pi_k}(Q_k - Q_k^\star) = P^{\pi_k}\Delta_k$, where the $\le$ is entry-wise, we apply recursion to (C.43) and obtain

$$\begin{aligned}
\Delta_{k+1} &\le \frac{k\alpha_k}{k+1}P^{\pi_k}\Delta_k + \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P_{k,i}-P)V_k + O\left(\frac{k\|h^\star\|_{\mathsf{sp}}}{T_k}\right) \\
&\le \underbrace{\frac{1}{k+1}\prod_{l=1}^{k}(\alpha_l P^{\pi_l})\Delta_1}_{=:E_{k,k,0}^3}
\end{aligned}$$

$$+\sum_{j=1}^{k}\prod_{l=j+1}^{k}(\alpha_l P^{\pi_l})\frac{j}{k+1}\underbrace{\eta_j\sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i}(P_{j,i}-P)V_j}_{=:E^3_{k,k,j}}+O\left(\frac{\|h^\star\|_{\mathsf{sp}}}{(k+1)c_N^2}\right), \tag{C.44}$$

where $\alpha_k \coloneqq \eta_k\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}\le 1$, and the last inequality uses $T_k\ge N_k\ge c_N k^2$. Noting that $\|\Delta_1\|_\infty\le 1+\|h^\star\|_{\mathsf{sp}}$ (cf. (C.39)) and the ℓ_1 norm of probability transition matrix $\|P^{\pi_l}\|_1=1$, we have

$$\|E^3_{k,k,0}\|_\infty\le\frac{1+\|h^\star\|_{\mathsf{sp}}}{k+1}. \tag{C.45}$$

Next, we shall control $E^3_{k,k,j}$. Note that

$$\begin{aligned}\|V_k-J^\star\|_{\mathsf{sp}}&\le\|V_k-V_k^\star\|_{\mathsf{sp}}+\|V_k^\star-J^\star\|_{\mathsf{sp}}\\&\overset{\text{(i)}}{=}\|V_k-V_k^\star\|_{\mathsf{sp}}+\frac{\|V^\star\|_{\mathsf{sp}}}{k}\\&\le\|\Delta_k\|_\infty+\frac{\|h^\star\|_{\mathsf{sp}}}{k},\end{aligned} \tag{C.46}$$

where (i) comes from the fact that $V_k^\star-J^\star=\frac{V^\star}{k}$ (cf. (C.9)). Moreover, notice that all entries in $J^\star$ are identical, the variance of the second term $E^3_{k,k,j}$ is bounded by

$$\begin{aligned}\mathsf{Var}\left(E^3_{k,k,j}\right)&=\mathsf{Var}\left(\eta_j\sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i}(P_{j,i}-P)(V_j-J^\star)\right)\\&\le\frac{\eta_j^2}{1-(1-\eta_j)^2}\mathsf{Var}\left((P_{j,i}-P)(V_j-J^\star)\right)\\&\le\eta_j\|V_j-J^\star\|_{\mathsf{sp}}^2\\&\overset{\text{(i)}}{\le}\eta_j\left(\|\Delta_j\|_\infty+\frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)^2,\end{aligned} \tag{C.47}$$

where (i) comes from (C.46). By applying Bernstein's inequality, with probability at least $1-\delta/(2j^2)$, the following holds; moreover, by a union bound over all $j\in[K]$,

the overall probability is at least $1-\delta$:

$$\begin{aligned}\|E^3_{k,k,j}\|_\infty &\lesssim \sqrt{\mathsf{Var}(E^3_{k,k,j})\log\frac{2SAj^2}{\delta}} + \eta_j\|V_j - J^\star\|_\infty \log\frac{2SAj^2}{\delta} \\ &\lesssim \sqrt{\eta_j}\left(\|\Delta_j\|_\infty + \frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)\sqrt{\log\frac{SAj}{\delta}} \qquad \text{(C.48)} \\ &\overset{\text{(i)}}{\lesssim} \sqrt{\frac{\log T_j}{N_j}}\left(\|\Delta_j\|_\infty + \frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)\sqrt{\log\frac{SAj}{\delta}} \\ &\overset{\text{(ii)}}{\lesssim} \frac{1}{j\log(j+1)}\left(\|\Delta_j\|_\infty + \frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)\frac{1}{\sqrt{c_N}}, \qquad \text{(C.49)}\end{aligned}$$

where (i) comes from the definition of η_j (cf. (3.10b)):

$$\eta_j = \frac{1}{1+\frac{N_j}{4\log(\sum_{i=1}^j N_i)}} = \frac{1}{1+\frac{N_j}{4\log T_j}} \le \frac{4\log T_j}{N_j}, \qquad \text{(C.50)}$$

(ii) arises from

$$\begin{aligned}N_j &= c_N j^2\log^5(j+1)\log^3\left(\frac{SA}{\delta}\right) \ge c_N j^2\log^2(j+1)\left(\log(j+1)\log\left(\frac{SA}{\delta}\right)\right)^2 \\ &\gtrsim c_N j^2\log^2(j+1)\log T_j\log\frac{SAj}{\delta}, \qquad \text{(C.51)}\end{aligned}$$

by virtue of the fact that $\log T_j \lesssim \log(j^3\log^3(j+1)\log^3(SA/\delta)) \lesssim \log(j+1) + \log(SA/\delta) \lesssim \log(j+1)\log(SA/\delta)$. Inequality (C.48) holds since

$$\eta_k \le \frac{4\log T_j}{N_j} \lesssim \frac{\log T_j}{c_N j^2\log^2(j+1)\log T_j\log\frac{SAj}{\delta}} \lesssim \frac{1}{\log\frac{SAj}{\delta}}.$$

Substituting (C.45) and (C.49) into (C.44), provided that $\|h^\star\|_{\mathsf{span}} \gtrsim 1$, we have

$$\begin{aligned}\Delta_{k+1} &\lesssim \frac{\|h^\star\|_{\mathsf{sp}}}{k+1} + \frac{1}{k+1}\sum_{j=1}^k \frac{\theta}{\log(j+1)}\left(\|\Delta_j\|_\infty + \frac{\|h^\star\|_{\mathsf{sp}}}{j}\right) + \frac{\|h^\star\|_{\mathsf{sp}}}{(k+1)c_N^2} \\ &\overset{\text{(i)}}{\lesssim} \frac{\|h^\star\|_{\mathsf{sp}}}{k+1} + \frac{1}{k+1}\sum_{j=1}^k \frac{\theta\|\Delta_j\|_\infty}{\log(j+1)} + \frac{\theta\|h^\star\|_{\mathsf{sp}}\log\log(k+1)}{k+1}\end{aligned}$$

$$\asymp \frac{\|h^\star\|_{\mathsf{sp}} \log\log(k+1)}{k+1} + \frac{1}{k+1}\sum_{j=1}^{k} \frac{\theta\|\Delta_j\|_\infty}{\log(j+1)}, \tag{C.52}$$

where $\theta := 1/\sqrt{c_N} \ll 1$, and (i) uses the fact that $\sum_{j=1}^{k} j^{-1}\log^{-1}(j+1) \le 2\log\log(k+1)$.

Similarly, we can derive the following lower bound for Δ_{k+1}:

$$\Delta_{k+1} \gtrsim -\frac{\|h^\star\|_{\mathsf{sp}} \log\log(k+1)}{k+1} - \frac{1}{k+1}\sum_{j=1}^{k} \frac{\theta\|\Delta_j\|_\infty}{\log(j+1)}.$$

Step 3: solving recursion. Combining the upper and the lower bound, we control the infinite norm of Δ_{k+1} as

$$\begin{aligned}\frac{\|\Delta_{k+1}\|_\infty}{\log(k+2)} &\le \frac{c\|h^\star\|_{\mathsf{sp}} \log\log(k+1)}{(k+1)\log(k+2)} + \frac{c\theta}{(k+1)\log(k+2)}\sum_{j=1}^{k}\frac{\|\Delta_j\|_\infty}{\log(j+1)} \\ &\le \frac{c\|h^\star\|_{\mathsf{sp}}}{k+1} + \frac{c\theta\Delta_k^{\mathsf{sum}}}{(k+1)\log(k+2)},\end{aligned} \tag{C.53}$$

where $\Delta_k^{\mathsf{sum}} := \sum_{j=1}^{k} \frac{\|\Delta_j\|_\infty}{\log(j+1)}$.

Now, we solve the recursive relation (C.53). We first claim that for sequence a_k satisfying

$$a_{k+1} = \lambda_k s_k + \beta_k,$$

where $\lambda_k, \beta_k, a_1 \ge 0$ and $s_k = \sum_{j=1}^{k} a_j$, we have

$$a_{k+1} = \beta_k + \lambda_k \sum_{i=1}^{k-1}\prod_{j=i+1}^{k-1}(1+\lambda_j)\beta_i + \lambda_k\prod_{i=1}^{k-1}(1+\lambda_i)a_1.$$

Taking $a_k := \|\Delta_k\|/\log(k+1)$, $\lambda_k := c\theta/(k+1)/\log(k+2)$, and $\beta_k := c\|h^\star\|_{\mathsf{sp}}/(k+1)$, we have

$$\frac{\|\Delta_{k+1}\|_\infty}{\log(k+2)} \le \frac{c\|h^\star\|_{\mathsf{sp}}}{k+1} + \frac{c\theta}{(k+1)\log(k+2)}\sum_{i=1}^{k-1}\prod_{j=i+1}^{k-1}\left(1+\frac{c\theta}{(j+1)\log(j+2)}\right)\frac{c\|h^\star\|_{\mathsf{sp}}}{i+1}$$

$$+\frac{c\theta}{(k+1)\log(k+2)}\prod_{i=1}^{k-1}\left(1+\frac{c\theta}{(i+1)\log(i+2)}\right)\frac{\|\Delta_1\|_\infty}{\log 2}. \tag{C.54}$$

For $\theta \le 1/2$, the product in the right-hand-side of (C.54) satisfies

$$\begin{aligned}\prod_{j=i+1}^{k-1}\left(1+\frac{\theta}{(j+1)\log(j+2)}\right) &\le \exp\left(\sum_{j=i+1}^{k-1}\frac{\theta}{(j+1)\log(j+2)}\right)\\ &\le \exp(2\theta\log\log(k+1)) \le \log k.\end{aligned}$$

Substituting it into (C.54), we have

$$\begin{aligned}\frac{\|\Delta_{k+1}\|_\infty}{\log(k+2)} &\le \frac{c\|h^\star\|_{\mathsf{sp}}}{k+1}+\frac{c^2\theta\|h^\star\|_{\mathsf{sp}}\log k}{(k+1)\log(k+2)}\sum_{i=1}^{k-1}\frac{1}{i+1}+\frac{c\theta\log k}{(k+1)\log(k+2)}\frac{\|\Delta_1\|_\infty}{\log 2}\\ &\lesssim \frac{(1+\theta\log k)\|h^\star\|_{\mathsf{sp}}}{k+1} \lesssim \frac{\|h^\star\|_{\mathsf{sp}}\log k}{k+1}.\end{aligned} \tag{C.55}$$

Recalling the definition of Δ_K, and by virtue of $\|J^\star - Q_K^\star\|_\infty = \|Q^\star\|_\infty/K \lesssim \|h^\star\|_{\mathsf{sp}}/(K+1)$ (cf. (C.12) and (C.6)), we get the final result that

$$\|Q_{K,N_K} - J^\star\|_\infty \le \|Q_{K,N_K} - Q_K^\star\|_\infty + \|J^\star - Q_K^\star\|_\infty \le \|\Delta_K\|_\infty + \frac{\|h^\star\|_{\mathsf{sp}}}{K} \lesssim \frac{\|h^\star\|_{\mathsf{sp}}\log K}{K}. \tag{C.56}$$

C.3 Analysis for the federated setting (Theorem 3)

C.3.1 Analysis for the first group of parameters

The proof will follow similar steps as in the proof of the single-agent case in Section C.2.1. We omit some of the repetitive derivations and only highlight the key differences here. Similar to the single-agent case, we split the error into two components, one due to horizon mismatch and the other due to the optimality gap in the discounted MDP, as follows:

$$\|Q_{k,t} - J^\star\|_\infty \le \|Q_{\gamma_k}^\star - J^\star\|_\infty + \|Q_{k,t} - Q_{\gamma_k}^\star\|_\infty. \tag{C.57}$$

Since the first term can be bounded by $\widetilde{O}((1-\gamma_k)\|h^\star\|_{\mathsf{sp}})$ as (C.21) for the single-agent case, we only need to focus on proving the bound of the second error term, that is, $\|Q_{k,t} - Q^\star_{\gamma_k}\|_\infty \lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}}$ for the first parameter group defined under the federated setup. Since γ_k remains fixed within stage k, the analysis follows a similar structure to the discounted MDP case in Woo et al. (2023). The key difference is that we derive error bounds in terms of both the discount factor and the span norm of the bias, ensuring that the two decomposed error terms in (C.20) converge at the same rate, thereby balancing the convergence bottleneck.

Step 1: error decomposition. Unlike the single-agent case, where the error is defined in terms of a single Q-estimate, in the federated setting, the error at iteration t of stage k is defined in terms of averaged Q-estimates of all agents as follows:

$$\Delta_{k,t} = \sum_{m=1}^{M} Q^m_{k,t} - Q^\star_{\gamma_k}.$$

Recalling the error decomposition for the single Q-estimate in (C.23) shown for the single-agent case, we decompose $\Delta_{k,t}$ as follows:

$$\begin{aligned}
\Delta_{k,t} &= \prod_{i=1}^{t}(1-\eta_{k,i})\Delta_{k,0} + \frac{\gamma_k}{M}\sum_{i=1}^{t}\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})\sum_{m=1}^{M}\left(P^m_{k,i}(V_{k,\iota(k,i)} - PV^\star_{\gamma_k}\right) \\
&= \underbrace{\prod_{i=1}^{t}(1-\eta_{k,i})\Delta_{k,0} + \frac{\gamma_k}{M}\sum_{i=1}^{t-g_k}\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})\sum_{m=1}^{M}P^m_{k,i}\left(V_{k,\iota(k,i)} - V^\star_{\gamma_k}\right)}_{=:E^1_{k,t}} \\
&\quad + \underbrace{\frac{\gamma_k}{M}\sum_{i=1}^{t}\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})\sum_{m=1}^{M}\left(P^m_{k,i} - P\right)V^\star_{\gamma_k}}_{=:E^2_{k,t}} \\
&\quad + \underbrace{\frac{\gamma_k}{M}\sum_{i=t-g_k+1}^{t}\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})\sum_{m=1}^{M}P^m_{k,i}\left(V_{k,\iota(k,i)} - V^\star_{\gamma_k}\right)}_{=:E^3_{k,t}}.
\end{aligned} \tag{C.58}$$

Step 2: bounding the decomposed errors. With the proposed learning rates (3.14a), for any $k \in [K]$ and sufficiently large $2g_k \leq t \leq N_k$, we can derive the bound of the error terms as follows:

$$\|E_{k,t}^1\|_\infty \leq 2(1-\gamma_k)^2, \tag{C.59}$$

$$\|E_{k,t}^2\|_\infty \leq 24(1-\gamma_k)^2\|h^\star\|_{\mathsf{sp}} \log\frac{2|\mathcal{S}||\mathcal{A}|T_K}{\delta}, \tag{C.60}$$

$$\|E_{k,t}^3\|_\infty \leq \gamma_k \max_{\iota(k,t-g_k+1)\leq i<t} \|\Delta_{k,i}\|_\infty. \tag{C.61}$$

Note that given $N_k \geq 1000$ for any $k \in [K]$, $N_k \geq 2g_k$ since $(\eta_k)^{-1} \leq 2(N_k)^{2/3}M^{-1/3}$ and $(MN_k)^{-1/3} \leq (1-\gamma_k) \leq 1$. We omit the detailed derivations for the bounds of $E_{k,t}^1$ and $E_{k,t}^3$ since they follow similarly as in the single-agent case. See Section C.2.1 step 2 for reference. The bound of the transition variance term $E_{k,t}^2$ is derived by similarly applying the Hoeffding bound, but the intermediate values are different from the single-agent case and are expressed in terms of the number of agents M, which appears only in the federated setting. We provide the detailed proof below.

Proof of (C.61)**.** Rewrite the second error term as the sum of random vector $E_{k,t}^2 = \sum_{i=1}^t z_{k,i}$, where $z_{k,i} = \eta_{k,i}\prod_{j=i+1}^t(1-\eta_{k,j})\left(P_{k,i}-P\right)V_{\gamma_k}^\star$ is an independent random vector with zero mean. Due to the properties of the learning rates (C.14), we can derive the following bounds for $z_{k,i}$,

$$\frac{1}{M^2}\sum_{m=1}^M\sum_{i=1}^t\left(\eta_{k,i}\prod_{j=i+1}^t(1-\eta_{k,j})\right)^2\|V_{\gamma_k}^\star\|_{\mathsf{sp}}^2 \leq \frac{1}{M}\eta_{k,t}\|V_{\gamma_k}^\star\|_{\mathsf{sp}}^2 \leq 16(1-\gamma_k)^2(6(1-\gamma_k)\|h^\star\|_{\mathsf{sp}})^2, \tag{C.62}$$

where $\eta_{k,t} \leq 16M(1-\gamma_k)^2$ for $t \geq \frac{N_k}{2}$ and $\|V_{\gamma_k}^\star\|_{\mathsf{sp}} = |\max_s V_{\gamma_k}^\star(s) - \min_s V_{\gamma_k}^\star(s)| \leq 6(1-\gamma_k)\|h^\star\|_{\mathsf{sp}}$ according to Lemma 17 and the following properties of the learning rates

$$\eta_{k,t} = \frac{1}{1+\frac{(Mt)^{2/3}}{8M\log(4Mt)}} \leq \frac{8M\log(4Mt)}{(Mt)^{2/3}} \leq 16M\frac{\log(2MN_k)}{(MN_k)^{2/3}} \leq 16M(1-\gamma_k)^2. \tag{C.63}$$

Since each $z_{k,i}$ is an independent random vector with zero mean, we apply the Hoeffding inequality as follows:

$$\begin{aligned}\left|E_{k,t}^2\right| &\leq \sqrt{\frac{1}{M^2}\sum_{m=1}^{M}\sum_{i=1}^{t}\left(\eta_{k,i}\prod_{j=i+1}^{t}(1-\eta_{k,j})\right)^2\|V_{\gamma_k}^\star\|_{\mathsf{sp}}^2\log\frac{2|\mathcal{S}||\mathcal{A}|T_K}{\delta}} \\ &\leq \sqrt{576(1-\gamma_k)^4\|h^\star\|_{\mathsf{sp}}^2\log\frac{2|\mathcal{S}||\mathcal{A}|T_K}{\delta}} \\ &\leq 24(1-\gamma_k)^2\|h^\star\|_{\mathsf{sp}}\log\frac{2|\mathcal{S}||\mathcal{A}|T_K}{\delta} \end{aligned} \tag{C.64}$$

where $T_K = \sum_{k=1}^{K} N_k$, and this completes the proof.

Step 3: solving recursion. Combining the error bounds obtained from the previous steps, we solve the recursive relation as follows:

$$\begin{aligned}\|\Delta_{k,N_k}\|_\infty &\lesssim \frac{1}{1-\gamma_k}(1-\gamma_k)^2\|h^\star\|_{\mathsf{sp}} + (\gamma_k)^L \max_{N_k-2Lg_k\leq i<N_k}\|\Delta_{k,i}\|_\infty \\ &\lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}} + e^{-(1-\gamma_k)L} \\ &\lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}} \end{aligned} \tag{C.65}$$

for $L = \lceil \frac{-\log(1-\gamma_k)}{(1-\gamma_k)} \rceil$ and $\frac{N_k}{2} \geq 2(\frac{\log(4/(1-\gamma_k))}{(1-\gamma_k)})\frac{\log(1/(1-\gamma_k)^2)}{\eta_{k,N_k}} \geq 2Lg_k$ because the learning rates and discount factors in (3.14a) satisfy $(1-\gamma_k)^{-1} \leq MN_k$ and

$$\frac{2\log(4MN_k)}{(MN_k)^{1/3}} \leq (1-\gamma_k) \quad \text{and} \quad \frac{4M\log(MN_k)}{(MN_k)^{2/3}} \leq \eta_{k,N_k}. \tag{C.66}$$

Finally, by plugging (C.65) into (C.57), we conclude that after K stages, the error of averaged Q-estimate is bounded as

$$\|Q_{K,N_K} - J^\star\|_\infty \lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}} \lesssim \frac{\|h^\star\|_{\mathsf{sp}}}{(MN_K)^{1/3}}. \tag{C.67}$$

To achieve $\|Q_{K,N_K} - J^\star\|_\infty \leq \varepsilon$, K should be large enough to satisfy

$$N_K = \max(1000, 2^K) \gtrsim \frac{1}{M}\left(\frac{\|h^\star\|_{\mathsf{sp}}}{\varepsilon}\right)^3 \tag{C.68}$$

and the total number of samples is bounded as

$$SA\sum_{k=1}^{K} N_k = \widetilde{O}\left(\frac{SA}{M}\left(\frac{\|h^\star\|_{\mathsf{sp}}}{\varepsilon}\right)^3\right). \tag{C.69}$$

Analysis of the number of communication rounds. When communication intervals at stage k are set as g_k, i.e., $\min_{t_1\neq t_2\in\mathcal{C}(k)}|t_1-t_2| = g_k$, the number of communication rounds is bounded as

$$\mathcal{C}(k) \lesssim \frac{N_k}{g_k} \leq N_k\eta_{k,N_k} \leq (MN_k)^{\frac{1}{3}} \tag{C.70}$$

Thus, the total number of communication rounds for K stages is bounded by

$$\sum_{k=1}^{K}|\mathcal{C}(k)| \lesssim M^{1/3}\sum_{k=1}^{K}(N_k)^{1/3} \lesssim M^{1/3}K^{2/3}(\sum_{k=1}^{K}N_k)^{1/3} \lesssim (MT_K)^{1/3}, \tag{C.71}$$

since $2^K \leq T_K = \sum_{k=1}^{K} N_k \leq 10^4 + 2^{K+1}$ and $K = O(\log(T_k))$.

C.3.2 Analysis for the second group of parameters

The proof is similar to the argument in Section C.2.2. We omit some of the repetitive derivations and only highlight the key differences here.

Step 1: error decomposition. Similar to the derivation of (C.36), we split the error $\Delta_{k+1} \coloneqq Q^m_{k+1,0} - Q^\star_{k+1}$ as follows:

$$\begin{aligned} Q_{k+1,0} - Q^\star_{k+1} &= (1-\eta_k)^{N_k}(Q_{k,0} - Q^\star_k) + (1-\eta_k)^{N_k}(Q^\star_k - Q^\star_{k+1}) \\ &\quad + \frac{1}{M}\frac{k\eta_k}{k+1}\sum_{m=1}^{M}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P^m_{k,i}V_k - PV^\star_k). \end{aligned} \tag{C.72}$$

Using a similar recursive approach as in (C.37), we obtain the following counterpart:

$$\Delta_{k+1} = \underbrace{\prod_{j=1}^{k}(1-\eta_j)^{N_j}\Delta_1}_{=:E_k^1} + \underbrace{\sum_{j=1}^{k}\prod_{l=j}^{k}(1-\eta_l)^{N_l}(Q_j^\star - Q_{j+1}^\star)}_{=:E_k^2}$$
$$+ \sum_{j=1}^{k} \underbrace{\prod_{l=j+1}^{k}(1-\eta_l)^{N_l}\frac{j\eta_j}{(j+1)M}\sum_{m=1}^{M}\sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i}(P_{j,i}^m V_j - PV_j^\star)}_{=:E_{k,j}^3}. \tag{C.73}$$

We claim that for sufficiently large c_N, the following inequalities hold, whose proof is postponed to the end of this section:

$$N_k \geq \log(MT_k), \qquad k \geq 1, \tag{C.74}$$
$$MN_k \gtrsim c_N k^2 \log^2(k+1)\log\frac{SA}{\delta}\log(MT_k), \qquad k \geq 1. \tag{C.75}$$

Based on this claim, we have $(1-\eta_k)^{N_k} \leq \exp\left(-\frac{N_k}{\frac{N_k}{2\log(MT_k)}}\right) = \frac{1}{M^2T_k^2}$ as the analysis of (C.38). Thus we have $\|E_k^1\|_\infty \lesssim \frac{\|h^\star\|_{\mathsf{sp}}}{M^2T_k^2}$, $\|E_k^2\|_\infty \lesssim \frac{k\|h^\star\|_{\mathsf{sp}}}{M^2T_k^2}$, and $\|E_{k,j}^3\|_\infty \lesssim \frac{k\|h^\star\|_{\mathsf{sp}}}{M^2T_k^2}$ for $j<k$. Then we obtain

$$\Delta_{k+1} = \frac{k\eta_k}{k+1}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}P(V_k - V_k^\star)$$
$$+ \frac{k\eta_k}{M(k+1)}\sum_{m=1}^{M}\sum_{i=0}^{N_k-1}(1-\eta_k)^{N_k-1-i}(P_{k,i}^m - P)V_k + O\left(\frac{k\|h^\star\|_{\mathsf{sp}}}{M^2T_k^2}\right). \tag{C.76}$$

Step 2: bounding the decomposed errors. Repeating the argument in (C.44), we obtain

$$\Delta_{k+1}$$
$$\leq \underbrace{\frac{1}{k+1}\prod_{l=1}^{k}(\alpha_l P^{\pi_l})\Delta_1}_{E_{k,k,0}^3}$$

$$+\sum_{j=1}^{k}\prod_{l=j+1}^{k}(\alpha_l P^{\pi_l})\frac{j}{k+1}\underbrace{\frac{\eta_j}{M}\sum_{m=1}^{M}\sum_{i=0}^{N_j-1}(1-\eta_j)^{N_j-1-i}(P_{j,i}^m-P)V_j}_{E_{k,k,j}^3}+O\left(\frac{\|h^\star\|_{\mathsf{sp}}}{(k+1)c_N^2}\right) \tag{C.77}$$

Since $\|E_{k,k,0}^3\|_\infty$ can be bounded by $\widetilde{O}(\frac{\|h^\star\|_{\mathsf{sp}}}{k+1})$, as shown in (C.45) for the single-agent case, it suffices to focus on establishing the bound for the second error term, $E_{k,k,j}^3$. To this end, we compute the variance of $E_{k,k,j}^3$ as described in (C.47) for the single-agent scenario:

$$\mathsf{Var}\left(E_{k,k,j}^3\right)\le\frac{\eta_j}{M}\left(\|\Delta_j\|_\infty+\frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)^2.$$

By applying Bernstein's inequality, with probability at least $1-\delta/(2j^2)$, we have

$$\begin{aligned}
E_{k,k,j}^3 &\lesssim \sqrt{\frac{\eta_j}{M}}\left(\|\Delta_j\|_\infty+\frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)\sqrt{\log\frac{SAj}{\delta}}+\frac{\eta_j}{M}\|V_j-J^\star\|_\infty\log\frac{2SAj^2}{\delta}\\
&\overset{\text{(i)}}{\lesssim}\sqrt{\frac{\log T_j}{MN_j}}\left(\|\Delta_j\|_\infty+\frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)\sqrt{\log\frac{SAj}{\delta}}\\
&\overset{\text{(ii)}}{\lesssim}\frac{1}{j\log(j+1)}\left(\|\Delta_j\|_\infty+\frac{\|h^\star\|_{\mathsf{sp}}}{j}\right)\frac{1}{\sqrt{c_N}},
\end{aligned} \tag{C.78}$$

where (i) arises from the definition of η_j (cf. (3.14b)):

$$\eta_j=\frac{1}{1+\frac{N_j}{4\log(MT_j)}}=\frac{1}{1+\frac{N_j}{4\log(MT_j)}}\le\frac{4\log(MT_j)}{N_j},$$

and (ii) arises from (C.75). By substituting (C.78) into (C.76), we derive the recursive formula presented in (C.52) and (C.53).

Step 3: solving recursion. Repeating argument in Step 3 in Section C.2.2, we have

$$\|Q_{K,N_K}^m-J^\star\|_\infty\le\|Q_{K,N_K}^m-Q_K^\star\|_\infty+\|J^\star-Q_K^\star\|_\infty\le\|\Delta_K\|_\infty+\frac{\|h^\star\|_{\mathsf{sp}}}{K}\lesssim\frac{\|h^\star\|_{\mathsf{sp}}\log K}{K}, \tag{C.79}$$

which is consistent with the result in single-agent setting. The sample complexity and the number of communications are obtained immediately.

Proof of (C.74) **and** (C.75)**.** Let

$$i_{\min} = \arg\min \left\{ i : M \log\left(M \log \frac{SA}{\delta}\right) \log(i+1) \geq \frac{i^2}{M} \log^5(i+1) \log^3\left(\frac{SA}{\delta}\right) \right\}.$$

We then have $N_k = c_N \log\left(M \log \frac{SA}{\delta}\right) \log(k+1)$ for $j < i_{\min}$ and $N_k = c_N \frac{k^2}{M} \log^5(k+1) \log^3 \frac{SA}{\delta}$ for $j \geq i_{\min}$. Next, we proceed by proving (C.74) for the cases $k < i_{\min}$ and $k \geq i_{\min}$ separately. For the case $k < i_{\min}$, assuming c_N is sufficiently large, we have

$$\begin{aligned} N_k &= c_N \log\left(M \log \frac{SA}{\delta}\right) \log(k+1) \geq \log c_N + 2\log M + \log\log\log \frac{SA}{\delta} + 2\log(k+1) \\ &\geq \log\left(c_N k M \log\left(M \log \frac{SA}{\delta}\right) \log(k+1)\right) = \log(MT_k). \end{aligned} \tag{C.80}$$

In the case where $k \geq i_{\min}$, we have the following inequality:

$$\begin{aligned} N_k &= \frac{c_N}{M} k^2 \log^5(k+1) \log^3 \frac{SA}{\delta} \geq c_N \log\left(M \log \frac{SA}{\delta}\right) \log(k+1) \\ &\geq \log c_N + 8\log(k+1) + 3\log\log \frac{SA}{\delta} \\ &\geq \log\left(c_N k^3 \log^5(k+1) \log^3 \frac{SA}{\delta}\right) = \log(MkN_k) \geq \log(MT_k). \end{aligned} \tag{C.81}$$

Similarly, (C.75) is derived from the following expression:

$$MN_k \geq c_N k^2 \log^5(k+1) \log^3 \frac{SA}{\delta} \geq c_N k^2 \log^2(k+1) \log \frac{SA}{\delta} \left(\log^3(k+1) \log^2 \frac{SA}{\delta}\right), \tag{C.82}$$

and

$$\begin{aligned} \log(MT_k) \leq \log(kMN_k) &= \log\left(c_N k M \log\left(M \log \frac{SA}{\delta}\right) \log(k+1)\right) \\ &\lesssim \log(c_N M) + \log k + \log\log\log \frac{SA}{\delta} \leq M \log^3(k+1) \log^2 \frac{SA}{\delta}. \end{aligned}$$

C.4 Analysis for optimal policy learning (Theorem 4)

We first decompose the estimation error of the optimal policy as

$$0 \leq J^\star - J^{\widehat{\pi}} \leq J^\star - V^\star_{\gamma_K} + V^\star_{\gamma_K} - V^{\widehat{\pi}}_{\gamma_K} + V^{\widehat{\pi}}_{\gamma_K} - J^{\widehat{\pi}}. \tag{C.83}$$

According to Lemma 6 - Lemma 8 in Wang et al. (2022), we have

$$\|J^\star - V^\star_{\gamma_K}\|_\infty \leq 4(1-\gamma_K)\|h^\star\|_{\mathsf{sp}}, \tag{C.84}$$

$$\|J^{\widehat{\pi}} - V^{\widehat{\pi}}_{\gamma_K}\|_\infty \leq \|V^{\widehat{\pi}}_{\gamma_K}\|_{\mathsf{sp}} \leq \|V^{\widehat{\pi}}_{\gamma_K} - V^\star_{\gamma_K}\|_\infty + \|V^\star_{\gamma_K}\|_{\mathsf{sp}} \leq 4(1-\gamma_K)\|h^\star\|_{\mathsf{sp}} + \|V^\star_{\gamma_K} - V^{\widehat{\pi}}_{\gamma_K}\|_\infty. \tag{C.85}$$

Moreover, by definition, we control the estimation error $\|V^\star_{\gamma_K} - V^{\widehat{\pi}}_{\gamma_K}\|_\infty$ by using the estimation error for $Q^\star_{\gamma_K}$ as follows:

$$\begin{aligned}
\|V^\star_{\gamma_K} - V^{\widehat{\pi}}_{\gamma_K}\|_\infty &= \max_s \frac{Q^\star_{\gamma_K}(s, \pi^\star_{\gamma_K}(s)) - Q^\star_{\gamma_K}(s, \widehat{\pi}(s))}{1-\gamma_K} \\
&= \frac{1}{1-\gamma_K} \max_s \big(Q^\star_{\gamma_K}(s, \pi^\star_{\gamma_K}(s)) - Q_{K,N_K}(s, \pi^\star_{\gamma_K}(s)) \\
&\qquad + Q_{K,N_K}(s, \pi^\star_{\gamma_K}(s)) - Q_{K,N_K}(s, \widehat{\pi}(s)) + Q_{K,N_K}(s, \widehat{\pi}(s)) - Q^\star_{\gamma_K}(s, \widehat{\pi}(s))\big) \\
&\overset{\text{(i)}}{\leq} \frac{2\|Q^\star_{\gamma_K} - Q_{K,N_K}\|_\infty}{1-\gamma_K},
\end{aligned} \tag{C.86}$$

where (i) holds since $Q_{K,N_K}(s, \pi^\star_{\gamma_K}(s)) - Q_{K,N_K}(s, \widehat{\pi}(s)) \leq 0$ by the definition of $\widehat{\pi}$. Substituting (C.84), (C.85), and (C.86) into (C.83), we have

$$\begin{aligned}
\|J^\star - J^{\widehat{\pi}}\|_\infty &\leq \|J^\star - V^\star_{\gamma_K}\|_\infty + \|V^\star_{\gamma_K} - V^{\widehat{\pi}}_{\gamma_K}\|_\infty + \|V^{\widehat{\pi}}_{\gamma_K} - J^{\widehat{\pi}}\|_\infty \\
&\leq 8(1-\gamma_K)\|h^\star\|_{\mathsf{sp}} + 2\|V^\star_{\gamma_K} - V^{\widehat{\pi}}_{\gamma_K}\|_\infty \leq 8(1-\gamma_K)\|h^\star\|_{\mathsf{sp}} + \frac{4\|Q^\star_{\gamma_K} - Q_{K,N_K}\|_\infty}{1-\gamma_K}.
\end{aligned}$$

Following similar analysis as that in federated setting, we claim that

$$\|Q^\star_{\gamma_K} - Q_{K,N_K}\|_\infty \lesssim \sqrt{\frac{\log \frac{SAMT_K}{\delta}}{(1-\gamma_K)MN_K}}\|h^\star\|_{\mathsf{sp}} \log(N_K M), \tag{C.87}$$

whose proof is postponed to the end of this section. Recalling the choices of γ_K, we have

$$\begin{aligned} J^\star - J^{\widehat{\pi}} &\lesssim \frac{\|h^\star\|_{\mathsf{sp}}}{(MN_K)^{1/5}} + \sqrt{\frac{\log \frac{SAMT_K}{\delta}}{(1-\gamma_K)^3 MN_K}} \|h^\star\|_{\mathsf{sp}} \log(N_K M) \\ &\lesssim \frac{\|h^\star\|_{\mathsf{sp}}}{(MN_K)^{1/5}} \log^{\frac{1}{2}} \frac{SAMT_K}{\delta} \log(N_K M) \end{aligned}$$

Provided that

$$N_K \gtrsim \frac{\|h^\star\|_{\mathsf{sp}}^5}{\varepsilon^5 M} \log^5(N_K M) \log^{\frac{5}{2}} \frac{SAMT_K}{\delta},$$

we have $J^\star - J^{\widehat{\pi}} \leq \varepsilon$. Recalling that $N_k = c_N 2^k$, we have $T_K = \sum_{k=1}^K N_k \leq 2N_K$ and complete the proof.

For the number of communication rounds, we have

$$|\mathcal{C}_k| = \frac{4\log(1-\gamma_k)}{\log((1+\gamma_k)/2)} + 1 \leq \frac{8\log(\frac{1}{1-\gamma_k})}{1-\gamma_k} = \frac{8}{5}(N_k M)^{1/5} \log(N_k M) \leq \frac{8}{5}(N_k M)^{1/5} \log(T_K M).$$

Thus we have

$$\sum_{k=1}^K |\mathcal{C}_k| \leq \frac{8\log(T_K M)}{5} \sum_{k=1}^K (N_k M)^{1/5} \lesssim (N_K M)^{1/5} \log(T_K M).$$

Proof of (C.87)**.** According to the update rule of $Q_{k,t}$, we have

$$\begin{aligned} Q_{k,t+1}^m - Q_{\gamma_k}^\star &= \prod_{j=1}^t (1-\eta_{k,j})(Q_{k,0} - Q_{\gamma_k}^\star) + \gamma_k \sum_{i=1}^t \eta_{k,i} \prod_{j=i+1}^t (1-\eta_{k,j})(P_{k,i}^m V_{k,\iota(k,i)} - PV_{\gamma_k}^\star) \\ &= \prod_{j=1}^t (1-\eta_{k,j})(Q_{k,0} - Q_{\gamma_k}^\star) + \gamma_k \sum_{i=1}^{\iota(k,t)} \eta_{k,i} \prod_{j=i+1}^t (1-\eta_{k,j})(P_{k,i}^m V_{k,\iota(k,i)} - PV_{\gamma_k}^\star) \\ &\quad + \gamma_k \sum_{i=\iota(k,t)+1}^t \eta_{k,i} \prod_{j=i+1}^t (1-\eta_{k,j})(P_{k,i}^m V_{k,\iota(k,i)} - PV_{\gamma_k}^\star). \end{aligned}$$

Denote by t_h the h-th smallest value in the set $\mathcal{C}_k$ for $h \geq 1$, and let $t_0 = 0$. Note that for $h \geq 2$, we have

$$\prod_{j=t_{h-1}+1}^{t_h} (1 - \eta_{k,j}) \leq \exp\left(- \sum_{j=t_{h-1}+1}^{t_h} \eta_{k,j} \right) \leq \exp\left(-(t_h - t_{h-1})\eta_{k,t_h}\right). \tag{C.88}$$

Considering that for $h \geq 2$, we have $t_h - t_{h-1} = (1-\gamma_k)t_h/2$, and

$$\eta_{k,t_h} = \left(1 + \frac{t_h(1-\gamma_k)}{8\log(MN_k)}\right)^{-1}, \quad \forall h \geq 1.$$

If $t_h \geq 8\log(MN_k)/(1-\gamma_k)$, then we have $\eta_{k,t_h} \geq \frac{4\log(MN_k)}{t_h(1-\gamma_k)}$ and

$$\prod_{j=t_{h-1}+1}^{t_h} (1 - \eta_{k,j}) \leq \exp\left(-\frac{(1-\gamma_k)t_h}{2} \frac{4\log(MN_k)}{t_h(1-\gamma_k)} \right) \leq \frac{1}{M^2 N_k^2}. \tag{C.89}$$

We are now ready to introduce an error sequence

$$\Delta_{k,h} \coloneqq \|Q_{k,t_h} - Q^\star_{\gamma_k}\|_\infty,$$

where Q_{k,t_h} denotes the result after the h-th communication round, shared identically among all agents. From (C.89), and noting that $Q_{k,t}, Q^\star_{\gamma_k} \leq 1$, we obtain

$$\Delta_{k,h} \leq \frac{2}{M^2 N_k} + \|\mathcal{E}^{(1)}_{k,h}\|_\infty + \|\mathcal{E}^{(2)}_{k,h}\|_\infty, \quad h \geq 1, \tag{C.90}$$

where

$$\mathcal{E}^{(1)}_{k,h} \coloneqq \frac{\gamma_k}{M} \sum_{m=1}^{M} \sum_{i=t_{h-1}+1}^{t_h} \eta_{k,i} \prod_{j=i+1}^{t_h} (1-\eta_{k,j})(P^m_{k,i} - P)V^\star_{\gamma_k}), \tag{C.91}$$

$$\mathcal{E}^{(2)}_{k,h} \coloneqq \frac{\gamma_k}{M} \sum_{m=1}^{M} \sum_{i=t_{h-1}+1}^{t_h} \eta_{k,i} \prod_{j=i+1}^{t_h} (1-\eta_{k,j}) P^m_{k,i}(V_{k,t_{h-1}} - V^\star_{\gamma_k}. \tag{C.92}$$

For $h \geq 2$, direct calculation yields

$$\frac{\gamma_k}{M} \sum_{m=1}^{M} \sum_{i=t_{h-1}+1}^{t_h} \eta_{k,i} \prod_{j=i+1}^{t_h} (1-\eta_{k,j}) \leq \gamma_k,$$

$$\frac{\gamma_k^2}{M^2} \sum_{m=1}^{M} \sum_{i=t_{h-1}+1}^{t_h} \eta_{k,i}^2 \leq \frac{\gamma_k^2}{M} \sum_{i=t_{h-1}+1}^{t_h} \eta_{k,i}^2 \leq \frac{\gamma_k^2 (t_h - t_{h-1})}{M} \eta_{k,t_{h-1}}^2$$

$$\leq \frac{\gamma_k^2 (1-\gamma_k) t_h}{2M} \frac{64 \log^2(MN_k)}{t_{h-1}^2 (1-\gamma_k)^2} \lesssim \frac{\gamma_k^2 \log^2(MN_k)}{(1-\gamma_k) M t_h},$$

where we have used $t_h / t_{h-1} \leq 2$. Recall that $\|V_{\gamma_k}^\star\|_{\mathsf{sp}} \lesssim (1-\gamma_k)\|h^\star\|_{\mathsf{sp}}$. By Hoeffding's inequality, with probability at least $1-\delta$, we have

$$\left\|\mathcal{E}_{k,h}^{(1)}\right\|_\infty \lesssim \gamma_k \sqrt{\frac{\log \frac{SAMT_K}{\delta}}{(1-\gamma_k) M t_h}} (1-\gamma_k) \|h^\star\|_{\mathsf{sp}} \log(MN_k)$$

$$\asymp \gamma_k \sqrt{\frac{(1-\gamma_k) \log \frac{SAMT_K}{\delta}}{M t_h}} \|h^\star\|_{\mathsf{sp}} \log(MN_k)$$

Moreover, we have $\left\|\mathcal{E}_{k,h}^{(2)}\right\|_\infty \leq \gamma_k \Delta_{k,h-1}$. Substituting these into (C.90), for $h \geq 2$, we obtain

$$\begin{aligned}
\Delta_{k,h} &\leq \gamma_k \Delta_{k,h-1} + C\gamma_k \|h^\star\|_{\mathsf{sp}} \sqrt{\frac{(1-\gamma_k) \log \frac{SAMT_K}{\delta}}{M t_h}} \log(MN_k) \\
&\leq \gamma_k^{h-1} \Delta_{k,1} + C\gamma_k \|h^\star\|_{\mathsf{sp}} \log(MN_k) \sqrt{\frac{(1-\gamma_k) \log \frac{SAMT_K}{\delta}}{M}} \sum_{i=1}^{h} \gamma_k^{h-i} \sqrt{\frac{1}{t_i}} \\
&\leq \gamma_k^{h-1} + C\gamma_k \|h^\star\|_{\mathsf{sp}} \log(MN_k) \sqrt{\frac{(1-\gamma_k) \log \frac{SAMT_K}{\delta}}{M}} \sum_{i=1}^{h} \left(\frac{2\gamma_k}{1+\gamma_k}\right)^{h-i} \sqrt{\frac{1}{t_h}} \\
&\leq \gamma_k^{h-1} + C\gamma_k \|h^\star\|_{\mathsf{sp}} \log(MN_k) \sqrt{\frac{(1-\gamma_k) \log \frac{SAMT_K}{\delta}}{M}} \sum_{i=1}^{h} \left(\frac{1+\gamma_k}{2}\right)^{h-i} \sqrt{\frac{1}{t_h}}. \quad \text{(C.93)}
\end{aligned}$$

Let

$$H = \left\lceil \frac{\log((1-\gamma_k)^2)}{\log(\frac{1+\gamma_k}{2})} \right\rceil.$$

Then we have

$$\Delta_{k,H} \leq \gamma_k^{H-1} + 2C\gamma_k \|h^\star\|_{\mathsf{sp}} \log(MN_k) \sqrt{\frac{\log \frac{SAMT_K}{\delta}}{(1-\gamma_k)MN_k}} \lesssim \|h^\star\|_{\mathsf{sp}} \log(MN_k) \sqrt{\frac{\log \frac{SAMT_K}{\delta}}{(1-\gamma_k)MN_k}},$$

which completes the proof.

Appendix D

Analysis of Federated Q-Learning for Offline RL

In this section, we will outline useful properties of FedLCB-Q and the key steps of the proof of Theorem 7, deferring the details, such as proofs of supporting lemmas, to Appendix D.3 and D.4.

Throughout the paper, we adopt the following shorthand notation

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

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

$$P^m_{k,h}(s) = \mathbb{I}(s = s^m_{k,h+1}), \quad \text{for all } s \in \mathcal{S}. \tag{D.2}$$

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 $(m,k,h) \in [M]\times[K]\times[H]$,

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

- $L_{k,h}^m(s,a)$: 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 $\mathcal{C}(K)$. For any positive integer k and u,

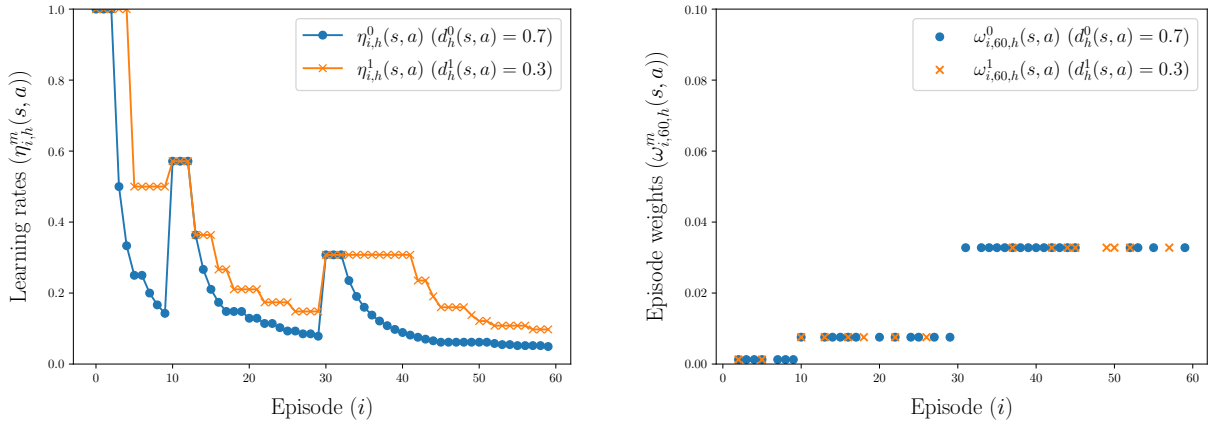
- t_u: the index of episodes, after which the uth synchronization occurs.
- τ_u: the number of local updates (episodes) taken between the $(u-1)$th and the uth synchronizations.
- $\iota(k)$: the most recent episode where the aggregation occurs before the kth episode.
- $\phi(k)$: the minimum index of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

D.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.4.1.

Lemma 19 (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 $k \in \mathcal{C}(K)$, the error between Q_h^π and $Q_{k,h}$ is decomposed as follows:*

$$\begin{aligned} Q_h^\pi(s,a) - Q_{k,h}(s,a) = {} & \underbrace{\omega_{0,k,h}(s,a)(Q_h^\pi(s,a) - Q_{0,h}(s,a))}_{=:D_1^\pi(s,a,k,h):\ \textit{initialization error}} \\ & + \underbrace{\sum_{m=1}^M \sum_{i\in L_{k,h}^m(s,a)} \omega_{i,k,h}^m(s,a)(P_{h,s,a} - P_{i,h}^m)V_{i-1,h+1}^m}_{=:D_2(s,a,k,h):\ \textit{transition variance}} \\ & + \underbrace{\sum_{u=1}^{\phi(k)} B_{t_u,h}(s,a) \prod_{u'=u+1}^{\phi(k)} \lambda_{u',h}(s,a)}_{=:D_3(s,a,k,h):\ \textit{global penalty}} \\ & + \underbrace{\sum_{m=1}^M \sum_{i\in L_{k,h}^m(s,a)} \omega_{i,k,h}^m(s,a) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m)}_{=:D_4^\pi(s,a,k,h):\ \textit{recursion}}, \end{aligned} \tag{D.3}$$



(a) Rescaled learning rates

(b) Episode weights

Figure D.1: 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 $\mathcal{C}(60)=\{10,30,60\}$.

where $L^m_{k,h}(s,a) \coloneqq \{1\le i\le k : (s^m_{i,h},a^m_{i,h})=(s,a)\}$ *and* $l^m_{k,h}(s,a)\coloneqq\{\iota(k)<i\le k : (s^m_{i,h},a^m_{i,h})=(s,a)\}$. *And, for simplicity, we use the shortened notations defined as*

$$\lambda_{v,h}(s,a)=\begin{cases}1 & \text{if } N_{k,h}(s,a)=0\\ \frac{N_{\iota(k),h}(s,a)}{N_{k,h}(s,a)+Hn_{k,h}(s,a)} & \text{otherwise}\end{cases},\quad v=\phi(k), \tag{D.4a}$$

$$\omega^m_{0,k,h}(s,a)=\begin{cases}1 & \text{if } N_{k,h}(s,a)=0\\ 0 & \text{otherwise}\end{cases}, \tag{D.4b}$$

$$\omega^m_{i,k,h}(s,a)=\frac{H+1}{N_{k,h}(s,a)+Hn_{k,h}(s,a)}\left(\prod_{x=\phi(i)}^{\phi(k)-1}\frac{N_{t_x,h}(s,a)}{N_{t_x,h}(s,a)+Hn_{t_x,h}(s,a)}\right),\quad i\in L^m_{k,h}(s,a). \tag{D.4c}$$

Equally favoring episodes within the same local update round. According to the decomposition (D.3) in Lemma 19, 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

$\omega^m_{i,k,h}(s,a)$ 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 D.1a, ensures that the weight $\omega^m_{i,k,h}(s,a)$ within the same local update round is always equal for all episodes and agents, as shown in (D.4c) and Figure D.1b. 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 interval 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 interval.

Bounded visitation counters. We next introduce the following lemma regarding the visitation counters, whose proof is provided in Appendix D.4.2.

Lemma 20 (Concentration bound on the visitation counters)**.** *Consider any $\delta \in (0,1)$ and some universal constant $c_1 > 0$, and let*

$$\zeta_0 \coloneqq \log\left(\frac{2|\mathcal{S}||\mathcal{A}|KH}{\delta}\right) \quad \textit{and} \quad K_0(s,a,h) \coloneqq \frac{4\zeta_0}{c_1 M d_h^{\mathsf{avg}}(s,a)}. \tag{D.5}$$

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

$$\textit{when } k \geq K_0(s,a,h): \qquad \frac{1}{2}kMd_h^{\mathsf{avg}}(s,a) \leq N_{k,h}(s,a) \leq 2kMd_h^{\mathsf{avg}}(s,a), \tag{D.6a}$$

$$\textit{when } k \leq K_0(s,a,h): \qquad N_{k,h}(s,a) \leq 8\zeta_0/c_1 \tag{D.6b}$$

with probability at least $1-\delta$.

Monotonic and pessimistic global value updates. Note that the global value estimate is always monotonically non-decreasing, i.e., for $k', k \in \mathcal{C}(K)$ it holds

$$\forall s \in \mathcal{S}: \qquad V_{k,h}(s) \geq V_{k',h}(s) \quad \text{when } k' \leq k, \tag{D.7}$$

which follows directly from the update rule (5.10). Moreover, we have the following important lemma regarding the pessimistic property of the value estimate, whose proof is provided in Appendix D.4.3.

Lemma 21 (Pessimistic global value)**.** *Recall $Q_{k,h}$, $V_{k,h}$, and $\pi_{k,h}$ in Algorithm 5. Let $\pi_k = \{\pi_{k,h}\}_{h\in[H]}$. Given any $\delta \in (0,1)$, for all $(k,h) \in \mathcal{C}(K)\times[H]$, it holds with probability at least $1-\delta$ that*

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad |D_2(s,a,k,h)| \le D_3(s,a,k,h) \le \sqrt{\frac{4c_B\zeta_1^2 H^4}{\max\{N_{k,h}(s,a),1\}}}, \tag{D.8a}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad Q_{k,h}(s,a) \le Q_h^{\pi_k}(s,a) \le Q_h^\star(s,a), \tag{D.8b}$$

$$\forall s \in \mathcal{S}: \qquad V_{k,h}(s) \le V_h^{\pi_k}(s) \le V_h^\star(s). \tag{D.8c}$$

In words, Lemma 21 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.

D.2 Proof outline of Theorem 7

Now we are ready to provide the proof of Theorem 7, 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 5 after K episodes and the optimal policy $\pi^\star$ can be bounded as follows:

$$\begin{aligned} V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &= \mathbb{E}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathbb{E}_{s_1\sim\rho}\left[V_1^{\pi_K}(s_1)\right] \\ &\overset{\text{(i)}}{\le} \mathbb{E}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathbb{E}_{s_1\sim\rho}\left[V_{K,1}(s_1)\right] \\ &\overset{\text{(ii)}}{\le} \frac{1}{K}\sum_{v=1}^{\phi(K)} \tau_v\big(\mathbb{E}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathbb{E}_{s_1\sim\rho}\left[V_{t_v,1}(s_1)\right]\big) \end{aligned}$$

$$= \frac{1}{K} \sum_{v=1}^{\phi(K)} \tau_v \sum_{s \in \mathcal{S}} \underbrace{d_1^{\pi^\star}(s)}_{=\rho(s)} \left(V_1^\star(s) - V_{t_v,1}(s)\right)$$

$$\leq \frac{1}{K} \max_{h \in [H]} \sum_{v=1}^{\phi(K)} \tau_v \sum_{s \in \mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_{t_v,h}(s)\right), \tag{D.9}$$

where (i) follows from Lemma 21, and (ii) follows from the monotonicity property in (D.7) and $\sum_{v=1}^{\phi(K)} \tau_v = K$.

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

$$\sum_{s \in \mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_{k,h}(s)\right) = \sum_{s \in \mathcal{S}} d_h^{\pi^\star}(s, \pi_h^\star(s)) \left(V_h^\star(s) - V_{k,h}(s)\right)$$

$$\leq \sum_{s \in \mathcal{S}} d_h^{\pi^\star}(s, \pi_h^\star(s)) \big(Q_h^\star(s, \pi_h^\star(s)) - Q_{k,h}(s, \pi_h^\star(s))\big), \tag{D.10}$$

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 (5.10).

To continue, applying Lemma 19 by setting $\pi = \pi^\star$, the Q-estimate error after k episodes is decomposed as follows:

$$Q_h^\star(s,a) - Q_{k,h}(s,a) = D_1^{\pi^\star}(s,a,k,h) + D_2(s,a,k,h) + D_3(s,a,k,h) + D_4^{\pi^\star}(s,a,k,h)$$

$$\leq D_1^{\pi^\star}(s,a,k,h) + D_4^{\pi^\star}(s,a,k,h) + 2D_3(s,a,k,h), \tag{D.11}$$

where the second line follows from Lemma 21. Finally, inserting the decomposition (D.11) and (D.10) back into (D.9), we control the performance gap with the following terms:

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

$$\leq \frac{1}{K} \max_{h \in [H]} \sum_{v=1}^{\phi(K)} \tau_v \sum_{s \in \mathcal{S}} d_h^{\pi^\star}(s) \left[D_1^{\pi^\star}(s, \pi_h^\star(s), t_v, h) + D_4^{\pi^\star}(s, \pi_h^\star(s), t_v, h) + 2D_3(s, \pi_h^\star(s), t_v, h)\right]$$

$$=: \frac{1}{K} \max_{h \in [H]} \left(D_{1,h} + D_{4,h} + 2D_{3,h}\right), \tag{D.12}$$

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

notation:

$$D_{i,h} \coloneqq \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) D_i^{\pi^\star}(s, \pi_h^\star(s), t_v, h) \qquad \text{for } i \in \{1, 4\},$$
$$D_{3,h} \coloneqq \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) D_3(s, \pi_h^\star(s), t_v, h). \tag{D.13}$$

Step 2: Bounding the decomposed terms. Here, we derive the bound of the decomposed terms separately as follows under the event that (D.6) holds, which is denoted as $\mathcal{E}_0$ and holds with probability at least $1 - \delta$.

- **Bounding $D_{1,h}$.** Using the fact that $0 \le Q_h^\star(s, \pi_h^\star(s)) - Q_{0,h}(s, \pi_h^\star(s)) \le H$, which follows from Lemma 21, it follows

$$\begin{aligned} D_{1,h} &= \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s, \pi_h^\star(s)) \omega_{0,t_v,h}(s, \pi_h^\star(s)) (Q_h^\star(s, \pi_h^\star(s)) - Q_{0,h}(s, \pi_h^\star(s))) \\ &\le \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s, \pi_h^\star(s)) \omega_{0,t_v,h}(s, \pi_h^\star(s)) H \\ &= H \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s, \pi_h^\star(s)) \sum_{v=1}^{\phi(K)} \tau_v \mathbf{1}\{N_{t_v,h}(s, \pi_h^\star(s)) = 0\}, \end{aligned} \tag{D.14}$$

  where the last line follows from (D.4b). To continue, note that

$$\begin{aligned} \sum_{v=1}^{\phi(K)} \tau_v \mathbf{1}\{N_{t_v,h}(s, \pi_h^\star(s)) = 0\} &= \sum_{v\in[\phi(K)]: t_v \le K_0(s,\pi_h^\star(s),h)} \tau_v \mathbf{1}\{N_{t_v,h}(s, \pi_h^\star(s)) = 0\} \\ &\le K_0(s, \pi_h^\star(s), h), \end{aligned}$$

  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 (D.14) leads to

$$\begin{aligned} D_{1,h} &\le H \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s, \pi_h^\star(s)) K_0(s, \pi_h^\star(s), h) \\ &= H \sum_{s\in\mathcal{S}} \frac{\min\{d_h^{\pi^\star}(s, \pi_h^\star(s)), 1/S\}}{d_h^{\mathsf{avg}}(s, \pi_h^\star(s))} \left(\frac{12\zeta_0}{M}\right) \frac{d_h^{\pi^\star}(s, \pi_h^\star(s))}{\min\{d_h^{\pi^\star}(s, \pi_h^\star(s)), 1/S\}} \end{aligned}$$

$$\lesssim \frac{HC^{\star}_{\mathsf{avg}}S}{M}, \tag{D.15}$$

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

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

- **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.4.4.

  Lemma 22. *For any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ and $k\in\mathcal{C}(K)$, 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:*

$$D_3(s,a,k,h) \in \left[\sqrt{\frac{c_B\zeta_1^2H^4}{N_{k,h}(s,a)}}, \sqrt{\frac{4c_B\zeta_1^2H^4}{N_{k,h}(s,a)}}\right]. \tag{D.16}$$

  With the above lemma in hand, recalling (D.13) gives

$$\begin{aligned} D_{3,h} &= \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s)) D_3(s,\pi_h^\star(s),t_v,h) \\ &\le \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s)) \sum_{v=1}^{\phi(K)} \tau_v \sqrt{\frac{4c_B\zeta_1^2H^4}{\max\{N_{t_v,h}(s,\pi_h^\star(s)),1\}}}. \end{aligned} \tag{D.17}$$

  According to Lemma 20, $N_{t_v,h}(s,a)\ge \frac{1}{2}t_vMd_h^{\mathsf{avg}}(s,a)$ holds if $t_v\ge K_0(s,a,h)$ under the event $\mathcal{E}_0$. Therefore,

$$\begin{aligned} \sum_{v=1}^{\phi(K)} \tau_v \sqrt{\frac{H^4}{\max\{N_{t_v,h}(s,a),1\}}} &\le \sum_{v:t_v\le K_0(s,a,h)} \tau_v H^2 + \sum_{v:t_v>K_0(s,a,h)} \tau_v \sqrt{\frac{H^4}{\max\{N_{t_v,h}(s,a),1\}}} \\ &\lesssim H^2K_0(s,a,h) + \sum_{v:t_v>K_0(s,a,h)} \tau_v \sqrt{\frac{H^4}{\max\{N_{t_v,h}(s,a),1\}}} \\ &\lesssim H^2K_0(s,a,h) + \sum_{v=1}^{\phi(K)} \tau_v \sqrt{\frac{H^4}{Mt_vd_h^{\mathsf{avg}}(s,a)}}. \end{aligned} \tag{D.18}$$

Plugging the above inequality and the definitions of $K_0(s, \pi_h^\star(s), h)$ (cf. (D.5)) and $C_{\mathsf{avg}}^\star$ to (D.17), we obtain

$$\begin{aligned}
D_{3,h} &\lesssim \frac{H^2}{M} \sum_{s\in\mathcal{S}} \frac{d_h^{\pi^\star}(s, \pi_h^\star(s))}{d_h^{\mathsf{avg}}(s, \pi_h^\star(s))} + \sum_{v=1}^{\phi(K)} \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s, \pi_h^\star(s)) \tau_v \sqrt{\frac{H^4}{M t_v d_h^{\mathsf{avg}}(s, \pi_h^\star(s))}} \\
&\lesssim \frac{H^2 C_{\mathsf{avg}}^\star}{M} \sum_{s\in\mathcal{S}} \frac{d_h^{\pi^\star}(s, \pi_h^\star(s))}{\min\{d_h^{\pi^\star}(s, \pi_h^\star(s)), 1/S\}} \\
&\qquad + \sum_{v=1}^{\phi(K)} \sqrt{\frac{H^4 C_{\mathsf{avg}}^\star \tau_v^2}{M t_v}} \sum_{s\in\mathcal{S}} \sqrt{\frac{(d_h^{\pi^\star}(s, \pi_h^\star(s)))^2}{\min\{d_h^{\pi^\star}(s, \pi_h^\star(s)), 1/S\}}} \\
&\overset{\text{(i)}}{\lesssim} \frac{H^2 C_{\mathsf{avg}}^\star S}{M} + \sqrt{\frac{H^4 C_{\mathsf{avg}}^\star S}{M}} \sum_{v=1}^{\phi(K)} \sqrt{\tau_v} \sqrt{\frac{\tau_v}{t_v}} \\
&\overset{\text{(ii)}}{\lesssim} \frac{H^2 C_{\mathsf{avg}}^\star S}{M} + \sqrt{\frac{H^4 S K C_{\mathsf{avg}}^\star}{M}},
\end{aligned} \tag{D.19}$$

where (i) holds due to the 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 24 (see Appendix D.3).

- **Bounding $D_{4,h}$.** In the following lemma, whose proof is provided in Appendix D.4.5, we extract the recursive formulation of $D_{4,h}$ as follows.

  Lemma 23. *Consider any $\delta \in (0,1)$. For any $h \in [H]$, the following holds with probability at least $1-\delta$:*

$$\begin{aligned}
&\sum_{v=1}^{\phi(K)} \tau_v \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{m=1}^{M} \sum_{i\in L_{t_v,h}^m(s,a)} \omega_{i,t_v,h}^m(s,a) P_{h,s,a}(V_{h+1}^\star - V_{\iota(i),h+1}) \\
&\lesssim \sigma_{\mathsf{aux}} + \left(1 + \frac{1}{H}\right) \sum_{u=1}^{\phi(K)} \tau_u \sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s) (V_{h+1}^\star(s) - V_{t_{u-1},h+1}(s)),
\end{aligned} \tag{D.20}$$

where $\sigma_{\mathsf{aux}} = \sqrt{\frac{H^2 K S C^\star_{\mathsf{avg}}}{M}} + \frac{H^2 S C^\star_{\mathsf{avg}}}{M}$.

Step 3: Recursion. Combining the bounds of the decomposed errors (cf. (D.15), (D.19), and (D.20)), for any $h \in [H]$, we obtain the following recursive relation:

$$\begin{aligned}
&\sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_h^\star(s) - V_{t_v,h}(s)\right) \\
&\lesssim \theta_K + \left(1+\frac{1}{H}\right)\sum_{u=1}^{\phi(K)} \tau_u \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_{h+1}^\star(s) - V_{t_{u-1},h+1}(s)\right) \\
&\overset{\text{(i)}}{\lesssim} (\theta_K + H\tau_1) + \left(1+\frac{1}{H}\right)\sum_{u=1}^{\phi(K)-1} \tau_{u+1} \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_{h+1}^\star(s) - V_{t_u,h+1}(s)\right) \\
&\overset{\text{(ii)}}{\lesssim} (\theta_K + H\tau_1) + \left(1+\frac{2}{H}\right)^2\sum_{u=1}^{\phi(K)-1} \tau_{u} \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_{h+1}^\star(s) - V_{t_u,h+1}(s)\right), \qquad \text{(D.21)}
\end{aligned}$$

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 21, and we denote

$$\theta_k \coloneqq \frac{H C^\star_{\mathsf{avg}} S}{M} + \frac{H^2 C^\star_{\mathsf{avg}} S}{M} + \sqrt{\frac{H^4 S C^\star_{\mathsf{avg}} k}{M}} + \sqrt{\frac{H^2 k S C^\star_{\mathsf{avg}}}{M}} + \frac{H^2 S C^\star_{\mathsf{avg}}}{M} \qquad \text{(D.22)}$$

for any $k \in [K]$. Then, by invoking the recursion $(H - h + 1)$ times, it follows that

$$\begin{aligned}
&\sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_h^\star(s) - V_{t_v,h}(s)\right) \\
&\lesssim (\theta_K + H\tau_1) + \left(1+\frac{2}{H}\right)^2(\theta_{t_{\phi(K)-1}} + H\tau_1) \\
&\qquad\qquad + \left(1+\frac{2}{H}\right)^4\sum_{u=1}^{\phi(K)-2} \tau_{u} \sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_{h+2}^\star(s) - V_{t_u,h+2}(s)\right) \\
&\lesssim (\theta_K + H\tau_1) + \left(1+\frac{2}{H}\right)^2(\theta_{t_{\phi(K)-1}} + H\tau_1) + \cdots + \left(1+\frac{2}{H}\right)^{2(H-h+1)}(\theta_{t_{\phi(K)-H+h-1}} + H\tau_1) \\
&\lesssim H\theta_K + H^2\tau_1 \qquad \text{(D.23)}
\end{aligned}$$

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

Finally, by plugging the above bound into (D.9), we obtain the bound of the performance gap as follows:

$$\begin{aligned} V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &\le \frac{1}{K} \max_{h\in[H]} \sum_{v=1}^{\phi(K)} \tau_v \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_{t_v,h}(s)\right) \\ &\lesssim \frac{1}{K}(H\theta_K + H^2\tau_1) \\ &\lesssim \frac{H^3 SC^\star_{\mathsf{avg}}}{MK} + \sqrt{\frac{H^6 SC^\star_{\mathsf{avg}}}{MK}} + \frac{H^2\tau_1}{K} \overset{T=HK}{\lesssim} \sqrt{\frac{H^7 SC^\star_{\mathsf{avg}}}{MT}} + \frac{H^4 SC^\star_{\mathsf{avg}}}{MT} \end{aligned} \tag{D.24}$$

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

D.3 Technical lemmas

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

Lemma 24. *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:

$$\begin{aligned} X(Z) = X(Z-1) + \frac{x_Z}{X_Z} &\le 1 + \log X_{Z-1} + 1 - \frac{X_{Z-1}}{X_Z} \\ &\le 1 + \log X_{Z-1} - \log\left(\frac{X_{Z-1}}{X_Z}\right) = 1 + \log X_Z, \end{aligned} \tag{D.25}$$

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 (D.4c).

Lemma 25. *For any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $k' \le k \in \mathcal{C}(K)$, where we denote $u = \phi(k)$, and $i \in L^m_{k,h}(s,a)$. Then, it follows that:*

$$\omega^m_{i,k,h}(s,a) \le \frac{2H}{N_{k,h}(s,a) + Hn_{k,h}(s,a)}, \tag{D.26a}$$

$$\sum_{m=1}^{M} \sum_{j \in L^m_{k,h}(s,a)} \omega^m_{j,k,h}(s,a) \le 1, \tag{D.26b}$$

$$\sum_{m=1}^{M} \sum_{j \in l^m_{k',h}(s,a)} \omega^m_{j,k,h}(s,a) \le \frac{(H+1)n_{k',h}}{N_{k,h} + Hn_{k,h}}, \tag{D.26c}$$

$$\sum_{m=1}^{M} \sum_{j \in L^m_{k,h}(s,a)} (\omega^m_{i,k,h}(s,a))^2 \le \frac{2H}{N_{k,h}(s,a) + Hn_{k,h}(s,a)}, \tag{D.26d}$$

$$\sum_{v \ge u}^{\infty} n_{t_v,h}(s,a) \sum_{m=1}^{M} \sum_{i \in l^m_{k,h}(s,a)} \omega^m_{i,t_v,h}(s,a) \le n_{k,h}(s,a) \left(1 + \frac{1}{H}\right). \tag{D.26e}$$

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

Proof of (D.26a)**.** Recalling the definition of $\omega^m_{i,k,h}$ in (D.4c) and using the fact that $H \ge 1$,

$$\omega^m_{i,k,h} = \frac{H+1}{N_{k,h} + Hn_{k,h}} \left(\prod_{x=\phi(i)}^{\phi(k)-1} \frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}} \right) \le \frac{2H}{N_{k,h} + Hn_{k,h}}. \tag{D.27}$$

Proof of (D.26b)**.** By rearranging the terms,

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

$$
\begin{aligned}
&= \sum_{v=1}^{\phi(k)} \left(1 - \frac{N_{t_{v-1},h}}{N_{t_v,h} + Hn_{t_v,h}}\right) \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \\
&= \sum_{v=1}^{\phi(k)} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}} - \prod_{x=v}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \\
&= 1 - \prod_{x=1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}} \leq 1. \qquad \text{(D.28)}
\end{aligned}
$$

Proof of (D.26c)**.** Let $v = \phi(k')$, i.e., $k' = t_v$. Similarly to the proof of (D.26b), by arranging some terms, we obtain the upper bound as follows:

$$
\begin{aligned}
\sum_{m=1}^{M} \sum_{j \in l^m_{k',h}(s,a)} \omega^m_{j,k,h}(s,a) &= \sum_{m=1}^{M} \sum_{j \in l^m_{t_v,h}(s,a)} \frac{H+1}{N_{t_v,h} + Hn_{t_v,h}} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \\
&= \frac{(H+1)n_{t_v,h}}{N_{t_v,h} + Hn_{t_v,h}} \left(\prod_{x=v+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \\
&= \frac{(H+1)n_{t_v,h}}{N_{k,h} + Hn_{k,h}} \left(\prod_{x=v}^{\phi(k)-1} \frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}}\right) \\
&\leq \frac{(H+1)n_{k',h}}{N_{k,h} + Hn_{k,h}}. \qquad \text{(D.29)}
\end{aligned}
$$

Proof of (D.26d)**.** Using the bound in (D.26a) and (D.26b),

$$
\sum_{m=1}^{M} \sum_{j \in L^m_{k,h}} (\omega^m_{j,k,h})^2 = \left(\max_{m \in [M], j \in L^m_{k,h}} \omega^m_{j,k,h}\right) \sum_{m=1}^{M} \sum_{j \in L^m_{k,h}} \omega^m_{j,k,h} \leq \max_{m \in [M], j \in L^m_{k,h}} \omega^m_{j,k,h} \leq \frac{2H}{N_{k,h} + Hn_{k,h}}. \qquad \text{(D.30)}
$$

Proof of (D.26e)**.** Recall that $k = t_u$. Then, reusing the intermediate result derived in (D.29),

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

$$
\begin{aligned}
&= (H+1)n_{t_u,h}\sum_{v\geq u}^{\infty}\frac{n_{t_v,h}}{N_{t_v,h}+Hn_{t_v,h}}\left(\prod_{x=u}^{v-1}\beta_{x,h}\right)\\
&= (H+1)n_{t_u,h}\sum_{v\geq u}^{\infty}\frac{1}{H}(1-\beta_{v,h})\left(\prod_{x=u}^{v-1}\beta_{x,h}\right)\\
&\leq n_{k,h}\left(1+\frac{1}{H}\right).
\end{aligned}
\tag{D.31}
$$

□

D.4 Proofs for Offline Federated Q-Learning (Theorem 7)

D.4.1 Proof of Lemma 19

For any $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$ and $k\in\mathcal{C}(K)$, according to the pessimistic aggregation update rule in (5.9), the estimate error of Q function at the k-th iteration can be written as follows:

$$
\begin{aligned}
Q_h^{\pi}(s,a)-Q_{k,h}(s,a) &= Q_h^{\pi}(s,a)-\left(\sum_{m=1}^{M}\alpha_{k,h}^m(s,a)Q_{k,h}^m(s,a)\right)+B_{k,h}(s,a)\\
&= \sum_{m=1}^{M}\alpha_{k,h}^m(s,a)\left(Q_h^{\pi}(s,a)-Q_{k,h}^m(s,a)\right)+B_{k,h}(s,a),
\end{aligned}
\tag{D.32}
$$

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 (5.7), 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:

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

$$+ \eta_{i,h}^m(s,a) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) + \eta_{i,h}^m(s,a)(P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m, \tag{D.33}$$

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:

$$\begin{aligned} Q_h^\pi(s,a) - Q_{k,h}^m(s,a) = & \prod_{i \in l_{k,h}^m(s,a)} (1 - \eta_{i,h}^m(s,a)) \left(Q_h^\pi(s,a) - Q_{\iota(k),h}(s,a) \right) \\ & + \sum_{i \in l_{k,h}^m(s,a)} \eta_{i,h}^m(s,a) \prod_{\{j > i : j \in l_{k,h}^m(s,a)\}} (1 - \eta_{j,h}^m(s,a)) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) \\ & + \sum_{i \in l_{k,h}^m(s,a)} \eta_{i,h}^m(s,a) \prod_{\{j > i : j \in l_{k,h}^m(s,a)\}} (1 - \eta_{j,h}^m(s,a)) (P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m \end{aligned} \tag{D.34}$$

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 (D.34) to (D.32) and letting $v = \phi(k)$, we obtain the following recursive relation for u-th local updates:

$$\begin{aligned} & Q_h^\pi(s,a) - Q_{k,h}(s,a) \\ & = \underbrace{\left(\sum_{m=1}^M \alpha_{k,h}^m(s,a) \prod_{i \in l_{k,h}^m(s,a)} (1 - \eta_{i,h}^m(s,a)) \right)}_{:= \lambda_{v,h}(s,a)} \left(Q_h^\pi(s,a) - Q_{\iota(k),h}(s,a) \right) + B_{k,h}(s,a) \\ & \quad + \sum_{m=1}^M \sum_{i \in l_{k,h}^m(s,a)} \left(\alpha_{k,h}^m(s,a) \eta_{i,h}^m(s,a) \prod_{\{j > i : j \in l_{k,h}^m(s,a)\}} (1 - \eta_{j,h}^m(s,a)) \right) P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) \\ & \quad + \sum_{m=1}^M \sum_{i \in l_{k,h}^m(s,a)} \left(\alpha_{k,h}^m(s,a) \eta_{i,h}^m(s,a) \prod_{\{j > i : j \in l_{k,h}^m(s,a)\}} (1 - \eta_{j,h}^m(s,a)) \right) (P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m \\ & = \lambda_{v,h}(s,a) \left(Q_h^\pi(s,a) - Q_{\iota(k),h}(s,a) \right) + B_{k,h}(s,a) \\ & \quad + \frac{(H+1)}{N_{t_v,h}(s,a) + H n_{t_v,h}(s,a)} \sum_{m=1}^M \sum_{i \in l_{k,h}^m(s,a)} P_{h,s,a}(V_{h+1}^\pi - V_{i-1,h+1}^m) \\ & \quad + \frac{(H+1)}{N_{t_v,h}(s,a) + H n_{t_v,h}(s,a)} \sum_{m=1}^M \sum_{i \in l_{k,h}^m(s,a)} (P_{h,s,a} - P_{i,h}^m) V_{i-1,h+1}^m. \end{aligned} \tag{D.35}$$

Here, the last line holds by invoking the definitions in (5.13) and (5.14) and observing with abuse of notation (omit (s,a) when it is clear)

$$
\begin{aligned}
&\alpha^m_{k,h}(s,a)\eta^m_{i,h}(s,a)\prod_{\{j>i:j\in l^m_{k,h}(s,a)\}}(1-\eta^m_{j,h}(s,a))\\
&=\frac{1}{M}\frac{N_{\iota(k),h}+M(H+1)n^m_{k,h}}{N_{k,h}+Hn_{k,h}}\frac{M(H+1)}{N_{\iota(i),h}+M(H+1)n^m_{i,h}}\left(\prod_{j=1}^{n^m_{k,h}-n^m_{i,h}}\left(\frac{N_{\iota(i),h}+M(H+1)(n^m_{i,h}+j-1)}{N_{\iota(i),h}+M(H+1)(n^m_{i,h}+j)}\right)\right)\\
&=\frac{1}{M}\frac{N_{\iota(k),h}+M(H+1)n^m_{k,h}}{N_{k,h}+Hn_{k,h}}\frac{M(H+1)}{N_{\iota(i),h}+M(H+1)n^m_{i,h}}\frac{N_{\iota(i),h}+M(H+1)n^m_{i,h}}{N_{\iota(i),h}+M(H+1)n^m_{k,h}}\\
&=\frac{(H+1)}{N_{k,h}+Hn_{k,h}}=\frac{(H+1)}{N_{t_v,h}+Hn_{t_v,h}}
\end{aligned}
\tag{D.36}
$$

where the last line holds since $\iota(i)=\iota(k)$ for $i\in l^m_{k,h}(s,a)$ 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:

$$
\begin{aligned}
&Q^\pi_h(s,a)-Q_{k,h}(s,a)\\
&=\underbrace{\prod_{u=1}^{\phi(k)}\lambda_{u,h}(s,a)}_{:=\omega_{0,k,h}(s,a)}(Q^\pi_h(s,a)-Q_{0,h}(s,a))+\sum_{u=1}^{\phi(k)}B_{t_u,h}(s,a)\prod_{x=u+1}^{\phi(k)}\lambda_{x,h}(s,a)\\
&\quad+\sum_{u=1}^{\phi(k)}\sum_{m=1}^{M}\sum_{i\in l^m_{t_u,h}(s,a)}\underbrace{\left(\frac{H+1}{N_{t_u,h}+Hn_{t_u,h}}\prod_{x=u+1}^{\phi(k)}\lambda_{x,h}(s,a)\right)}_{:=\omega_{i,k,h}(s,a)}(P_{h,s,a}-P^m_{i,h})V^m_{i-1,h+1}\\
&\quad+\sum_{u=1}^{\phi(k)}\sum_{m=1}^{M}\sum_{i\in l^m_{t_u,h}(s,a)}\left(\frac{H+1}{N_{t_u,h}+Hn_{t_u,h}}\prod_{x=u+1}^{\phi(k)}\lambda_{x,h}(s,a)\right)P_{h,s,a}(V^\pi_{h+1}-V^m_{i-1,h+1})\\
&=\omega_{0,k,h}(s,a)(Q^\pi_h(s,a)-Q_{0,h}(s,a))\\
&\quad+\sum_{m=1}^{M}\sum_{i\in L^m_{k,h}(s,a)}\omega^m_{i,k,h}(s,a)(P_{h,s,a}-P^m_{i,h})V^m_{i-1,h+1}\\
&\quad+\sum_{u=1}^{\phi(k)}B_{t_u,h}(s,a)\prod_{x=u+1}^{\phi(k)}\lambda_{x,h}(s,a)
\end{aligned}
$$

$$+\sum_{m=1}^{M}\sum_{i\in L_{k,h}^{m}(s,a)}\omega_{i,k,h}^{m}(s,a)P_{h,s,a}(V_{h+1}^{\pi}-V_{i-1,h+1}^{m}). \tag{D.37}$$

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 (D.4a), (D.4b), and (D.4c), respectively, which will be proved momentarily. For notational simplicity, we omit (s,a) in the derivations.

Proof of (D.4a)**.** 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

$$\begin{aligned}\lambda_{v,h}&=\sum_{m=1}^{M}\frac{1}{M}\frac{(H+1)Mn_{k,h}^{m}}{(H+1)n_{k,h}}\prod_{j=1}^{n_{k,h}^{m}}\left(\frac{M(H+1)(j-1)}{M(H+1)j}\right)\\&=\sum_{m\in[M]:n_{k,h}^{m}=0}^{M}\underbrace{\frac{n_{k,h}^{m}}{n_{k,h}}}_{=0}+\sum_{m\in[M]:n_{k,h}^{m}>0}^{M}\frac{n_{k,h}^{m}}{n_{k,h}}\underbrace{\prod_{j=1}^{n_{k,h}^{m}}\left(\frac{(H+1)(j-1)}{(H+1)j}\right)}_{=0}=0.\end{aligned} \tag{D.38}$$

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

$$\begin{aligned}\lambda_{v,h}&=\sum_{m=1}^{M}\frac{1}{M}\frac{N_{\iota(k),h}+M(H+1)n_{k,h}^{m}}{N_{\iota(k),h}+(H+1)n_{k,h}}\prod_{j=1}^{n_{k,h}^{m}}\left(\frac{N_{\iota(k),h}+M(H+1)(j-1)}{N_{\iota(k),h}+M(H+1)j}\right)\\&=\sum_{m=1}^{M}\frac{1}{M}\frac{N_{\iota(k),h}+M(H+1)n_{k,h}^{m}}{N_{\iota(k),h}+(H+1)n_{k,h}}\frac{N_{\iota(k),h}}{N_{\iota(k),h}+M(H+1)n_{k,h}^{m}}=\frac{N_{\iota(k),h}}{N_{k,h}+Hn_{k,h}}.\end{aligned} \tag{D.39}$$

Proof of (D.4b)**.** According to (D.4a), 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 (D.4c)**.** For i such that $\phi(i) = u$, by rearranging terms and applying (D.4a),

$$\begin{aligned}\omega_{i,k,h}^m &= \frac{(H+1)}{N_{t_u,h} + Hn_{t_u,h}} \left(\prod_{x=u+1}^{\phi(k)} \frac{N_{t_{x-1},h}}{N_{t_x,h} + Hn_{t_x,h}} \right) \\ &= \frac{H+1}{N_{k,h} + Hn_{k,h}} \left(\prod_{x=u}^{\phi(k)-1} \frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}} \right). \end{aligned} \tag{D.40}$$

D.4.2 Proof of Lemma 20

Consider any given $\delta \in (0,1)$ and $(k,s,a,h) \in [K] \times \mathcal{S} \times \mathcal{A} \times [H]$. Note that $N_{k,h}^m(s,a) \sim \mathsf{Binomial}(k, d_h^m(s,a))$ for all $m \in [M]$. Then recall the definition of $N_{k,h}(s,a)$ in Section 5.3, we can view $N_{k,h}(s,a) = \sum_{m=1}^M N_{k,h}^m(s,a)$ as a sum of kM independent Bernoulli variables with expectation $\nu := \mathbb{E}[N_{k,h}(s,a)] = kMd_h^{\mathsf{avg}}(s,a)$. Therefore, applying Chernoff bound (see Mitzenmacher and Upfal (2005, Theorem 4.4)) yields:

$$\forall t \in [0,1] \quad : \quad \mathbb{P}(\left|N_{k,h}^m(s,a) - \nu\right| \geq \nu t) \leq \exp\left(-c_1 \nu t^2\right), \tag{D.41a}$$

$$\forall t \geq 1 \quad : \quad \mathbb{P}(N_{k,h}^m(s,a) - \nu \geq t\nu) \leq \exp\left(-c_1 \nu t\right), \tag{D.41b}$$

for some universal constant $c_1 > 0$.

Armed with above facts and notations, now we are ready to prove (D.6). First, applying (D.41a) with $t = \frac{1}{2}$, we arrive at:

$$\mathbb{P}\left(\left|N_{k,h}^m(s,a) - \nu\right| \geq \frac{\nu}{2}\right) \leq \exp\left(-\frac{c_1\nu}{4}\right) \leq \delta, \tag{D.42}$$

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

To continue, when $\nu = kMd_h^{\mathsf{avg}}(s,a) \leq \frac{4}{c_1}\log(1/\delta)$, applying (D.41b) with $t = \frac{4\log(1/\delta)}{\nu c_1} \geq 1$ gives:

$$\mathbb{P}\left(N_{k,h}^m(s,a) - \nu \geq \frac{4\log(1/\delta)}{c_1}\right) \leq \exp(-4\log(1/\delta)) \leq \delta. \tag{D.43}$$

Summing up (D.42) and (D.43) 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:

$$\text{when } k \geq \frac{4}{c_1 M d_h^{\mathsf{avg}}} \log\left(\frac{|\mathcal{S}||\mathcal{A}|KH}{\delta}\right) \;:\; \frac{kMd_h^{\mathsf{avg}}}{2} = \frac{\nu}{2} \leq N_{k,h}^m(s,a) \leq \frac{3\nu}{2} \leq 2kMd_h^{\mathsf{avg}},$$
$$\text{when } k \leq \frac{4}{c_1 M d_h^{\mathsf{avg}}} \log\left(\frac{|\mathcal{S}||\mathcal{A}|KH}{\delta}\right) \;:\; N_{k,h}^m(s,a) \leq \frac{8}{c_1} \log\left(\frac{|\mathcal{S}||\mathcal{A}|KH}{\delta}\right)$$

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

D.4.3 Proof of Lemma 21

Proof of (D.8a)

Noticing that the (D.8a) involves two terms of interest, and we start with bounding $D_2(s,a,k,h)$. For any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and any $k \in \mathcal{C}(K)$, we can rewrite $D_2(s,a,k,h)$ as

$$D_2(s,a,k,h) = \sum_{i=1}^{k}\sum_{m=1}^{M} X_{i,k,h}^m(s,a), \tag{D.44}$$

where $X_{i,k,h}^m(s,a) = \omega_{i,k,h}^m(s,a)(P_{h,s,a} - P_{i,h}^m)V_{i-1,h+1}^m \mathbf{1}\{(s_{i,h}^m, a_{i,h}^m) = (s,a)\}$. To continue, we first introduce Lemma 26, whose proof is provided in Appendix D.4.3.

Lemma 26. *For any $(k,s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and $N \in [1, MK]$, let*

$$\widetilde{X}_{i,k,h}^m(s,a;N) = \widetilde{\omega}_{i,k,h}^m(s,a;N)(P_{h,s,a} - P_{i,h}^m)V_{i-1,h+1}^m \mathbf{1}\{(s_{i,h}^m, a_{i,h}^m) = (s,a)\}, \tag{D.45}$$

where

$$\widetilde{\omega}_{i,k,h}^m(s,a;N) \coloneqq \frac{H+1}{N + Hn_{k,h}(s,a)} \left(\prod_{x=\phi(i)}^{\phi(k)-1} \frac{N_{t_x,h}(s,a)}{N_{t_x,h}(s,a) + Hn_{t_x,h}(s,a)}\right) I_{i,h}^m(s,a;N), \tag{D.46}$$

and $I_{i,h}^m(s,a;N) \coloneqq \mathbf{1}\{\sum_{m'=1}^{M} N_{i-1,h}^{m'}(s,a) + \sum_{m'=1}^{m} \mathbf{1}\{(s_{i,h}^{m'}, a_{i,h}^{m'}) = (s,a)\} \leq N\}$. Then, for

any $\delta \in (0,1)$*, the following holds:*

$$\left|\sum_{i=1}^{k}\sum_{m=1}^{M}\widetilde{X}_{i,k,h}^{m}(s,a;N)\right| \leq \sqrt{\frac{81H^4\zeta_1^2}{N}} \tag{D.47}$$

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

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

$$\text{when } N_{k,h}(s,a) > 0: \quad |D_2(s,a,k,h)| \leq \left|\sum_{i=1}^{k}\sum_{m=1}^{M}\widetilde{X}_{i,k,h}^{m}(s,a;N_{k,h}(s,a))\right| \leq \sqrt{\frac{81H^4\zeta_1^2}{N_{k,h}(s,a)}} \tag{D.48}$$

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

$$|D_2(s,a,k,h)| \leq \left|\sum_{i=1}^{k}\sum_{m=1}^{M}\widetilde{X}_{i,k,h}^{m}(s,a;N_{k,h}(s,a))\right| \leq \sqrt{\frac{81H^4\zeta_1^2}{N_{k,h}(s,a)}}. \tag{D.49}$$

Finally, combining the results for $D_2(s,a,k,h)$ (cf. (D.49)) and $D_3(s,a,k,h)$ (cf. (D.16) in Lemma 22), we conclude that for any $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]$ with $k \in \mathcal{C}(K)$, it holds with probability at least $1-\delta$ that

$$|D_2(s,a,k,h)| \leq \sqrt{\frac{81H^4\zeta_1^2}{N_{k,h}(s,a)}} = \sqrt{\frac{c_B\zeta_1^2H^4}{N_{k,h}(s,a)}} \leq D_3(s,a,k,h). \tag{D.50}$$

Proof of (D.8b) **and** (D.8c)

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

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

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

- **Base case.** When $h = H + 1$, for all $(s, a, k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{C}(K)$, 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 $(s, a, k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{C}(K)$. First, we will verify the Q-estimates at step h are pessimistic. For any $(s, a, k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{C}(K)$, applying Lemma 19,

$$Q^{\pi_k}_h(s, a) - Q_{k,h}(s, a) = D^{\pi_k}_1(s, a, k, h) + D_2(s, a, k, h) + D_3(s, a, k, h) + D^{\pi_k}_4(s, a, k, h). \tag{D.51}$$

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 (D.8a), $|D_2(s, a, k, h)| \leq D_3(s, a, k, h)$. And it is clear that $D_4 \geq 0$ due to

$$V^{\pi_k}_{h+1} \geq V_{k,h+1} \geq V_{\iota(i),h+1}, \tag{D.52}$$

where the first inequality holds by the induction assumption, and the last inequality arises from the monotonicity of the global value update in (5.10). Therefore, it is clear that for any $(s, a, k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{C}(K)$, the Q-estimates at step h are pessimistic, i.e.,

$$Q^{\pi_k}_h(s, a) - Q_{k,h}(s, a) \geq 0. \tag{D.53}$$

Next, to show that value estimates at step h are pessimistic, recalling the global update in (5.10),

$$V^{\pi_k}_h(s) - V_{k,h}(s) = Q^{\pi_k}_h(s, \pi_{k,h}(s)) - \max\{\max_a Q_{k,h}(s, a), V_{\iota(k),h}(s)\}$$

$$= Q_h^{\pi_k}(s, \pi_{k,h}(s)) - \max_a Q_{k_0,h}(s,a)$$
$$= Q_h^{\pi_k}(s, \pi_{k_0,h}(s)) - Q_{k_0,h}(s, \pi_{k_0,h}(s)) \geq 0, \tag{D.54}$$

where k_0 denotes the most recent episode satisfying $V_{k,h}(s) = \max_a Q_{k_0,h}(s,a)$ and $k \geq k_0 \in \mathcal{C}(K)$, and the last inequality holds because $\pi_{k,h}(s) = \pi_{k_0,h}(s)$ and $Q_h^{\pi_k}(s,a) - Q_{k_0,h}(s,a) \geq 0$ can be similarly verified using (D.51) and (D.52) for k_0. Now, we verify that $Q_h^{\pi_k}(s,a) \geq Q_{k,h}(s,a)$ and $V_h^{\pi_k}(s) \geq V_{k,h}(s)$ holds at step h for any $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{C}(K)$, and this directly completes the induction argument.

Proof of Lemma 26

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

$$\forall (i,m) \in [k] \times [M]: \quad \mathbb{E}_{(i,m)}[\cdot] = \mathbb{E}\big[\, \cdot \mid \big\{s_{j,h}^{m'}, a_{j,h}^{m'}, V_{j,h+1}^{m}\big\}_{j<i, m' \in [M]},\ \big\{s_{i,h}^{m'}, a_{i,h}^{m'}\big\}_{m' \leq m}\big]. \tag{D.55}$$

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

Consequently, we can apply Freedman's inequality (see the user-friendly version provided in Theorem 10) 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:

$$\sum_{i=1}^{k} \sum_{m=1}^{M} \widetilde{X}_{i,k,h}^m(s,a;N) \overset{\text{(i)}}{\leq} \sqrt{8B_1\zeta_1} + \frac{4}{3}B_2\zeta_1 \overset{\text{(ii)}}{\leq} \sqrt{\frac{32H^4\zeta_1}{N}} + \frac{3H^2\zeta_1}{N} \leq \sqrt{\frac{81H^4\zeta_1^2}{N}} \tag{D.56}$$

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

$$B_1 \coloneqq \sum_{i=1}^{k} \sum_{m=1}^{M} \mathbb{E}_{(i,m)} \left[\left(\widetilde{X}_{i,k,h}^m(s,a;N) \right)^2 \right] \leq \frac{4H^4}{N}, \tag{D.57}$$

$$B_2 \coloneqq \max_{(i,m)\in[k]\times[M]} \left|\widetilde{X}^m_{i,k,h}(s,a;N)\right| \le \frac{2H^2}{N} \tag{D.58}$$

where the proofs of (D.57) and (D.58) are provided as below, respectively.

Proof of (D.57)**.** In view of that the events happen at any time step h are 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

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{m=1}^{M} \mathbb{E}_{(i,m)}[(\widetilde{X}^m_{i,k,h}(s,a;N))^2] &= \sum_{i=1}^{k}\sum_{m=1}^{M} \mathbb{E}_{(i,m)}[(\widetilde{\omega}^m_{i,k,h}(s,a;N))^2]\mathsf{Var}_{P_{h,s,a}}(V^m_{i-1,h+1}) \\
&\le H^2 \sum_{i=1}^{k}\sum_{m=1}^{M} \mathbb{E}_{(i,m)}[(\widetilde{\omega}^m_{i,k,h}(s,a;N))^2] \\
&\le H^2 N \left(\frac{2H}{N}\right)^2 = \frac{4H^4}{N},
\end{aligned} \tag{D.59}$$

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

Proof of (D.58)**.** For any $(i,m,h) \in [k]\times[M]\times[H]$ and fixed $N \in [1, MK]$, it is observed that

$$\begin{aligned}
\left|\widetilde{X}^m_{i,k,h}(s,a;N)\right| &= \left|\widetilde{\omega}^m_{i,k,h}(s,a;N)(P_{h,s,a} - P^m_{i,h})V^m_{i-1,h+1}\mathbf{1}\{(s^m_{i,h}, a^m_{i,h}) = (s,a)\}\right| \\
&\le |\widetilde{\omega}^m_{i,k,h}(s,a;N)| \cdot \|P_{h,s,a} - P^m_{i,h}\|_1 \cdot \|V^m_{i-1,h+1}\|_\infty \le \frac{2H^2}{N},
\end{aligned} \tag{D.60}$$

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

D.4.4 Proof of Lemma 22

With slight abuse of notation, we will omit (s,a) from some notation when it is clear from the context for simplicity in this proof. Recall the definition of $D_3(s,a,k,h)$ in (D.3)

and the global penalty defined in (5.12). When $N_{k,h}(s,a) = 0$, the global penalties are all 0, which yields $D_3(s,a,k,h) = 0$. Therefore, it suffices to focus on the case when $N_{k,h}(s,a) > 0$ and show that for $c_B = 81$, $c_u = 4$ and $c_l = 1$,

$$D_3(s,a,k,h) = \sum_{u=1}^{\phi(k)} B_{t_u,h}(s,a) \prod_{u'=u+1}^{\phi(k)} \lambda_{u',h}(s,a) \in \left[\sqrt{\frac{c_l c_B \zeta_1^2 H^4}{N_{k,h}(s,a)}}, \sqrt{\frac{c_u c_B \zeta_1^2 H^4}{N_{k,h}(s,a)}} \right]. \quad \text{(D.61)}$$

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

$$\begin{aligned} \sum_{u=1}^{z} B_{t_u,h} \prod_{u'=u+1}^{z} \lambda_{u',h} &= \sum_{u=1}^{z} \frac{(H+1)n_{t_u,h}}{N_{k,h} + Hn_{t_u,h}} \sqrt{\frac{c_B \zeta_1^2 H^4}{N_{t_u,h}}} \prod_{u'=u+1}^{z} \lambda_{u',h} \\ &= \sqrt{c_B \zeta_1^2 H^4} \sum_{u=1}^{z} \sqrt{\frac{1}{N_{t_u,h}}} (1 - \lambda_{u,h}) \prod_{u'=u+1}^{z} \lambda_{u',h} \\ &= \sqrt{c_B \zeta_1^2 H^4} Y(z) \end{aligned} \quad \text{(D.62)}$$

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

$$Y(z) \coloneqq \sum_{u=1}^{z} \sqrt{\frac{1}{N_{t_u,h}}} (1 - \lambda_{u,h}) \prod_{u'=u+1}^{z} \lambda_{u',h}. \quad \text{(D.63)}$$

As a result, to show (D.61), it suffices to verify that

$$Y(z) \in \left[\sqrt{\frac{c_l}{N_{t_z,h}(s,a)}}, \sqrt{\frac{c_u}{N_{t_z,h}(s,a)}} \right], \quad \text{(D.64)}$$

which we proceed by an induction argument.

Proof of (D.64) **by induction.** To begin with, for the basic case $z = 1$, it is easily verified that

$$Y(1) = \begin{cases} \sqrt{\frac{1}{N_{t_1,h}}} & \text{if } \ n_{t_1,h} > 0 \\ 0 & \text{if } \ n_{t_1,h} = 0 \end{cases}, \tag{D.65}$$

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 (D.64) holds for $z-1$, namely,

$$Y(z-1) \in \left[\sqrt{\frac{c_l}{N_{t_{z-1},h}}}, \sqrt{\frac{c_u}{N_{t_{z-1},h}}}\right], \tag{D.66}$$

we hope to show (D.64) holds for z. Towards this, we first show the upper bound in (D.64) holds for z as follows:

$$\begin{aligned}
Y(z) &= Y(z-1)\lambda_{z,h} + \sqrt{\frac{1}{N_{t_z,h}}}(1-\lambda_{z,h}) \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{c_u}{N_{t_{z-1},h}}}\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}} + \sqrt{\frac{1}{N_{t_z,h}}}\frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}} \\
&\leq \sqrt{\frac{c_u}{N_{t_z,h}}}\sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}}} + \sqrt{\frac{1}{N_{t_z,h}}}\frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}} \\
&= \sqrt{\frac{c_u}{N_{t_z,h}}}\left(\sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}}} + \sqrt{\frac{1}{c_u}}\frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}}\right) \\
&= \sqrt{\frac{c_u}{N_{t_z,h}}}\left(\sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}}} + \sqrt{\frac{1}{c_u}}\left(1-\sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}}}\right)\left(1+\sqrt{\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}}}\right)\right) \\
&\leq \sqrt{\frac{c_u}{N_{t_z,h}}},
\end{aligned} \tag{D.67}$$

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 (D.67), the lower bound of $Y(z)$ is derived as below:

$$\begin{aligned}
Y(z) &= Y(z-1)\lambda_{z,h} + \sqrt{\frac{1}{N_{t_z,h}}}(1-\lambda_{z,h}) \\
&\geq \sqrt{\frac{c_l}{N_{t_{z-1},h}}}\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}} + \sqrt{\frac{1}{N_{t_z,h}}}\frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}} \\
&\geq \sqrt{\frac{c_l}{N_{t_z,h}}}\frac{N_{t_{z-1},h}}{N_{t_z,h}+Hn_{t_z,h}} + \sqrt{\frac{1}{N_{t_z,h}}}\frac{(H+1)n_{t_z,h}}{N_{t_z,h}+Hn_{t_z,h}} \\
&\geq \sqrt{\frac{c_l}{N_{t_z,h}}}, \qquad \text{(D.68)}
\end{aligned}$$

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, (D.64) holds for any $z \in \phi(K)$, and this completes the proof.

D.4.5 Proof of Lemma 23

Recall the definition of $D_{4,h}$ (see (D.13) and (D.3)), $D_{4,h}$ can be rewritten as follows:

$$\begin{aligned}
D_{4,h} &= \sum_{v=1}^{\phi(K)} \tau_v \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{m=1}^{M} \sum_{i\in L^m_{t_v,h}(s,a)} \omega^m_{i,t_v,h}(s,a) P_{h,s,a}(V^\star_{h+1} - V_{\iota(i),h+1}) \\
&\overset{\text{(i)}}{=} \sum_{v=1}^{\phi(K)} \tau_v \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1}) \underbrace{\sum_{m=1}^{M}\left(\sum_{i\in l^m_{t_u,h}(s,a)} \omega^m_{i,t_v,h}(s,a)\right)}_{=:\psi_{u,v,h}(s,a)} \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \sum_{t_{v-1}<j\leq t_v} d_h^{\pi^\star}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V^\star_{h+1} - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a), \qquad \text{(D.69)}
\end{aligned}$$

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_v.

To further control (D.69), we introduce the following lemma that bounds the expectation form (D.69) by an empirical version; the proof is postponed to Appendix D.4.5.

Lemma 27. *Consider any $\delta \in (0,1)$. For any $h \in [H]$, the following holds:*

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \sum_{t_{v-1}<j\le t_v} d_h^{\pi^\star}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a)$$
$$\lesssim \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} n_{t_v,h}(s,a) \sum_{u=1}^{v} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) + \sigma_{\mathsf{aux},1} \tag{D.70}$$

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

$$\sigma_{\mathsf{aux},1} \lesssim \sqrt{\frac{H^2 K S C_{\mathsf{avg}}^\star}{M}} + \frac{H^2 S C_{\mathsf{avg}}^\star}{M} \tag{D.71}$$

Then, applying concentration bounds, $D_{4,h}$ is bounded as follows:

$$\begin{aligned} D_{4,h} &\overset{\text{(i)}}{\lesssim} \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{v=1}^{\phi(K)} \sum_{u=1}^{v} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} n_{t_v,h}(s,a) P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1})\psi_{u,v,h}(s,a) + \sigma_{\mathsf{aux},1} \\ &= \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{u=1}^{\phi(K)} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1}) \sum_{v=u}^{\phi(K)} n_{t_v,h}(s,a)\psi_{u,v,h}(s,a) + \sigma_{\mathsf{aux},1} \\ &\overset{\text{(ii)}}{\le} \frac{1}{M} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{u=1}^{\phi(K)} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1}) n_{t_u,h}(s,a)\Big(1+\frac{1}{H}\Big) + \sigma_{\mathsf{aux},1} \end{aligned} \tag{D.72}$$

where (i) follows from Lemma 27, and (ii) holds because

$$\sum_{v\ge u}^{\infty} n_{t_v,h}(s,a) \sum_{m=1}^{M} \sum_{i\in l_{t_u,h}^m(s,a)} \omega_{i,t_v,h}^m(s,a) \le n_{t_u,h}(s,a)\Big(1+\frac{1}{H}\Big) \tag{D.73}$$

according (D.26e) in Lemma 25.

To continue, we introduce the following lemma that transfers the distribution at time step h to the distribution at time step $h+1$; the proof is provided in Appendix D.4.5.

Lemma 28. *Consider any $\delta \in (0,1)$. For any $h \in [H]$, the following holds:*

$$\begin{aligned}
&\sum_{u=1}^{\phi(K)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_{t_u,h}(s,a)}{M d_h^{\mathsf{avg}}(s,a)} d_h^{\pi^\star}(s,a) P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1}) \\
&\lesssim \sum_{u=1}^{\phi(K)} \tau_u \sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)(V_{h+1}^\star(s) - V_{t_{u-1},h+1}(s)) + \sigma_{\mathsf{aux},2} \qquad \text{(D.74)}
\end{aligned}$$

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

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

Armed with the above lemma, rearranging the terms in (D.72) and applying Lemma 28,

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

and this completes the proof.

Proof of Lemma 27

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

$$G_{v,h}(s,a) \coloneqq \sum_{u=1}^{v} P_{h,s,a}(V_{h+1}^\star - V_{t_{u-1},h+1}) \psi_{u,v,h}(s,a). \qquad \text{(D.75)}$$

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

$$Y_{j,h}^m \coloneqq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big(d_h^{\mathsf{avg}}(s,a) - \mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}\Big) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} G_{v,h}(s,a) \qquad \text{(D.76)}$$

$$\widetilde{Y}_{j,h}^m \coloneqq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big(d_h^{m}(s,a) - \mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}\Big) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} \widetilde{G}_{v,h}^{-j,m}(s,a), \qquad \text{(D.77)}$$

where we define

$$\widetilde{G}_{v,h}^{-j,m}(s,a) \coloneqq \begin{cases} \widetilde{\psi}_{v,v,h}^{-j,m}(s,a) P_{h,s,a}(V_{h+1}^{\star} - V_{t_{v-1},h+1}) + (1 - \widetilde{\psi}_{v,v,h}^{-j,m}(s,a)) G_{v-1,h}(s,a) & \text{if } v > 1 \\ P_{h,s,a}(V_{h+1}^{\star} - V_{0,h+1}) & \text{if } v = 1 \end{cases} \tag{D.78}$$

and

$$\begin{aligned} \widetilde{\psi}_{v,v,h}^{-j,m}(s,a) &\coloneqq \frac{(H+1)(n_{t_v,h}(s,a) - \mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\})}{N_{t_{v-1},h}(s,a) + (H+1)(n_{t_v,h}(s,a) - \mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\})} \\ &= \frac{(H+1)(\sum_{(m',j') \in [M] \times (t_{v-1}, t_v] \setminus \{(j,m)\}} \mathbf{1}\{(s,a) = (s_{j',h}^{m'}, a_{j',h}^{m'})\})}{N_{t_{v-1},h}(s,a) + (H+1)(\sum_{(m',j') \in [M] \times (t_{v-1}, t_v] \setminus \{(j,m)\}} \mathbf{1}\{(s,a) = (s_{j',h}^{m'}, a_{j',h}^{m'})\})} \end{aligned} \tag{D.79}$$

We replaced $G_{v,h}(s,a)$ with a surrogate $\widetilde{G}_{v,h}^{-j,m}(s,a)$, 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 surrogate is carefully designed to remove the dependency on the event $\mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\}$ 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 the above defined auxiliary terms whose proofs are provided in Appendix D.4.5: for any $v \in [\phi(K)]$,

$$G_{v,h}(s,a) = \begin{cases} \psi_{v,v,h}(s,a) P_{h,s,a}(V_{h+1}^{\star} - V_{t_{v-1},h+1}) + (1 - \psi_{v,v,h}(s,a)) G_{v-1,h}(s,a) & \text{if } v > 1 \\ P_{h,s,a}(V_{h+1}^{\star} - V_{0,h+1}) & \text{if } v = 1 \end{cases}, \tag{D.80a}$$

$$0 \le \widetilde{G}_{v,h}^{-j,m}(s,a),\ G_{v,h}(s,a) \le H, \tag{D.80b}$$

$$|\widetilde{G}_{v,h}^{-j,m}(s,a) - G_{v,h}(s,a)| \le \min\left\{H, \frac{2H^2}{N_{t_v,h}(s,a)}\right\}. \tag{D.80c}$$

Now, we are ready to prove (D.70). Towards this, we first observe that moving the first term in the right-hand side of (D.70) to the left-hand side, and multiplying by a factor of

M, yields

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

$$\overset{\text{(i)}}{=}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{v=1}^{\phi(K)}\left(\sum_{m=1}^{M}\sum_{t_{v-1}<j\le t_v} d_h^{\mathsf{avg}}(s,a)-\sum_{m=1}^{M}n_{t_v,h}^m(s,a)\right)\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)}G_{v,h}(s,a)$$

$$\overset{\text{(ii)}}{=}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{m=1}^{M}\left(\sum_{j=1}^{K}d_h^{\mathsf{avg}}(s,a)-\sum_{j=1}^{K}\mathbf{1}\{(s,a)=(s_{j,h}^m,a_{j,h}^m)\}\right)\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)}G_{v,h}(s,a)$$

$$=\sum_{j=1}^{K}\sum_{m=1}^{M}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(d_h^{\mathsf{avg}}(s,a)-\mathbf{1}\{(s,a)=(s_{j,h}^m,a_{j,h}^m)\}\right)\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)}G_{v,h}(s,a)=\sum_{j=1}^{K}\sum_{m=1}^{M}Y_{j,h}^m \tag{D.81}$$

where (i) holds by plugging in (D.75) and $n_{t_v,h}(s,a)=\sum_{m=1}^{M}n_{t_v,h}^m(s,a)$, (ii) follows from $\sum_{v=1}^{\phi(K)}\sum_{t_{v-1}<j\le t_v}1=K$ and $\sum_{v=1}^{\phi(K)}n_{t_v,h}^m(s,a)=\sum_{j=1}^{K}\mathbf{1}\{(s,a)=(s_{j,h}^m,a_{j,h}^m)\}$, and the last equality arise from the definition of $Y_{j,h}^m$ in (D.4.5).

Therefore, the above fact shows that to prove (D.70), it is suffices to show:

$$\left|\sum_{j=1}^{K}\sum_{m=1}^{M}Y_{j,h}^m\right|\le\left|\sum_{j=1}^{K}\sum_{m=1}^{M}\widetilde{Y}_{j,h}^m\right|+\left|\sum_{j=1}^{K}\sum_{m=1}^{M}\left(Y_{j,h}^m-\widetilde{Y}_{j,h}^m\right)\right|\lesssim M\sigma_{\mathsf{aux},1}. \tag{D.82}$$

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 (D.78)) is independent of agent m's visits on (s,a) at the 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 (D.79)). It follows that $\mathbb{E}_{j-1}[\widetilde{Y}_{j,h}^m]=0$, where we denote

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

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

following holds:

$$\left|\sum_{j=1}^{K}\sum_{m=1}^{M}\widetilde{Y}_{j,h}^{m}\right| \le \sqrt{8W\log\frac{2H}{\delta}} + \frac{8}{3}B\log\frac{2H}{\delta}$$
$$\lesssim \sqrt{H^2MKSC_{\mathsf{avg}}^{\star}} + HSC_{\mathsf{avg}}^{\star} \tag{D.83}$$

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

$$\left|\widetilde{Y}_{j,h}^{m}\right| \le 2C_{\mathsf{avg}}^{\star}(1+d_h^{\pi^\star}(s,\pi^\star(s))S)\max_{s\in\mathcal{S}}\widetilde{G}_{\phi(j),h}^{-j,m}(s,\pi^\star(s)) \le 4SC_{\mathsf{avg}}^{\star}H =: B \tag{D.84}$$

$$\begin{aligned}
\sum_{j=1}^{K}\sum_{m=1}^{M}\mathbb{E}_{j-1}\left[\left(\widetilde{Y}_{j,h}^{m}\right)^2\right] &\le \sum_{j=1}^{K}\sum_{m=1}^{M}\mathbb{E}_{(s_{j,h}^m,a_{j,h}^m)\sim d_h^m}\left[\left(\frac{d_h^{\pi^\star}(s_{j,h}^m,a_{j,h}^m)}{d_h^{\mathsf{avg}}(s_{j,h}^m,a_{j,h}^m)}\widetilde{G}_{\phi(j),h}^{-j,m}(s_{j,h}^m,a_{j,h}^m)\right)^2\right] \\
&\le \sum_{j=1}^{K}\sum_{m=1}^{M}\sum_{s\in\mathcal{S}} d_h^m(s,\pi^\star(s))\left(\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{d_h^{\mathsf{avg}}(s,\pi^\star(s))}\widetilde{G}_{\phi(j),h}^{-j,m}(s,\pi^\star(s))\right)^2 \\
&\le H^2C_{\mathsf{avg}}^{\star}\sum_{j=1}^{K}\sum_{s\in\mathcal{S}}\sum_{m=1}^{M} d_h^m(s,\pi^\star(s))\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{d_h^{\mathsf{avg}}(s,\pi^\star(s))}(1+d_h^{\pi^\star}(s,\pi^\star(s))S) \\
&\le H^2C_{\mathsf{avg}}^{\star}\sum_{j=1}^{K}\sum_{s\in\mathcal{S}} Md_h^{\pi^\star}(s,\pi^\star(s))(1+d_h^{\pi^\star}(s,\pi^\star(s))S) \\
&\le 2H^2SC_{\mathsf{avg}}^{\star}MK =: W
\end{aligned} \tag{D.85}$$

using the fact that $|\widetilde{G}_{\phi(j),h}^{-j,m}(s_{j,h}^m,a_{j,h}^m)| \le H$ shown in (D.80b) 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:

$$\left|\sum_{j=1}^{K}\sum_{m=1}^{M}\left(\widetilde{Y}_{j,h}^{m}-Y_{j,h}^{m}\right)\right|$$

$$
\begin{aligned}
&= \Big| \sum_{v=1}^{\phi(K)} \sum_{m=1}^{M} \sum_{t_{v-1}<j\le t_v} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^m(s,a) - \mathbf{1}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\}\right) \\
&\qquad\qquad \times \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} (\widetilde{G}_{v,h}^{-j,m}(s,a) - G_{v,h}(s,a)) \Big| \\
&\overset{\text{(i)}}{=} \sum_{v=1}^{\phi(K)} \sum_{m=1}^{M} \sum_{t_{v-1}<j\le t_v} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbf{1}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\} \\
&\qquad\qquad \times (1-d_h^m(s,a)) \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} \left|\widetilde{G}_{v,h}^{-j,m}(s,a) - G_{v,h}(s,a)\right| \\
&\overset{\text{(ii)}}{\le} \sum_{v=1}^{\phi(K)} \sum_{m=1}^{M} \sum_{t_{v-1}<j\le t_v} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbf{1}\{(s,a)=(s_{j,h}^m, a_{j,h}^m)\} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mathsf{avg}}(s,a)} \min\left\{\frac{2H^2}{N_{t_v,h}(s,a)}, H\right\} \\
&\overset{\text{(iii)}}{\le} C_{\mathsf{avg}}^\star \sum_{s\in\mathcal{S}} \sum_{v=1}^{\phi(K)} n_{t_v,h}(s,\pi^\star(s)) \frac{d_h^{\pi^\star}(s,\pi^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi^\star(s)), 1/S\}} \min\left\{\frac{2H^2}{N_{t_v,h}(s,\pi^\star(s))}, H\right\} \\
&\overset{\text{(iv)}}{\le} 2H^2 C_{\mathsf{avg}}^\star \sum_{s\in\mathcal{S}} (1+d_h^{\pi^\star}(s,\pi^\star(s))S) \sum_{v=1}^{\phi(K)} \min\left\{\frac{n_{t_v,h}(s,\pi^\star(s))}{N_{t_v,h}(s,\pi^\star(s))}, n_{t_v,h}(s,\pi^\star(s))\right\} \\
&\overset{\text{(v)}}{\lesssim} C_{\mathsf{avg}}^\star H^2 S
\end{aligned} \tag{D.86}
$$

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 (D.80a), (ii) follows from (D.80c), (iii) naturally holds according to the definition of $C_{\mathsf{avg}}^\star$, (iv) holds because $\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{\min\{d_h^{\pi^\star}(s,\pi^\star(s)),1/S\}} \le 1 + d_h^{\pi^\star}(s,\pi^\star(s))S$, and (v) holds because for any $z \in [\phi(K)]$,

$$
\sum_{v=1}^{z} \frac{n_{t_v,h}(s,\pi^\star(s))}{N_{t_v,h}(s,\pi^\star(s))} \le 1 + \log\left(N_{t_z,h}(s,\pi^\star(s))\right), \tag{D.87}
$$

according to Lemma 24.

Now, combining the bounds obtained above (cf. (D.83) and (D.86)) into (D.82), we conclude that

$$
\left|\sum_{j=1}^{K}\sum_{m=1}^{M} Y_{j,h}^m\right| \lesssim \sqrt{H^2 MKSC_{\mathsf{avg}}^\star} + H^2 SC_{\mathsf{avg}}^\star = M\left(\sqrt{\frac{H^2 KSC_{\mathsf{avg}}^\star}{M}} + \frac{H^2 SC_{\mathsf{avg}}^\star}{M}\right) \tag{D.88}
$$

which completes the proof.

Proof of (D.80)

Proof of (D.80a)**.** We will proof (D.80a) by considering different cases separately. When $v = 1$, we have

$$G_{v,h}(s,a) = P_{h,s,a}(V^\star_{h+1} - V_{t_{v-1},h+1})\psi_{1,1,h}(s,a)$$
$$= P_{h,s,a}(V^\star_{h+1} - V_{0,h+1}) \sum_{m=1}^{M} \left(\sum_{i \in l^m_{t_1,h}(s,a)} \omega^m_{i,t_1,h}(s,a) \right) = P_{h,s,a}(V^\star_{h+1} - V_{0,h+1}) \tag{D.89}$$

where the second equality follows from the definition of $\psi_{u,v,h}(s,a)$ in (D.69), and the last equality holds since

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

When $v > 1$, invoking the definition of $\omega^m_{i,t_v,h}$ in (D.4c) yields that for any $u < v$,

$$\begin{aligned}
\psi_{u,v,h}(s,a) &= \sum_{m=1}^{M} \sum_{i \in l^m_{t_u,h}(s,a)} \omega^m_{i,t_v,h}(s,a) \\
&= \frac{(H+1)n_{t_u,h}}{N_{t_v,h} + Hn_{t_v,h}} \left(\prod_{x=u}^{v-1} \frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}} \right) \\
&= \frac{(H+1)n_{t_u,h}}{N_{t_{v-1},h} + Hn_{t_{v-1},h}} \left(\prod_{x=u}^{v-2} \frac{N_{t_x,h}}{N_{t_x,h} + Hn_{t_x,h}} \right) \frac{N_{t_{v-1},h}}{N_{t_v,h} + Hn_{t_v,h}} \\
&= \psi_{u,v-1,h}(s,a)(1 - \psi_{v,v,h}(s,a)).
\end{aligned} \tag{D.90}$$

where the second equality holds by $\phi(i) = u$ for all $i \in l^m_{t_u,h}(s,a)$ and the fact $\sum_{m=1}^{M} \sum_{i \in l^m_{t_u,h}(s,a)} 1 = n_{t_u,h}$, 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 the above fact back into (D.75) complete the proof by showing

that

$$
\begin{aligned}
&G_{v,h}(s,a)\\
&=\sum_{u=1}^{v} P_{h,s,a}(V^{\star}_{h+1}-V_{t_{u-1},h+1})\psi_{u,v,h}(s,a)\\
&=P_{h,s,a}(V^{\star}_{h+1}-V_{t_{v-1},h+1})\psi_{v,v,h}(s,a)+\sum_{u=1}^{v-1} P_{h,s,a}(V^{\star}_{h+1}-V_{t_{u-1},h+1})\psi_{u,v,h}(s,a)\\
&=P_{h,s,a}(V^{\star}_{h+1}-V_{t_{v-1},h+1})\psi_{v,v,h}(s,a)+(1-\psi_{v,v,h}(s,a))\sum_{u=1}^{v-1} P_{h,s,a}(V^{\star}_{h+1}-V_{t_{u-1},h+1})\psi_{u,v-1,h}(s,a)\\
&=P_{h,s,a}(V^{\star}_{h+1}-V_{t_{v-1},h+1})\psi_{v,v,h}(s,a)+(1-\psi_{v,v,h}(s,a))G_{v-1,h}(s,a). \qquad \text{(D.91)}
\end{aligned}
$$

Proof of (D.80b)**.** First, applying (D.8c) in Lemma 21 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

$$
\begin{aligned}
G_{v,h}(s,a)&=\sum_{u=1}^{v} P_{h,s,a}(V^{\star}_{h+1}-V_{t_{u-1},h+1})\psi_{u,v,h}(s,a)\\
&\leq P_{h,s,a}(V^{\star}_{h+1}-V_{0,h+1})\sum_{u=1}^{v}\psi_{u,v,h}(s,a)\\
&\leq H\sum_{u=1}^{v}\psi_{u,v,h}(s,a)\\
&=H\sum_{u=1}^{v}\sum_{m=1}^{M}\left(\sum_{i\in l^{m}_{t_u,h}(s,a)}\omega^{m}_{i,t_v,h}(s,a)\right)\leq H, \qquad \text{(D.92)}
\end{aligned}
$$

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 (5.10) 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 (D.69), and the last inequality follows from (D.26b) in Lemma 25.

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

Proof of (D.80c)**.** Consider $v = \phi(j)$. If $v = 1$, combing (D.80a) and (D.78) directly gives $\widetilde{G}_{v,h}^{-j,m}(s,a) = G_{v,h}(s,a)$. Then we turn to the case when $v > 1$ and bound the term of interest in two different cases, respectively.

- When $(s_{j,h}^m, a_{j,h}^m) \neq (s,a)$. In this case, invoking the definition in (D.79) gives

$$\widetilde{\psi}_{v,v,h}^{-j,m}(s,a) = \frac{(H+1)n_{t_v,h}(s,a)}{N_{t_{v-1},h}(s,a) + (H+1)n_{t_v,h}(s,a)} = \psi_{v,v,h}^{-j,m}(s,a), \tag{D.93}$$

  which indicates (see the definition in (D.78))

$$\widetilde{G}_{v,h}^{-j,m}(s,a) = G_{v,h}(s,a) \tag{D.94}$$

- When $(s_{j,h}^m, a_{j,h}^m) = (s,a)$. In view of (D.80a) and (D.78), it holds that:

$$\begin{aligned}
&|\widetilde{G}_{v,h}^{-j,m}(s,a) - G_{v,h}(s,a)| \\
&= \left|(\widetilde{\psi}_{v,v,h}^{-j,m}(s,a) - \psi_{v,v,h}(s,a))P_{h,s,a}(V_{h+1}^\star - V_{t_{v-1},h+1}) + (\psi_{v,v,h}(s,a) - \widetilde{\psi}_{v,v,h}^{-j,m}(s,a))G_{v-1,h}(s,a)\right| \\
&= \left|(\psi_{v,v,h}(s,a) - \widetilde{\psi}_{v,v,h}^{-j,m}(s,a))(G_{v-1,h}(s,a) - P_{h,s,a}(V_{h+1}^\star - V_{t_{v-1},h+1})\right| \\
&\leq \left|\psi_{v,v,h}(s,a) - \widetilde{\psi}_{v,v,h}^{-j,m}(s,a)\right| \max\left\{G_{v-1,h}(s,a), \|P_{h,s,a}\|_1 \left\|V_{h+1}^\star - V_{t_{v-1},h+1}\right\|_\infty\right\} \\
&\overset{\text{(i)}}{\leq} H\left|\psi_{v,v,h}(s,a) - \widetilde{\psi}_{v,v,h}^{-j,m}(s,a)\right| \\
&\overset{\text{(ii)}}{\leq} \min\left\{H, \frac{2H^2}{N_{t_v,h}(s,a)}\right\},
\end{aligned} \tag{D.95}$$

  where (i) holds by (D.80b), $\|P_{h,s,a}\|_1 = 1$, and $\left\|V_{h+1}^\star - V_{t_{v-1},h+1}\right\|_\infty \leq H$. Here, (ii) can be verified by

$$\begin{aligned}
0 &\overset{\text{(iii)}}{\leq} \psi_{v,v,h}(s,a) - \widetilde{\psi}_{v,v,h}^{-j,m}(s,a) \\
&= \frac{(H+1)n_{t_v,h}(s,a)}{N_{t_{v-1},h}(s,a) + (H+1)n_{t_v,h}(s,a)} \\
&\qquad - \frac{(H+1)(n_{t_v,h}(s,a) - \mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\})}{N_{t_{v-1},h}(s,a) + (H+1)(n_{t_v,h}(s,a) - \mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\})} \\
&= \frac{(H+1)n_{t_v,h}(s,a)}{N_{t_{v-1},h}(s,a) + (H+1)n_{t_v,h}(s,a)} - \frac{(H+1)(n_{t_v,h}(s,a) - 1)}{N_{t_{v-1},h}(s,a) + (H+1)(n_{t_v,h}(s,a) - 1)}
\end{aligned}$$

$$\leq \frac{(H+1)}{N_{t_{v-1},h}(s,a) + (H+1)n_{t_v,h}(s,a)}$$
$$\leq \min\left\{1, \frac{2H}{N_{t_v,h}(s,a)}\right\}. \tag{D.96}$$

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

Proof of Lemma 28

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

$$Z_{j,h}^m \coloneqq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(\mathbf{1}\{(s,a) = (s_{j,h}^m, a_{j,h}^m)\} - d_h^m(s,a)\right) \frac{d_h^{\pi^\star}(s,a)}{M d_h^{\mathsf{avg}}(s,a)} P_{h,s,a}(V_{h+1}^\star - V_{t_{\phi(j)-1},h+1}). \tag{D.97}$$

Then, to prove Lemma 28, it suffices to show $\left|\sum_{j=1}^K \sum_{m=1}^M Z_{j,h}^m\right| \lesssim \sigma_{\mathsf{aux},2}$.

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

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

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

$$\left|\sum_{j=1}^K \sum_{m=1}^M Z_{j,h}^m\right| \leq \sqrt{8W \log\frac{2H}{\delta}} + \frac{8}{3} B \log\frac{2H}{\delta} \lesssim \sqrt{\frac{H^2 K S C_{\mathsf{avg}}^\star}{M}} + \frac{H S C_{\mathsf{avg}}^\star}{M} \tag{D.98}$$

using the following properties:

$$|Z_{j,h}^m| \leq \frac{2C_{\mathsf{avg}}^\star H}{M}\left(\sum_{s\in\mathcal{S}}(1 + d_h^{\pi^\star}(s,\pi^\star(s))S)\right) \leq \frac{4HSC_{\mathsf{avg}}^\star}{M} =: B \tag{D.99}$$

$$\sum_{j=1}^K \sum_{m=1}^M \mathbb{E}_{j-1}[(Z_{j,h}^m)^2] \leq \sum_{j=1}^K \sum_{m=1}^M \mathbb{E}_{(s_{j,h}^m, a_{j,h}^m)\sim d_h^m}\left[\left(\frac{d_h^{\pi^\star}(s_{j,h}^m, a_{j,h}^m)}{M d_h^{\mathsf{avg}}(s_{j,h}^m, a_{j,h}^m)} P_{h,s,a}(V_{h+1}^\star - V_{t_{\phi(j)-1},h+1})\right)^2\right]$$
$$\leq H^2 \sum_{j=1}^K \sum_{m=1}^M \sum_{s\in\mathcal{S}} d_h^m(s,\pi^\star(s)) \left(\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{M d_h^{\mathsf{avg}}(s,\pi^\star(s))}\right)^2$$

$$
\begin{aligned}
&\leq \frac{H^2 C^\star_{\mathsf{avg}}}{M} \sum_{s\in\mathcal{S}} \sum_{j=1}^{K} \left(\frac{d_h^{\pi^\star}(s,\pi^\star(s))}{M d_h^{\mathsf{avg}}(s,\pi^\star(s))} \right) (1 + d_h^{\pi^\star}(s,\pi^\star(s))S) \sum_{m=1}^{M} d_h^m(s,\pi^\star(s)) \\
&= \frac{H^2 C^\star_{\mathsf{avg}}}{M} \sum_{s\in\mathcal{S}} \sum_{j=1}^{K} d_h^{\pi^\star}(s,\pi^\star(s))(1 + d_h^{\pi^\star}(s,\pi^\star(s))S) \\
&= \frac{2H^2 K S C^\star_{\mathsf{avg}}}{M} =: W,
\end{aligned}
\tag{D.100}
$$

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

D.4.6 Proof of Corollary 1

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

$$
MT \gtrsim H^5 S C^\star_{\mathsf{avg}} \quad \text{and} \quad H \leq \sqrt{\frac{HSC^\star_{\mathsf{avg}} T}{M}}, \tag{D.101}
$$

as long as $\varepsilon \leq H$ and $\varepsilon \leq \frac{H^3 S C^\star_{\mathsf{avg}}}{M}$. 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

$$
V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{H^4 S C^\star_{\mathsf{avg}}}{MT} + \sqrt{\frac{H^7 S C^\star_{\mathsf{avg}}}{MT}} + \frac{H^3}{T}\sqrt{\frac{HSC^\star_{\mathsf{avg}} T}{M}} \lesssim \sqrt{\frac{H^7 S C^\star_{\mathsf{avg}}}{MT}}. \tag{D.102}
$$

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

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

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

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{H^4 SC^\star_{\mathsf{avg}}}{MT} + \sqrt{\frac{H^7 SC^\star_{\mathsf{avg}}}{MT}} + \frac{H^3}{T}\sqrt{\frac{HSC^\star_{\mathsf{avg}}T}{M}} \lesssim \sqrt{\frac{H^7 SC^\star_{\mathsf{avg}}}{MT}}. \tag{D.103}$$

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

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

since

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

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

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

we obtain

$$\phi(K) = \left\lceil \frac{\log\left(\frac{K}{H^2} + 1\right)}{\log\left(1 + \frac{1}{H}\right)} \right\rceil \leq 1 + (1 + H)\log\Big(\frac{K}{H^2} + 1\Big) \lesssim H \tag{D.104}$$

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

Appendix E

Analysis of Federated Value Iteration with Heterogeneous Rewards

E.1 Notations

Throughout the analysis, $X_{k,h}$ for $X \in \{N, \widehat{P}, B^m, Q^m, V^m\}$ denotes the value of X at the *beginning of episode* k (i.e., before episode k's execution step). All these values are computed using data from episodes $1, \ldots, k-1$. In particular:

- $X_{1,h}$ refers to the initialization (no data collected yet);
- $N_{k,h}(s,a) = \sum_{m=1}^{M} \sum_{i=1}^{k-1} \mathbb{I}\{(s_{i,h}^m, a_{i,h}^m) = (s,a)\}$;
- $\widehat{P}_{k,h}(s'|s,a) = N_{k,h}(s,a,s') / \max\{1, N_{k,h}(s,a)\}$.

The policy executed in episode k is $\pi_{k,h}^m(s) = \arg\max_a Q_{k,h}^m(s,a)$.

E.2 Proof of Theorem 8

For each agent m at episode k, we can decompose the regret as follows:

$$\begin{aligned}
&\mathbb{E}_{s_1 \sim \rho}[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)] \\
&\leq \mathbb{E}_{s_1 \sim \rho}[V_{k,1}^m(s_1) - V_1^{m,\pi_k^m}(s_1)]
\end{aligned}$$

$$= \mathbb{E}_{s_1 \sim \rho, \pi_k^m}[\sum_{h=1}^{H} \Big(Q_{k,h}^m(s_h, \pi_{k,h}^m(s_h)) - Q_h^{m,\pi_k^m}(s_h, \pi_{k,h}^m(s_h)) \Big)]. \tag{E.1}$$

Here, the first inequality holds since $Q_{k,h}^m(s,a) \geq Q_h^{m,\star}(s,a)$ for all (s,a,h) by the optimism property of the algorithm, which we will prove in the following Step 1.

Step 1: Optimism. We want to show that $V_{k,h}^m(s) - V_h^{m,\star}(s) \geq 0$ for all k, h, m, s. We can prove this by induction on h. For $h = H+1$, $V_{k,H+1}^m(s) = V_{H+1}^{m,\star}(s) = 0$ for all m, k, s. We can then use backward induction to show that the optimism holds for all $h \in [H]$. For any (s,h) at episode k of agent m and $a = \pi_h^{m,\star}(s)$, we can decompose the estimation error as follows:

$$\begin{aligned}
V_{k,h}^m(s) - V_h^{m,\star}(s) &\geq Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a) \\
&= \Big(r_h^m(s,a) + \widehat{P}_{k,h,s,a} V_{k,h+1}^m + B_{k,h}^m(s,a) \Big) - \Big(r_h^m(s,a) + P_{h,s,a} V_{h+1}^{m,\star} \Big) \\
&= \big(\widehat{P}_{k,h,s,a} - P_{h,s,a} \big) V_{k,h+1}^m + P_{h,s,a} \big(V_{k,h+1}^m - V_{h+1}^{m,\star} \big) + B_{k,h}^m(s,a) \\
&\geq -B_{k,h}^m(s,a) - \frac{2}{H} P_{h,s,a} \Delta_{k,h+1}^m + P_{h,s,a} \Delta_{k,h+1}^m + B_{k,h}^m(s,a) \\
&= \left(1 - \frac{2}{H} \right) P_{h,s,a} \Delta_{k,h+1}^m \geq 0.
\end{aligned} \tag{E.2}$$

where the first inequality holds by the transition estimation error bound in Lemma 29 and the induction hypothesis $\Delta_{k,h+1}^m(s) \geq 0$ for all s.

Lemma 29. *(Transition Estimation Error Bound) Let $\Delta_{k,h}^m(s) = V_{k,h}^m(s) - V_h^{m,\star}(s)$ For any episode k, step h, state-action pair (s,a), and agent m, with probability at least $1-\delta$, we have*

$$-B_{k,h}^m(s,a) - \frac{2}{H} P_{h,s,a} \Delta_{k,h+1}^m \leq \big(\widehat{P}_{k,h,s,a} - P_{h,s,a} \big) V_{k,h+1}^m \leq B_{k,h}^m(s,a) + \frac{2}{H} P_{h,s,a} \Delta_{k,h+1}^m \tag{E.3}$$

Proof. By applying the empirical Bernstein inequality (Maurer and Pontil, 2009), we can

show the upper bound as follows:

$$
\begin{aligned}
&\big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)V^m_{k,h+1} \\
&= \big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)V^{m,\star}_{h+1} + \big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)(V^m_{k,h+1} - V^{m,\star}_{h+1}) \\
&\le \sqrt{\frac{2\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V^{m,\star}_{h+1})\log(4MSAKH/\delta)}{N_{k,h}(s,a)}} + \frac{7H\log(4MSAKH/\delta)}{3N_{k,h}(s,a)} \\
&\qquad + |(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta^m_{k,h+1}| \\
&\le \sqrt{\frac{2\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V^{m,\star}_{h+1})\log(4MSAKH/\delta)}{N_{k,h}(s,a)}} + \frac{11SH^2\log(4MSAKH/\delta)}{N_{k,h}(s,a)} \\
&\qquad + \frac{1}{H}P_{h,s,a}\Delta^m_{k,h+1} \\
&\le B^m_{k,h}(s,a) + \frac{2}{H}P_{h,s,a}\Delta^m_{k,h+1}
\end{aligned} \tag{E.4}
$$

Here, the last inequality transitions from the optimal variance $\mathsf{Var}_{\widehat{P}}(V^\star)$ to the empirical variance $\mathsf{Var}_{\widehat{P}}(V_k)$ used in $B^m_{k,h}$. By standard deviation properties, $\sqrt{\mathsf{Var}_{\widehat{P}}(V^\star)} \le \sqrt{\mathsf{Var}_{\widehat{P}}(V_k)} + \sqrt{H\widehat{P}\Delta_k}$. Since $\widehat{P}\Delta_k \le 2P\Delta_k + \mathcal{O}(1/N)$ via Bernstein, applying Young's inequality bounds the difference by $\frac{1}{H}P\Delta_k + \mathcal{O}(1/N)$, yielding the additional $\frac{1}{H}P\Delta_k$ term.

Similarly, we can also show the lower bound as follows:

$$
\big(\widehat{P}_{k,h,s,a} - P_{h,s,a}\big)V^m_{k,h+1} \ge -B^m_{k,h}(s,a) - \frac{2}{H}P_{h,s,a}\Delta^m_{k,h+1}. \tag{E.5}
$$

Bounding $|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta^m_{k,h+1}|$: For any fixed Δ_ϵ that approximates $\Delta^m_{k,h+1}$ by the error of ϵ, by applying the Bernstein inequality, we can show that

$$
|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_\epsilon| \le \sqrt{\frac{2\mathsf{Var}_{P_{h,s,a}}(\Delta_\epsilon)\ln(2/\delta')}{N_{k,h}(s,a)}} + \frac{2H\log(2/\delta')}{3N_{k,h}(s,a)}. \tag{E.6}
$$

To cover all possible $\Delta^m_{k,h+1}$ with an error of ϵ, we need to consider all possible Δ_ϵ, splitting $[0,H]$ into $\lceil H/\epsilon \rceil$ intervals for all states, incurring covering number $\mathcal{N}(\epsilon) \le \big(1+\frac{H}{\epsilon}\big)^S$. By taking a union bound over $\mathcal{N}(\epsilon)$, (h,s,a,m), and possible visitation counter $N_{k,h}(s,a) =$

$N \in [MK]$, we can replace δ' with $\delta/(MSAKH\mathcal{N}(\epsilon))$ in the above inequality, yielding

$$
\begin{aligned}
&|(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta^m_{k,h+1}| \\
&= |(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_\epsilon| + |(\widehat{P}_{k,h,s,a} - P_{h,s,a})(\Delta^m_{k,h+1} - \Delta_\epsilon)| \\
&\le |(\widehat{P}_{k,h,s,a} - P_{h,s,a})\Delta_\epsilon| + 2\epsilon \\
&\le \sqrt{\frac{2\mathsf{Var}_{P_{h,s,a}}(\Delta_\epsilon)\log(2MSAKH\mathcal{N}(\epsilon)/\delta)}{N_{k,h}(s,a)}} + \frac{2H\log(2MSAKH\mathcal{N}(\epsilon)/\delta)}{3N_{k,h}(s,a)} + 2\epsilon \\
&\overset{(i)}{\le} \sqrt{\frac{2H(P\Delta^m_{k,h+1} + \epsilon)\log(2MSAKH\mathcal{N}(\epsilon)/\delta)}{N_{k,h}(s,a)}} + \frac{2H\log(2MSAKH\mathcal{N}(\epsilon)/\delta)}{3N_{k,h}(s,a)} + 2\epsilon \\
&\overset{(ii)}{\le} \sqrt{\frac{6SHP\Delta^m_{k,h+1}\log(2MSAKH/\delta)}{N_{k,h}(s,a)}} + \frac{6HS\log(2MSAKH/\delta)}{N_{k,h}(s,a)} \\
&\overset{(iii)}{\le} \frac{1}{H}P_{h,s,a}\Delta^m_{k,h+1} + \frac{8SH^2\log(2MSAKH/\delta)}{N_{k,h}(s,a)}
\end{aligned} \tag{E.7}
$$

where (i) holds because

$$
\mathsf{Var}_{P_{h,s,a}}(\Delta_\epsilon) = P_{h,s,a}\Delta^2_\epsilon - (P_{h,s,a}\Delta_\epsilon)^2 \le P_{h,s,a}\Delta^2_\epsilon \le HP_{h,s,a}\Delta_\epsilon \le HP_{h,s,a}\Delta^m_{k,h+1} + H\epsilon, \tag{E.8}
$$

(ii) holds when choosing very small $\epsilon = 1/(HMK)^2$, and (iii) follows by applying the Young's inequality that $xy \le cx^2 + \frac{y^2}{4c}$ for any $x, y > 0$ and $c > 0$, which yields

$$
\sqrt{\frac{6SHP_{h,s,a}\Delta^m_{k,h+1}\log(2MSAKH/\delta)}{N_{k,h}(s,a)}} \le \frac{1}{H}P_{h,s,a}\Delta^m_{k,h+1} + \frac{2SH^2\log(2MSAKH/\delta)}{N_{k,h}(s,a)} \tag{E.9}
$$

when $c = 1/H$, $x = \sqrt{P\Delta^m_{k,h+1}}$, $y = \sqrt{\frac{6SH\log(2MSAKH/\delta)}{N_{k,h}(s,a)}}$.

□

Step 2: Solving the Recursion. Now, we can obtain the regret bound by bounding the Q-estimation error as follows:

$$\mathbb{E}_{s_1\sim\rho}[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)] \leq \sum_{h=1}^{H} \mathbb{E}_{s_1\sim\rho,\pi_k^m}[\Big(Q_{k,h}^m(s_h,\pi_k^m(s_h)) - Q_h^{m,\pi_k^m}(s_h,\pi_k^m(s_h))\Big)] \tag{E.10}$$

Denoting $\Delta_{k,h+1}^{m,\pi_k^m} = V_{k,h+1}^m - V_{h+1}^{m,\pi_k^m}$ and $\Delta_{k,h+1}^m = V_{k,h+1}^m - V_{h+1}^{m,\star}$, for any step h and state-action pair (s_h, a_h) induced by ρ and π_k^m, we can decompose the estimation error as follows:

$$\begin{aligned}
&Q_{k,h}^m(s_h,a_h) - Q_h^{m,\pi_k^m}(s_h,a_h) \\
&= \Big(r_h^m(s_h,a_h) + \widehat{P}_{k,h,s_h,a_h} V_{k,h+1}^m + B_{k,h}^m(s_h,a_h)\Big) - \Big(r_h^m(s_h,a_h) + P_{h,s_h,a_h} V_{h+1}^{m,\pi_k^m}\Big) \\
&= B_{k,h}^m(s_h,a_h) + \big(\widehat{P}_{k,h,s_h,a_h} - P_{h,s_h,a_h}\big)V_{k,h+1}^m + P_{h,s_h,a_h}\Delta_{k,h+1}^{m,\pi_k^m} \\
&\overset{(i)}{\leq} 2B_{k,h}^m(s_h,a_h) + \frac{2}{H}P_{h,s_h,a_h}\Delta_{k,h+1}^m + P_{h,s_h,a_h}\Delta_{k,h+1}^{m,\pi_k^m} \\
&\overset{(ii)}{\leq} 2B_{k,h}^m(s_h,a_h) + \left(1+\frac{2}{H}\right)P_{h,s_h,a_h}\Delta_{k,h+1}^{m,\pi_k^m}
\end{aligned} \tag{E.11}$$

where (i) holds due to Lemma 29, which yields:

$$\big(\widehat{P}_{k,h,s_h,a_h} - P_{h,s_h,a_h}\big)V_{k,h+1}^m \leq B_{k,h}^m(s_h,a_h) + \frac{2}{H}P_{h,s_h,a_h}\Delta_{k,h+1}^m, \tag{E.12}$$

and (ii) holds because $V_{h+1}^{m,\star} \geq V_{h+1}^{m,\pi_k^m}$ and $\Delta_{k,h+1}^m \leq \Delta_{k,h+1}^{m,\pi_k^m}$.

Taking the expectation over the trajectory induced by π_k^m, the transition operator P_{h,s_h,a_h} naturally evaluates the expected value of the next state $s_{h+1} \sim P(\cdot|s_h,a_h)$:

$$\begin{aligned}
&\mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[Q_{k,h}^m(s_h,a_h) - Q_h^{m,\pi_k^m}(s_h,a_h)\right] \\
&= \mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[\Delta_{k,h}^{m,\pi_k^m}(s_h)\right] \\
&\leq \mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[2B_{k,h}^m(s_h,a_h) + \left(1+\frac{2}{H}\right)\Delta_{k,h+1}^{m,\pi_k^m}(s_{h+1})\right].
\end{aligned} \tag{E.13}$$

Unrolling this recursion from step $h = 1$ to H, and using the analytical fact that $(1 + 2/H)^{h-1} \leq (1 + 2/H)^H \leq e^2$, we can explicitly bound the expected regret for episode k:

$$\mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\right] \leq \mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[\sum_{h=1}^{H}\left(1+\frac{2}{H}\right)^{h-1} 2B_{k,h}^m(s_h,a_h)\right]$$
$$\leq 2e^2 \cdot \mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[\sum_{h=1}^{H} B_{k,h}^m(s_h,a_h)\right]. \tag{E.14}$$

Step 3: Bounding the Total Regret via Martingale Concentration. Summing over all agents $m \in [M]$ and all episodes $k \in [K]$, the total expected regret of the federated learning system is bounded by:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\right] \leq 2e^2 \cdot \sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{\pi_k^m}\left[\sum_{h=1}^{H} B_{k,h}^m(s_h,a_h)\right]. \tag{E.15}$$

Crucially, the exploration bonus $B_{k,h}^m$ is computed using the *global* visitation count $N_{k,h}(s,a)$, which aggregates the experiences from all M agents. To rigorously bound the expected sum of bonuses using the empirical observations collected along the actual trajectories $(s_{k,h}^m, a_{k,h}^m)$, we construct a formal Martingale Difference Sequence (MDS).

We define a global rollout index τ that strictly flattens the multi-agent episodes into a step-by-step sequential timeline:

$$\tau = (k-1)MH + (m-1)H + h \quad \text{for} \quad k \in [K],\ m \in [M],\ h \in [H], \quad \tau \in [1, MKH]. \tag{E.16}$$

Let $\mathcal{F}_\tau$ be the filtration (a σ-algebra) capturing all historical information available exactly strictly before the start of the τ-th step. Formally, $\mathcal{F}_\tau$ is defined as:

$$\mathcal{F}_\tau = \sigma\left(\bigcup_{i=1}^{\tau-1}\left\{s^{(i)}, a^{(i)}, r^{(i)}\right\}\right), \quad \mathcal{F}_1 = \{\emptyset, \Omega\}, \tag{E.17}$$

where the superscript (i) refers to the state-action pair at the specific global step index i. Since the global estimator $\widehat{P}$ and the counter N are updated strictly using past data

contained in the episode-start filtration $\mathcal{F}_{(k-1)MH+1} \subset \mathcal{F}_\tau$, the policy π_k^m and the bonus function $B_{k,h}^m$ evaluated during step τ are $\mathcal{F}_\tau$-measurable.

Let $W_{k,h}^m(s) = \mathbb{E}_{\pi_k^m}\left[\sum_{i=h}^H B_{k,i}^m(s_i, a_i) \mid s_h = s\right]$ be the expected cumulative bonus from step h. We define the sequence of random variables $\{D_\tau\}_{\tau=1}^{MKH}$ as the step-wise expected deviation of the future bonuses:

$$D_\tau = \mathbb{E}_{s' \sim P(\cdot|s_h, a_h)}\left[W_{k,h+1}^m(s') \mid \mathcal{F}_\tau\right] - W_{k,h+1}^m(s_{h+1}). \tag{E.18}$$

By the tower property of conditional expectation, it strictly holds that:

$$\mathbb{E}\big[D_\tau \mid \mathcal{F}_\tau\big] = \mathbb{E}_{s'}\left[W_{k,h+1}^m(s') \mid \mathcal{F}_\tau\right] - \mathbb{E}_{s'}\left[W_{k,h+1}^m(s') \mid \mathcal{F}_\tau\right] = 0. \tag{E.19}$$

Thus, $\{D_\tau\}_{\tau=1}^{MKH}$ forms a valid Martingale Difference Sequence adapted to the filtration $\{\mathcal{F}_\tau\}$. Because $W_{k,h}^m(s_h) = B_{k,h}^m(s_h, a_h) + \mathbb{E}[W_{k,h+1}^m \mid \mathcal{F}_\tau]$, summing D_τ over a single episode yields a telescoping series that exactly bridges the expected and empirical bonuses: $\sum_{h=1}^H D_\tau = W_{k,1}^m(s_1) - \sum_{h=1}^H B_{k,h}^m(s_h, a_h)$.

Furthermore, the expected bonus sums are naturally bounded in $[0, H]$. Due to the algorithmic truncation $Q_{k,h}^m \leq H$, the value updates inherently act as if they are computed with an effective exploration bonus bounded by H. Thus, without loss of generality, we can formally treat the cumulative future bonus as bounded: $W_{k,h+1}^m(s) \leq H$ for all (k, h, s, a). Consequently, the step-wise deviation D_τ is bounded by H almost surely:

$$W_{k,h+1}^m(\cdot) \in [0, H] \implies \big|D_\tau\big| \leq H \quad \text{almost surely.} \tag{E.20}$$

Applying the Azuma-Hoeffding inequality to the bounded MDS $\{D_\tau\}_{\tau=1}^{MKH}$, we guarantee that with probability at least $1 - \delta$:

$$\sum_{\tau=1}^{MKH} D_\tau \leq \sqrt{\frac{1}{2}\sum_{\tau=1}^{MKH}\big(2H\big)^2 \log(1/\delta)} = \sqrt{2H^3 MK \log(1/\delta)}. \tag{E.21}$$

Substituting the telescoping sum of D_τ back into this concentration bound, we tightly

relate the expected sum of bonuses to the empirical realization:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[\sum_{h=1}^{H}B_{k,h}^m(s_h,a_h)\right]\le\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}B_{k,h}^m(s_{k,h}^m,a_{k,h}^m)+\sqrt{2H^3MK\log(1/\delta)}. \tag{E.22}$$

Step 4: Variance Decomposition. Recall the definition of the exploration bonus $B_{k,h}^m(s,a)$, which utilizes the empirical variance of $V_{k,h+1}^m$:

$$B_{k,h}^m(s,a)=\sqrt{\frac{4\mathsf{Var}_{\widehat{P}_{k,h,s,a}}(V_{k,h+1}^m)\iota}{N_{k,h}(s,a)}}+\frac{cSH^2\iota}{N_{k,h}(s,a)}, \tag{E.23}$$

where $\iota=\log(4MSAKH\mathcal{N}(\epsilon)/\delta)$ bounds the logarithmic terms including the covering number, and c absorbs lower-order term constants.

To aggregate the total expected regret, we must translate the empirical variance $\mathsf{Var}_{\widehat{P}}$ into the true variance Var_P. Using Lemma 30, we can upper bound the empirical variance by the true variance as:

$$B_{k,h}^m(s,a)\le\sqrt{\frac{16\mathsf{Var}_{P_{h,s,a}}(V_{h+1}^{m,\star})\iota}{N_{k,h}(s,a)}}+\frac{1}{4e^2H}P_{h,s,a}\Delta_{k,h+1}^m+\frac{c'SH^2\iota}{N_{k,h}(s,a)}, \tag{E.24}$$

where c' is a large enough constant to absorb the additional lower-order term from the variance gap. Here, to transition to the optimal variance, we applied the gap $\mathsf{Var}_P(V_k)\le 2\mathsf{Var}_P(V^\star)+2HP\Delta_k$ and extracted the error linearly via Young's inequality $\left(\sqrt{xy}\le\alpha x+\frac{y}{4\alpha}\right)$ with $\alpha=\frac{1}{4e^2H}$.

Lemma 30. *(Empirical Variance Gap) For any (s,a,m,h) and $N_{k,h}(s,a)\le MK$ any fixed value function $V_{h+1}^{m,\star}\in[0,H]$, with probability at least $1-\delta$, the deviation between the empirical and true variance is bounded by:*

$$\left|\mathsf{Var}_{\widehat{P}_{s,a,k,h}}(V_{h+1}^{m,\star})-\mathsf{Var}_{P_{h,s,a}}(V_{h+1}^{m,\star})\right|\le\mathsf{Var}_{P_{h,s,a}}(V_{h+1}^{m,\star})+\frac{10H^2\iota}{N_{k,h}(s,a)}. \tag{E.25}$$

Proof. For notational simplicity, we omit the subscripts and superscripts in $\widehat{P}_{k,h,s,a}$ and

$P_{h,s,a}$, and simply denote them as $\widehat{P}$ and P respectively. We also denote $V = V_{h+1}^{m,\star}$ for brevity. First, we decompose the variance difference as follows:

$$\begin{aligned}
\left|\mathsf{Var}_{\widehat{P}}(V) - \mathsf{Var}_P(V)\right| &= \left|(\widehat{P}V^2 - (\widehat{P}V)^2) - (PV^2 - (PV)^2)\right| \\
&= \left|(\widehat{P}V^2 - PV^2) + ((PV)^2 - (\widehat{P}V)^2)\right| \\
&\le \left|(\widehat{P} - P)V^2\right| + \left|(P + \widehat{P})V \cdot (\widehat{P} - P)V\right| \\
&\le \left|(\widehat{P} - P)V^2\right| + 2H\left|(\widehat{P} - P)V\right| \qquad \text{(E.26)}
\end{aligned}$$

Then, applying the Bernstein inequality to both terms, we can show that

$$\begin{aligned}
\left|(\widehat{P} - P)V^2\right| &\le \sqrt{\frac{2\mathsf{Var}_P(V^2)\iota}{N}} + \frac{2H^2\iota}{3N} \quad (V^2 \in [0, H^2]) \\
&\le \sqrt{\frac{8H^2\mathsf{Var}_P(V)\iota}{N}} + \frac{2H^2\iota}{3N}, \qquad \text{(E.27)} \\
\left|(\widehat{P} - P)V\right| &\le \sqrt{\frac{2\mathsf{Var}_P(V)\iota}{N}} + \frac{2H\iota}{3N} \le \sqrt{\frac{2\mathsf{Var}_P(V)\iota}{N}} + \frac{2H\iota}{3N}, \qquad \text{(E.28)}
\end{aligned}$$

using the fact that

$$\begin{aligned}
\mathsf{Var}_P(V^2) &= \frac{1}{2}\mathbb{E}_{x,y}\left[(V(x)^2 - V(y)^2)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{x,y}\left[(V(x) - V(y))^2(V(x) + V(y))^2\right] \le 4H^2\mathsf{Var}_P(V). \qquad \text{(E.29)}
\end{aligned}$$

Combining the above bounds, we have:

$$\left|\mathsf{Var}_{\widehat{P}}(V) - \mathsf{Var}_P(V)\right| \le \sqrt{\frac{32H^2\mathsf{Var}_P(V)\iota}{N}} + \frac{2H^2\iota}{N}. \qquad \text{(E.30)}$$

Applying the Young's inequality,

$$\sqrt{\frac{32H^2\mathrm{Var}_P(V)\iota}{N}} = 2 \cdot \sqrt{\mathrm{Var}_P(V)} \cdot \sqrt{\frac{8H^2\iota}{N}} \le \mathrm{Var}_P(V) + \frac{8H^2\iota}{N} \qquad \text{(E.31)}$$

□

Step 5: Finalizing the Regret Bound. Now, we substitute the upper bound of the empirical exploration bonus into the total expected regret. The dominant variance term across all M agents, K episodes, and H steps is given by:

$$(\mathrm{I}) = \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{16\mathsf{Var}_{P_{h,s,a}}(V_{h+1}^{m,\star})\iota}{N_{k,h}(s_{k,h}^m, a_{k,h}^m)}}. \tag{E.32}$$

We apply the Cauchy-Schwarz inequality globally over all $M \times K \times H$ elements simultaneously:

$$(\mathrm{I}) \leq \sqrt{16\iota\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathsf{Var}_{P_{h,s,a}}(V_{h+1}^{m,\star})(s_{k,h}^m, a_{k,h}^m)} \times \sqrt{\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{N_{k,h}(s_{k,h}^m, a_{k,h}^m)}}. \tag{E.33}$$

We bound the two factors in Eq. (E.33) separately:

1. Bounding the Total Variance: For each agent m, the sum of true variances along a trajectory is bounded by the Law of Total Variance. Over K episodes, applying the standard variance lemma (e.g., Azar et al., 2017) yields:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathsf{Var}_{P_{h,s,a}}(V_{h+1}^{m,\star})(s_{k,h}^m, a_{k,h}^m) \leq H^2K + \mathcal{O}\left(H^{3/2}\sqrt{K\iota}\right) \leq 2H^2K. \tag{E.34}$$

Summing this strictly decoupled bound over all M agents gives:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathsf{Var}_{P_{h,s,a}}(V_{h+1}^{m,\star})(s_{k,h}^m, a_{k,h}^m) \leq 2H^2MK. \tag{E.35}$$

2. Bounding the Harmonic Sum of Counts: The counter $N_{k,h}(s,a)$ tracks the cumulative visitations. Since the sequence of visits made by all M agents across K episodes directly increments this global counter, we can apply the Pigeonhole Principle (or the integral bound for harmonic series). Bounding the sum of reciprocals of visitation counts

over the state-action-step space gives:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{N_{k,h}(s_{k,h}^m,a_{k,h}^m)}\le\sum_{h=1}^{H}\sum_{s,a}\sum_{n=1}^{N_K(s,a)}\frac{1}{n}\le 2SAH\log(MK). \tag{E.36}$$

Combining the Bounds: Substituting the variance bound and the harmonic sum bound back into Eq. (E.33), we obtain the tightest bound for the dominant term:

$$\begin{aligned}\text{(I)}&\le\sqrt{16\iota\big(2H^2MK\big)}\times\sqrt{2SAH\log(MK)}\\&=\sqrt{64H^3SAMK\iota\log(MK)}\\&=\widetilde{\mathcal{O}}\left(\sqrt{H^3SAMK}\right).\end{aligned} \tag{E.37}$$

Similarly, the lower-order term (II) is bounded using the same Pigeonhole Principle:

$$\text{(II)}=\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{c'SH^2\iota}{N_{k,h}(s_{k,h}^m,a_{k,h}^m)}\le c'SH^2\iota\big(2SAH\log(MK)\big)=\widetilde{\mathcal{O}}\left(S^2AH^3\right). \tag{E.38}$$

Step 6: Putting All Together.

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1)-V_1^{m,\pi_k^m}(s_1)\right]\\&\le\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[\sum_{h=1}^{H}\left(Q_{k,h}^m(s_h,\pi_{k,h}^m(s_h))-Q_h^{m,\pi_k^m}(s_h,\pi_{k,h}^m(s_h))\right)\right]\quad\textbf{(step 1)}\\&\le 2e^2\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho,\pi_k^m}\left[\sum_{h=1}^{H}B_{k,h}^m(s_h,a_h)\right]\quad\textbf{(step 2)}\\&\le 2e^2\left(\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}B_{k,h}^m(s_{k,h}^m,a_{k,h}^m)+\sqrt{2MKH^3\log(1/\delta)}\right)\quad\textbf{(step 3)}\\&\le 2e^2\Big(\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{16\mathsf{Var}_P(V_{h+1}^{m,\star})\iota}{N_{k,h}(s_{k,h}^m,a_{k,h}^m)}}+\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{c'SH^2\iota}{N_{k,h}(s_{k,h}^m,a_{k,h}^m)}+\sqrt{2H^3MK\log(1/\delta)}\\&\qquad+\sum_{m,k,h}\frac{1}{4e^2H}P_{h,s,a}\Delta_{k,h+1}^m\Big)\quad\textbf{(step 4)}\\&\le\mathcal{O}\left(\sqrt{H^3SAMK\iota^2}+S^2AH^3\iota^2\right)+\frac{1}{2H}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1)-V_1^{m,\pi_k^m}(s_1)\right]\quad\textbf{(step 5)}\end{aligned}$$

By moving the estimation error term to the left-hand side, we absorb the variance substitution difference completely:

$$\left(1-\frac{1}{2H}\right)\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1)-V_1^{m,\pi_k^m}(s_1)\right]\leq\mathcal{O}\left(\sqrt{H^3SAMK\iota^2}+S^2AH^3\iota^2\right) \tag{E.39}$$

Since the dominant term $\widetilde{\mathcal{O}}(\sqrt{H^3SAMK})$ scales with $\sqrt{K}$ and strictly dominates the lower-order term $\widetilde{\mathcal{O}}(S^2AH^3)$ for sufficiently large K, the final total regret bound across all agents is tight.

E.3 Proof of Theorem 9

The proof proceeds in four major steps, demonstrating how the federated global coverage eliminates the need for action-wise exploration bonuses.

Step 1: Global Concentration via Reward Heterogeneity. Under Assumption 2, the heterogeneity of the agents' optimal policies guarantees that the global trajectory covers the state-action space. Although individual agents act purely greedily, the federated global count $N_{k,h}(s,a)$ grows linearly with the episode k.

Lemma 31 (Global concentration via heterogeneity)**.** *Suppose Assumption 2 holds. Let $N_{k,h}(s,a)$ be the global visitation count for any state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$ at step $h\in[H]$ until episode k. With probability at least $1-\frac{\delta}{2}$, for all $k\geq 1$, $h\in[H]$, and $(s,a)\in\mathcal{S}\times\mathcal{A}$:*

$$N_{k+1,h}(s,a)\geq kM\lambda_0-\sqrt{2kM\log\left(\frac{2SAKH}{\delta}\right)} \tag{E.40}$$

Consequently, there exists a critical initial phase $K_{th}=\left\lceil\frac{8\log(2SAKH/\delta)}{M\lambda_0^2}\right\rceil+1$, such that for all $k>K_{th}$, the global visitation count is strictly lower bounded by a linear rate: $N_{k+1,h}(s,a)\geq\frac{1}{2}kM\lambda_0$.

This implies that the environment is thoroughly explored by the collective behaviors of the heterogeneous agents, completely bypassing the need for forced exploration.

Proof. Let $\mathcal{F}_{\tau-1}$ be the filtration representing all historical interactions up to the end of episode $\tau-1$. At the beginning of episode τ, the central server broadcasts the empirical transition kernel $\widehat{P}_{\tau,h}$, which is fully determined by $\mathcal{F}_{\tau-1}$.

Crucially, we clarify the relationship between the estimated dynamics $\widehat{P}_{\tau,h}$ and the coverage guarantee. While each agent m computes its greedy policy π_τ^m using the estimated transition kernel $\widehat{P}_{\tau,h} \in \mathcal{F}_{\tau-1}$, the marginal visitation probability $d_h^{\pi_\tau^m}(s,a)$ is always evaluated under the *true* transition kernel P_h, not under $\widehat{P}_{\tau,h}$. Since $\widehat{P}_{\tau,h}$ is a valid stochastic matrix and therefore belongs to the set of all valid transition kernels $\mathcal{P}$, Assumption 2 applies directly with $\widetilde{P} = \widehat{P}_{\tau,h}$: regardless of how inaccurate $\widehat{P}_{\tau,h}$ may be relative to the true dynamics P_h, the diversity of the agents' reward functions $\{r^m\}_{m=1}^M$ structurally ensures that the resulting greedy policies $\{\pi_\tau^m\}_{m=1}^M$ collectively cover the state-action space under the true environment, yielding $\sum_{m=1}^M d_h^{\pi_\tau^m}(s,a) \geq M\lambda_0$ for every episode τ.

Conditioned on the filtration $\mathcal{F}_{\tau-1}$, each agent $m \in [M]$ computes its purely greedy policy $\pi_{\tau,h}^m$ based on its local reward r_h^m and the shared estimated dynamics $\widehat{P}_{\tau,h}$. Let $d_h^{\pi_\tau^m}(s,a)$ be the exact marginal probability of visiting (s,a) at step h under this executed policy π_τ^m.

We define the sequence of random variables $X_\tau^m = d_h^{\pi_\tau^m}(s,a) - \mathbb{I}(s_{\tau,h}^m = s, a_{\tau,h}^m = a)$. By definition, X_τ^m is a martingale difference sequence with respect to $\mathcal{F}_{\tau-1}$, since $\mathbb{E}[X_\tau^m \mid \mathcal{F}_{\tau-1}] = 0$. Furthermore, the absolute value is strictly bounded: $|X_\tau^m| \leq 1$.

By applying the Azuma-Hoeffding inequality over the $k \times M$ random variables, and taking a union bound over all S states, A actions, H steps and K episodes, with probability at least $1 - \frac{\delta}{2}$:

$$\sum_{\tau=1}^{k}\sum_{m=1}^{M} \mathbb{I}(s_{\tau,h}^m = s, a_{\tau,h}^m = a) \geq \sum_{\tau=1}^{k}\sum_{m=1}^{M} d_h^{\pi_\tau^m}(s,a) - \sqrt{2kM\log\left(\frac{2SAKH}{\delta}\right)} \tag{E.41}$$

Crucially, because the empirical transition $\widehat{P}_{\tau,h} \in \mathcal{P}$ is a valid transition kernel (being an empirical average of valid states), we can directly invoke Assumption 2. Regardless

of how inaccurate $\widehat{P}_{\tau,h}$ might be compared to the true dynamics P_h, the ensemble of the agents' greedy policies structurally guarantees a uniform coverage lower bound. Thus, for every single episode τ:

$$\sum_{m=1}^{M} d_h^{\pi_\tau^m}(s,a) \geq M\lambda_0 \tag{E.42}$$

Substituting this lower bound into the martingale concentration inequality, we obtain the cumulative lower bound for the global visitation count:

$$N_{k+1,h}(s,a) \geq \sum_{\tau=1}^{k}(M\lambda_0) - \sqrt{2kM\log\left(\frac{2SAKH}{\delta}\right)} = kM\lambda_0 - \sqrt{2kM\log\left(\frac{2SAKH}{\delta}\right)} \tag{E.43}$$

To ensure $N_{k+1,h}(s,a) \geq \frac{1}{2}kM\lambda_0$, we require $kM\lambda_0 - \sqrt{2kM\log(2SAKH/\delta)} \geq \frac{1}{2}kM\lambda_0$, which simplifies to $k \geq \frac{8\log(2SAKH/\delta)}{M\lambda_0^2} + 1$. We define this threshold as K_{th}. This completes the proof. □

Step 2: Shrinking Bellman Error. With the linearly growing global visitation counts for $k > K_{th}$, the empirical transition dynamics $\widehat{P}_{k,h}$ rapidly converge to the true dynamics P_h as follows:

Lemma 32 (Shrinking Bellman Error)**.** *Suppose the result of Lemma 31 holds. Let $Q_{k,h}^m(s,a)$ be the action-value function computed by agent m at episode k using the empirical transition $\widehat{P}_{k,h}$, and let $Q_h^{m,\star}(s,a)$ be the true optimal action-value function. With probability at least $1-\delta$, for all $k > K_{th}$, $m \in [M]$, $h \in [H]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$:*

$$\left|Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a)\right| \leq \mathcal{O}\left(\sqrt{\frac{H^4\log(4KMHSA/\delta)}{M\lambda_0 k}} + \frac{H^2\log(4KMHSA/\delta)}{M\lambda_0 k}\right) \tag{E.44}$$

Proof. We bound the absolute estimation error of the Q-value by backward induction on h. For the base case $h = H+1$, the error is 0.

For any step $h \in [H]$, taking the absolute difference between the empirical Q-value update and the true optimal Bellman equation, we apply Hölder's inequality:

$$\left|Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a)\right| \leq \left|\langle \widehat{P}_{k,h}(s,a) - P_h(s,a), V_{h+1}^{m,\star}\rangle\right| + \left|\langle \widehat{P}_{k,h}(s,a), V_{k,h+1}^m - V_{h+1}^{m,\star}\rangle\right|$$

Let $X_i = V_{h+1}^{m,\star}(s_{i,h+1}) - P_h(s,a)V_{h+1}^{m,\star}$. Then, the first term is the sample mean of X_i over $N_{k,h}$ samples centered at 0. Invoking Bernstein's inequality, with at least probability $1-\delta'$, the deviation is bounded by:

$$\langle \widehat{P}_{k,h} - P_h, V_{h+1}^{m,\star}\rangle \leq \sqrt{\frac{2\mathsf{Var}_{P_h}(V_{h+1}^{m,\star})\log(\frac{2}{\delta'})}{N_{k,h}(s,a)}} + \frac{2\|V_{h+1}^{m,\star}\|_\infty \log(\frac{2}{\delta'})}{3N_{k,h}(s,a)} \tag{E.45}$$

For all $k > K_{th}$, by substituting the lower bound $N_{k,h}(s,a) \geq \frac{1}{2}kM\lambda_0$ from Lemma 31 and $\|V_{h+1}^{m,\star}\|_\infty \leq H$, we have:

$$\langle \widehat{P}_{k,h} - P_h, V_{h+1}^{m,\star}\rangle \leq \mathcal{O}\left(\sqrt{\frac{\mathsf{Var}_{P_h}(V_{h+1}^{m,\star})\log(\frac{2}{\delta'})}{M\lambda_0 k}} + \frac{H\log(\frac{2}{\delta'})}{M\lambda_0 k}\right) \tag{E.46}$$

The second term can be recursively unpacked as the maximum estimation error of the value function at step $h+1$. As the errors propagate backward until reaching the last step H, the maximum error accumulated over H steps is bounded by:

$$\begin{aligned}\max_{s,a}\left|Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a)\right| &\leq \mathcal{O}\left(\sqrt{\frac{\log(\frac{2}{\delta'})}{M\lambda_0 k}}\sum_{h'=h}^{H}\sqrt{\mathsf{Var}_{P_{h'}}(V_{h'+1}^{m,\star})} + \frac{H^2\log(\frac{2}{\delta'})}{M\lambda_0 k}\right) \\ &\leq \mathcal{O}\left(\sqrt{\frac{\log(\frac{2}{\delta'})}{M\lambda_0 k}}H^2 + \frac{H^2\log(\frac{2}{\delta'})}{M\lambda_0 k}\right)\end{aligned} \tag{E.47}$$

Here, we use the fact that the variance of the optimal value function is bounded by H^2, implying $\sqrt{\mathsf{Var}_{P_{h'}}(V_{h'+1}^{m,\star})} \leq H$. Summing this worst-case bound over H steps yields H^2, which is then absorbed into the square root as H^4.

We require this bound to hold uniformly for all $k \in [K]$, $m \in [M]$, $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$. To apply the union bound (allocating the remaining $\delta/2$ failure probability), we

set $\delta' = \frac{\delta}{2KMHSA}$. Then,

$$\max_{s,a} \left| Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a) \right| \leq \mathcal{O}\left(\sqrt{\frac{H^4 \log(4KMHSA/\delta)}{M\lambda_0 k}} + \frac{H^2 \log(4KMHSA/\delta)}{M\lambda_0 k} \right) \tag{E.48}$$

This completes the proof. □

Step 3: Burn-in Costs. Because PF-EFVI does not inject any UCB bonuses, an agent m will strictly choose the optimal action $\pi_h^{m,\star}(s)$ if its Q-value estimation error is strictly less than half of the minimum suboptimality gap, i.e., $\left| Q_{k,h}^m - Q_h^{m,\star} \right| < \Delta_{\min}^m/2$. We derive the critical burn-in episode K_0^m for agent m as follows:

Lemma 33 (The burn-in cost of pure exploitation)**.** *There exists a critical burn-in episode K_0^m such that for all episodes $k > \max(K_{th}, K_0^m)$, the purely greedy policy $\pi_{k,h}^m(s) = \arg\max_a Q_{k,h}^m(s,a)$ exactly matches the true optimal policy $\pi_h^{m,\star}(s)$, yielding zero instantaneous regret. This tipping point is bounded by:*

$$K_0^m = \mathcal{O}\left(\frac{H^4 \log(KMHSA/\delta)}{M\lambda_0 (\Delta_{\min}^m)^2} \right) \tag{E.49}$$

Proof. Let $\epsilon_k = \max_{s,a,h} |Q_{k,h}^m(s,a) - Q_h^{m,\star}(s,a)|$ be the maximum absolute estimation error at episode k. By Lemma 32, we have $\epsilon_k \leq \mathcal{O}\left(\sqrt{\frac{H^4 \log(KMHSA/\delta)}{kM\lambda_0}} \right)$.

Because the agent acts greedily without any UCB bonuses, it will strictly prefer the optimal action $a^\star$ over any suboptimal action a' if:

$$Q_{k,h}^m(s,a^\star) > Q_{k,h}^m(s,a') \tag{E.50}$$

Using the absolute error bound, the estimated values satisfy $Q_{k,h}^m(s,a^\star) \geq Q_h^{m,\star}(s,a^\star) - \epsilon_k$ and $Q_{k,h}^m(s,a') \leq Q_h^{m,\star}(s,a') + \epsilon_k$. Therefore, the optimal action is guaranteed to be chosen

if:

$$Q_h^{m,\star}(s,a^\star) - \epsilon_k > Q_h^{m,\star}(s,a') + \epsilon_k \implies 2\epsilon_k < Q_h^{m,\star}(s,a^\star) - Q_h^{m,\star}(s,a') = \Delta_h^m(s,a') \tag{E.51}$$

To ensure this holds for all suboptimal actions, we require $2\epsilon_k < \Delta_{\min}^m$. Substituting the upper bound of ϵ_k and solving for k, we find that the condition holds for all $k > K_0^m$. For $k > \max(K_{th}, K_0^m)$, no suboptimal actions are ever chosen. □

Step 4: Summing the Regret. To compute the total regret for agent m, we split the learning process into the initial uncontrolled phase ($k \leq K_{th}$) and the shrinking phase ($K_{th} < k \leq K_0^m$). If a suboptimal action is chosen, its gap is bounded by $2\epsilon_k$. Since the instantaneous regret is strictly zero for all $k > \max(K_{th}, K_0^m)$ (by Lemma 33), we sum the regret across these two phases.

During the uncontrolled phase ($P1$), the instantaneous regret is trivially bounded by the horizon H. Thus:

$$Regret_{P1}^m(K) \leq \sum_{k=1}^{K_{th}} H = H \cdot K_{th} = \mathcal{O}\left(\frac{H \log(SAKH/\delta)}{M\lambda_0^2}\right) \tag{E.52}$$

During the shrinking phase ($P2$), we integrate the bounded error ϵ_k:

$$\begin{aligned}
Regret_{P2}^m(K) &= \sum_{k=K_{th}+1}^{K_0^m} \mathbb{E}_{s_1\sim\rho}\left[V_1^{m,\star}(s_1) - V_1^{m,\pi_k^m}(s_1)\right] \\
&\overset{(i)}{=} \sum_{k=K_{th}+1}^{K_0^m} \mathbb{E}_{\pi_k^m, s_1\sim\rho}\left[\sum_{h=1}^{H}\left(Q_h^{m,\star}(s_h,\pi_h^{m,\star}(s_h)) - Q_h^{m,\star}(s_h,\pi_{k,h}^m(s_h))\right)\right] \\
&\leq \sum_{k=1}^{K_0^m} \mathcal{O}\left(\sqrt{\frac{H^6\log(KMHSA/\delta)}{kM\lambda_0}}\right) \\
&\leq \mathcal{O}\left(\sqrt{\frac{H^6\log(KMHSA/\delta)}{M\lambda_0}}\sum_{k=1}^{K_0^m}\frac{1}{\sqrt{k}}\right)
\end{aligned}$$

$$= \mathcal{O}\left(\sqrt{\frac{H^6 \log(KMHSA/\delta)}{M\lambda_0}} \cdot 2\sqrt{K_0^m}\right) \tag{E.53}$$

where (i) follows from the Performance Difference Lemma. Since the estimated Q-value error at any step is bounded by ϵ_k, the suboptimality gap of the chosen greedy action at each step is at most $2\epsilon_k$. Summing this over the H steps of the trajectory yields an instantaneous regret bound of $2H\epsilon_k$.

We substitute the explicit value of $\sqrt{K_0^m} = \mathcal{O}\left(\frac{H^2\sqrt{\log(KMHSA/\delta)}}{\sqrt{M\lambda_0}\Delta_{\min}^m}\right)$ into Equation (E.53):

$$\begin{aligned} Regret_{P2}^m(K) &\leq \mathcal{O}\left(\sqrt{\frac{H^6 \log(KMHSA/\delta)}{M\lambda_0}} \cdot \frac{H^2\sqrt{\log(KMHSA/\delta)}}{\sqrt{M\lambda_0}\Delta_{\min}^m}\right) \\ &= \mathcal{O}\left(\frac{H^5 \log(KMHSA/\delta)}{M\lambda_0\Delta_{\min}^m}\right) \end{aligned} \tag{E.54}$$

Summing these independent bounds over all M agents yields the final total system regret:

$$\begin{aligned} Regret(K) &= \sum_{m=1}^{M} \left(Regret_{P1}^m(K) + Regret_{P2}^m(K)\right) \\ &\leq M \cdot \mathcal{O}\left(\frac{H\log(SAH/\delta)}{M\lambda_0^2}\right) + \sum_{m=1}^{M} \mathcal{O}\left(\frac{H^5 \log(KMHSA/\delta)}{M\lambda_0\Delta_{\min}^m}\right) \\ &= \mathcal{O}\left(\frac{H\log(SAH/\delta)}{\lambda_0^2}\right) + \mathcal{O}\left(\frac{H^5 \log(KMHSA/\delta)}{M\lambda_0} \sum_{m=1}^{M} \frac{1}{\Delta_{\min}^m}\right). \end{aligned} \tag{E.55}$$

Bibliography

Bai, Q., Mondal, W. U., and Aggarwal, V. (2024). Regret analysis of policy gradient algorithm for infinite horizon average reward Markov decision processes. *International Conference on Artificial Intelligence*.

Bastani, H., Bayati, M., and Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1992.

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

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

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

Bravo, M. and Contreras, J. P. (2024). Stochastic halpern iteration in normed spaces and applications to reinforcement learning. *arXiv preprint arXiv:2403.12338*.

Chen, Z. (2025). Non-asymptotic guarantees for average-reward Q-learning with adaptive stepsizes. *arXiv preprint arXiv:2504.18743*.

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

Civitavecchia, L. and Papini, M. (2025). Exploration-free reinforcement learning with

linear function approximation. In *Reinforcement Learning Conference*.

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

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

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

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

Ganesh, S., Mondal, W. U., and Aggarwal, V. (2024). A sharper global convergence analysis for average reward reinforcement learning via an actor-critic approach. *arXiv preprint arXiv:2407.18878*.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes.* John Wiley and Sons.

Jiao, Y., Woo, J., Li, G., Joshi, G., and Chi, Y. (2026). Sample complexity of average-reward q-learning: From single-agent to federated reinforcement learning. *arXiv preprint arXiv:2601.13642*.

Jin, Y., Gummadi, R., Zhou, Z., and Blanchet, J. (2024). Feasible Q-learning for average reward reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1630–1638. PMLR.

Jin, Y. and Sidford, A. (2021). Towards tight bounds on the sample complexity of average-reward MDPs. *International Conference on Machine Learning*.

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

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems

in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Kannan, S., Morgenstern, J., Roth, A., Waggoner, B., and Wu, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

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

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

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*.

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

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

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc.

Kumar, N., Murthy, Y., Shufaro, I., Levy, K. Y., Srikant, R., and Mannor, S. (2025). On the global convergence of policy gradient in average reward Markov decision processes. *International Conference on Learning Representations*.

Labbi, S., Tiapkin, D., Mancini, L., Mangold, P., and Moulines, É. (2024). Federated ucbvi: Communication-efficient federated regret minimization with heterogeneous agents. In *International Conference on Artificial Intelligence and Statistics*.

Lee, J., Bravo, M., and Cominetti, R. (2025). Near-optimal sample complexity for MDPs

via anchoring. In *Forty-second International Conference on Machine Learning*.

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

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

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022a). Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*.

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

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2022b). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473.

Li, T., Wu, F., and Lan, G. (2024b). Stochastic first-order methods for average-reward Markov decision processes. *Mathematics of Operations Research*.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. In *Proc. Computational Learning Theory Conference*.

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

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

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

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic program-*

ming. John Wiley & Sons.

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

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

Salgia, S. and Chi, Y. (2024). The sample-communication complexity trade-off in federated Q-learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

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

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018). Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM.

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

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

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

Tuynman, A., Degenne, R., and Kaufmann, E. (2024). Finding good policies in average-

reward Markov decision processes without prior knowledge. *Neural Information Processing Systems*.

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

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

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

Wan, Y., Naik, A., and Sutton, R. S. (2021). Learning and planning in average-reward Markov decision processes. *International Conference on Machine Learning*.

Wan, Y., Yu, H., and Sutton, R. S. (2024). On convergence of average-reward Q-learning in weakly communicating markov decision processes. *arXiv preprint arXiv:2408.16262*.

Wang, H., He, S., Zhang, Z., Miao, F. M., and Anderson, J. (2024a). Momentum for the win: collaborative federated reinforcement learning across heterogeneous environments. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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

Wang, J., Wang, M., and Yang, L. F. (2022). Near Sample-Optimal Reduction-based Policy Learning for Average Reward MDP. *arXiv preprint arXiv:2212.00603*.

Wang, M. (2017). Primal-dual π learning: Sample complexity and sublinear run time for ergodic Markov decision problems. *arXiv:1710.06100*.

Wang, S., Blanchet, J., and Glynn, P. (2024b). Optimal sample complexity for average reward Markov decision processes. *International Conference on Learning Representations*.

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

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. *International Conference on Machine Learning*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3-4).

Woo, J., Joshi, G., and Chi, Y. (2023). The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37157–37216. PMLR.

Woo, J., Joshi, G., and Chi, Y. (2025). The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *Journal of Machine Learning Research*, 26(26):1–85.

Woo, J., Shi, L., Joshi, G., and Chi, Y. (2024). Federated offline reinforcement learning: Collaborative single-policy coverage suffices. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 53165–53201. PMLR.

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

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021a). Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021b). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

Xiong, G., Wang, S., Jiang, D., and Li, J. (2025). On the linear speedup of personalized federated reinforcement learning with shared representations. In *The Thirteenth International Conference on Learning Representations*.

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

Yang, T., Cen, S., Wei, Y., Chen, Y., and Chi, Y. (2024a). Federated natural policy gradient and actor critic methods for multi-task reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yang, T., Cen, S., Wei, Y., Chen, Y., and Chi, Y. (2024b). Federated natural policy gradient and actor critic methods for multi-task reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yin, M. and Wang, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in Neural Information Processing Systems*, 34.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. (2020). MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*.

Zhang, C. and Azizan, N. (2026). Personalized collaborative learning with affinity-based variance reduction. In *The Fourteenth International Conference on Learning Representations*.

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

Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward Markov decision processes. *Conference on Learning Theory*.

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

Zurek, M. and Chen, Y. (2024). Span-based optimal sample complexity for weakly communicating and general average reward MDPs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zurek, M. and Chen, Y. (2025a). The plug-in approach for average-reward and discounted

MDPs: Optimal sample complexity analysis. In *Conference on Algorithmic Learning Theory*.

Zurek, M. and Chen, Y. (2025b). Span-agnostic optimal sample complexity and oracle inequalities for average-reward rl. In Haghtalab, N. and Moitra, A., editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 6156–6209.