Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents

Safwan Labbi[1] **Daniil Tiapkin**[1,2] **Lorenzo Mancini**[1] **Paul Mangold**[1] **Eric Moulines**[1,3]
[1] CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
[2] Université Paris-Saclay, CNRS, LMO, 91405, Orsay, France
[3] Mohamed bin Zayed University of Artificial Intelligence, UAE

Abstract

In this paper, we present the Federated Upper Confidence Bound Value Iteration algorithm (`Fed-UCBVI`), a novel extension of the `UCBVI` algorithm (Azar et al., 2017) tailored for the federated learning framework. We prove that the regret of `Fed-UCBVI` scales as $\widetilde{\mathcal{O}}(\sqrt{H^3|\mathcal{S}||\mathcal{A}|T/M})$, with a small additional term due to heterogeneity, where $|\mathcal{S}|$ is the number of states, $|\mathcal{A}|$ is the number of actions, H is the episode length, M is the number of agents, and T is the number of episodes. Notably, in the single-agent setting, this upper bound matches the minimax lower bound up to polylogarithmic factors, while in the multi-agent scenario, `Fed-UCBVI` has linear speed-up. To conduct our analysis, we introduce a new measure of heterogeneity, which may hold independent theoretical interest. Furthermore, we show that, unlike existing federated reinforcement learning approaches, `Fed-UCBVI`'s communication complexity only marginally increases with the number of agents.

1 INTRODUCTION

Federated reinforcement learning (FRL, Zhuo et al., 2019; Qi et al., 2021) adapts the principles of federated learning (FL, McMahan et al., 2017) to the domain of reinforcement learning (RL, Sutton and Barto, 2018). It enables multiple agents, evolving in independent environments, to learn a policy collaboratively without directly exchanging their states/actions. To learn together, agents communicate under the supervision of a central server (CS), aiming to maximize the expected rewards averaged across all agents. Consequently, agents participating in FRL may learn better policies with fewer interactions with the environment. FRL appears to be a promising solution for reducing the cost of training. However, the efficient implementation of FRL faces significant challenges. Similarly to FL, agents typically evolve in different environments and often have limited computational power and communication bandwidth. Furthermore, the traditional challenges of RL, such as balancing exploration and exploitation, remain. Thus, there is a growing demand for methods tailored for FRL, aiming to reduce communication complexity (i.e., the number of communications) while maintaining efficient exploration and learning.

FRL has attracted considerable attention in recent years, with a strong focus put on federated versions of Q-Learning. This research often relies on one of two following assumptions: either (1) all agents operate in *identical environments* (Chen et al., 2023; Zheng et al., 2024, 2025), or (2) a generative model is available, allowing access to sampling from any state-action pair without exploration (Jin et al., 2022; Wang et al., 2024). Another notable category of methods, called distributed reinforcement learning (Bai et al., 2019; Zhang et al., 2020), enables agents to address RL problems collaboratively. However, these methods require centralizing observational data on a single server, which may not be feasible in real applications.

Unfortunately, the aforementioned approaches do not address the exploration-exploitation trade-off in heterogeneous environments. Furthermore, their high communication complexity poses a major challenge for their use, even with homogeneous agents.

In this paper, we introduce the algorithm `Fed-UCBVI` for tabular episodic FRL and we analyze its *federated regret*, i.e., the regret averaged across all agents, in the presence of environmental heterogeneity. The tabular FRL problem involves M agents, each inter-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Table 1: Comparison with related algorithms in the online setting

Type	Algorithm	Heterogeneity	Communication complexity	Regret
Model-based	`Concurrent UCBVI` (Azar et al. 2017)	✗	$\mathcal{O}(T)$	$\tilde{\mathcal{O}}(\sqrt{H^3\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert T/M})$
	`Byzan-UCBVI` (Chen et al. 2023)	✗	$\mathcal{O}(M\cdot\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H\cdot\log(T))$	$\tilde{\mathcal{O}}(\sqrt{H^4\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert T/M})$
	`Fed-UCBVI` (our work)	✓	$\mathcal{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H\cdot\log(T)\right)$	$\tilde{\mathcal{O}}(\sqrt{H^3\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert T/M})$
Model-free	`Concurrent UCB-Advantage` (Zhang et al. 2020)	✗	$\mathcal{O}(T)$	$\tilde{\mathcal{O}}(\sqrt{H^3\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert T/M})$
	`FedQ-Bernstein` (Zheng et al. 2024)	✗	$\mathcal{O}(M\cdot\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^3\cdot\log(T))$	$\tilde{\mathcal{O}}(\sqrt{H^4\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert T/M})$
	`FedQ-Advantage` (Zheng et al. 2025)	✗	$\mathcal{O}(M\cdot\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^2\log(H)\cdot\log(T))$	$\tilde{\mathcal{O}}(\sqrt{H^3\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert T/M})$
	`Lower Bound` (Jin et al. 2018; Domingues et al. 2021b)	✗	?	$\tilde{\mathcal{O}}(\sqrt{H^3\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert T/M})$

* The results are derived in a homogeneous setting. For all bounds, only the leading term with respect to the dependence on T is shown. H: number of steps per episode; T: total episodes collected per agent; $|\mathcal{S}|$: number of states; $|\mathcal{A}|$; number of actions; M: number of agents.

acting with its own environment, modeled as a finite-horizon Markov Decision Process (MDP). For an agent $i \in [M]$, a finite-horizon MDP is defined by a tuple $\mathcal{M}^i := (\mathcal{S}, \mathcal{A}, H, \{\mathsf{P}_h^i\}_{h\in[H]}, \{\mathsf{r}_h^i\}_{h\in[H]})$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, H is the number of steps in one episode (also referred to as a planning horizon), $\mathsf{P}_h^i(s'|s,a)$ denotes the probability of transitioning from a state $s \in \mathcal{S}$ to the next state $s' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$ at step h for agent i, and $\mathsf{r}_h^i(s,a)$, is a bounded deterministic reward function that satisfies $\mathsf{r}_h^i(s,a) \in [0,1]$ for all $(s,a,h,i) \in \mathcal{S}\times\mathcal{A}\times[H]\times[M]$. Note that both the transition probabilities (kernel) and the reward function can vary depending on the decision-making step $h \in [H]$. The learning process is divided into T episodes, each of length H. Both the transition kernel and the reward function are assumed to be *unknown* to all agents and the central server (CS).

`Fed-UCBVI` is a model-based approach where each agent independently estimates its local state-action transition kernel. These local estimates are then used to compute state-action value functions, which are aggregated by a CS using an adaptive scheme that accounts for each agent's level of uncertainty. Communication complexity is managed through an adaptive communication strategy triggered by the optimization process's progress and ensures efficient coordination. Overall, our contributions are:

- We propose `Fed-UCBVI`, an FRL algorithm designed to aggregate the local estimators of each agent. We prove that the federated regret of `Fed-UCBVI` scales as $\mathcal{O}(\sqrt{H^3|\mathcal{S}||\mathcal{A}|T/M})$, up to a heterogeneity term which scales proportionally to our heterogeneity measure. This shows that `Fed-UCBVI` achieves a linear speedup and effectively accelerates training compared to single-agent RL. To our knowledge, `Fed-UCBVI` is the first provably efficient algorithm for regret minimization in heterogeneous environments.
- To analyze `Fed-UCBVI`, we introduce a new measure of heterogeneity that quantifies the divergence of each agent's state-transition kernel from a baseline kernel, which may be of independent interest.
- We develop a novel method for reducing the communication cost. We prove that the communication complexity of `Fed-UCBVI` is $\mathcal{O}(M\log\log T + \log T)$. This is a significant improvement over existing methods (e.g., Zheng et al., 2024), that require $\mathcal{O}(M\log T)$ communication rounds.
- We validate our theoretical results through numerical experiments on FRL problems, demonstrating that our algorithm outperforms existing FRL baselines with theoretical guarantees. In particular, our simulations show a significant improvement in regret compared to Fed-Q-learning (Zheng et al., 2024) for different degrees of heterogeneity.

The paper is organized as follows: we review the related work in Section 2, and introduce the necessary mathematical background in Section 3. In Section 4, we introduce and analyze the `Fed-UCBVI` algorithm. Then, we present numerical experiments in Section 5.

2 RELATED WORK

Reinforcement Learning. Two main approaches have been proposed for regret minimization in the single-agent, finite-horizon tabular setting: (i) model-based algorithms (Azar et al., 2017; Dann et al., 2017; Zanette and Brunskill, 2019; Zhang et al., 2024b), and (ii) model-free algorithms (Jin et al., 2018; Zhang et al., 2020; Li et al., 2021).

Both approaches offer algorithms that achieve the minimax optimal lower bound up to poly-logarithmic factors, specifically $\Omega(\sqrt{H^3|\mathcal{S}||\mathcal{A}|T})$ (Jin et al., 2018; Domingues et al., 2021b). Among these, `UCBVI` (Azar et al., 2017), which is based on the principle of optimism in the face of uncertainty, was the first algorithm to achieve the minimax bound.

Federated Reinforcement Learning. The FRL method most closely related to ours is the Byzantine robust distributed `UCBVI` algorithm (Chen et al., 2023), which assumes homogeneous agents. This algorithm achieves a regression bound of $\tilde{\mathcal{O}}(\sqrt{H^4|\mathcal{S}|^2|\mathcal{A}|T/M})$

and a communication complexity that scales logarithmically with the number of episodes T.

In contrast, our method achieves a regret of $\tilde{\mathcal{O}}(\sqrt{H^3|\mathcal{S}||\mathcal{A}|T/M})$, which is optimal in single-agent environments. Moreover, we also provide guarantees in heterogeneous environments.

Other FRL approaches are based on model-free methods. Zhang et al. (2020) proposed a federated variant of Q-learning, with regret $\tilde{\mathcal{O}}(\sqrt{H^3|\mathcal{S}||\mathcal{A}|T/M})$, and communication complexity linear in T. Zheng et al. (2024) later reduced the communication cost to $O(M \log T)$, but introduced an additional factor of H in the regret bound. More recently, Zheng et al. (2025) improved both regret and communication cost. However, their method still requires homogeneous agents, and the communication complexity remains $O(M \log T)$. FRL's communication complexity has been studied in several related settings: finding the optimal Q-function Salgia and Chi (2024), policy optimization Lan et al. (2024); Chen et al. (2021) and policy evaluation Mangold et al. (2025).

3 SETTING

3.1 Federated Reinforcement Learning

Policy and Value Functions. A deterministic policy π is a set of functions $\pi_h \colon \mathcal{S} \to \mathcal{A}$ where $\pi_h(s) \in \mathcal{A}$, $h \in [H]$. The value function $\mathcal{V}_h^{i,\pi}$, is defined as:

$$\mathcal{V}_h^{i,\pi}(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} \mathsf{r}_{h'}^i(s_{h'}^i, a_{h'}^i) \middle| s_h^i = s \right] , \qquad (1)$$

where for all $h \le h' \le H$, $a_{h'}^i \sim \pi_{h'}^i(.|s_{h'}^i)$ and for all $h \le h' \le H-1$, $s_{h'+1}^i \sim \mathsf{P}_{h'}^i(.|s_{h'}^i, a_{h'}^i)$. Similarly, the Q-function of a policy π for agent i at step h is

$$\mathcal{Q}_h^{i,\pi}(s,a) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} \mathsf{r}_{h'}^i(s_{h'}^i, a_{h'}^i) \middle| s_h^i = s, a_h^i = a \right] ,$$

and satisfies the Bellman equations

$$\begin{aligned} \mathcal{Q}_h^{i,\pi}(s,a) &= \mathsf{r}_h^i(s,a) + \mathsf{P}_h^i \mathcal{V}_{h+1}^{i,\pi}(s,a) , \\ \mathcal{V}_{h+1}^{i,\pi}(s) &= \mathcal{Q}_h^{i,\pi}(s, \pi_h(s)) . \end{aligned} \qquad (2)$$

Additionally, the optimal Q-value satisfies the optimal Bellman equations

$$\begin{aligned} \mathcal{Q}_h^{i,\star}(s,a) &= \mathsf{r}_h^i(s,a) + \mathsf{P}_h^i \mathcal{V}_{h+1}^{i,\star}(s,a) , \\ \mathcal{V}_h^{i,\star}(s) &= \max_{a \in \mathcal{A}} \mathcal{Q}_h^{i,\star}(s,a) . \end{aligned} \qquad (3)$$

Learning Protocol. At the beginning of each episode $t \in [T]$, all agents select a common policy π_t, which is computed based on the information *exchanged* prior to episode t. Subsequently, each agent generates an independent trajectory of length H. At each step h, an agent observes its state $s_{t,h}^i \in \mathcal{S}$ and takes an action $a_{t,h}^i = \pi_{t,h}(s_{t,h}^i) \in \mathcal{A}$. The agent then observes the next state $s_{t,h+1}^i$ according to the transition probabilities $\mathsf{P}_h^i(\cdot|s_{t,h}^i, a_{t,h}^i)$ and receives a deterministic reward $\mathsf{r}_{t,h}^i = \mathsf{r}_h^i(s_{t,h}^i, a_{t,h}^i)$. After generating these trajectories, agents *may* exchange information through the central server.

Federated Regret. The performance of the learning algorithm is evaluated using the *federated regret*, defined as

$$\mathfrak{R}(T) := \max_\pi \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \mathcal{V}_1^{i,\pi}(s_{t,1}^i) - \mathcal{V}_1^{i,\pi_t}(s_{t,1}^i) . \quad (4)$$

This regret measures the cumulative difference, in expectation, between the average value of the optimal collaborative policy and the policies used throughout the training procedure.

Communication Complexity and Cost. The *communication complexity*, denoted by $\mathfrak{C}(T)$, is defined as the number of episodes where communication between the CS and the agents occurs. The *communication cost* refer to the total number of bits exchanged between the central server and the agents during the learning process. The objective of the FRL algorithm is to simultaneously minimize both the regret $\mathfrak{R}(T)$ and the communication complexity $\mathfrak{C}(T)$.

3.2 Environmental Heterogeneity

The environments in which agents evolve may differ from one to another. However, since agents aim to learn a shared policy, environmental heterogeneity must be small. To measure this, we introduce a new notion of heterogeneity, decomposing each agent's state-action transition kernel into a common part, shared by all agents, and an individual part that reflects unique environmental characteristics. Formally, this is captured by the following assumption.

A-1. *There exists a non-homogeneous transition kernel* $\{\mathsf{P}_h^{\mathsf{c}}\}_{h\in[H]}$, *$M$ individual non-homogeneous transition kernels* $\{\mathsf{P}_h^{\mathsf{ind},i}\}_{h\in[H]}$ *for any* $i \in [M]$, *and a constant* $\varepsilon_{\mathsf{p}} \in [0,1)$, *such that for any* $i \in [M]$ *and* $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$,

$$\mathsf{P}_h^i(s'|s,a) = (1-\varepsilon_{\mathsf{p}})\mathsf{P}_h^{\mathsf{c}}(s'|s,a) + \varepsilon_{\mathsf{p}} \mathsf{P}_h^{\mathsf{ind},i}(s'|s,a) .$$

This assumption can be interpreted as follows: the environment of all agents reacts similarly to a given action, but occasionally, due to local specificities, an individual environment reacts differently. This assumption is related to the ε-contamination model proposed by Huber (1964). Likewise, we assume that agents receive comparable rewards for a given state-action pair.

A-2. *There exists a constant* $\varepsilon_{\mathsf{r}} \in [0,1)$ *such that for all* $(i,j) \in [M]$, *and for all* $h \in [H]$ *it holds that*

$$\|\mathsf{r}_h^i - \mathsf{r}_h^j\|_\infty \leq \varepsilon_{\mathsf{r}} \, .$$

Note that **A**-1 implies the following bound on the difference between the common transition kernel and each agent's transition kernel, measured in L_1-norm,

$$\max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \|\mathsf{P}_h^{\mathsf{c}}(\cdot|s,a) - \mathsf{P}_h^i(\cdot|s,a)\|_1 \leq \varepsilon_{\mathsf{p}} \, . \quad (5)$$

We prove this inequality in Appendix F. Thus, (1) is slightly stronger than (5), which is the typical assumption in other FRL settings, such as FedSARSA (Zhang et al., 2024a) or policy optimization with access to a simulator (Jin et al., 2022; Wang et al., 2024). The motivation for using **A**-1 over (5) lies in the need to control how *samples from* P_h^i relate to *samples from* $\mathsf{P}_h^{\mathsf{c}}$. This is crucial in RL without a generative model, as the data generation process is not independent and identically distributed, forcing agents to exploit all the samples they have. In Section 4, we discuss in detail the necessity of this assumption for our analysis.

4 FED-UCBVI ALGORITHM

In this section, we present the `Fed-UCBVI` algorithm, which extends the `UCBVI` algorithm proposed by Azar et al. (2017) to the federated learning framework. The process involves multiple communication rounds with a CS. The number of episodes in each communication round (or epoch) r is random, and each epoch is decomposed into three phases:

(i) *Data collection*: During this phase, each agent interacts with its environment using the policy $\pi_{(r)}$ provided by the CS, gathering trajectory data.

(ii) *Synchronization*: Once any agent meets the synchronization conditions, it sends a synchronization signal to the central server, which then broadcasts this information to all other agents.

(iii) *Policy update*: In this phase, all agents engage in H sequential communications with the CS. At each step $h = H$ to 1, agents send their local estimates of the Q-values and other related information related to step h to the CS. In return, they receive a global estimate of the V-values, along with an updated policy and related information for that step.

The following sections provide a detailed overview of each of these stages.

Data Collection. At the beginning of round r, each agent $i \in [M]$ follows the policy $\pi_{(r)}$ to collect new trajectories. For $\ell \in \mathbb{N}$, denote by $n^i_{(r,\ell),h}(s,a)$ and $n^i_{(r,\ell),h}(s,a,s')$ the number of visits to a state-action pair (s,a) and the number of transitions from (s,a) to s' at step h after ℓ episodes in the round r.

Synchronization. At the start of epoch r, all agents receive the current global counters

$$N_{(r),h}(s,a) := \textstyle\sum_{i=1}^M n^i_{(r),h}(s,a), \quad (6)$$

where $n^i_{(r),h}(s,a) := n^i_{(r,0),h}(s,a)$ is the number of visits of a state-action pair by agent i prior to round r.

During epoch r, after ℓ episodes, agent i sends a synchronization signal if a newly visited state-action-step triplet (s,a,h) is identified and one of two synchronization conditions is met. These conditions depend on whether the total number of visits $N_{(r),h}(s,a)$ exceeds a threshold $\nu(\delta,T) = \widetilde{\mathcal{O}}(\varepsilon_{\mathsf{p}} THM + M)$ (see Equation (57) in Appendix E for the full expression).

1) *Local Doubling Condition.* If $N_{(r),h}(s,a) \leq \nu(\delta,T)$, an agent i sends the synchronization signal if

$$n^i_{(r,\ell),h}(s,a) > 2n^i_{(r),h}(s,a) \, . \quad (7)$$

2) *Globally Estimated Doubling Condition.* If $N_{(r),h}(s,a) > \nu(\delta,T)$, agent i sends the synchronization signal if

$$\hat{N}^i_{(r,\ell),h}(s,a) > 2N_{(r),h}(s,a) \, , \quad (8)$$

where $\hat{N}^i_{(r,\ell),h}(s,a)$ is an estimate of $\sum_{i=1}^M n^i_{(r,\ell),h}(s,a)$ based on the information available to agent i.

Policy Update. Upon receiving the synchronization signal, each agent computes its local estimates of transition probabilities as

$$\widehat{\mathsf{P}}^i_{(r+1),h}(s'|s,a) := \frac{n^i_{(r+1),h}(s,a,s')}{n^i_{(r+1),h}(s,a)} \quad (9)$$

if $n^i_{(r+1),h}(s,a) > 0$, otherwise $\widehat{\mathsf{P}}^i_{(r+1),h}(s'|s,a) := 1/|\mathcal{S}|$. Additionally, each agent builds a local estimate of the reward function, denoted as $\widehat{\mathsf{r}}^{\,i}$. This estimate is initially set to a null $H\times\mathcal{S}\times\mathcal{A}$ tensor. The entry $\widehat{\mathsf{r}}^i_h(s,a)$ is updated with the received reward whenever agent i selects action a in state s at step h.

Next, the agents and the central server exchange their Q- and V-value estimates. For $h = H,\dots,1$, each agent computes the local Q-value estimate

$$\hat{\mathcal{Q}}^i_{(r+1),h}(s,a) := \left[\widehat{\mathsf{r}}^i_h + \widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}_{(r+1),h+1}\right](s,a), \quad (10)$$

using the global value estimate $\hat{\mathcal{V}}_{(r+1),h+1}$ previously received from the CS; note that for $h = H$, this value is set to zero and does not require communication.

Then, the CS collects the local Q-value estimates from all agents, along with additional information

Algorithm 1: Fed-UCBVI

Initialization: $t = 1$; $r = 1$; $\widehat{\mathsf{r}}^i_h(s,a) = 0$; $N_{(1),h}(s,a) = 0$; $n^i_{(1,0),h}(s,a) = 0$; $\hat{\mathcal{Q}}_{(1),h}(s,a) = \hat{\mathcal{V}}_{(1),h}(s) = H$ for all $(s,a,h,i) \in \mathcal{S} \times \mathcal{A} \times [H] \times [M]$; $\pi_{(1)} = \{\pi_{(1),h}\}_h$ for some policy $\pi_{(1)}$; and $\nu(\delta, T)$ set as in (57).

while $t \leq T$ **do**

  for *each agent* $i = 1$ ***to*** M ***in parallel*** **do**

    Set $l = 1$; $n^i_{(r),h}(s,a) = n^i_{(r,0),h}(s,a)$; and $\hat{N}^i_{(r,0),h}(s,a) = N_{(r),h}(s,a)$

    while *no synchronization signal* **do**

      Collect $(s^i_{t,h}, a^i_{t,h}, \mathsf{r}^i_{t,h}, s^i_{t,h+1})_{1 \leq h \leq H}$ using $\pi_{(r)}$

      for $h = H$ ***to*** 1 **do**

        Set $n^i_{(r,\ell),h}(s,a) = n^i_{(r,\ell-1),h}(s,a) + 1_{(s,a)}(s^i_{t,h}, a^i_{t,h})$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$ and

        $n^i_{(r,\ell),h}(s,a,s') = n^i_{(r,\ell-1),h}(s,a,s') + 1_{(s,a,s')}(s^i_{t,h}, a^i_{t,h}, s^i_{t,h+1})$ for $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$

        Set $\hat{N}^i_{(r,\ell),h}(s,a) = \hat{N}^i_{(r,\ell-1),h}(s,a) + M 1_{(s,a)}(s^i_{t,h}, a^i_{t,h})\}$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$

        Set $\widehat{\mathsf{r}}^i_h(s^i_{t,h}, a^i_{t,h}) = \mathsf{r}^i_{t,h}$ and $\ell = \ell + 1$

      if $\big(N_{(r),h}(s^i_{t,h}, a^i_{t,h}) \leq \nu(\delta, T)$ **and** $n^i_{(r,\ell),h}(s^i_{t,h}, a^i_{t,h}) > 2n^i_{(r),h}(s^i_{t,h}, a^i_{t,h})\big)$

        or $\big(N_{(r),h}(s^i_{t,h}, a^i_{t,h}) > \nu(\delta, T)$ **and** $\hat{N}^i_{(r,\ell),h}(s^i_{t,h}, a^i_{t,h}) > N_{(r),h}(s^i_{t,h}, a^i_{t,h})\big)$ **then**

        Send synchronization signal

    Set $t = t + \ell$; $n^i_{(r+1),h}(s,a) = n^i_{(r,\ell),h}(s,a)$ and update the transition kernels using (9)

  Set $\hat{\mathcal{V}}_{(r+1),H+1}(s) = 0$ for all $s \in \mathcal{S}$ and broadcast it to all the clients

  for $h = H$ ***to*** 1 **do**

    for *agent* $i = 1$ ***to*** M ***in parallel*** **do**

      Compute $\hat{\mathcal{Q}}^i_{(r+1),h}$ using (10)

      Send $n^i_{(r+1),h}, \widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}_{(r+1),h+1}$, $\widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}^2_{(r+1),h+1}$, and $\hat{\mathcal{Q}}^i_{(r+1),h}$ to the central server

    Compute $N_{(r+1),h}$, $\hat{\mathcal{Q}}_{(r+1),h}$, $\hat{\mathcal{V}}_{(r+1),h}(s)$, and $\pi_{(r+1),h}$ using (6), (11), (13), and (14) and broadcast them to all the clients

  Set $r = r + 1$

necessary to compute a Bernstein-like bonus function $b_{(r+1),h}(s,a)$ (see (39) in Appendix for an exact expression). The aggregated Q-value is computed as

$$\hat{\mathcal{Q}}_{(r+1),h}(s,a) := \min\big([\mathcal{T}^{\omega}_{(r+1),h} + b_{(r+1),h}](s,a), H\big)\,, \quad (11)$$

with $\mathcal{T}^{\omega}_{(r+1),h}(s,a) = \sum_{i=1}^{M} \omega^i_{(r+1),h}(s,a)\hat{\mathcal{Q}}^i_{(r+1),h}(s,a)$, and

$$\omega^i_{(r+1),h}(s,a) := \frac{n^i_{(r+1),h}(s,a)}{N_{(r+1),h}(s,a)}\,. \quad (12)$$

Finally, the central server updates the value function and policy according to the equations

$$\hat{\mathcal{V}}_{(r+1),h}(s) := \max_{a \in \mathcal{A}} \hat{\mathcal{Q}}_{(r+1),h}(s,a)\,, \quad (13)$$

$$\pi_{(r+1),h}(s) := \arg\max_{a \in \mathcal{A}} \hat{\mathcal{Q}}_{(r+1),h}(s,a)\,. \quad (14)$$

These updated values are distributed to all agents, and the process continues for all $h = H, \ldots, 1$. Once $h = 1$ is reached, the new epoch $r + 1$ begins.

Communication Complexity. Our algorithmic design shares similarities with previous work on reinforcement learning with low switching cost (Bai et al., 2019; Zhang et al., 2020; Qiao et al., 2022). In particular, the number of times the local data collection policy changes—known as the switching cost—directly corresponds to the number of communication rounds in our framework, which we define as the communication complexity. In its simplest form, the doubling condition in this context can be expressed as:

$$\exists (s,a,h) : N_{(r,\ell),h}(s,a) > 2N_{(r),h}(s,a),\,, \quad (15)$$

where $N_{(r,\ell),h}(s,a) := \sum_{i=1}^{M} n^i_{(r,\ell),h}(s,a)$ represents the cumulative count across agents.

However, this condition cannot be directly verified in a federated learning setting, as the value of $N_{(r,\ell),h}(s,a)$ is not accessible to any individual agent. One potential solution is to use a weaker local doubling condition, as defined in (7). However, this approach results in communication complexity scaling linearly with the number of agents M, which is impractical for large-scale federated learning environments. Instead, we propose to construct an estimate of the global counter $\hat{N}^i_{(r,\ell),h}(s,a)$ to serve as a plug-in estimate on the left-hand side of (15). This is the core idea behind the

condition in (8). While such estimates may be inaccurate during the initial stages of training, they become reliable once the number of visits exceeds a threshold $\nu(\delta, T) = \widetilde{\mathcal{O}}(\varepsilon_{\mathsf{p}} THM + M)$, defined in (57). At that point, $\hat{N}^i_{(r,\ell),h}(s,a)$ can be effectively used as a plug-in estimate. Using this approach, we establish a bound on the communication complexity of `Fed-UCBVI`.

Lemma 4.1 (Communication Complexity)**.** *With probability at least $1-\delta$, the number of communication rounds of `Fed-UCBVI` is bounded by*

$$\mathfrak{C}(T) \leq \mathcal{O}\big(|\mathcal{S}||\mathcal{A}|H \log T + M|\mathcal{S}||\mathcal{A}|H \log\log T + M|\mathcal{S}||\mathcal{A}|H \log(1+\varepsilon_{\mathsf{p}} T)\big) \,,$$

where logarithmic dependence in $|\mathcal{S}|, |\mathcal{A}|$, H, $1/\delta$ and M is ignored.

Sketch of the proof: To prove the result, we consider a fixed triplet (s,a,h) and count how many synchronizations this triplet can trigger. Let $k^{\min}_{s,a,h}$ represent the index of the last round where $N_{(r),h}(s,a) \leq \nu(\delta,T)$. To bound the number of synchronizations that occur between the first round and round $k^{\min}_{s,a,h}$, note that agents send an abort signal only when their local visit count of (s,a) at time h has doubled. This can happen at most $\log_2(\nu(\delta,T))$ times for an individual agent, and for all agents combined, the total is upper bounded by $M\log_2(\nu(\delta,T))$.

Next, we bound the number of synchronizations between round $k^{\min}_{s,a,h}$ and the final round. By applying a Bernstein-type concentration inequality, we can show that the synchronization rule (8) implies the equivalent of (15), although with a coefficient of 8/7 instead of 2 on the right-hand side. Using a similar argument as above, we obtain $\mathcal{O}(\log(M))$ synchronizations triggered by a single state-action-step triplet. We complete the proof by summing these bounds over all (s,a,h) and using the expression of $\nu(\delta,T)$. □

A complete proof of Lemma 4.1 is provided in Appendix E. Importantly, we observe that, in the homogeneous setting, the linear dependence on M vanishes. Moreover, we can estimate the communication cost, i.e., the number of bits exchanged, by noting that in each communication round, each agent transmits objects of size at most $|\mathcal{S}||\mathcal{A}|H$.

Computational and Space Complexity. First, we remark that, at all times, agents store objects of size $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|H)$. At every episode, agents perform $\mathcal{O}(1)$ operations, while they perform $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|H)$ operations at communication times. By Lemma 4.1, we deduce that the computation complexity of this algorithm is $\mathcal{O}(T+|\mathcal{S}|^3|\mathcal{A}|^2H^2\log T + M|\mathcal{S}||\mathcal{A}|H\log\log T + M|\mathcal{S}||\mathcal{A}|H\log(1+\varepsilon_{\mathsf{p}}T))$ for all T episodes.

Regret Bound. We now state our main result, which bounds the federated regret of `Fed-UCBVI`.

Theorem 4.1. *With probability at least $1-\delta$, the following bound on the regret of `Fed-UCBVI` holds*

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(\sqrt{H^3|\mathcal{S}||\mathcal{A}|T/M} + H^3|\mathcal{S}|^2|\mathcal{A}|\right) + \widetilde{\mathcal{O}}\left(TH(H\varepsilon_{\mathsf{p}} + \varepsilon_{\mathsf{r}})\right) \,.$$

We give a sketch of the proof below, and postpone the detailed proof to Appendix D.

In the homogeneous setting, where $\varepsilon_{\mathsf{p}} = \varepsilon_{\mathsf{r}} = 0$, we recover the expected linear speedup in number of agents and achieve a minimax optimal regret bound up to logarithmic factors (see Table 1 for comparisons). In contrast, in the heterogeneous setting, an additional term, that scales linearly with the degree of heterogeneity, emerges. We show in Lemma F.9 in Appendix that this is expected, and comes from the fact that, in some cases, a policy optimal for one agent is suboptimal by at least $\varepsilon_{\mathsf{p}}H^2$ for another agent. This illustrates the trade-off involved in cooperation between heterogeneous agents: if the degree of heterogeneity is too large, cooperation can become counterproductive.

Sketch of the proof: As a first step of the proof, we reduce the problem of minimizing the federated regret (4) to the problem of minimizing a *common* regret. We introduce the common MDP $\mathcal{M}^{\mathsf{c}}$ as follows

$$\mathcal{M}^{\mathsf{c}} := (\mathcal{S}, \mathcal{A}, H, \{\mathsf{r}^{\mathsf{c}}_h := \tfrac{1}{M}\textstyle\sum_{i=1}^M \mathsf{r}^i_h\}_h, \{\mathsf{P}^{\mathsf{c}}_h\}_h) \,, \quad (16)$$

where $\{\mathsf{P}^{\mathsf{c}}_h\}_h$ is defined in **A**-1. We set $\mathcal{V}^{\mathsf{c},\pi}_h$ and $\mathcal{V}^{\mathsf{c},\star}_h$ the value-function of a policy π and optimal value-function in $\mathcal{M}^{\mathsf{c}}$. The common regret is defined as

$$\mathfrak{R}^{\mathsf{c}}(T) := \frac{1}{M}\sum_{t=1}^T\sum_{i=1}^M \mathcal{V}^{\mathsf{c},\star}_1(s^i_{t,1}) - \mathcal{V}^{\mathsf{c},\pi_t}_1(s^i_{t,1}) \,. \quad (17)$$

Adapting the performance-difference lemma of Russo (2019) under **A**-1, it may be shown that

$$\mathfrak{R}(T) = \max_\pi \frac{1}{M}\sum_{t=1}^T\sum_{i=1}^M \mathcal{V}^{i,\pi}_1(s^i_{t,1}) - \mathcal{V}^{i,\pi_t}_1(s^i_{t,1}) \leq \mathfrak{R}^{\mathsf{c}}(T) + 2T\varepsilon_{\mathsf{P}} H^2 + 2T\varepsilon_{\mathsf{r}} H \,.$$

As shown in Lemma F.9, the scaling $\mathcal{O}(T(\varepsilon_{\mathsf{p}}H^2 + \varepsilon_{\mathsf{r}}H))$ with H^2 is unavoidable.

The remainder of the proof involves three key steps, outlined below. The first step focuses on estimating the common transition kernel and introduces the primary technical innovations of this work. It also provides justification of **A**-1.

Step 1: Estimation of the common transition kernel. First, we prove that the weighted average kernel,

$$\widehat{\mathsf{P}}_{(r),h}(s'|s,a) := \sum_{i=1}^{M} \omega^i_{(r),h}(s,a) \cdot \widehat{\mathsf{P}}^i_{(r),h}(s'|s,a) \;,$$

where weights are defined in (12), forms a well-defined (biased) estimator of the common transition kernel $\mathsf{P}^{\mathsf{c}}_h$ using data from all agents. Importantly, neither the agents nor the CS have direct access to this quantity.

The analysis of $\widehat{\mathsf{P}}_{(r),h}$, under **A**-1 poses significant challenges compared to both the generative model setting and the case involving homogeneous agents. To illustrate, the kernel can be reformulated as follows, incorporating all samples from the agents:

$$\widehat{\mathsf{P}}_{(r),h}(s'|s,a) = \tfrac{1}{N_{(r),h}(s,a)} \sum_{i=1}^{M} n^i_{(r),h}(s,a,s') \;.$$

In the homogeneous scenario, where $\varepsilon_{\mathsf{p}} = 0$, as explored in prior work (Zheng et al., 2024, 2025), the estimate is derived from an i.i.d. sequence of categorical random variable samples from $\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)$, simplifying the analysis. Moreover, within the generative model framework, such as in (Jin et al., 2022; Wang et al., 2024), we can ensure an equal sample count from each agent's transition kernel P^i_h, resulting in $\widehat{\mathsf{P}}_{(r),h}$ as a simple mean of independent biased estimates of the common kernel.

However, in our setting, the estimator $\widehat{\mathsf{P}}_{(r),h}$ incorporates a random and non-stationary number of samples from each agent, making standard techniques of conditioning on a total sample size $N_{(r),h}(s,a)$ inapplicable. Using union-bound arguments to account for the variability in sample sizes across agents results in an exponential number of configurations with respect to M, constraining any possibility of linear speed-up.

Using **A**-1, every kernel P^i_h is a mixture of $\mathsf{P}^{\mathsf{c}}_h$ and $\mathsf{P}^{\mathsf{ind},i}_h$. The samples obtained by agent i as a mixture of samples coming from the two latter kernels: sample $s^i_{t,h}$ is with probability $1-\varepsilon_{\mathsf{p}}$ generated from $\mathsf{P}^{\mathsf{c}}_h$, and with probability ε_{p} from $\mathsf{P}^{\mathsf{ind},i}_h$. We define a *virtual estimate of the common kernel*, $\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}$, for each communication round r, representing the estimate we would have obtained if all samples were drawn solely from $\mathsf{P}^{\mathsf{c}}_h$. This estimate is subject to a bias resulting from the heterogeneity.

$$\left\| (\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h})(\cdot|s,a) \right\|_1 = \widetilde{\mathcal{O}}\left(\varepsilon_{\mathsf{P}} + \tfrac{1}{N}\right) \;, \qquad (18)$$

where $N = N_{(r),h}(s,a)$, that holds for any $(r,s,a,h) \in [\mathfrak{C}(T)] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

Step 2: Optimism. In our setting, our estimates are not optimistic due to the presence of heterogeneity; however, we can show the analog of the required properties $\hat{\mathcal{V}}_{(r),h}(s) \geq \mathcal{V}^{\mathsf{c},\star}_h(s) - (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H + 1 - h))$, for any r and $(s,h) \in \mathcal{S} \times [H]$. The key ingredients are concentration inequalities, an inequality (18) and Lemma 14 of (Zhang et al., 2021); see also Lemma D.1 in Appendix. The proof is carried out by induction on h. Applying the update rule (20), combined with a simple rearranging of the terms, yields

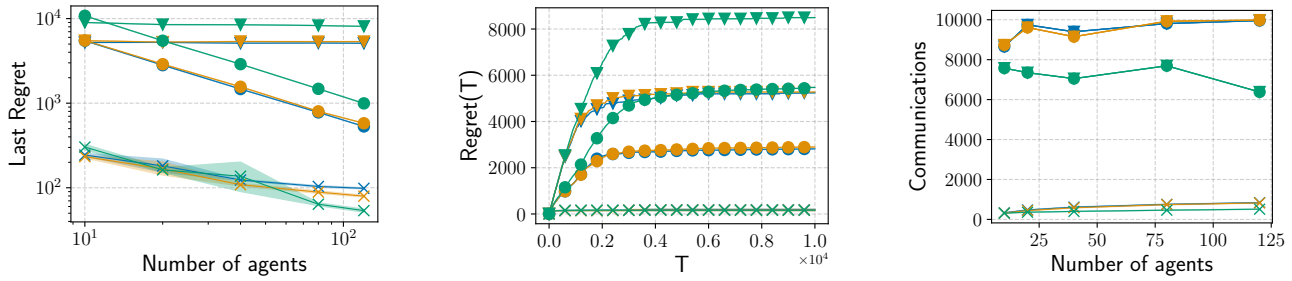
$$\begin{aligned}
\hat{\mathcal{Q}}_{(r),h}(s,a) \geq \mathcal{Q}^{\mathsf{c},\star}_h(s,a) &+ \underbrace{(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h)\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)}_{\textbf{(IV)}\text{: concentration error}} \\
&+ \underbrace{\sum_{i=1}^{M} \omega^i_{(r),h}(s,a)\widehat{\mathsf{r}}^i_h(s,a) - \frac{1}{M}\sum_{i=1}^{M} \mathsf{r}^i_h(s,a)}_{\textbf{(I)}\text{: reward heterogeneity error}} \\
&+ \underbrace{\widehat{\mathsf{P}}_{(r),h}(\hat{\mathcal{V}}_{(r),h+1}(s,a) - \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a))}_{\textbf{(II)}\text{: correction error}} \\
&+ \underbrace{(\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h})\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)}_{\textbf{(III)}\text{: transition heterogeneity error}} + b_{(r),h}(s,a) \;.
\end{aligned}$$

Terms (**II**) and (**IV**), which represent the correction and the concentration errors, are standard and are controlled using respectively induction hypothesis, Lemma D.1 and standard deviation inequalities. We control (**I**) by applying **A**-2 and noticing that the convex combination of $\widehat{\mathsf{r}}^i_h(s,a)$ is also a convex combination of the true rewards $\mathsf{r}^i_h(s,a)$. Finally, to control (**III**) we combine Holder's inequality and inequality (18). An appropriate choice of the exploration bonus concludes the statement.

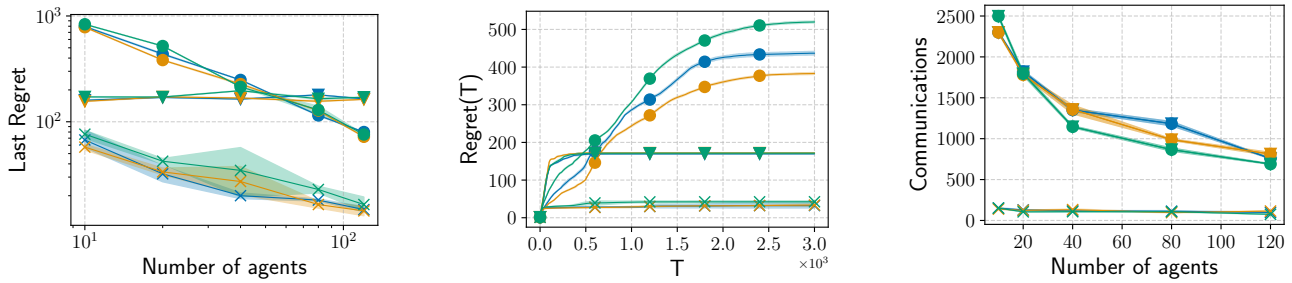
Step 3: Bounding the regret. For each quantity indexed by the number of communication rounds r (e.g. $\hat{\mathcal{V}}_{(r),h}$), we introduce a corresponding quantity indexed by the episode number t (e.g. $\hat{\mathcal{V}}_{t,h}$), defined as the value of the former at the last communication round before t (see (27) in Appendix for formal definitions). Next, following the approach of Azar et al. (2017), we define $\delta^i_{t,h} = \hat{\mathcal{V}}_{t,h}(s^i_{t,h}) - \mathcal{V}^{\mathsf{c},\pi_t}_1(s^i_{t,h})$ and analyze this term independently

$$\begin{aligned}
\delta^i_{t,h} \leq \delta^i_{t,h+1} &+ \underbrace{[\widehat{\mathsf{P}}_{t,h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}]\hat{\mathcal{V}}_{t,h+1}(s^i_{t,h},a^i_{t,h})}_{\textbf{(A)}\text{: heterogeneity error}} + \zeta^i_{t,h} \\
&+ \underbrace{[\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h} - \mathsf{P}^{\mathsf{c}}_h]\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\star}_{h+1}\right](s^i_{t,h},a^i_{t,h})}_{\textbf{(B)}\text{: correction error}} + 2\varepsilon_{\mathsf{r}} \\
&+ \underbrace{[\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h} - \mathsf{P}^{\mathsf{c}}_h]\mathcal{V}^{\mathsf{c},\star}_{h+1}(s^i_{t,h},a^i_{t,h})}_{\textbf{(C)}\text{: concentration error}} + b_{t,h}(s^i_{t,h},a^i_{t,h}) \\
&+ \underbrace{[\mathsf{P}^{\mathsf{c}}_h - \mathsf{P}^i_h]\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h},a^i_{t,h})}_{\textbf{(D)}\text{: heterogeneity error}} \;,
\end{aligned}$$

where $\zeta^i_{t,h}$ is a martingale term defined in (54). The analysis of (**C**) and $\zeta^i_{t,h}$ is standard in the literature.



(a) GridWorld, $T = 3 \cdot 10^4$ (b) GridWorld (c) GridWorld, $T = 3 \cdot 10^4$



(d) Synthetic, $T = 3 \cdot 10^3$ (e) Synthetic (f) Synthetic, $T = 3 \cdot 10^3$

Figure 1: Comparaison of different properties of `Fed-UCBVI`(represented by crosses), `FedQ-Bernstein` (represented by circles) and `FedQ-Advantage` (represented by triangles) in two environments: `synthetic` (below) and `GridWorld` (above). From left to right: plot of the common regret *(lower is better)*, as a function of M for different ε_{p} in a log-log scale; plot of the common regret *(lower is better)* for $M = 20$ agents as a function of T for different ε_{p}; plot of the number of communication *(lower is better)* as a function of M for different ε_{p}.

To bound (**A**) we employ a combination of (18) and Holder's inequality. The bound on (**D**) also combines Holder's inequality and Lemma F.1. The standard recursion argument concludes the proof. □

5 EXPERIMENTS

In this section, we study the empirical performance of `Fed-UCBVI`, and compare it with the `FedQ-Bernstein` (Zheng et al., 2024) and `FedQ-Advantage` (Zheng et al., 2025) algorithms on two environments.[1]

Environments. We consider two environments specifically designed to satisfy **A**-1 and **A**-2. In both environments, transitions are defined using two distinct kernels: with probability $1 - \varepsilon_{\mathsf{p}}$, the agent follow the global kernel, and with probability ε_{p}, it follows an individualized kernel. The first environment is based on `GridWorld` (Domingues et al., 2021a), where the agent navigates a grid to reach a target. Upon reaching the target, the agent receives a reward of +1; otherwise, the reward is 0. At each step, the agent selects one of four possible directions (up, down, left, or right). Under the global transition kernel, the agent moves to the intended square with a probability of 0.8, and to a random neighboring square with the remaining probability. In the individual transition kernels, the agent's movement to neighboring squares follows a probability distribution unique to each agent. We use a 3×3 grid with a wall located at coordinate $(1, 1)$, resulting in $|\mathcal{S}| = 8$ possible states. The planning horizon is set to $H = 10$, with the agent starting at coordinate $(0, 0)$ and aiming to reach the target at $(2, 2)$.

The second environment is a `synthetic` setting, modeled after Zheng et al. (2024), with $|\mathcal{S}| = 5$, $|\mathcal{A}| = 5$, and $H = 5$. All agents share the same reward function $\mathsf{r}_h(s, a)$, with rewards drawn uniformly from $[0, 1]$ for each $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. For each s, a, h, the common and individual transition kernels are drawn uniformly at random from the $|\mathcal{S}|$-dimensional simplex.

In all results, we report the common regret instead of the federated regret to simplify computations. Experiments were conducted on a computer with an Intel Xeon 6534 and 196GB RAM. We report the average over 5 runs and the standard deviation in all the plots. The code is provided in the supplementary material.

`Fed-UCBVI` has linear speed-up. In Figures 1a and 1d, we evaluate the regret after T iterations of training with varying numbers of agents M across different levels of heterogeneity. As shown in Theorem 4.1, the regret decreases as M increases. Notably, this trend persists even in high-heterogeneity settings, highlighting the robust empirical performance

[1] Our code is available online on GitHub: https://github.com/Labbi-Safwan/Fed-UCBVI

of our approach. This phenomenon also holds for `FedQ-Bernstein`. However, `FedQ-Advantage` does not have linear speed-up.

Impact of Heterogeneity. In Figures 1b and 1e, we present the regret of `Fed-UCBVI` for various values of ε_{p}. `Fed-UCBVI`'s regret is significantly lower than that of `FedQ-Bernstein` and `FedQ-Advantage`, reflecting similar performance gaps as observed in the single-agent setting. Moreover, as predicted by our theoretical analysis, increasing ε_{p} only incurs a slight increase in `Fed-UCBVI`'s regret, due to the additional term scaling linearly with T.

`Fed-UCBVI`'s communication complexity is small. In Figures 1c and 1f, we observe that the communication complexity of `Fed-UCBVI` is significantly lower than the number of iterations T and increases only marginally with the number of agents M. This aligns with the results of Lemma 4.1. In contrast, `FedQ-Bernstein` and `FedQ-Advantage` exhibit very similar and consistently high communication complexity. The reduced communication in `Fed-UCBVI` results from our novel method for triggering communication rounds based on local estimates of global counters, validating the effectiveness of this approach.

6 CONCLUSION

In this paper, we presented `Fed-UCBVI`, a federated reinforcement learning method based on a new aggregation strategy that reduces communication cost and handles heterogeneous agents. We introduced a novel measure of heterogeneity, under which we provide a formal analysis of `Fed-UCBVI`'s regret, showing that it nearly matches minimax optimal regret bounds. To our knowledge, this is the first federated regret analysis with guarantees in heterogeneous environments. Furthermore, our method provably removes the linear dependence of the communication complexity on $M \log T$. A promising direction for future work is to reduce the communication cost further, by developing new methods that correct for heterogeneity.

ACKNOWLEDGEMENTS

The work of S. Labbi, L. Mancini and P. Mangold has been supported by Technology Innovation Institute (TII), project Fed2Learn. The work of D.Tiapkin has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF. The work of E. Moulines has been partly funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR.

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

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

Chen, Y., Zhang, X., Zhang, K., Wang, M., and Zhu, X. (2023). Byzantine-robust online and offline distributed reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3230–3269. PMLR.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30.

Domingues, O. D., Flet-Berliac, Y., Leurent, E., Ménard, P., Shang, X., and Valko, M. (2021a). rlberry - A Reinforcement Learning Library for Research and Education.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021b). Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. (2021c). Kernel-based reinforcement learning: A finite-time analysis. In *International Conference on Machine Learning*, pages 2783–2792. PMLR.

Doob, J. (1953). *Stochastic Processes*. Probability and Statistics Series. Wiley.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? *Advances in neural information processing systems*, 31.

Jin, H., Peng, Y., Yang, W., Wang, S., and Zhang, Z. (2022). Federated reinforcement learning with environment heterogeneity. In *International Conference*

on Artificial Intelligence and Statistics, pages 18–37. PMLR.

Jonsson, A., Kaufmann, E., Ménard, P., Darwiche Domingues, O., Leurent, E., and Valko, M. (2020). Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263.

Lan, G., Han, D.-J., Hashemi, A., Aggarwal, V., and Brinton, C. G. (2024). Asynchronous federated reinforcement learning with policy gradient updates: Algorithm design and convergence analysis. *arXiv preprint arXiv:2404.08003*.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. (2021). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17762–17776.

Mangold, P., Samsonov, S., Labbi, S., Levin, I., Alami, R., Naumov, A., and Moulines, E. (2025). Scafflsa: Taming heterogeneity in federated linear stochastic approximation and td learning. *Advances in Neural Information Processing Systems*, 37:13927–13981.

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Qi, J., Zhou, Q., Lei, L., and Zheng, K. (2021). Federated reinforcement learning: techniques, applications, and open challenges. *Intelligence & Robotics*, 1(1).

Qiao, D., Yin, M., Min, M., and Wang, Y.-X. (2022). Sample-efficient reinforcement learning with loglog (t) switching cost. In *International Conference on Machine Learning*, pages 18031–18061. PMLR.

Ross, S. and Bagnell, D. (2010). Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32.

Salgia, S. and Chi, Y. (2024). The sample-communication complexity trade-off in federated q-learning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 39694–39747. Curran Associates, Inc.

Sobel, M. J. (1982). The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Talebi, M. S. and Maillard, O.-A. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805. PMLR.

Tiapkin, D., Belomestny, D., Calandriello, D., Moulines, E., Munos, R., Naumov, A., Perrault, P., Tang, Y., Valko, M., and Menard, P. (2023). Fast rates for maximum entropy exploration. In *International Conference on Machine Learning*, pages 34161–34221. PMLR.

Wang, M., Yang, P., and Su, L. (2024). On the convergence rates of federated q-learning across heterogeneous environments. In *International Workshop on Federated Foundation Models in Conjunction with NeurIPS 2024*.

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR.

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

Zhang, Z., Chen, Y., Lee, J. D., and Du, S. S. (2024b). Settling the sample complexity of online reinforcement learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5213–5219. PMLR.

Zhang, Z., Ji, X., and Du, S. (2021). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207.

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*.

Zheng, Z., Zhang, H., and Xue, L. (2025). Federated q-learning with reference-advantage decomposition: Almost optimal regret and logarithmic communication cost. In *The Thirteenth International Conference on Learning Representations*.

Zhuo, H. H., Feng, W., Lin, Y., Xu, Q., and Yang, Q. (2019). Federated deep reinforcement learning. *arXiv preprint arXiv:1901.08277*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Answer: **Yes**
   Justification: **The mathematical setting, the assumptions, and the algorithm are extensively described in Section 3 and Section 4.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Answer: **Yes**
   Justification: **All properties associated with the algorithm are detailed in Section 4, along with detailed proof.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Answer: **Yes**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Answer: **Yes**
   Justification: **All the assumptions required to conduct the proof are described in Section 3.**
   (b) Complete proofs of all theoretical results.
   Answer: **Yes**
   Justification: **A proof sketch for all the theorems is provided in Section 4, with the full proofs included in the appendix.**
   (c) Clear explanations of any assumptions.
   Answer: **Yes**
   Justification: **The assumptions are clearly stated and explained in Section 3.2.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   Answer: **Yes, our code is available online on GitHub:** https://github.com/Labbi-Safwan/Fed-UCBVI
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Answer: **Yes**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   Answer: **Yes**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   Answer: **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Answer: **Yes**
   (b) The license information of the assets, if applicable.
   Answer: **Yes**
   (c) New assets either in the supplemental material or as a URL, if applicable.
   Answer: **Yes**
   (d) Information about consent from data providers/curators.
   Answer: **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   Answer: **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   Answer: **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   Answer: **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   Answer: **Not Applicable**

Supplementary Materials

A NOTATION

For clarity, we summarize here the notations that we use

Symbols	Meaning	Definition
$\mathfrak{C}(T)$	Number of communication rounds performed in average	Section 3
$\mathfrak{R}(T)$	Federated regret of the algorithm	Equation (4)
$\mathcal{S}$	State space	Section 3
$\mathcal{A}$	Action space	Section 3
M	Number of agents	Section 3
T	Total number of collected episodes per agent	Section 3
H	Length of an episode	Section 3
$R_{\max}$	Maximum number of communication rounds	Equation (58)
P_h^i	Transition kernel at step h of agent i	Section 3
$\mathsf{P}_h^{\mathsf{c}}$	Common transition kernel at step h	**A**-1
$\mathsf{P}_h^{\mathsf{ind},i}$	Individual transition kernel at step h	**A**-1
ε_{p}	Degree of heterogeneity on the transition kernels	**A**-1
ε_{r}	Degree of heterogeneity on the rewards	**A**-2
r_h^i	Reward at step h of agent i	Section 3
$\mathsf{r}_h^{\mathsf{c}}$	Reward function of the common MDP	Equation (16)
$\mathcal{Q}_h^{i,\pi}$	Q-function of a policy π at step h of agent i	Equation (2)
$\mathcal{V}_h^{i,\pi}$	Value function of a policy π at step h in the i-th MDP	Equation (2)
$\mathcal{Q}_h^{i,\star}$	Optimal Q-function at step h of agent i in the i-th environment	Equation (3)
$\mathcal{V}_h^{i,\star}$	Optimal value function at step h of agent i in the i-th environment	Equation (3)
$\nu(\delta, T)$	Threshold for defining the condition on initiating the aggregation signal	Equation (57)
$\widehat{\mathsf{P}}^i_{(r),h}$	Estimated transition kernel during the round r by agent i at step h	Equation (9)
$\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}$	Virtual estimate of common transition kernel by agent i at step h	Equation (24)
$\widehat{\mathsf{r}}_h^i$	Estimated reward at step h of agent i	`Fed-UCBVI`
$n^i_{(r,\ell),h}$	Local counter of the cumulative number of visits at the level of agent i	`Fed-UCBVI`
$N_{(r),h}$	Global counter of the cumulative number of visits over all the agents	Equation (6)
$\check{N}^i_{(r,\ell),h}$	Local estimator of agent i of the *true* cumulative number of visits	`Fed-UCBVI`
$b_{(r),h}$	Bonus function used in round r and step H	Equation (39)
$\hat{\mathcal{Q}}^i_{(r),h}(s,a)$	Estimator of the Q-function at the level of agent i	Equation (10)
$\widehat{\mathsf{P}}_{(r),h}$	Weighted average of $\{\widehat{\mathsf{P}}^i_{(r),h}\}_i$ during the round r at step h	Equation (24)
$\hat{\mathcal{Q}}_{(r),h}(s,a)$	Global estimator of the Q-function	Equation (11)
$\hat{\mathcal{V}}_{(r),h}(s)$	Global estimator of the value function	Equation (13)
$\mathrm{Var}_{\widehat{\mathsf{P}}_{(r),h}}(f)(s,a)$	Variance of a function f with respect to $\widehat{\mathsf{P}}_{(r),h}(\cdot\|s,a)$	Equation (19)
$\mathrm{Var}_{\widehat{\mathsf{P}}^c_{(r),h}}(f)(s,a)$	Variance of a function f with respect to $\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot\|s,a)$	Equation (19)

Table 2: Summary of the notations.

Let $(X, \mathcal{X})$ be a measurable space. For any probability measures P and Q on $(X, \mathcal{X})$, and for any $f : X \to \mathbb{R}$ we define

$$\mathsf{P}f := \mathbb{E}_{s\sim\mathsf{P}}[f(s)] \,, \quad \mathrm{Var}_{\mathsf{P}}\, f := \mathbb{E}_{s\sim\mathsf{P}}[(f(s) - \mathsf{P}f)^2] \,. \tag{19}$$

For any probability measures P and Q on $(X, \mathcal{X})$, the Kullback-Leibler divergence $\mathrm{KL}(\mathsf{P}\|\mathsf{Q})$ is given by

$$\mathrm{KL}\left(\mathsf{P}\middle\|\mathsf{Q}\right) := \begin{cases} \mathbb{E}_{\mathsf{P}}\left[\log \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}\mathsf{Q}}\right], & \mathsf{P} \ll \mathsf{Q} \,, \\ +\infty, & \text{otherwise} \quad . \end{cases}$$

Let A be an element of the σ-algebra $\mathcal{X}$. We define the indicator function of A as

$$\begin{aligned} \mathbb{1}_A(\cdot) \colon X &\longrightarrow \{0, 1\} \\ x &\longmapsto \begin{cases} 1 & \text{if } x \in A \,, \\ 0, & \text{otherwise} \quad . \end{cases} \end{aligned}$$

We define the indicator function of an element $x \in X$ as

$$\begin{aligned} \mathbb{1}_x(\cdot) \colon X &\longrightarrow \{0, 1\} \\ y &\longmapsto \begin{cases} 1 & \text{if } x = y \,, \\ 0, & \text{otherwise} \quad . \end{cases} \end{aligned}$$

We write $f(|\mathcal{S}|, |\mathcal{A}|, H, T, M) = \mathcal{O}(g(|\mathcal{S}|, |\mathcal{A}|, H, T, M, \delta))$ if there exists $S_0, A_0, H_0, T_0, \delta_0$ and a constant C such that for any $|\mathcal{S}| \geq S_0, |\mathcal{A}| \geq A_0, H \geq H_0, T \geq T_0$, and $\delta \leq \delta_0$, we have $f(|\mathcal{S}|, |\mathcal{A}|, H, T, M) \leq C \cdot g(|\mathcal{S}|, |\mathcal{A}|, H, T, M, \delta)$. We say that $f(|\mathcal{S}|, |\mathcal{A}|, H, T, M) = \widetilde{\mathcal{O}}(g(|\mathcal{S}|, |\mathcal{A}|, H, T, M, \delta))$ if in the previous bound C is a poly-logarihmic function with respect to the variables $|\mathcal{S}|, |\mathcal{A}|, H, T, M, \delta$.

For $a \in \mathbb{N}$, define $[a]$ as the set of all natural numbers from 1 to a:

$$[a] := \{k \in \mathbb{N} \mid 1 \leq k \leq a\}.$$

Additionally, for $(a, b) \in \mathbb{N} \times \bar{\mathbb{N}}$, where $\bar{\mathbb{N}} = \mathbb{N} \cup \{+\infty\}$, such that $a \leq b$, define the set $[\![a, b]\!]$ as the set of all natural numbers between a and b, inclusive:

$$[\![a, b]\!] := \{k \in \mathbb{N} \mid a \leq k \leq b\}.$$

B PSEUDO CODE

For clarity of exposition, we provide the complete pseudo-code of the server-side and client-side algorithms in Algorithm 2 and Algorithm 3.

C CONCENTRATION EVENTS

Before we proceed, let us define several essential quantities.

Change of epoch notation We notice that the set of all regular episodes $t \in [T]$ is separated into a sequence of different *random* epochs $\mathrm{E}_1, \mathrm{E}_2, \ldots$. To define them properly, let us define the epoch-changing timestamps as follows

$$T_1 := 0, \quad T_{r+1} := \min\{t > T_r \mid \mathrm{Sync}_r(t) = \texttt{True}\}. \tag{20}$$

where the epoch-switching predicate is defined as

$$\mathrm{Sync}_r(t) = \begin{cases} \exists i \in [M] : n^i_{(r,\ell),h}(s^i_{t,h}, a^i_{t,h}) \geq 2n^i_{(r),h}(s^i_{t,h}, a^i_{t,h}) & \text{if } N_{(r),h}(s^i_{t,h}, a^i_{t,h}) < \nu(\delta, T) \\ \exists i \in [M] : \hat{N}^i_{(r,\ell),h}(s^i_{t,h}, a^i_{t,h}) \geq 2N_{(r),h}(s^i_{t,h}, a^i_{t,h}) & \text{if } N_{(r),h}(s^i_{t,h}, a^i_{t,h}) \geq \nu(\delta, T) \end{cases} , \tag{21}$$

for $\ell = t - T_r - 1$ and $\nu(\delta, T)$ is defined in (57). In particular, this condition exactly corresponds to the synchronization condition used by `Fed-UCBVI`. Then, the epoch E_r is defined as $\mathrm{E}_r := [\![T_r + 1; T_{r+1}]\!]$. In particular, for any $t \in [T]$, we define r_t as a unique index r such that $t \in \mathrm{E}_r$:

$$r_t = \min\{r \geq 1 \mid t > T_r\}. \tag{22}$$

Algorithm 2: `Fed-UCBVI (Central Server)`

Initialize: $t = 1$, $r = 1$, $N_{(1),h}(s,a) = 0$, $\hat{\mathcal{Q}}_{(1),h}(s,a) = \hat{\mathcal{V}}_{(1),h}(s) = H$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $\pi_{(1)} = \{\pi_{(1),h}\}_{1 \le h \le H}$ an arbitrary deterministic policy

while $t \le T$ **do**

Broadcast $\pi_{(r)} = \{\pi_{(r),h}\}_{1 \le h \le H}$, $\{N_{(r),h}\}_{1 \le h \le H}$, r, and t to all clients;

Wait until receiving the synchronization signal and an updated episode number t and forward the abortion signal to all clients;

Set $\hat{\mathcal{V}}_{(r+1),H+1}(s) = 0$ for all $s \in \mathcal{S}$ and send it to all clients;

for $h = H$ **to** 1 **do**

Receive $\{\hat{\mathcal{Q}}^i_{(r+1),h}\}_i$, $\{n^i_{(r+1),h}\}_i$,$\{\widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}_{(r+1),h+1}(s,a)\}_{s,a,i}$, and $\{\widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}^2_{(r+1),h+1}(s,a)\}_{s,a,i}$ from the different clients;

for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

Compute $N = N_{(r+1),h}(s,a) = \sum_{i=1}^M n^i_{(r+1),h}(s,a)$;

Set $n^i = n^i_{(r+1),h}(s,a)$ for $i \in [M]$;

$$\text{Set } V := \mathrm{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a) = \frac{1}{N}\sum_{i=1}^M n^i \widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}^2_{(r+1),h+1}(s,a) - \left(\frac{1}{N}\sum_{i=1}^M n^i \widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}_{(r+1),h+1}(s,a)\right)^2$$

$$\text{Compute } b_{(r),h}(s,a) = \begin{cases} \frac{28\beta^\star(\delta)H + 11\beta^{\mathsf{c}}(\delta,N)}{N} + \sqrt{\frac{8\beta^\star(\delta)}{N} \cdot V}\,, & N \ge 2\,, \\ H\,, & N \le 1\,; \end{cases}$$

$$\text{Set } \hat{\mathcal{Q}}_{(r+1),h}(s,a) = \begin{cases} \min\left(\sum_{i=1}^M \frac{n^i}{N}\hat{\mathcal{Q}}^i_{(r+1),h}(s,a) + b_{(r+1),h}(s,a), H\right) & \text{if } N > 0\,, \\ H & \text{otherwise;} \end{cases}$$

for $s \in \mathcal{S}$ **do**

Compute $\hat{\mathcal{V}}_{(r+1),h}(s) = \max_{a \in \mathcal{A}} \hat{\mathcal{Q}}_{(r+1),h}(s,a)$;

Compute $\pi_{(r+1),h}(s) = \arg\max_{a \in \mathcal{A}} \hat{\mathcal{Q}}_{(r+1),h}(s,a)$;

Broadcast $\hat{\mathcal{V}}_{(r+1),h}$ to all clients;

Set $r = r + 1$;

Send a signal to inform the clients of the end of training.

Definitions First of all, let us recall that by **A**-1 the transition kernel P^i_h for the agent i is a mixture of common kernel $\mathsf{P}^{\mathsf{c}}_h$ and individual kernel $\mathsf{P}^{\mathsf{ind},i}_h$, thus any sample $s^i_{t,h+1} \sim \mathsf{P}^i_h(s^i_{t,h}, a^i_{t,h})$ for $(t,h,i) \in [T] \times [H] \times [M]$ can be represented via the following experiment

$$s^i_{t,h+1} = \begin{cases} s^{\mathsf{c},i}_{t,h+1} \sim \mathsf{P}^{\mathsf{c}}_h(s^i_{t,h}, a^i_{t,h})\,, & \xi^i_{t,h} = 0\,, \\ s^{\mathsf{ind},i}_{t,h+1} \sim \mathsf{P}^{\mathsf{ind},i}_h(s^i_{t,h}, a^i_{t,h})\,, & \xi^i_{t,h} = 1\,, \end{cases} \tag{23}$$

where $\xi^i_{t,h} \sim \mathrm{Ber}(\varepsilon_{\mathsf{p}})$ is a choice of component of the mixture. Using this representation, we can define a *virtual estimate of the common kernel* for a step t as follows

$$\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(s'|s,a) := \frac{1}{N_{(r),h}(s,a)}\sum_{i=1}^M \sum_{t=1}^{T_r} \mathbb{1}_{(s,a,s')}(s^i_{t,h}, a^i_{t,h}, s^{\mathsf{c},i}_{t,h+1})\,, \tag{24}$$

where T_r is defined in (20). We emphasize that $\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}$ is never computed explicitly by the algorithm since the values of $\xi^i_{t,h}$ are never observed, however we are very interested in the analysis of it.

Algorithm 3: `Fed-UCBVI (i-th Client Side)`

Initialize: $n^i_{(1,0),h}(s,a)=0$, $\widehat{\mathsf{r}}^i_h(s,a)=0$, $n^i_{(1,0),h}(s,a,s')=0$ for all $(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$;
Compute $\nu(\delta,T)=14\varepsilon_{\mathsf{p}}THM+182M\beta^{\mathsf{c}}(\delta,T)$;
while *signal of end of training not received* **do**
  Receive $\{\pi_{(r),h}\}_{1\le h\le H}$, $(N_{(r),h})_{1\le h\le H}$, r, and t from the central server;
  Set $\ell=1$;
  Set $\hat{N}^i_{(r,0),h}(s,a)=N_{(r),h}(s,a)$ for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$;
  Set $n^i_{(r,0),h}(s,a)=n^i_{(r),h}(s,a)$ for all $(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$;
  while *no synchronization signal from central server and* $t\le T$ **do**
    synchronize = `False`;
    while *synchronize =* `False` **do**
      Collect a new trajectory $(s^i_{t,h},a^i_{t,h},\mathsf{r}^i_{t,h})_{1\le h\le H}$ using the policy $\pi_{(r)}$;
      for $h=1$ **to** H **do**
        Set $\widehat{\mathsf{r}}^i_h(s^i_{t,h},a^i_{t,h})=\mathsf{r}^i_{t,h}$;
        Set $n^i_{(r,\ell),h}(s,a)=n^i_{(r,\ell-1),h}(s,a)+1_{(s,a)}(s^i_{t,h},a^i_{t,h})$ for $(s,a)\in\mathcal{S}\times\mathcal{A}$;
        $n^i_{(r,\ell),h}(s,a,s')=n^i_{(r,\ell-1),h}(s,a,s')+1_{(s,a,s')}(s^i_{t,h},a^i_{t,h},s^i_{t,h+1})$ for $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$;
        Set $\hat{N}^i_{(r,\ell),h}(s,a)=\hat{N}^i_{(r,\ell-1),h}(s,a)+M1_{(s,a)}(s^i_{t,h},a^i_{t,h})$ for $(s,a)\in\mathcal{S}\times\mathcal{A}$;
        if $N_{(r),h}(s^i_{t,h},a^i_{t,h})<\nu(\delta,T)$ *and* $n^i_{(r,l),h}(s^i_{t,h},a^i_{t,h})\ge 2n^i_{(r),h}(s^i_{t,h},a^i_{t,h})$ **then**
          synchronize = `True`;
        else if $N_{(r),h}(s^i_{t,h},a^i_{t,h})\ge\nu(\delta,T)$ *and* $\hat{N}^i_{(r,l),h}(s^i_{t,h},a^i_{t,h})\ge 2N_{(r),h}(s^i_{t,h},a^i_{t,h})$ **then**
          synchronize = `True`;
      Set $\ell=\ell+1$ and $t=t+1$;
      ;
    Send an abortion signal and an episode number t to the central server;
  Set $n^i_{(r+1),h}(s,a)=n^i_{(r,\ell),h}(s,a)$ and $n^i_{(r+1),h}(s,a,s')=n^i_{(r,\ell),h}(s,a,s')$ for all $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$;
  Set $\widehat{\mathsf{P}}^i_{(r+1),h}(s'|s,a)=\begin{cases}\frac{n^i_{(r+1),h}(s,a,s')}{n^i_{(r+1),h}(s,a)} & \text{if }, n^i_{(r+1),h}(s,a)>0\\ \frac{1}{|\mathcal{S}|} & \text{else;}\end{cases}$
  for $h=H$ **to** 1 **do**
    Receive $\hat{\mathcal{V}}_{(r+1),h+1}$ from the central server;
    Compute $\hat{\mathcal{Q}}^i_{(r+1),h}(s,a)=\widehat{\mathsf{r}}^i_h(s,a)+\widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}_{(r+1),h+1}(s,a)$ for all $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$;
    Send $\hat{\mathcal{Q}}^i_{(r+1),h}$, $n^i_{(r+1),h}$,$\{\widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}_{(r+1),h+1}(s,a)\}_{s,a}$, and $\{\widehat{\mathsf{P}}^i_{(r+1),h}\hat{\mathcal{V}}^2_{(r+1),h+1}(s,a)\}_{s,a}$ to the central server.

Additionally, let us define the weighted average kernel

$$\widehat{\mathsf{P}}_{(r),h}(s'|s,a):=\sum_{i=1}^{N}\frac{n^i_{(r),h}(s,a)}{N_{(r),h}(s,a)}\widehat{\mathsf{P}}^i_{(r),h}(s'|s,a)=\frac{N_{(r),h}(s,a,s')}{N_{(r),h}(s,a)}\,, \tag{25}$$

where $N_{(r),h}(s,a)=\sum_{i=1}^{M}n^i_{(r),h}(s,a)$ was defined in (6), and $\widehat{\mathsf{P}}^i_{(r),h}$ was defined in (9) as

$$\widehat{\mathsf{P}}^i_{(r),h}(s'|s,a)=\begin{cases}\frac{n^i_{(r),h}(s,a,s')}{n^i_{(r),h}(s,a)} & \text{if } n^i_{(r),h}(s,a)>0\\ \frac{1}{|\mathcal{S}|} & \text{else}\end{cases}\,. \tag{26}$$

Notably, the kernel $\widehat{\mathsf{P}}_{(r),h}$ is never revealed to any agent or to a central server, but it is very useful in the analysis. Also, for any time t we define r_t as an index of the previous epoch. For convenience and ease of reading, we introduce the transition kernels and counters in the *regular timescale*

$$\widehat{\mathsf{P}}^i_{t,h}:=\widehat{\mathsf{P}}^i_{(r_t),h}\,,\quad \widehat{\mathsf{P}}_{t,h}:=\widehat{\mathsf{P}}_{(r_t),h}\quad n^i_{t,h}=n^i_{(r_t),h}\,,\text{ and }N^i_{t,h}=N^i_{(r_t),h}\,, \tag{27}$$

where r_t is defined in (22).

Let $\beta^{\mathrm{KL}}, \beta^{\mathsf{c}}, \beta^{\mathrm{Var}} \colon (0,1) \times \mathbb{N} \to \mathbb{R}_+$ and $\beta^\star, \beta, \beta^{\max} \colon (0,1) \to \mathbb{R}_+$ be some functions defined later on in Lemma C.1, and $R_{\max}$ be the maximal number of communications defined (58). We define the following favorable events

$$\mathcal{E}^{\mathrm{KL}}(\delta) := \left\{ \forall r \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad \mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(s,a) \middle\| \mathsf{P}^{\mathsf{c}}_h(s,a) \right) \le \frac{|\mathcal{S}| \beta^{\mathrm{KL}}(\delta, N_{(r),h}(s,a))}{N_{(r),h}(s,a)} \right\},$$

$$\mathcal{E}^{\mathsf{c}}(\delta) := \left\{ \forall r \in [R_{\max}], \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \left\| \widehat{\mathsf{P}}_{(r),h}(s,a) - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(s,a) \right\|_1 \le \frac{9}{8} \varepsilon_{\mathsf{p}} + \frac{11 \beta^{\mathsf{c}}(\delta, N_{(r),h}(s,a))}{N_{(r),h}(s,a)} \right\},$$

$$\mathcal{E}^\star(\delta) := \Bigg\{ \forall r \in [R_{\max}], \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad \left| [\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h] \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a) \right|$$
$$\le 1_{[\![2;+\infty]\!]}(N_{(r),h}(s,a)) \left(\sqrt{\frac{2 \operatorname{Var}_{\widehat{\mathsf{P}}^{c}_{(r),h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s,a) \beta^\star(\delta)}{N_{(r),h}(s,a) - 1}} + \frac{7 \beta^\star(\delta)}{N_{(r),h}(s,a) - 1} \right) + H 1_{[\![0;2]\!]}(N_{(r),h}(s,a)) \Bigg\},$$

$$\mathcal{E}^{\mathrm{Var}}(\delta) := \left\{ \forall t \in [T] : \sum_{(t' \ge 1, h \ge 1, i \ge 1)}^{(t,H,M)} \operatorname{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{i,\pi_{t'}}_{h+1})(s^i_{t',h}, a^i_{t',h}) \le \sqrt{2H^5 M t \beta^{\mathrm{Var}}(\delta,t)} + 3H^3 \beta^{\mathrm{Var}}(\delta,t) + H^2 M t \right\},$$

$$\mathcal{E}^{\mathrm{count}}(\delta) := \Bigg\{ \forall t \in [T], \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \forall i \in [M] : \quad |\widetilde{N}^M_{t,h}(s,a) - \hat{N}^i_{t,h}(s,a)|$$
$$\le \frac{2}{7} \hat{N}^i_{t,h}(s,a) + 2\varepsilon_{\mathsf{p}} THM + 26 M \beta^{\mathsf{c}}(\delta, T) \Bigg\},$$

$$\mathcal{E}(\delta) := \Bigg\{ \forall h \in [H] : \sum_{(t \ge 1, h' \ge h, i \ge 1)}^{(T,H,M)} \gamma_{h'-1} \left(\mathsf{P}^i_{h'} \left[\hat{\mathcal{V}}_{t,h'+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h'+1} \right] (s^i_{t,h'}, a^i_{t,h'}) - \left[\hat{\mathcal{V}}_{t,h'+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h'+1} \right] (s^i_{t,h'+1}) \right)$$
$$\le \sqrt{8 \mathrm{e}^2 H^2 \cdot THM \cdot \beta(\delta)}, \quad \gamma_h := \left(1 + \frac{1}{H} \right)^{H-h} , \text{ and}$$
$$\sum_{(t \ge 1, h' \ge h, i \ge 1)}^{(T,H,M)} \left(\mathsf{P}^i_{h'} \left[\hat{\mathcal{V}}_{t,h'+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h'+1} \right] (s^i_{t,h'}, a^i_{t,h'}) - \left[\hat{\mathcal{V}}_{t,h'+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h'+1} \right] (s^i_{t,h'+1}) \right) \le \sqrt{8H^2 \cdot THM \cdot \beta(\delta)} \Bigg\}.$$

We also introduce the intersection of these events, $\mathcal{G}(\delta) := \mathcal{E}^{\mathrm{KL}}(\delta) \cap \mathcal{E}^{\mathsf{c}}(\delta) \cap \mathcal{E}^\star(\delta) \cap \mathcal{E}^{\mathrm{Var}}(\delta) \cap \mathcal{E}^{\mathrm{count}}(\delta) \cap \mathcal{E}(\delta)$. We prove that for the right choice of the functions $\beta^{\mathrm{KL}}, \beta^{\mathsf{c}}, \beta^\star, \beta^{\mathrm{Var}}$, and β the above events hold with high probability.

Lemma C.1. *For any $\delta \in (0,1)$ and for the following choices of functions β,*

$$\beta^{\mathrm{KL}}(\delta, n) := \log(6|\mathcal{S}||\mathcal{A}|H/\delta) + \log\left(\mathrm{e}(1+n)\right), \qquad \beta^{\mathsf{c}}(\delta, n) := \log(6|\mathcal{S}||\mathcal{A}|H/\delta) + \log\left(6\mathrm{e}(2n+1)\right),$$
$$\beta^\star(\delta) := \log(12|\mathcal{S}||\mathcal{A}|H/\delta), \qquad \beta^{\mathrm{Var}}(\delta, t) := \log\left(24\mathrm{e}(2Mt+1)/\delta\right),$$
$$\beta(\delta) := \log(48H/\delta).$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \ge 1 - \delta/6, \quad \mathbb{P}[\mathcal{E}^{\mathsf{c}}(\delta)] \ge 1 - \delta/6, \quad \mathbb{P}[\mathcal{E}^\star(\delta)] \ge 1 - \delta/6, \quad \mathbb{P}[\mathcal{E}^{\mathrm{Var}}(\delta)] \ge 1 - \delta/6,$$
$$\mathbb{P}[\mathcal{E}^{\mathrm{count}}(\delta)] \ge 1 - \delta/6, \quad \mathbb{P}[\mathcal{E}(\delta)] \ge 1 - \delta/6.$$

In particular, $\mathbb{P}[\mathcal{G}(\delta)] \ge 1 - \delta$.

Proof. First, let us define an appropriate filtration for martingale and optional skipping-based arguments. A natural federated online filtration is defined as

$$\mathcal{F}^i_{t,h} = \sigma\left(\{ s^{i'}_{t',h'} a^{i'}_{t',h'} \}_{(t',h',i') \preceq (t,h,i)} \right), \tag{28}$$

where the order over triplets (t', h', i') is lexicographic. With respect to this filtration, for any fixed state-action-step triplet $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ we define the partial global counters that form a sequence of excursion times on an extended time $(t, i) \in \mathbb{N} \times [M]$ and a first extended timestamp to reach a particular partial counter value $j \in [T \cdot M]$

$$\widetilde{N}^i_{t,h}(s,a) = \sum_{(t',i') \preceq (t,i)} 1_{(s,a)}(s^i_{t,h}, a^i_{t,h}), \quad (t_{s,a,h,j}, i_{s,a,h,j}) := \min\{(t,i) \in \mathbb{N} \times [M] \mid \widetilde{N}^i_{t,h}(s,a) = j\}. \tag{29}$$

For a given time t, we also define $\psi_t := T_{r_t}$ representing the number of episodes visited before r_t. In particular, we have $N_{(r),h}(s,a) = \widetilde{N}^M_{T_r,h}(s,a)$.

Event $\mathcal{E}^{\mathrm{KL}}(\delta)$ To analyze it, we need first to represent the virtual estimate of the common transition kernel as follows

$$\begin{aligned}\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(s'|s,a) &= \frac{1}{N_{(r),h}(s,a)} \sum_{t=1}^{T_r} \sum_{i=1}^{M} 1_{(s,a)}(s^i_{t,h}, a^i_{t,h}) 1_{s'}(s^{\mathsf{c},i}_{t,h+1}) \\ &= \frac{1}{N_{(r),h}(s,a)} \sum_{j=1}^{N_{(r),h}(s,a)} 1_{s'}(s^{\mathsf{c},i_{s,a,h,j}}_{t_{s,a,h,j},h+1}). \end{aligned} \tag{30}$$

By the optional skipping argument (see, e.g., Doob, 1953, Chapter III, p. 145), the sampled states $\{\tilde{s}^{\mathsf{c}}_{s,a,h,j}\}_{j \in [TM]} := \{s^{\mathsf{c},i_{s,a,h,j}}_{t_{s,a,h,j},h+1}\}_{j \in [TM]}$, conditioned on the value of $N_{(r),h}(s,a)$, form an i.i.d. sequence of categorical random variables from the distribution $\mathsf{P}^{\mathsf{c}}_h(s,a)$. Thus, we have for any fixed $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ by Lemma F.4

$$\mathbb{P}\left[\exists r \geq 1 : \mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(s,a) \middle\| \mathsf{P}^{\mathsf{c}}_h(s,a)\right) \leq \frac{\log(6|\mathcal{S}||\mathcal{A}|H/\delta) + |\mathcal{S}| \log(\mathrm{e}(1+n))}{N_{(r),h}(s,a)}\right] \leq \frac{\delta}{6|\mathcal{S}||\mathcal{A}|H}.$$

By a union bound argument and noticing that $\log(6|\mathcal{S}||\mathcal{A}|H/\delta) + |\mathcal{S}| \log(\mathrm{e}(1+n)) \leq |\mathcal{S}|\beta^{KL}(\delta, n)$, we conclude the first statement.

Event $\mathcal{E}^{\mathsf{c}}(\delta)$ By a union bound argument, it is enough to show that each of the following events

$$\overline{\mathcal{E}}^{\mathsf{c}}(\delta, s, a, h) = \left\{\exists r \in [R_{\max}] : \left\|\widehat{\mathsf{P}}_{(r),h}(\cdot|s,a) - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot|s,a)\right\|_1 \geq \frac{9}{8}\varepsilon_{\mathsf{p}} + \frac{11\beta^{\mathsf{c}}(\delta, N_{(r),h}(s,a))}{N_{(r),h}(s,a)}\right\}$$

holds with probability less or equal than $\delta' := \delta/(6|\mathcal{S}||\mathcal{A}|H)$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. To do it, let us analyze the difference between kernels. By the definitions (24)-(25):

$$[\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}](s'|s,a) = \frac{1}{N_{(r),h}(s,a)} \sum_{i=1}^{M} \sum_{t=1}^{T_r} \left(1_{(s,a,s')}(s^i_{t,h}, a^i_{t,h}, s^i_{t,h}) - 1_{(s,a,s')}(s^i_{t,h}, a^i_{t,h}, s^{\mathsf{c},i}_{t,h})\right).$$

Next, we notice that, using representation (23), we can rewrite the first indicator as follows

$$1_{(s,a,s')}(s^i_{t,h}, a^i_{t,h}, s^i_{t,h}) = (1 - \xi^i_{t,h}) \cdot 1_{(s,a,s')}(s^i_{t,h}, a^i_{t,h}, s^{\mathsf{c},i}_{t,h}) + \xi^i_{t,h} 1_{(s,a,s')}(s^i_{t,h}, a^i_{t,h}, s^{\mathsf{ind},i}_{t,h}),$$

thus, we have the following expression for the difference between kernels

$$[\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}](s'|s,a) = \frac{1}{N_{(r),h}(s,a)} \sum_{i=1}^{M} \sum_{t=1}^{T_r} \xi^i_{t,h} 1_{(s,a)}(s^i_{t,h}, a^i_{t,h}) \cdot \left(1_{s'}(s^{\mathsf{ind},i}_{t,h}) - 1_{s'}(s^{\mathsf{c},i}_{t,h})\right),$$

and thus, using $|1_{s'}(s^{\mathsf{ind},i}_{t,h}) - 1_{s'}(s^{\mathsf{c},i}_{t,h})| \leq 1$ and Definition (29), we obtain

$$\left\|[\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}](s,a)\right\|_1 \leq \frac{1}{N_{(r),h}(s,a)} \sum_{t=1}^{T_r} \sum_{i=1}^{M} \xi^i_{t,h} 1_{(s,a)}(s^i_{t,h}, a^i_{t,h}) = \frac{1}{N_{(r),h}(s,a)} \sum_{j=1}^{N_{(r),h}(s,a)} \xi^{i_{s,a,h,j}}_{t_{s,a,h,j}}. \tag{31}$$

Again, by the optional skipping argument, conditioned on the event $N_{(r),h}(s,a) = N$ the sequence $\{\tilde{\xi}_{s,a,h,j}\}_{j \in [TM]} := \{\xi^{i_{s,a,h,j}}_{t_{s,a,h,j}}\}_{j \in [TM]}$ is i.i.d., thus Corollary F.1 implies

$$\mathbb{P}[\overline{\mathcal{E}}^{\mathsf{c}}(\delta, s, a, h)] \leq \mathbb{P}\left[\exists N \geq 1 : \sum_{j=1}^{N} \tilde{\xi}_{s,a,h,j} > \frac{9}{8} N \varepsilon_{\mathsf{p}} + 11 \log\left(\frac{24|\mathcal{S}||\mathcal{A}|H\mathrm{e}(2N+1)}{\delta}\right)\right] \leq \frac{\delta}{6|\mathcal{S}||\mathcal{A}|H}.$$

Event $\mathcal{E}^\star(\delta)$ To analyze this event, we use the representation of the kernel (30) and optional skipping argument conditioned on $N_{(r),h}(s,a) = N$ where $N \geq 2$

$$[\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h]\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a) = \frac{1}{N}\sum_{j=1}^{N} \mathcal{V}^{\mathsf{c},\star}_{h+1}(s^{\mathsf{c}}_{s,a,h,j}) - \mathsf{P}^{\mathsf{c}}_h \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)\,.$$

Thus, we have a sum of centered i.i.d. random variables, and thus we can apply Lemma F.7

$$\mathbb{P}\left[\left|[\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h]\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)\right| \geq \sqrt{\frac{2\operatorname{Var}_{\widehat{\mathsf{P}}^{c}_{(r),h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s,a)\beta^\star(\delta)}{N-1}} + \frac{7\beta^\star(\delta)}{N-1}\middle| N_{(r),h}(s,a) = N\right] \leq \frac{\delta}{6|\mathcal{S}||\mathcal{A}|HT},$$

where $\beta^\star(\delta) = \log(12|\mathcal{S}||\mathcal{A}|H/\delta)$. We then conclude by a union bound over $(s,a,h,N) \in \mathcal{S}\times\mathcal{A}\times[H]\times\{2,\dots,MT\}$. If $N_{(r),h}(s,a) \leq 1$, we have the trivial bound $\left|[\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h]\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)\right| \leq H$.

Event $\mathcal{E}^{\mathrm{Var}}(\delta)$ For any $t' \in [T]$, define

$$X^i_{t'} = \sum_{h=1}^{H} \operatorname{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{i,\pi_{t'}}_{h+1})\left(s^i_{t',h}, a^i_{t',h}\right) - \sigma\mathcal{V}^{i,\pi_{t'}}_1\left(s^i_{t',1}\right)\,,$$

where $\sigma\mathcal{V}^{i,\pi_{t'}}_1$ is defined in (59). This sequence forms a martingale-difference sequence with respect to the following filtration

$$\mathcal{F}^i_t = \sigma\left(\{s^{i'}_{t',h}a^{i'}_{t',h}\}_{(t',i')\preceq(t,i),h\in[H]}\right)\,,$$

where the order over the pairs (t',i') is lexicographic. Applying Theorem F.1 yields

$$\mathbb{P}\left[\exists t \geq 1, \sum_{(t'\geq 1, i\geq 1)}^{(t,M)} X^i_{t'} \leq \sqrt{2\sum_{(t'\geq 1,i\geq 1)}^{(t,M)} \mathbb{E}_\pi[(X^i_{t'})^2 \mid \mathcal{F}^{i_{\mathrm{prev}}}_{t'_{\mathrm{prev}}}]\log\left(24\mathrm{e}(2Mt+1)/\delta\right)} + 3H^3\log\left(24\mathrm{e}(2Mt+1)/\delta\right)\right] \leq \frac{\delta}{6}\,,$$

as we have $|X^i_{t'}| \leq H^3$ and where $(t'_{\mathrm{prev}}, i_{\mathrm{prev}})$ is a previous element in a lexicographic order with respect to (t',i). Now, we bound the conditional second-order moment of $X^i_{t'}$ as follows

$$\mathbb{E}_{\pi_{t'}}[(X^i_{t'})^2 \mid \mathcal{F}^{i_{\mathrm{prev}}}_{t'_{\mathrm{prev}}}] \leq \mathbb{E}_{\pi_{t'}}\left[\left(\sum_{h=1}^{H} \operatorname{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{i,\pi_{t'}}_{h+1})\left(s^i_{t',h}, a^i_{t',h}\right)\right)^2 \middle| \mathcal{F}^{i-1}_t\right] \leq H^3\mathbb{E}_{\pi_{t'}}\left[\sum_{h=1}^{H} \operatorname{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{i,\pi_{t'}}_{h+1})\left(s^i_{t',h}, a^i_{t',h}\right)\right]\,.$$

By Lemma F.3, we have

$$\mathbb{E}_{\pi_{t'}}\left[\sum_{h=1}^{H} \operatorname{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{i,\pi_{t'}}_{h+1})\left(s^i_{t',h}, a^i_{t',h}\right)\right] = \mathbb{E}_{\pi_{t'}}\left[\left(\sum_{h=1}^{H} \mathsf{r}^i_h(s^i_h,a^i_h) - \mathcal{V}^{i,\pi_{t'}}_1(s^i_1)\right)^2\right] \leq \mathbb{E}_{\pi_{t'}}\left[\left(\sum_{h=1}^{H}\mathsf{r}^i_h(s^i_h,a^i_h)\right)^2\right] \leq H^2\,.$$

By combining the previous inequalities, we obtain

$$\sum_{(t'\geq 1,i\geq 1)}^{(t,M)} X^i_{t'} \leq \sqrt{2H^5Mt\log\left(24\mathrm{e}(2Mt+1)/\delta\right)} + 3H^3\log\left(24\mathrm{e}(2Mt+1)/\delta\right)$$

Now using Lemma F.3 again we get

$$\begin{aligned}\sum_{(t'\geq 1,i\geq 1)}^{(t,M)} \sum_{h=1}^{H} \operatorname{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{i,\pi_{t'}}_{h+1})\left(s^i_{t',h}, a^i_{t',h}\right) &= \sum_{(t'\geq 1,i\geq 1)}^{(t,M)} X^i_{t'} + \sigma\mathcal{V}^{i,\pi_{t'}}_1\left(s^i_{t',1}\right) \\ &\leq \sqrt{2H^5Mt\log\left(24\mathrm{e}(2Mt+1)/\delta\right)} + 3H^3\log\left(24\mathrm{e}(2Mt+1)/\delta\right) + H^2Mt\,.\end{aligned}$$

Event $\mathcal{E}^{\text{count}}(\delta)$ For any fixed $(s,a,h,i,t_1) \in \mathcal{S}\times\mathcal{A}\times[H]\times[M]\times[T]$, we have by Corollary F.1

$$\mathbb{P}\big[\exists t_2 \in \mathbb{N} : \left|\sum_{t'=t_1}^{t_2} 1_{(s,a)}(s^i_{t',h}, a^i_{t',h}) - d_h^{i,\pi_{t'}}(s,a)\right| \geq \frac{1}{8}\sum_{t'=t_1}^{t_2} d_h^{i,\pi_{t'}}(s,a) + 11\beta^{\mathsf{c}}(\delta, t_2-t_1+1)\big] \leq \frac{\delta}{6|\mathcal{S}||\mathcal{A}|MTH},$$

holds with probability less or equal than $\delta' := \delta/(6|\mathcal{S}||\mathcal{A}|MTH)$. Thus, by a union bound argument, the following event

$$\overline{\mathcal{E}}^{\text{dev}}(\delta) := \Big\{\forall (t_1,t_2)\in[T]^2, \forall h\in[H], \forall (s,a)\in\mathcal{S}\times\mathcal{A}, \forall i\in[M] : \left|\sum_{t'=t_1}^{t_2} 1_{(s,a)}(s^i_{t',h}, a^i_{t',h}) - d_h^{i,\pi_{t'}}(s,a)\right| \geq \frac{1}{8}\sum_{t'=t_1}^{t_2} d_h^{i,\pi_{t'}}(s,a) + 11\beta^{\mathsf{c}}(\delta, t_2-t_1+1)\Big\},$$

holds with probability less or equal to $\delta/6$. Now, to conclude the proof, it is enough to show $\mathcal{E}^{\text{count}}(\delta) \subset \mathcal{E}^{\text{dev}}(\delta)$. Let's recall the definition of the estimated counter by agent i

$$\hat{N}^i_{t,h}(s,a) = N_{(r_t),h}(s,a) + M\sum_{t'=\psi_t}^{t} 1_{(s,a)}(s^i_{t',h}, a^i_{t',h}) .$$

Using (29), the definition of $\hat{N}^i_{t,h}(s,a)$, and the triangular inequality, we have for any fixed $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
|\widetilde{N}^M_{t,h}(s,a) - \hat{N}^i_{t,h}(s,a)| &= \left|\sum_{j\neq i}\sum_{t'=\psi_t}^{t} 1_{(s,a)}(s^j_{t',h}, a^j_{t',h}) - 1_{(s,a)}(s^i_{t',h}, a^i_{t',h})\right| \\
&\leq \underbrace{\left|\sum_{j\neq i}\sum_{t'=\psi_t}^{t} 1_{(s,a)}(s^j_{t',h}, a^j_{t',h}) - d_h^{j,\pi_{t'}}(s,a)\right|}_{(1)} + \underbrace{\left|\sum_{j\neq i}\sum_{t'=\psi_t}^{t} d_h^{j,\pi_{t'}}(s,a) - d_h^{i,\pi_{t'}}(s,a)\right|}_{(2)} \\
&\quad + \underbrace{(M-1)\left|\sum_{t'=\psi_t}^{t} d_h^{i,\pi_{t'}}(s,a) - 1_{(s,a)}(s^i_{t',h}, a^i_{t',h})\right|}_{(3)} .
\end{aligned}$$

Term (2): Heterogeneity error Using Lemma F.2 combined with the triangular inequality, it holds that

$$(2) = \left|\sum_{j\neq i}\sum_{t'=\psi_t}^{t} d_h^{j,\pi_{t'}}(s,a) - d_h^{i,\pi_{t'}}(s,a)\right| \leq \varepsilon_{\mathsf{p}} HMT .$$

Terms (1) and (3): concentration error On the event $\mathcal{E}^{\text{dev}}(\delta)$, we can bound (1) as follows

$$\begin{aligned}
(1) &\leq \sum_{j\neq i}\sum_{t'=\psi_t}^{t}\left|1_{(s,a)}(s^j_{t',h}, a^j_{t',h}) - d_h^{j,\pi_{t'}}(s,a)\right| \\
&\leq \sum_{j\neq i}\frac{1}{8}\sum_{t'=\psi_t}^{t} d_h^{j,\pi_{t'}}(s,a) + 11M\beta^{\mathsf{c}}(\delta,T) \\
&\leq \underbrace{\frac{1}{8}\sum_{j\neq i}\left|\sum_{t'=\psi_t}^{t} d_h^{j,\pi_{t'}}(s,a) - d_h^{i,\pi_{t'}}(s,a)\right|}_{(2)} + 11M\beta^{\mathsf{c}}(\delta,T) + \frac{M}{8}\sum_{t'=\psi_t}^{t} d_h^{i,\pi_{t'}}(s,a) .
\end{aligned}$$

Now using the latter bound on (**2**) combined with the inequality $\frac{7}{8}\sum_{t'=\psi_t}^{t} d_h^{i,\pi_{t'}}(s,a) - 11\beta^{\mathsf{c}}(\delta,T) \leq \sum_{t'=\psi_t}^{t} 1_{(s,a)}(s^i_{t',h}, a^i_{t',h})$ that follows from $\mathcal{E}^{\mathrm{dev}}(\delta)$, we get

$$(\mathbf{1}) \leq \frac{1}{8}\varepsilon_{\mathsf{p}} HMT + \frac{88M}{7}\beta^{\mathsf{c}}(\delta,T) + \frac{1}{7}\hat{N}^i_{t,h}(s,a) \,.$$

We proceed similarly to bound (**3**)

$$\begin{aligned}(\mathbf{3}) &= (M-1)\left|\sum_{t'=\psi_t}^{t} d_h^{i,\pi_{t'}}(s,a) - 1_{(s,a)}(s^i_{t',h}, a^i_{t',h})\right| \\ &\leq \frac{M}{8}\sum_{t'=\psi_t}^{t} d_h^{j,\pi_{t'}}(s,a) + 11M\beta^{\mathsf{c}}(\delta,T) \leq \frac{1}{7}\hat{N}^i_{t,h}(s,a) + \frac{88M}{7}\beta^{\mathsf{c}}(\delta,T) \,.\end{aligned}$$

Finally combining the bounds on (**1**), (**2**) and (**3**) yields the desired result.

Event $\mathcal{E}(\delta)$ Notice that the two following sequences

$$\begin{aligned}X^i_{t,h} &:= \left(1+\frac{1}{H}\right)^{H-h'-1}\left(\mathsf{P}^i_h\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h}, a^i_{t,h}) - \left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h+1})\right) \,, \\ Y^i_{t,h} &:= \mathsf{P}^i_h\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h}, a^i_{t,h}) - \left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h+1}) \,,\end{aligned}$$

forms a martingale-difference sequence with respect to filtration $\mathcal{F}^i_{t,h}$ defined in (28). Thus, applying Azuma-Hoeffding inequality with a union bound over h and over the two events allows us to conclude the statement. □

Lemma C.2. *Conditioned on $\mathcal{E}^{\mathrm{KL}}(\delta)$, for any function $f : \mathcal{S} \mapsto [0,H], h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$, and any $r \in [R_{\max}]$, we have*

$$\begin{aligned}(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h) f(s,a) &\leq \frac{1}{H}\mathsf{P}^{\mathsf{c}}_h f(s,a) + \frac{2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta, N_{(r),h}(s,a))}{N_{(r),h}(s,a)} \,, \\ \|\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h\|_1 &\leq \sqrt{\frac{2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta, N_{(r),h}(s,a))}{N_{(r),h}(s,a)}} \,.\end{aligned}$$

Proof. Using Lemma F.5 with $\mathsf{P} = \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot|s,a)$ and $\mathsf{Q} = \mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)$ it holds that

$$(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h) f(\cdot|s,a) \leq \sqrt{2\operatorname{Var}_{\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)}(f)\mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot|s,a)\middle\|\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)\right)} + \frac{2}{3}H\mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot|s,a)\middle\|\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)\right) \,. \quad (32)$$

Now, since f's values are in $[0,H]$, we can write

$$\operatorname{Var}_{\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)}(f) \leq \mathsf{P}^{\mathsf{c}}_h(f^2)(s,a) \leq H\mathsf{P}^{\mathsf{c}}_h(f)(s,a) \,. \quad (33)$$

Combining the latter inequality with the fact that for all $a,b \geq 0$, $\sqrt{2ab} \leq a+b$, we obtain

$$\begin{aligned}\sqrt{2\operatorname{Var}_{\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)}(f)\mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot|s,a)\middle\|\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)\right)} &= \sqrt{\frac{2}{H}\mathsf{P}^{\mathsf{c}}_h(f)(s,a)\cdot H^2\mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot|s,a)\middle\|\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)\right)} \\ &\leq \frac{1}{H}\mathsf{P}^{\mathsf{c}}_h(f)(s,a) + H^2\mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(\cdot|s,a)\middle\|\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a)\right) \,. \quad (34)\end{aligned}$$

Furthermore, since $\mathcal{E}^{\mathrm{KL}}(\delta)$ holds, we have the inequality $\mathrm{KL}\left(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}(s,a)\middle\|\mathsf{P}^{\mathsf{c}}_h(s,a)\right) \leq \frac{|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,N_{(r),h}(s,a))}{N_{(r),h}(s,a)}$. Plugging this bound in (34), we can upper bound (32) as

$$(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h) f(\cdot|s,a) \leq \frac{1}{H}\mathsf{P}^{\mathsf{c}}_h(f)(s,a) + H^2\frac{|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,N_{(r),h}(s,a))}{N_{(r),h}(s,a)} + \frac{2H}{3}\frac{|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,N_{(r),h}(s,a))}{N_{(r),h}(s,a)} \,,$$

which gives the result. The second inequality follows from the combination of Pinsker inequality and the definition of $\mathcal{E}^{\mathrm{KL}}(\delta)$. □

D REGRET ANALYSIS

We define the common MDP $\mathcal{M}^c$ as

$$\mathcal{M}^c := (\mathcal{S}, \mathcal{A}, H, \{\mathsf{r}^c_h := \frac{1}{M}\sum_{i=1}^{M} \mathsf{r}^i_h\}_h, \{\mathsf{P}^c_h\}_h) \,. \tag{35}$$

We denote by $\mathcal{V}^{c,\star}_h$ and $\mathcal{Q}^{c,\star}_h$ the value function and the Q-function at step h in the common environment $\mathcal{M}^c$. In particular, these functions satisfy Bellman's equations and Bellman's optimality equations (Sutton and Barto, 2018)

$$\mathcal{Q}^{c,\pi}_h(s,a) = \mathsf{r}^c_h(s,a) + \mathsf{P}^c_h \mathcal{V}^{c,\pi}_{h+1}(s,a) \,, \qquad \mathcal{V}^{c,\pi}_h(s) = \mathcal{Q}^{c,\pi}_h(s,\pi_h(s)) \tag{36}$$

$$\mathcal{Q}^{c,\star}_h(s,a) = \mathsf{r}^c_h(s,a) + \mathsf{P}^c_h \mathcal{V}^{c,\star}_{h+1}(s,a) \,, \qquad \mathcal{V}^{c,\star}_h(s) = \max_{a\in\mathcal{A}} \mathcal{Q}^{c,\star}_h(s,a) \,, \tag{37}$$

D.1 Optimism

Let us define the following event

$$\mathcal{E}^{\text{optimism}} = \Bigg\{ \forall r \in [R_{\max}], \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : \hat{\mathcal{V}}_{(r),h}(s) \geq \mathcal{V}^{c,\star}_h(s) - (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}} H)(H+1-h),$$
$$\hat{\mathcal{Q}}_{(r),h}(s,a) \geq \mathcal{Q}^{c,\star}_h(s,a) - (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}} H)(H+1-h) \Bigg\} .$$

Then, we will show that this event holds on event $\mathcal{G}(\delta)$. To prove the optimism of our estimates, we use the same monotonicity arguments as in Zhang et al. (2021), see also Zhang et al. (2024b). Define

$$g(\mathsf{P}, f, \alpha) = \mathsf{P} f + \max(\sqrt{\alpha \operatorname{Var}_{\mathsf{P}}(f)}, \alpha H) \,, \tag{38}$$

where P be is a probability measure on $\mathcal{S}$, $f \in \mathbb{R}^{|\mathcal{S}|}$ is a non negative vector satisfying $\|f\|_\infty \leq H$, and α is a positive real number.

Lemma D.1 (Lemma 14 by Zhang et al. 2021)**.** *The function g is non-decreasing in each entry of f.*

For completeness, we provide the proof below.

Proof. To justify this claim, consider any $s \in \mathcal{S}$, and let us fix P, α and all but the s-th entries of f. It then suffices to observe that (i) g is a differentiable almost anywhere function, and (ii) except for at most two possible choices of $f(s)$ that obey $\sqrt{\alpha \operatorname{Var}_{\mathsf{P}}(f)} = \alpha H$, one can use the properties of P and f to calculate

$$\begin{aligned}
\frac{\partial g(\mathsf{P}, f, \alpha)}{\partial f(s)} &= \mathsf{P}(s) + \sqrt{\alpha}\mathbf{1}\left\{\sqrt{\alpha \operatorname{Var}_{\mathsf{P}}(f)} \geq \alpha H\right\} \frac{\mathsf{P}(s)(f(s) - \mathbb{E}_{s'\sim\mathsf{P}}[f(s')])}{\sqrt{\operatorname{Var}_{\mathsf{P}}(f)}} \\
&= \mathsf{P}(s) + \mathbf{1}\left\{\sqrt{\alpha \operatorname{Var}_{\mathsf{P}}(f)} \geq \alpha H\right\} \frac{\alpha H}{\sqrt{\alpha \operatorname{Var}_{\mathsf{P}}(f)}} \cdot \frac{\mathsf{P}(s)(f(s) - \mathbb{E}_{s'\sim\mathsf{P}}[f(s')])}{H} \\
&\geq \min\left\{\mathsf{P}(s) + \mathsf{P}(s)\frac{(f(s) - \mathbb{E}_{s'\sim\mathsf{P}}[f(s')])}{H}, \mathsf{P}(s)\right\} \\
&\geq \mathsf{P}(s)\min\left\{\frac{H + f(s) - \mathbb{E}_{s'\sim\mathsf{P}}[f(s')]}{H}, 1\right\} \geq 0 \,,
\end{aligned}$$

where in the end we used the fact that $\|f\|_\infty \leq H$. □

We define the bonus function as

$$b_{(r),h}(s,a) := \begin{cases} \frac{28\beta^\star(\delta)H + 11\beta^c(\delta,N)}{N} + \sqrt{\frac{8\beta^\star(\delta)}{N} \cdot \operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a)} \,, & N \geq 2 \\ H \,, & N \leq 1 \end{cases} \tag{39}$$

for $N = N_{(r),h}(s,a)$ and where $\beta^\star$ and β^{c} are defined in Lemma C.1.

Lemma D.2. *Under conditions of Lemma C.1, it holds $\mathcal{E}^{\text{optimism}} \subseteq \mathcal{G}(\delta)$ for any $\delta \in (0,1)$.*

Proof. We process the proof by backward induction over h.

Base case For $h = H+1$ and for all $(s,a,r) \in \mathcal{S} \times \mathcal{A} \times [R_{\max}]$, we have

$$\hat{\mathcal{V}}_{(r),h}(s) = 0 \geq \mathcal{V}_h^{\mathsf{c},\star}(s) - 0 = 0 \quad \text{and} \quad \hat{\mathcal{Q}}_{(r),h}(s,a) = 0 \geq \mathcal{Q}_h^{\mathsf{c},\star}(s,a) - 0 = 0 \;,$$

which gives the desired result.

Induction Let $h \in [H]$ such that for all $(s,a,r) \in \mathcal{S} \times \mathcal{A} \times [R_{\max}]$ and $h' \geq h$

$$\hat{\mathcal{V}}_{(r),h'}(s) \geq \mathcal{V}_{h'}^{\mathsf{c},\star}(s) - (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}} H)(H+1-h) \;, \text{ and} \tag{40}$$

$$\hat{\mathcal{Q}}_{(r),h'}(s,a) \geq \mathcal{Q}_{h'}^{\mathsf{c},\star}(s,a) - (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}} H)(H+1-h) \;. \tag{41}$$

First, let us consider a trivial case $\hat{\mathcal{Q}}_{(r),h}(s,a) = H$. The result is trivial since $H \geq \mathcal{Q}_h^{\mathsf{c},\star}(s,a)$.

Next, we assume that $\hat{\mathcal{Q}}_{(r),h}(s,a) < H$. In particular, by the definition of bonuses, it automatically follows that $N_{(r),h}(s,a) \geq 2$. In this case, according to the update rule (11), we have

$$\begin{aligned}
\hat{\mathcal{Q}}_{(r),h}(s,a) &\geq \sum_{i=1}^{M} \frac{n^i_{(r),h}(s,a)}{N_{(r),h}(s,a)} \hat{\mathcal{Q}}^i_{(r),h}(s,a) + b_{(r),h}(s,a) \\
&= \sum_{i=1}^{M} \frac{n^i_{(r),h}(s,a)}{N_{(r),h}(s,a)} \widehat{\mathsf{r}}^i_h(s,a) + \widehat{\mathsf{P}}_{(r),h} \hat{\mathcal{V}}_{(r),h+1}(s,a) + b_{(r),h}(s,a) \\
&= \frac{1}{M} \sum_{i=1}^{M} \mathsf{r}^i_h(s,a) + \mathsf{P}^{\mathsf{c}}_h \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a) + b_{(r),h}(s,a) + \underbrace{\sum_{i=1}^{M} \frac{n^i_{(r),h}(s,a)}{N_{(r),h}(s,a)} \widehat{\mathsf{r}}^i_h(s,a) - \frac{1}{M} \sum_{i=1}^{M} \mathsf{r}^i_h(s,a)}_{(\mathbf{I})} \\
&\quad + \underbrace{\widehat{\mathsf{P}}_{(r),h}(\hat{\mathcal{V}}_{(r),h+1}(s,a) - \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a))}_{(\mathbf{II})} + \underbrace{(\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h})\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)}_{(\mathbf{III})} + \underbrace{(\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h)\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)}_{(\mathbf{IV})} \;.
\end{aligned} \tag{42}$$

Terms (I) and (III): heterogeneity errors First, let us handle the terms that come from the presence of heterogeneity between agents. To analyse (**I**), recall that since for all $(s,a,i,h,r) \in \mathcal{S} \times \mathcal{A} \times [M] \times [H] \times [R_{\max}]$, either 1) $n^i_{(r),h}(s,a) = 0$ and the value of $\widehat{\mathsf{r}}^i_h(s,a)$ is ignored in the weighted sum, or 2) $n^i_{(r),h}(s,a) > 0$ and $\widehat{\mathsf{r}}^i_h(s,a) = \mathsf{r}^i_h(s,a)$. Thus, $\sum_{i=1}^{M} \frac{n^i_{(r),h}(s,a)}{N_{(r),h}(s,a)} \widehat{\mathsf{r}}^i_h(s,a)$ is a convex combination of the true rewards over i, which ensures that

$$(\mathbf{I}) = \sum_{i=1}^{M} \frac{n^i_{(r),h}(s,a)}{N_{(r),h}(s,a)} \widehat{\mathsf{r}}^i_h(s,a) - \frac{1}{M} \sum_{i=1}^{M} \mathsf{r}^i_h(s,a) \geq -2\varepsilon_{\mathsf{r}} \;. \tag{43}$$

Conditioned on $\mathcal{E}^{\mathsf{c}}(\delta)$, Hölder's inequality yields the following bound on (**III**)

$$(\mathbf{III}) = (\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h})\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a) \geq -\|\widehat{\mathsf{P}}_{(r),h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}\|_1 \cdot \|\mathcal{V}^{\mathsf{c},\star}_{h+1}\|_\infty \geq -2\varepsilon_{\mathsf{p}} H - \frac{11\beta^{\mathsf{c}}(\delta, N_{(r),h}(s,a))}{N_{(r),h}(s,a)} H \;. \tag{44}$$

Term (II): correction error To control this term, we aim to apply Lemma D.1. We first define the *shifted estimator* $\bar{\mathcal{V}}_{(r),h+1}$ as

$$\bar{\mathcal{V}}_{(r),h+1}(s) := \hat{\mathcal{V}}_{(r),h+1}(s) + (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}} H)(H-h) \;. \tag{45}$$

By the induction hypothesis (40), we know that $\bar{\mathcal{V}}_{(r),h+1}(s) \geq \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)$. We decompose further (**II**) as

$$\begin{aligned}
(\mathbf{II}) &= \widehat{\mathsf{P}}_{(r),h} \hat{\mathcal{V}}_{(r),h+1}(s,a) - \widehat{\mathsf{P}}_{(r),h} \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a) \\
&\geq \widehat{\mathsf{P}}_{(r),h} \bar{\mathcal{V}}_{(r),h+1}(s,a) + \max\left(\sqrt{\frac{4\beta^\star(\delta) \operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\bar{\mathcal{V}}_{(r),h+1}(s))(s,a)}{N_{(r),h}(s,a)}}, \frac{4\beta^\star(\delta) H)}{N_{(r),h}(s,a)} \right) - \widehat{\mathsf{P}}_{(r),h} \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)
\end{aligned}$$

$$- \sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\bar{\mathcal{V}}_{(r),h+1}(s))(s,a)}{N_{(r),h}(s,a)}} - \frac{4\beta^\star(\delta)H}{N_{(r),h}(s,a)} - (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H-h)\ ,$$

where we used in the last inequality that for any $a, b \in \mathbb{R}_+$, $\max(a,b) \le a+b$ and the fact that $\widehat{\mathsf{P}}_{(r),h}\hat{\mathcal{V}}_{(r),h+1}(s,a) - \widehat{\mathsf{P}}_{(r),h}\bar{\mathcal{V}}_{(r),h+1}(s,a) = -(2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H-h)$. Now by applying Lemma D.1, we get

$$\begin{aligned}(\mathbf{II}) &\ge \max\left(\sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}}, \frac{4\beta^\star(\delta)H}{N_{(r),h}(s,a)}\right) - \sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\bar{\mathcal{V}}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} - \frac{4\beta^\star(\delta)H}{N_{(r),h}(s,a)} \\ &\ge \underbrace{\sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}}}_{(\mathbf{1})} - \underbrace{\sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\bar{\mathcal{V}}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}}}_{(\mathbf{2})} - \frac{4\beta^\star(\delta)H}{N_{(r),h}(s,a)}\ . \end{aligned} \tag{46}$$

We want now to control the variance terms that appear in (**1**) and (**2**). Using inequalities (62) and (61) of Lemma F.6, we have

$$\begin{aligned}\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s,a) &\ge \operatorname{Var}_{\widehat{\mathsf{P}}^c_{(r),h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s,a) - 3H^2\varepsilon_{\mathsf{p}}\ , \\ \operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\bar{\mathcal{V}}_{(r),h+1})(s,a) &\le 2\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a) + 2\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h}|\bar{\mathcal{V}}_{(r),h+1} - \hat{\mathcal{V}}_{(r),h+1}| \\ &\le 2\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a) + 2(2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H-h)\ ,\end{aligned}$$

where in the last inequality we used the induction hypothesis. Besides, as for any $a, b, c \in \mathbb{R}_+$, we have $a \ge b - c \implies \sqrt{a} \ge \sqrt{b} - \sqrt{c}$, and also for any $d, f \in \mathbb{R}_+$ we have $\sqrt{d+f} \le \sqrt{d} + \sqrt{f}$, we get

$$(\mathbf{1}) := \sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} \ge \sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}^c_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} - \sqrt{\frac{12\varepsilon_{\mathsf{p}}H^2\beta^\star(\delta)}{N_{(r),h}(s,a)}}\ , \text{ and} \tag{47}$$

$$(\mathbf{2}) := \sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\bar{\mathcal{V}}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} \le \sqrt{\frac{8\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} + \sqrt{\frac{8\beta^\star(\delta)(3\varepsilon_{\mathsf{p}}H + 2\varepsilon_{\mathsf{r}})(H+1-h)}{N_{(r),h}(s,a)}} \tag{48}$$

Plugging the inequalities (47) and (48) in (46), we obtain

$$\begin{aligned}(\mathbf{II}) \ge &\sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}^c_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} - \sqrt{\frac{8\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} - \sqrt{\frac{12\varepsilon_{\mathsf{p}}H^2\beta^\star(\delta)}{N_{(r),h}(s,a)}} - \frac{4\beta^\star(\delta)H}{N_{(r),h}(s,a)} \\ &- \sqrt{\frac{8\beta^\star(\delta)(3\varepsilon_{\mathsf{p}}H + 2\varepsilon_{\mathsf{r}})(H+1-h)}{N_{(r),h}(s,a)}}\ .\end{aligned} \tag{49}$$

Finally, as for any $a, b \in \mathbb{R}+$ we have $\sqrt{2ab} \le a + b$ then

$$\begin{aligned}(\mathbf{II}) \ge &\sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}^c_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} - \sqrt{\frac{8\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} - \varepsilon_{\mathsf{p}}H - \frac{10\beta^\star(\delta)H}{N_{(r),h}(s,a)} \\ &- (3\varepsilon_{\mathsf{p}}H + 2\varepsilon_{\mathsf{r}})(H-h) - \frac{4\beta^\star(\delta)H}{N_{(r),h}(s,a)}\end{aligned} \tag{50}$$

Term (IV): concentration error Conditioned on $\mathcal{E}^\star(\delta)$, we have

$$(\mathbf{IV}) = (\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h)\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a) \ge -\left|[\widehat{\mathsf{P}}^{\mathsf{c}}_{(r),h} - \mathsf{P}^{\mathsf{c}}_h]\mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a)\right|$$

$$\geq -\sqrt{\frac{2\operatorname{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s,a)\beta^\star(\delta)}{N_{(r),h}(s,a)-1}} - \frac{7\beta^\star(\delta)}{N_{(r),h}(s,a)-1} \geq -\sqrt{\frac{4\operatorname{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{(r)}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s,a)\beta^\star(\delta)}{N_{(r),h}(s,a)}} - \frac{14\beta^\star(\delta)}{N_{(r),h}(s,a)}\,, \tag{51}$$

as for $n \geq 2$, we have $\frac{2}{n} \geq \frac{1}{n-1}$.

Combine everything together By plugging in the bounds on (**I**), (**II**), (**III**), and (**IV**) in (42), we get

$$\hat{\mathcal{Q}}_{(r),h}(s,a) \geq \frac{1}{M}\sum_{i=1}^{M} \mathsf{r}^i_h(s,a) + \mathsf{P}^{\mathsf{c}}_h \mathcal{V}^{\mathsf{c},\star}_{h+1}(s,a) + b_{(r),h}(s,a) - \frac{11\beta^{\mathsf{c}}(\delta, N_{(r),h}(s,a))}{N_{(r),h}(s,a)}H - \frac{28\beta^\star(\delta)H}{N_{(r),h}(s,a)}$$
$$- \sqrt{\frac{8\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a)}{N_{(r),h}(s,a)}} - (3\varepsilon_{\mathsf{p}}H + 2\varepsilon_{\mathsf{r}})(H+1-h) = \mathcal{Q}^{\mathsf{c},\star}_h(s,a) - (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H+1-h)\,,$$

where the last inequality is a consequence of the definition of the bonus (39) and optimal Bellman equations (37). □

D.2 Regret decomposition

We will start by writing down a regret decomposition. Let us define the essential technical quantities, such as common regret and partial common upper regret

$$\mathfrak{R}^{\mathsf{c}}(T) := \frac{1}{M}\sum_{t=1}^{T}\sum_{i=1}^{M} \mathcal{V}^{\mathsf{c},\star}_1(s^i_{t,1}) - \mathcal{V}^{\mathsf{c},\pi_t}_1(s^i_{t,1}), \qquad \overline{\mathfrak{R}}^{\mathsf{c}}_h(T) := \frac{1}{M}\sum_{t=1}^{T}\sum_{i=1}^{M} \hat{\mathcal{V}}_{t,h}(s^i_{t,h}) - \mathcal{V}^{\mathsf{c},\pi_t}_h(s^i_{t,h})$$

Lemma D.3. *Assume conditions of Lemma C.1. Then, on the event $\mathcal{G}(\delta)$, the following inequality for any partial upper common regret holds*

$$\overline{\mathfrak{R}}^{\mathsf{c}}_h(T) \leq U^T_h := A^T_h + B^T_h + C^T_h + 7\mathrm{e}TH^2\varepsilon_{\mathsf{p}} + 2\mathrm{e}TH\varepsilon_{\mathsf{r}} + \sqrt{8H^2\cdot TH\cdot\beta(\delta)/M} + \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H} 2\mathrm{e}H\mathbb{1}_{[\![0;1]\!]}(\bar{N}^i_{t,h'}),$$

where

$$\bar{N}^i_{t,h} := N_{(r_t),h}(s^i_{t,h}, a^i_{t,h})\,,$$
$$A^T_h := \frac{\mathrm{e}}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H}\sqrt{\frac{4\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h'}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h'}, a^i_{t,h'})}{\bar{N}^i_{t,h'}}}\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'})\,,$$
$$B^T_h := \frac{\mathrm{e}}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H}\sqrt{\frac{8\beta^\star(\delta)\operatorname{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}}(\hat{\mathcal{V}}_{t,h'+1})(s^i_{t,h'}, a^i_{t,h'})}{\bar{N}^i_{t,h'}}}\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'})\,,$$
$$C^T_h := \frac{\mathrm{e}}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H}\frac{22\beta^{\mathsf{c}}(\delta,\bar{N}^i_{t,h}) + 46H\beta^\star(\delta) + 2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}}\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'})\,.$$

Proof. Let us define $\delta^i_{t,h} = \hat{\mathcal{V}}_{t,h}(s^i_{t,h}) - \mathcal{V}^{\mathsf{c},\pi_t}_1(s^i_{t,h})$ and let us study this term separately. Since the policy is deterministic, i.e., $a^i_{t,h} = \pi_{t,h}(s^i_{t,h})$, and satisfies $\hat{\mathcal{V}}_{t,h}(s^i_{t,h}) = \hat{\mathcal{Q}}_{t,h}(s^i_{t,h}, a^i_{t,h})$, we have

$$\delta^i_{t,h} = \hat{\mathcal{Q}}_{t,h}(s^i_{t,h}, a^i_{t,h}) - \mathcal{Q}^{\mathsf{c},\star}_h(s^i_{t,h}, a^i_{t,h}) + \mathcal{Q}^{\mathsf{c},\star}_h(s^i_{t,h}, a^i_{t,h}) - \mathcal{Q}^{\mathsf{c},\pi_t}_h(s^i_{t,h}, a^i_{t,h}).$$

Next, for empirical Q-values, we have the following bound due to the clipping mechanism and **A**-1

$$\hat{\mathcal{Q}}_{t,h}(s,a) \leq \sum_{i=1}^{N}\frac{n^i_{t,h}(s,a)}{\bar{N}^i_{t,h}}\widehat{\mathsf{r}}^i_h(s,a) + \widehat{\mathsf{P}}_{t,h}\hat{\mathcal{V}}_{t,h+1}(s,a) + b_{t,h}(s,a)$$

$$\leq \mathsf{r}^{\mathsf{c}}_h(s,a) + \widehat{\mathsf{P}}_{t,h}\hat{\mathcal{V}}_{t,h+1}(s,a) + b_{t,h}(s,a) + 2\varepsilon_{\mathsf{r}}\,,$$

thus, applying Bellman equations (36) and optimal Bellman equations (37), we have after a simple rearranging

$$\begin{aligned}\delta^i_{t,h} &\leq \widehat{\mathsf{P}}_{t,h}\hat{\mathcal{V}}_{t,h+1}(s^i_{t,h},a^i_{t,h}) - \mathsf{P}^{\mathsf{c}}_h\mathcal{V}^{\mathsf{c},\star}_{h+1}(s^i_{t,h},a^i_{t,h}) + b_{t,h}(s^i_{t,h},a^i_{t,h}) + \mathsf{P}^{\mathsf{c}}_h\left[\mathcal{V}^{\mathsf{c},\star}_{h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h},a^i_{t,h}) + 2\varepsilon_{\mathsf{r}} \\ &= [\widehat{\mathsf{P}}_{t,h} - \mathsf{P}^{\mathsf{c}}_h]\hat{\mathcal{V}}_{t,h+1}(s^i_{t,h},a^i_{t,h}) + \mathsf{P}^{\mathsf{c}}_h\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h},a^i_{t,h}) + b_{t,h}(s^i_{t,h},a^i_{t,h}) + 2\varepsilon_{\mathsf{r}}\,.\end{aligned}$$

In the decomposition above, we further rearrange it, using a virtual estimate of $\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}$ defined in (24) and re-introducing again the kernel for i-th agent P^i_h

$$\delta^i_{t,h} \leq \underbrace{[\widehat{\mathsf{P}}_{t,h} - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}]\hat{\mathcal{V}}_{t,h+1}(s^i_{t,h},a^i_{t,h})}_{(\mathbf{A})} + \underbrace{[\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h} - \mathsf{P}^{\mathsf{c}}_h]\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\star}_{h+1}\right](s^i_{t,h},a^i_{t,h})}_{(\mathbf{B})} + \underbrace{[\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h} - \mathsf{P}^{\mathsf{c}}_h]\mathcal{V}^{\mathsf{c},\star}_{h+1}(s^i_{t,h},a^i_{t,h})}_{(\mathbf{C})} \tag{52}$$

$$+ \underbrace{[\mathsf{P}^{\mathsf{c}}_h - \mathsf{P}^i_h]\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h},a^i_{t,h})}_{(\mathbf{D})} + \underbrace{\mathsf{P}^i_h\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h},a^i_{t,h}) - \left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h+1})}_{=:\zeta^i_{t,h}} \tag{53}$$

$$+ \underbrace{\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h+1})}_{\delta^i_{t,h+1}} + b_{t,h}(s^i_{t,h},a^i_{t,h}) + 2\varepsilon_{\mathsf{r}}\,. \tag{54}$$

Next, we analyze each term separately. With a slight abuse of notation, let us define $\bar{N}^i_{t,h} = N_{(r_t),h}(s^i_{t,h},a^i_{t,h})$. In the sequel, we analyze only such $(t,i,h) \in [T]\times[M]\times[H]$ such that $\bar{N}^i_{t,h} \geq 2$. In the case where $\bar{N}^i_{t,h} \leq 1$, we have the trivial bound $\delta^i_{t,h} \leq H$.

Terms (A) and (D): heterogeneity errors First, let us handle the terms that come from the presence of heterogeneity between agents. To analyze (**A**), let us apply the definition of the event $\mathcal{E}^{\mathsf{c}}(\delta) \subseteq \mathcal{G}(\delta)$ combined with Holder's inequality

$$(\mathbf{A}) \leq H\|\widehat{\mathsf{P}}_{t,h}(s^i_{t,h},a^i_{t,h}) - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}(s^i_{t,h},a^i_{t,h})\|_1 \leq 2H\varepsilon_{\mathsf{p}} + \frac{11H\beta^{\mathsf{c}}(\delta,\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}}\,.$$

For (**D**) we apply Holder's inequality, **A**-1 and Corollary F.2

$$(\mathbf{D}) \leq 2H\|\mathsf{P}^{\mathsf{c}}_h(s^i_{t,h},a^i_{t,h}) - \mathsf{P}^i_h(s^i_{t,h},a^i_{t,h})\|_1 \leq 2H\varepsilon_{\mathsf{p}}\,.$$

Term (B): correction error To analyze this term, we apply Lemma C.2 with $f(s') := [\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\star}_{h+1}](s')$ and get

$$\begin{aligned}(\mathbf{B}) &\leq \frac{1}{H}\mathsf{P}^{\mathsf{c}}_h\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\star}_{h+1}\right](s^i_{t,h},a^i_{t,h}) + \frac{2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}} \\ &\overset{(1)}{\leq} \frac{1}{H}\mathsf{P}^{\mathsf{c}}_h\left[\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}\right](s^i_{t,h},a^i_{t,h}) + \frac{2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}} \\ &\overset{(2)}{\leq} \frac{1}{H}(\mathbf{D}) + \frac{1}{H}\delta^i_{t,h+1} + \frac{1}{H}\zeta^i_{t,h+1} + \frac{2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}}\,.\end{aligned}$$

where (1) follows from the definition of optimal policy, and (2) follows from a simple rearranging of terms, similar to the decomposition of $\delta^i_{t,h}$. Additional term (**D**) appeared compared to a standard decomposition.

Term (C): concentration error From the definition of the event $\mathcal{E}^\star(\delta) \subseteq \mathcal{G}(\delta)$ defined in Lemma C.1, and from the analysis of the case $\bar{N}^i_{t,h} \geq 2$ it follows that

$$(\mathbf{C}) \leq \sqrt{\frac{2\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h},a^i_{t,h})\beta^\star(\delta)}{\bar{N}^i_{t,h}-1}} + \frac{7\beta^\star(\delta)}{\bar{N}^i_{t,h}-1} \leq \sqrt{\frac{4\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h},a^i_{t,h})\beta^\star(\delta)}{\bar{N}^i_{t,h}}} + \frac{14\beta^\star(\delta)}{\bar{N}^i_{t,h}}\,.$$

Bounding the bonus From the definition of the bonus (39), we have for all $(t,i,h) \in [T] \times [M] \times [H]$ such that $\bar{N}^i_{t,h} \geq 2$

$$b_{t,h}(s^i_{t,h}, a^i_{t,h}) = \frac{28\beta^\star(\delta)H + 11\beta^{\mathsf{c}}(\delta, \bar{N}^i_{t,h})}{\bar{N}^i_{t,h}} + \sqrt{\frac{8\beta^\star(\delta)}{\bar{N}^i_{t,h}} \cdot \mathrm{Var}_{\widehat{\mathsf{P}}_{(r)}}(\hat{\mathcal{V}}_{(r),h+1})(s,a)}\,.$$

Using the inequality (62) of Lemma F.6, we have $\mathrm{Var}_{\widehat{\mathsf{P}}_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h}) \leq \mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h}) + 3H^2\varepsilon_{\mathsf{p}}$. Besides, as for any $a, b \in \mathbb{R}_+$, we have $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we get

$$\sqrt{\frac{8\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h})}{\bar{N}^i_{t,h}}} \leq \sqrt{\frac{8\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h})}{\bar{N}^i_{t,h}}} + \sqrt{\frac{24\varepsilon_{\mathsf{p}}H^2\beta^\star(\delta)}{\bar{N}^i_{t,h}}}\,.$$

Now as for any $a, b \in \mathbb{R}_+$, we have $\sqrt{2ab} \leq a + b$, we get

$$\sqrt{\frac{8\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h})}{\bar{N}^i_{t,h}}} \leq \sqrt{\frac{8\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h})}{\bar{N}^i_{t,h}}} + 3\varepsilon_{\mathsf{p}}H + \frac{4H\beta^\star(\delta)}{\bar{N}^i_{t,h}}\,.$$

Combine everything together After combining all the terms, we have for all $(t,i,h) \in [T] \times [M] \times [H]$ such that $\bar{N}^i_{t,h} \geq 2$

$$\begin{aligned}\delta^i_{t,h} \leq 7H^2\varepsilon_{\mathsf{p}} + 2\varepsilon_{\mathsf{r}} + \left(1 + \frac{1}{H}\right)\delta^i_{t,h+1} + \left(1 + \frac{1}{H}\right)\zeta^i_{t,h+1} + \frac{22\beta^{\mathsf{c}}(\delta, \bar{N}^i_{t,h}) + 46H\beta^\star(\delta) + 2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta, \bar{N}^i_{t,h})}{\bar{N}^i_{t,h}} \\ + \sqrt{\frac{4\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h}, a^i_{t,h})\beta^\star(\delta)}{\bar{N}^i_{t,h}}} + \sqrt{\frac{8\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h})}{\bar{N}^i_{t,h}}}\,.\end{aligned}$$

Let us define $\gamma_h = (1 + 1/H)^{H-h}$. Notice that for any $h \geq 0$ it holds $\gamma_h \leq \mathrm{e}$. Then by summing and expanding over $h \in [H]$ we have

$$\begin{aligned}\overline{\mathfrak{R}}^{\mathsf{c}}_h(T) \leq 7\mathrm{e}TH^2\varepsilon_{\mathsf{p}} + 2\mathrm{e}TH\varepsilon_{\mathsf{r}} + \frac{1}{M}\sum_{i=1}^M\sum_{t=1}^T\sum_{h'=h}^H \gamma_{h'-1}\zeta^i_{t,h'+1} + \frac{1}{M}\sum_{i=1}^M\sum_{t=1}^T\sum_{h'=h}^H 2\mathrm{e}H\mathbb{1}_{[\![0;1]\!]}(\bar{N}^i_{t,h'}) \\ + \frac{\mathrm{e}}{M}\sum_{i=1}^M\sum_{t=1}^T\sum_{h'=h}^H \sqrt{\frac{4\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h'}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h'}, a^i_{t,h'})}{\bar{N}^i_{t,h'}}}\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'}) &=: A^T_h \\ + \frac{\mathrm{e}}{M}\sum_{i=1}^M\sum_{t=1}^T\sum_{h'=h}^H \sqrt{\frac{8\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}^c_{t,h}}(\hat{\mathcal{V}}_{t,h'+1})(s^i_{t,h'}, a^i_{t,h'})}{\bar{N}^i_{t,h'}}}\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'}) &=: B^T_h \\ + \frac{\mathrm{e}}{M}\sum_{i=1}^M\sum_{t=1}^T\sum_{h'=h}^H \frac{22\beta^{\mathsf{c}}(\delta, \bar{N}^i_{t,h}) + 46H\beta^\star(\delta) + 2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta, \bar{N}^i_{t,h})}{\bar{N}^i_{t,h}}\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'})\,. &=: C^T_h\end{aligned}$$

To conclude the statement, we apply a definition of the event $\mathcal{E}(\delta)$ to the third term in the decomposition above. □

Lemma D.4. *Define $\bar{N}^i_{t,h} = N_{(r_t),h}(s^i_{t,h}, a^i_{t,h})$. Assume conditions of Lemma C.1. Then, on the event $\mathcal{G}(\delta)$, the following inequalities holds:*

$$\begin{aligned}\sum_{i=1}^M\sum_{t=1}^T\sum_{h=1}^H \frac{\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}} \leq 4|\mathcal{S}||\mathcal{A}|H\log\left(\frac{\mathrm{e}MTH}{|\mathcal{S}||\mathcal{A}|}\right)\,, \\ \sum_{i=1}^M\sum_{t=1}^T\sum_{h=1}^H \sqrt{\frac{\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}{N^{(r,l)}_{h,i}}} \leq 8H\sqrt{|\mathcal{S}||\mathcal{A}|MT}\,,\end{aligned}$$

$$\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} 1_{[\![0;1]\!]}(\bar{N}^i_{t,h}) \leq 4H|\mathcal{S}||\mathcal{A}| \ .$$

Proof. The quantity $\widetilde{N}^i_{t,h}(s,a)$ represents the exact number of visits of the pair (s,a) at step h until episode t, and after the first i agents executed the h-step. We want to bound $\bar{N}^i_{t,h}$ using $\widetilde{N}^i_{t,h}(s^i_{t,h},a^i_{t,h})$ so that we can compute the latter sums by applying the pigeon-hole principle. To derive such a bound, we distinguish two cases:

Case 1: $N_{r_t,h}(s,a) < \nu(\delta,T)$ In this case, by the synchronization rule described in Algorithm 2, we have $n^i_{t,h}(s,a) < 2n^i_{(r_t),h}(s,a)$. If we sum the latter inequality over all the agents, we obtain $\widetilde{N}^M_{t,h}(s,a) \leq 2N_{(r_t),h}(s,a)$. Now using definition of $\widetilde{N}^i_{t,h}(s,a)$ yields

$$N_{(r_t),h}(s,a) \leq \widetilde{N}^i_{t,h}(s,a) \leq 2N_{(r_t),h}(s,a) \ .$$

Case 2: $N_{r_t,h}(s,a) \geq \nu(\delta,T)$ In this case, the synchronization rule ensures $\hat{N}^i_{t,h}(s,a) \leq 2N_{r_t,h}(s,a)$. Conditioned on $\mathcal{E}^{\text{count}}(\delta)$, we have $\widetilde{N}^M_{t,h}(s,a) \leq \frac{10}{7}\hat{N}^i_{t,h}(s,a)$. Combining the two latter inequalities gives

$$N_{(r_t),h}(s,a) \leq \widetilde{N}^i_{t,h}(s,a) \leq 4N_{(r_t),h}(s,a) \ ,$$

where the lower bound follows directly from the definition of $\widetilde{N}^i_{t,h}$. Using the two previous inequalities, we derive the following bound

$$\bar{N}^i_{t,h} \leq \widetilde{N}^i_{t,h}(s^i_{t,h},a^i_{t,h}) \leq 4\bar{N}^i_{t,h} \ .$$

Applying the latter inequality in the first sum of the lemma yields

$$\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{1_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}} \leq \sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{4\cdot 1_{[\![1;+\infty]\!]}(\tilde{N}^i_{t,h})}{\tilde{N}^i_{t,h}} \ .$$

By construction, this counter is thus incremented by at most 1 every time and we can apply the pigeon-hole principle on this counter which yields

$$\begin{aligned}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{1_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}} &\leq \sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{4\cdot 1_{[\![1;+\infty]\!]}(\tilde{N}^i_{t,h})}{\tilde{N}^i_{t,h}} \leq 4\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_{T,h}(s,a)}\frac{1}{n} \\ &\leq 4\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}(\log(N_{T,h}(s,a))+1) \leq 4|\mathcal{S}||\mathcal{A}|H\log\left(\frac{\mathrm{e}MTH}{|\mathcal{S}||\mathcal{A}|}\right) \ ,\end{aligned}$$

where we used the concavity of the logarithm in the last inequality. Similarly, we have

$$\begin{aligned}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \sqrt{\frac{1_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}{\bar{N}^i_{t,h}}} &\leq \sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \sqrt{\frac{4\cdot 1_{[\![1;+\infty]\!]}(\tilde{N}^i_{t,h})}{\tilde{N}^i_{t,h}}} \\ &\leq 2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_{T,h}(s,a)}\sqrt{\frac{1}{n}} \leq 8\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sqrt{N_{T,h}(s,a)} \leq 8H\sqrt{|\mathcal{S}||\mathcal{A}|MT} \ ,\end{aligned}$$

where we used the concavity of the square root in the last inequality. Now as $\widetilde{N}^i_{t,h}(s^i_{t,h},a^i_{t,h}) \leq 4\bar{N}^i_{t,h}$, then we have $1_{[\![0;1]\!]}(\bar{N}^i_{t,h}) \leq 1_{[\![0;4]\!]}(\tilde{N}^i_{t,h})$. Plugging in the latter inequality in the last sum of the lemma yields

$$\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} 1_{[\![0;1]\!]}(\bar{N}^i_{t,h}) \leq \sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} 1_{[\![0;4]\!]}(\tilde{N}^i_{t,h}) \leq \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{n=1}^{4} 1 \leq 4H|\mathcal{S}||\mathcal{A}| \ .$$

□

For ease of reading, we define $\beta^{\max}(\delta)$ as

$$\beta^{\max}(\delta) := \max\left(\beta^{\mathrm{KL}}(\delta, MT), \beta^{\mathsf{c}}(\delta, MT), \beta^{\star}(\delta), \beta(\delta), \beta^{\mathrm{Var}}(\delta, T), \log\left(\frac{\mathrm{e}MTH}{|\mathcal{S}||\mathcal{A}|}\right)\right) . \tag{55}$$

Lemma D.5. *Assume conditions of Lemma C.1. Then, on the event $\mathcal{G}(\delta)$, the following inequality holds*

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\widehat{\mathsf{P}}^{c}_{t,h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h}, a^i_{t,h})\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) \le 2H^2T + 2H^2U_1^T + 11H^3T\varepsilon_{\mathsf{p}} + 6HT\varepsilon_{\mathsf{r}} + 30H^3\beta^{\max}(\delta)|\mathcal{S}||\mathcal{A}|^{1/2}T^{1/2}M^{-1/2} ,$$

and we also have

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\widehat{\mathsf{P}}^{c}_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h})\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) \le 2H^2T + 2H^2U_1^T + 17H^3T\varepsilon_{\mathsf{p}} + 10HT\varepsilon_{\mathsf{r}} + 30H^3\beta^{\max}(\delta)|\mathcal{S}||\mathcal{A}|^{1/2}T^{1/2}M^{-1/2} ,$$

where $\beta^{\max}(\delta)$ is defined in Lemma C.1 as a worst-case concentration logarithmic factor.

Proof. Using inequality (62) of Lemma F.6, we have

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\widehat{\mathsf{P}}^{c}_{t,h}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h}, a^i_{t,h})\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) \le \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h}, a^i_{t,h})\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}_{(\mathbf{W})} + \underbrace{3H^2\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\|(\mathsf{P}^i_h - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h})(s^i_{t,h}, a^i_{t,h})\|_1\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}_{(\mathbf{X})} .$$

Term (X): We have for any $(s,a) \in \mathcal{S}\times\mathcal{A}$

$$\|\mathsf{P}^i_h(s,a) - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}(s,a)\|_1 \le \|\mathsf{P}^i_h(s,a) - \mathsf{P}^{\mathsf{c}}_h(s,a)\|_1 + \|\mathsf{P}^{\mathsf{c}}_h(s,a) - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}(s,a)\|_1 \le \varepsilon_{\mathsf{p}} + \sqrt{\frac{2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta, \bar{N}^i_{t,h})}{\bar{N}^i_{t,h}}} , \tag{56}$$

where the bound on $\|\mathsf{P}^i_h(s,a) - \mathsf{P}^{\mathsf{c}}_h(s,a)\|_1$ is provided by Lemma F.1 and the bound on $\|\mathsf{P}^{\mathsf{c}}_h(s,a) - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}(s,a)\|_1$ is provided by the second inequality of Lemma C.2. Thus we get

$$3H^2\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\|(\mathsf{P}^{\mathsf{c}}_h - \widehat{\mathsf{P}}^{\mathsf{c}}_{t,h})(s^i_{t,h}, a^i_{t,h})\|_1\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) \le \frac{3H^2}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\varepsilon_{\mathsf{p}} + \sqrt{\frac{2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta, \bar{N}^i_{t,h})}{\bar{N}^i_{t,h}}}\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) .$$

Finally applying Lemma D.4 yields

$$(\mathbf{X}) \le 3H^3T\varepsilon_{\mathsf{p}} + 24\sqrt{\frac{2H^6\beta^{\mathrm{KL}}(\delta, MT)|\mathcal{S}|^2|\mathcal{A}|T}{M}} .$$

Term (W): Using inequality (61) of Lemma F.6, we have

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h}, a^i_{t,h})\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) \le \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}2\,\mathrm{Var}_{\mathsf{P}^i_h}(\mathcal{V}^{i,\pi_t}_{h+1})(s^i_{t,h}, a^i_{t,h})\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}_{(\mathbf{Y})} \le \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}2H\mathsf{P}^i_h|\mathcal{V}^{\mathsf{c},\star}_{h+1} - \mathcal{V}^{i,\pi_t}_{h+1}|(s^i_{t,h}, a^i_{t,h})\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h})}_{(\mathbf{Z})} ,$$

where we recall that $\mathcal{V}_{h+1}^{i,\pi_t}$ is the value function of the policy π_t in the environment of the i-th agent (1). Conditioned on $\mathcal{E}^{\mathrm{Var}}(\delta)$, we have

$$(\mathbf{Y}) \leq \sqrt{\frac{8H^5T\beta^{\mathrm{Var}}(\delta,T)}{M}} + \frac{6H^3\beta^{\mathrm{Var}}(\delta,T)}{M} + 2H^2T\,.$$

Now by Corollary F.2, we have conditioned on $\mathcal{E}^{\mathrm{optimism}}$ for all $s \in \mathcal{S}$

$$\begin{aligned}|\mathcal{V}_{h+1}^{\mathsf{c},\star}(s) - \mathcal{V}_{h+1}^{i,\pi_t}(s)| &\leq |\mathcal{V}_{h+1}^{\mathsf{c},\pi_t}(s) - \mathcal{V}_{h+1}^{i,\pi_t}(s)| + \mathcal{V}_{h+1}^{\mathsf{c},\star}(s) - \mathcal{V}_{h+1}^{\mathsf{c},\pi_t}(s)\\ &\leq \varepsilon_{\mathsf{p}}H^2 + \varepsilon_{\mathsf{r}}H + \hat{\mathcal{V}}_{t,h+1}(s) - \mathcal{V}_{h+1}^{\mathsf{c},\pi_t}(s) + (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H-h)\\ &\leq 4\varepsilon_{\mathsf{p}}H^2 + 3\varepsilon_{\mathsf{r}}H + \hat{\mathcal{V}}_{t,h+1}(s) - \mathcal{V}_{h+1}^{\mathsf{c},\pi_t}(s)\,.\end{aligned}$$

Using the definition of $\delta_{t,h}^i$ and $\zeta_{t,h+1}^i$ introduced in Lemma D.3, we have

$$\begin{aligned}(\mathbf{Z}) &\leq \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} 2H(4\varepsilon_{\mathsf{p}}H^2 + 3\varepsilon_{\mathsf{r}} + \delta_{t,h}^i + \zeta_{t,h+1}^i)\\ &\leq 8H^3T\varepsilon_{\mathsf{p}} + 6HT\varepsilon_{\mathsf{r}} + \sqrt{\frac{8H^5T\beta(\delta)}{M}} + 2H\sum_{h=1}^{H}\overline{\mathfrak{R}}_{h+1}^{\mathsf{c}}(T) \leq 8H^3T\varepsilon_{\mathsf{p}} + 6HT\varepsilon_{\mathsf{r}} + \sqrt{\frac{8H^5T\beta(\delta)}{M}} + 2H^2U_1^T\,,\end{aligned}$$

where the second inequality holds conditioned on $\mathcal{E}(\delta)$. Combining everything yields

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\mathsf{P}_h^i}(\mathcal{V}_{h+1}^{\mathsf{c},\star})(s_{t,h}^i,a_{t,h}^i)1_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i) &\leq (\mathbf{X}) + (\mathbf{Y}) + (\mathbf{Z})\\ &\leq 11H^3T\varepsilon_{\mathsf{p}} + 24\sqrt{\frac{2H^6\beta^{\mathrm{KL}}(\delta,MT)|\mathcal{S}|^2|\mathcal{A}|T}{M}} + \sqrt{\frac{8H^5T\beta^{\mathrm{Var}}(\delta,T)}{M}} + \frac{6H^3\beta^{\mathrm{Var}}(\delta,T)}{M} + 2H^2T\\ &\quad + 6HT\varepsilon_{\mathsf{r}} + \sqrt{\frac{8H^5T\beta(\delta)}{M}} + 2H^2U_1^T\end{aligned}$$

Now let's move to the second inequality of this lemma. Again by using inequality (62) of Lemma F.6, we have

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\widehat{\mathsf{P}}_{t,h}^c}(\hat{\mathcal{V}}_{t,h+1})(s_{t,h}^i,a_{t,h}^i)1_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i) &\leq \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\mathsf{P}_h^i}(\hat{\mathcal{V}}_{t,h+1})(s_{t,h}^i,a_{t,h}^i)1_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i)}_{(\mathbf{W}')}\\ &\quad + \underbrace{3H^2\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\|(\mathsf{P}_h^i - \widehat{\mathsf{P}}_{t,h}^{\mathsf{c}})(s_{t,h}^i,a_{t,h}^i)\|_1 1_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i)}_{(\mathbf{X})}\,.\end{aligned}$$

Term $(\mathbf{W}')$: Using inequality (61) of Lemma F.6, we have

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\mathsf{P}_h^i}(\hat{\mathcal{V}}_{t,h+1})(s_{t,h}^i,a_{t,h}^i)1_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i) &\leq \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}2\,\mathrm{Var}_{\mathsf{P}_h^i}(\mathcal{V}_{h+1}^{i,\pi_t})(s_{t,h}^i,a_{t,h}^i)1_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i)}_{(\mathbf{Y})}\\ &\leq \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}2H\mathsf{P}_h^i|\hat{\mathcal{V}}_{t,h+1} - \mathcal{V}_{h+1}^{i,\pi_t}|(s_{t,h}^i,a_{t,h}^i)1_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i)}_{(\mathbf{Z}')}\,.\end{aligned}$$

Now again by Corollary F.2, we have conditioned on $\mathcal{E}^{\mathrm{optimism}}$ for all $s \in \mathcal{S}$

$$|\hat{\mathcal{V}}_{t,h+1}(s) - \mathcal{V}_{h+1}^{i,\pi_t}(s)| \leq |\mathcal{V}_{h+1}^{\mathsf{c},\pi_t}(s) - \mathcal{V}_{h+1}^{i,\pi_t}(s)| + \hat{\mathcal{V}}_{t,h+1}(s) + (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H-h) - \mathcal{V}_{h+1}^{\mathsf{c},\pi_t} + (2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}}H)(H-h)$$

$$\leq \varepsilon_{\mathsf{p}} H^2 + \varepsilon_{\mathsf{r}} H + \hat{\mathcal{V}}_{t,h+1}(s) - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}(s) + 2(2\varepsilon_{\mathsf{r}} + 3\varepsilon_{\mathsf{p}} H)(H-h)$$
$$\leq 7\varepsilon_{\mathsf{p}} H^2 + 5\varepsilon_{\mathsf{r}} H + \hat{\mathcal{V}}_{t,h+1}(s) - \mathcal{V}^{\mathsf{c},\pi_t}_{h+1}(s) \, .$$

By combining the bounds that we have on $(\mathbf{X}), (\mathbf{Y})$, and $(\mathbf{Z}')$, we derive the following bound

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h}) \mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) \leq (\mathbf{X}) + (\mathbf{Y}) + (\mathbf{Z}')$$
$$\leq 17H^3 T\varepsilon_{\mathsf{p}} + 24\sqrt{\frac{2H^6\beta^{\mathrm{KL}}(\delta, MT)|\mathcal{S}|^2|\mathcal{A}|T}{M}} + \sqrt{\frac{8H^5 T\beta^{\mathrm{Var}}(\delta,T)}{M}} + \frac{6H^3\beta^{\mathrm{Var}}(\delta,T)}{M} + 10H^2 T$$
$$+ 10HT\varepsilon_{\mathsf{r}} + \sqrt{\frac{8H^5 T\beta(\delta)}{M}} + 2H^2 U_1^T \, .$$

Finally, as we have

$$24\sqrt{\frac{2H^6\beta^{\mathrm{KL}}(\delta, MT)|\mathcal{S}|^2|\mathcal{A}|T}{M}} \leq 48H^3\beta^{\max}(\delta)|\mathcal{S}||\mathcal{A}|^{1/2}T^{1/2}M^{-1/2}$$
$$\sqrt{\frac{8H^5 T\beta(\delta)}{M}} \leq 3H^3\beta^{\max}(\delta)|\mathcal{S}||\mathcal{A}|^{1/2}T^{1/2}M^{-1/2}$$
$$\sqrt{\frac{8H^5 T\beta^{\mathrm{Var}}(\delta,T)}{M}} \leq 3H^3\beta^{\max}(\delta)|\mathcal{S}||\mathcal{A}|^{1/2}T^{1/2}M^{-1/2}$$
$$\frac{6H^3\beta^{\mathrm{Var}}(\delta,T)}{M} \leq 6H^3\beta^{\max}(\delta)|\mathcal{S}||\mathcal{A}|^{1/2}T^{1/2}M^{-1/2} \, ,$$

then

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h}}(\hat{\mathcal{V}}_{t,h+1})(s^i_{t,h}, a^i_{t,h}) \mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h}) \leq 17H^3 T\varepsilon_{\mathsf{p}} + 10HT\varepsilon_{\mathsf{r}} + 2H^2 T + 2H^2 U_1^T$$
$$+ 60H^3\beta^{\max}(\delta)|\mathcal{S}||\mathcal{A}|^{1/2}T^{1/2}M^{-1/2} \, .$$

□

Lemma D.6. *Assume conditions of Lemma C.1. Then, on the event $\mathcal{G}(\delta)$, the following inequality holds*

$$A_1^T \leq 23\mathrm{e}\cdot\beta^{\max}(\delta)\cdot\sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}} + 23\mathrm{e}\beta^{\max}(\delta)\sqrt{H^3|\mathcal{S}||\mathcal{A}|U_1^T M^{-1}}$$
$$+ 16\mathrm{e}\beta^{\max}(\delta)T(6H^2\varepsilon_{\mathsf{p}} + 3H\varepsilon_{\mathsf{r}}) + 96\mathrm{e}H^3|\mathcal{S}|^{3/2}|\mathcal{A}|M^{-1/2}(\beta^{\max}(\delta))^2 \, ,$$
$$B_1^T \leq 46\mathrm{e}\cdot\beta^{\max}(\delta)\cdot\sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}} + 46\mathrm{e}\beta^{\max}(\delta)\sqrt{H^3|\mathcal{S}||\mathcal{A}|U_1^T M^{-1}}$$
$$+ 32\mathrm{e}\beta^{\max}(\delta)T(6H^2\varepsilon_{\mathsf{p}} + 3H\varepsilon_{\mathsf{r}}) + 192\mathrm{e}H^3|\mathcal{S}|^{3/2}|\mathcal{A}|M^{-1/2}(\beta^{\max}(\delta))^2 \, ,$$
$$C_1^T \leq 272|\mathcal{S}|^2|\mathcal{A}|M^{-1}H^3(\beta^{\max}(\delta))^2 \, .$$

Proof. **Term A_1^T.** To bound the term A_1^T, we start by applying Cauchy-Schwartz inequality

$$A_1^T = \frac{\mathrm{e}}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H}\sqrt{\frac{4\beta^\star(\delta)\,\mathrm{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h'}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h'}, a^i_{t,h'})}{\bar{N}^i_{t,h'}}}\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'})$$
$$\leq \frac{\mathrm{e}}{\sqrt{M}}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H}\mathrm{Var}_{\widehat{\mathsf{P}}^{\mathsf{c}}_{t,h'}}(\mathcal{V}^{\mathsf{c},\star}_{h+1})(s^i_{t,h'}, a^i_{t,h'})\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'})}\sqrt{\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H}\frac{4\beta^\star(\delta)\mathbb{1}_{[\![2;+\infty]\!]}(\bar{N}^i_{t,h'})}{\bar{N}^i_{t,h'}}} \, ,$$

Now, applying Lemma D.5, Lemma D.4, and the subadditivity of the square root, we obtain

$$A_1^T \leq \frac{16\beta^{\max}(\delta)\mathrm{e}}{\sqrt{M}}\Big(\sqrt{2H^3|\mathcal{S}||\mathcal{A}|T} + \sqrt{2H^3|\mathcal{S}||\mathcal{A}|U_1^T} + \sqrt{H^2|\mathcal{S}||\mathcal{A}|T\cdot(11H^2\varepsilon_{\mathsf{p}} + 6H\varepsilon_{\mathsf{r}})}$$

$$+ \sqrt{60H^4|\mathcal{S}|^2|\mathcal{A}|^{3/2}T^{1/2}M^{-1/2} \cdot \beta^{\max}(\delta)}\Big) \, .$$

Next, we analyze the last two terms in the upper bound above. For the third one, by a standard inequality $\sqrt{2ab} \leq a + b$ it holds that

$$\sqrt{H^2|\mathcal{S}||\mathcal{A}|T \cdot (11H^2\varepsilon_{\mathsf{p}} + 6H\varepsilon_{\mathsf{r}})} \leq 3HT\varepsilon_{\mathsf{r}} + 6H^2T\varepsilon_{\mathsf{p}} + H^2|\mathcal{S}||\mathcal{A}| \, .$$

Notably, the first two terms already appeared in the regret decomposition; see Lemma D.3. For the last term, the decomposition is more standard

$$\sqrt{60H^4|\mathcal{S}|^2|\mathcal{A}|^{3/2}T^{1/2}M^{-1/2}\beta^{\max}(\delta)} \leq 6\sqrt{H^3|\mathcal{S}||\mathcal{A}|T} + 5H^{5/2}|\mathcal{S}|^{3/2}|\mathcal{A}|M^{-1/2}\beta^{\max}(\delta) \, .$$

Thus, by a simple rearranging of the terms and applying inequalities $M \geq 1, H \geq 1$, we have

$$\begin{aligned} A_1^T &\leq 23\mathrm{e} \cdot \beta^{\max}(\delta) \cdot \sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}} + 23\mathrm{e}\beta^{\max}(\delta)\sqrt{H^3|\mathcal{S}||\mathcal{A}|U_1^TM^{-1}} \\ &\quad + 16\mathrm{e}\beta^{\max}(\delta)T(6H^2\varepsilon_{\mathsf{p}} + 3H\varepsilon_{\mathsf{r}}) + 96\mathrm{e}H^3|\mathcal{S}|^{3/2}|\mathcal{A}|M^{-1/2}(\beta^{\max}(\delta))^2 \, . \end{aligned}$$

Term B_1^T. Similarly, the bound for the term B_1^T is derived using a combination of Cauchy-Scwartz, Lemma D.5, Lemma D.4, and the subadditivity of the square root

$$\begin{aligned} B_1^T &\leq 46\mathrm{e} \cdot \beta^{\max}(\delta) \cdot \sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}} + 46\mathrm{e}\beta^{\max}(\delta)\sqrt{H^3|\mathcal{S}||\mathcal{A}|U_1^TM^{-1}} \\ &\quad + 32\mathrm{e}\beta^{\max}(\delta)T(6H^2\varepsilon_{\mathsf{p}} + 3H\varepsilon_{\mathsf{r}}) + 192\mathrm{e}H^3|\mathcal{S}|^{3/2}|\mathcal{A}|M^{-1/2}(\beta^{\max}(\delta))^2 \, . \end{aligned}$$

Term C_1^T. Finally to estimate C_1^T, we apply Lemma D.4

$$\begin{aligned} &\frac{\mathrm{e}}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h'=h}^{H}\frac{22\beta^{\mathsf{c}}(\delta,\bar{N}_{t,h}^i) + 46H\beta^{\star}(\delta) + 2H^2|\mathcal{S}|\beta^{\mathrm{KL}}(\delta,\bar{N}_{t,h}^i)}{\bar{N}_{t,h}^i}\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}_{t,h'}^i) \\ &\qquad \leq \frac{68H^2|\mathcal{S}|\beta^{\max}(\delta)}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{\bar{N}_{t,h}^i}\mathbf{1}_{[\![2;+\infty]\!]}(\bar{N}_{t,h}^i) \leq 272|\mathcal{S}|^2|\mathcal{A}|M^{-1}H^3(\beta^{\max}(\delta))^2 \, . \end{aligned}$$

□

D.3 Proof of Theorem 4.1

Hereafter, we establish the following bound on the regret

$$\mathfrak{R}(T) \leq 138\mathrm{e}\sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}(\beta^{\max}(\delta))^2} + 6020\mathrm{e}^2H^3|\mathcal{S}|^2|\mathcal{A}|(\beta^{\max}(\delta))^2 + \mathrm{e}\beta^{\max}(\delta)TH(595H\varepsilon_{\mathsf{p}} + 148H\varepsilon_{\mathsf{r}}) \, .$$

Proof. Let us start by moving from our regret to a regret connected to a common kernel, using a combination of Corollary F.2 and **A**-1

$$\mathfrak{R}(T) = \max_{\pi}\frac{1}{M}\sum_{t=1}^{T}\sum_{i=1}^{M}\mathcal{V}_1^{i,\pi}(s_{t,1}^i) - \mathcal{V}_1^{i,\pi_t}(s_{t,1}^i) \leq \underbrace{\frac{1}{M}\sum_{t=1}^{T}\sum_{i=1}^{M}\mathcal{V}_1^{\mathsf{c},\star}(s_{t,1}^i) - \mathcal{V}_1^{\mathsf{c},\pi_t}(s_{t,1}^i)}_{\mathfrak{R}^{\mathsf{c}}(T)} + 2T\varepsilon_{\mathsf{p}}H^2 + 2T\varepsilon_{\mathsf{r}}H \, .$$

Next, we assume that the event $\mathcal{G}(\delta)$, defined in Lemma C.1 holds. Then Lemma D.2 implies

$$\mathfrak{R}^{\mathsf{c}}(T) \leq \overline{\mathfrak{R}}_1^{\mathsf{c}}(T) + 3T\varepsilon_{\mathsf{p}}H^2 + 2T\varepsilon_{\mathsf{r}}H = \frac{1}{M}\sum_{t=1}^{T}\sum_{i=1}^{M}\hat{\mathcal{V}}_{t,1}(s_{t,1}^i) - \mathcal{V}_1^{\mathsf{c},\pi_t}(s_{t,1}^i) + 3T\varepsilon_{\mathsf{p}}H^2 + 2T\varepsilon_{\mathsf{r}}H \, .$$

By Lemma D.3 we have

$$\overline{\mathfrak{R}}_1^{\mathsf{c}}(T) \leq U_1^T = A_1^T + B_1^T + C_1^T + 7\mathrm{e}TH^2\varepsilon_{\mathsf{p}} + 2\mathrm{e}TH\varepsilon_{\mathsf{r}} + \sqrt{8H^3T \cdot \beta(\delta)/M} + 4\mathrm{e}H^2|\mathcal{S}||\mathcal{A}| \, ,$$

and, applying Lemma D.6, we have the following quadratic inequality on U_1^T

$$\begin{aligned}U_1^T &\le 69\mathrm{e}\cdot\beta^{\max}(\delta)\cdot\sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}} + 69\mathrm{e}\beta^{\max}(\delta)\sqrt{H^3|\mathcal{S}||\mathcal{A}|U_1^TM^{-1}}\\ &+ 144\mathrm{e}\beta^{\max}(\delta)T(2H^2\varepsilon_{\mathsf{p}}+\varepsilon_{\mathsf{r}}H) + 288\mathrm{e}H^3|\mathcal{S}|^{3/2}|\mathcal{A}|M^{-1/2}(\beta^{\max}(\delta))^2\\ &+ 272|\mathcal{S}|^2|\mathcal{A}|M^{-1}H^3(\beta^{\max}(\delta))^2 + \mathrm{e}T(7H^2\varepsilon_{\mathsf{p}}+2H\varepsilon_{\mathsf{r}}) + \sqrt{8H^3TM^{-1}\beta^{\max}(\delta)} + 4\mathrm{e}H^2|\mathcal{S}||\mathcal{A}|\,.\end{aligned}$$

After some rearranging of the terms, we have the following simplified version

$$\begin{aligned}U_1^T &\le 69\mathrm{e}\cdot\beta^{\max}(\delta)\sqrt{H^3|\mathcal{S}||\mathcal{A}|U_1^TM^{-1}} + 71\mathrm{e}\cdot\beta^{\max}(\delta)\cdot\sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}}\\ &+ \mathrm{e}\beta^{\max}(\delta)T(295H^2\varepsilon_{\mathsf{p}}+146H\varepsilon_{\mathsf{r}}) + 560H^3|\mathcal{S}|^2|\mathcal{A}|(\beta^{\max}(\delta))^2\,.\end{aligned}$$

Finally, using inequality $2ab\le a^2+b^2$, we have

$$69\mathrm{e}\cdot\beta^{\max}(\delta)\sqrt{H^3|\mathcal{S}||\mathcal{A}|U_1^TM^{-1}} \le \frac{1}{2}U_1^T + 2450\mathrm{e}^2H^3|\mathcal{S}||\mathcal{A}|M^{-1}(\beta^{\max}(\delta))^2\,,$$

thus

$$U_1^T \le 138\mathrm{e}\sqrt{H^3|\mathcal{S}||\mathcal{A}|TM^{-1}(\beta^{\max}(\delta))^2} + 6020\mathrm{e}^2H^3|\mathcal{S}|^2|\mathcal{A}|(\beta^{\max}(\delta))^2 + \mathrm{e}\beta^{\max}(\delta)TH(590H\varepsilon_{\mathsf{p}}+144H\varepsilon_{\mathsf{r}})\,.$$

□

E COMMUNICATION COMPLEXITY

In the sequel, we prove the bound on the communication complexity of `Fed-UCBVI` stated in Lemma 4.1.

Lemma 4.1 (Communication Complexity)**.** *With probability at least $1-\delta$, the number of communication rounds of `Fed-UCBVI` is bounded by*

$$\begin{aligned}\mathfrak{C}(T) \le \mathcal{O}\big(&|\mathcal{S}||\mathcal{A}|H\log T + M|\mathcal{S}||\mathcal{A}|H\log\log T\\ &+ M|\mathcal{S}||\mathcal{A}|H\log(1+\varepsilon_{\mathsf{p}}T)\big)\,,\end{aligned}$$

where logarithmic dependence in $|\mathcal{S}|, |\mathcal{A}|$, H, $1/\delta$ and M is ignored.

Proof. Let us fix $(s,a,h)\in|\mathcal{S}|\times\mathcal{A}\times[H]$ and bound the maximum number of abortion signals triggered by this triplet. We define R as the value of the variable r that indicates the current round of the communication, defined in `Fed-UCBVI`, during iteration T. Let us also define $k_{s,a,h,i}$ as the number of times agent i triggered the synchronization rule because of the triplet (s,a,h) and $k_{s,a,h}$ the number of times the synchronization rule was triggered because of the triplet (s,a,h). Recall that

$$\nu(\delta,T) = 14\varepsilon_{\mathsf{p}}THM + 182M\beta^{\mathsf{c}}(\delta,T)\,. \tag{57}$$

We distinguish two cases:

(1) $N_{(R),h}(s,a)\le\nu(\delta,T)$: Thus, it holds that $2^{k_{s,a,h,i}}\le n^i_{(R),h}\le\nu(\delta,T)$. Thereby $k_{s,a,h,i}\le\log(\nu(\delta,T))$. Hence, we have $k_{s,a,h}\le M\log(\nu(\delta,T))$.

(2) $N_{(R),h}(s,a)>\nu(\delta,T)$: In this case, we can define

$$k^{\min}_{s,a,h} = \min\{r\in[R] : N_{(r),h}(s,a)\le\nu(\delta,T) \text{ and } N_{(r+1),h}(s,a)>\nu(\delta,T)\}\,.$$

By the precedent case, we have $k^{\min}_{s,a,h}\le M\log(\nu(\delta,T))$. Now let us denote by $r_1,\dots,r_p$, where $p=k_{s,a,h}-k^{\min}_{s,a,h}$, the indices of the rounds where the synchronization rule was triggered because of the triplet (s,a,h) starting from round $k^{\min}_{s,a,h}$. Thus, for a certain $i\in[M]$, we have

$$\hat{N}^i_{(r_{t+1}),h}(s,a)\ge 2N_{(r_{t+1}-1),h}(s,a)\,.$$

Under the event $\mathcal{E}^{\text{count}}(\delta)$, we have for any $t \in [1; p]$,

$$\widetilde{N}^{M}_{(r_{t+1}),h}(s,a) \geq \frac{3}{7}\hat{N}^{i}_{(r_{t+1}),h}(s,a) + \frac{1}{7}\nu(\delta,T) \geq \frac{4}{7}\hat{N}^{i}_{(r_{t+1}),h}(s,a),$$

where $\widetilde{N}^{M}_{(r_{t+1}),h}(s,a)$ is defined in (29). Combining the two previous inequalities, it gives

$$\widetilde{N}^{M}_{(r_{t+1}),h}(s,a) \geq 2 \cdot (4/7) N_{(r_{t+1}-1),h}(s,a) \geq (8/7) N_{(r_t),h}(s,a) = (8/7)\widetilde{N}^{M}_{(r_t),h}(s,a) \ ,$$

where the second inequality comes from $r_{t+1} > \mathsf{r}_t$ and monotonicity of the counters. Unrolling the last recursion yields

$$TM \geq \widetilde{N}^{M}_{(r_p),h}(s,a) \geq (8/7)^{k_{s,a,h} - k^{\min}_{s,a,h}} \nu(\delta,T) \ .$$

Thus, we obtain

$$k_{s,a,h} \leq k^{\min}_{s,a,h} + \frac{\log(TM/\nu(\delta,T))}{\log(8/7)} \leq M \log\left(\nu(\delta,T)\right) + \frac{\log(TM/\nu(\delta,T))}{\log(8/7)} \ ,$$

which yields

$$\mathfrak{C}(T) \leq R_{\max} := M|\mathcal{S}||\mathcal{A}|H \log\left(\nu(\delta,T)\right) + |\mathcal{S}||\mathcal{A}|H \frac{\log(TM/\nu(\delta,T))}{\log(8/7)} \ . \tag{58}$$

□

F TECHNICAL LEMMAS

Lemma F.1 (ℓ_1-norm Bound)**.** *Assume* ***A*****-1***, then*

$$\max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \|\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a) - \mathsf{P}^{i}_h(\cdot|s,a)\|_1 \leq \varepsilon_{\mathsf{p}} \ .$$

Proof. Let $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. Using **A**-1, we get

$$\|\mathsf{P}^{\mathsf{c}}_h(\cdot|s,a) - \mathsf{P}^{i}_h(\cdot|s,a)\|_1 = \textstyle\sum_{s'\in\mathcal{S}} |\mathsf{P}^{\mathsf{c}}_h(s'|s,a) - \mathsf{P}^{i}_h(s'|s,a)| = \sum_{s'\in\mathcal{S}} \varepsilon_{\mathsf{p}} |\mathsf{P}^{\mathsf{c}}_h(s'|s,a) - \mathsf{P}^{\mathsf{ind},i}_h(s'|s,a)| \leq \varepsilon_{\mathsf{p}} \ .$$

□

Lemma F.2. *For any policy π, for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, and for any $(i,j) \in [M]^2$, we have*

$$|d^{i,\pi}_{s_0,h}(s,a) - d^{j,\pi}_{s_0,h}(s,a)| \leq \varepsilon_{\mathsf{p}} H \ .$$

Proof. Let us consider two following MDPs $\mathcal{M}_1 = (\mathcal{S},\mathcal{A},H,\{\mathsf{P}_{h',i}\}_{1\leq h'\leq H}, \{1_{(s,a)}(\cdot)1_h(h')\}_{1\leq h'\leq H})$ and $\mathcal{M}_2 = (\mathcal{S},\mathcal{A},H,\{\mathsf{P}_{h',j}\}_{1\leq h'\leq H}, \{1_{(s,a)}(\cdot)1_h(h')\}_{1\leq h'\leq H})$. Let's denote by $\tilde{V}^{i,\pi}_h$ and $\tilde{V}^{j,\pi}_h$ the values function associated with the policy π in these two respective environments. We have

$$\tilde{V}^{i,\pi}_h(s_0) = \mathbb{E}_\pi\left[\sum_{h'=1}^{H} 1_{(s,a)}(s^i_h,a^i_h)1_h(h')\right] = \mathbb{E}_\pi\left[1_{(s,a)}(s^i_h,a^i_h)\right] = d^{i,\pi}_{s_0,h}(s,a) \ .$$

Similarly, we have $\tilde{V}^{j,\pi}_h(s_0) = d^{j,\pi}_{s_0,h}(s,a)$. Finally applying Lemma F.8 combined with Holder's inequality, and the fact that $\|\tilde{V}^{j,\pi}_h\|_\infty \leq 1$ yields

$$|d^{i,\pi}_{s_0,h}(s,a) - d^{j,\pi}_{s_0,h}(s,a)| \leq \varepsilon_{\mathsf{p}} H \ .$$

□

F.1 Bellman type equations for the variance

For a deterministic policy π and an agent i, we recall the following definitions of the Bellman-type equations for the variances as follows

$$\begin{aligned}\sigma\mathcal{Q}_h^{i,\pi}(s,a) &:= \mathrm{Var}_{\mathsf{P}_h^i}(\mathcal{V}_{h+1}^{i,\pi})(s,a) + \mathsf{P}_h^i \sigma\mathcal{V}_{h+1}^{i,\pi}(s,a) \\ \sigma\mathcal{V}_h^{i,\pi}(s) &:= \sigma\mathcal{Q}_h^{i,\pi}(s,\pi(s)) \\ \sigma\mathcal{V}_{H+1}^{i,\pi}(s) &:= 0\,, \end{aligned} \tag{59}$$

where $\mathrm{Var}_{\mathsf{P}_h^i}(f)(s,a) := \mathbb{E}_{s'\sim\mathsf{P}_h^i(\cdot|s,a)}\left[\left(f(s') - \mathsf{P}_h^i f(s,a)\right)^2\right]$ denotes the variance operator. Unrolling the precedent relation yields

$$\sigma\mathcal{V}_1^{i,\pi}(s) = \sum_{h=1}^{H}\sum_{s',a'} d_{s,h}^{i,\pi}(s',a')\,\mathrm{Var}_{\mathsf{P}_h^i}(\mathcal{V}_{h+1}^{i,\pi})(s',a')\,,$$

where $d_{s,h}^{i,\pi}(s',a')$ is the probability of visiting a pair (s',a') in the i-th environment while following the policy π and starting from a state s. Next, we state the well-known Bellman equation for variances (see, e.g., Sobel 1982; Azar et al. 2017).

Lemma F.3. *For any deterministic policy π, for all $h\in[H]$, and for all $i\in[M]$,*

$$\mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H}\mathsf{r}_{h'}^i(s_{h'}^i,a_{h'}^i) - \mathcal{Q}_h^{i,\pi}(s_h^i,a_h^i)\right)^2\middle|(s_h^i,a_h^i)=(s,a)\right] = \mathcal{Q}_h^{i,\pi}(s,a)\,. \tag{60}$$

In particular,

$$\mathbb{E}_\pi\left[\left(\sum_{h=1}^{H}\mathsf{r}_h^i(s_h^i,a_h^i) - \mathcal{V}_1^{i,\pi}(s_1^i)\right)^2\right] = \sigma\mathcal{V}_1^{i,\pi}(s_1^i) = \sum_{h=1}^{H}\sum_{s,a} d_h^{i,\pi}(s,a)\,\mathrm{Var}_{\mathsf{P}_h^i}(\mathcal{V}_{h+1}^{i,\pi})(s,a)\,.$$

F.2 Concentration inequalities

Lemma F.4 (Deviation inequality for categorical distribution, Jonsson et al. 2020)**.** *Let $(X_t)_{t\in\mathbb{N}^*}$ be i.i.d. samples from a probability measure P supported on $\{1,\dots,m\}$. We denote by $\widehat{\mathsf{P}}_n$ the empirical vector of probabilities, i.e., for all $k\in\{1,\dots,m\}$,*

$$\widehat{\mathsf{P}}_n(k) := \frac{1}{n}\sum_{\ell=1}^{n}\mathbf{1}_{\{k\}}(X_\ell).$$

For all P and for all $\delta\in(0,1)$,

$$\mathbb{P}\left(\exists n\in\mathbb{N}^\star, n\mathrm{KL}\big(\widehat{\mathsf{P}}_n\big\|\mathsf{P}\big) > \log(1/\delta) + (m-1)\log(e(1+n/(m-1)))\right) \le \delta\,.$$

Lemma F.5 (Corollary 11 by Talebi and Maillard 2018)**.** *Let P,Q two probability distributions on $\mathcal{S}$. For all functions $f:\mathcal{S}\mapsto[0,H]$,*

$$\mathsf{P}f - \mathsf{Q}f \le \sqrt{2\,\mathrm{Var}_{\mathsf{Q}}(f)\mathrm{KL}(\mathsf{P}\,\|\mathsf{Q})} + \frac{2}{3}H\mathrm{KL}(\mathsf{P},\mathsf{Q})\,.$$

where we have defined $\mathsf{P}f := \mathbb{E}_{s\sim\mathsf{P}}[f(s)]$.

Lemma F.6 (Lemma H.9 by Tiapkin et al. 2023)**.** *For any two probability measures P,Q on $\mathcal{S}$, for $f,g:\mathcal{S}\mapsto[0,b]$ two functions defined on $\mathcal{S}$, we have that*

$$\mathrm{Var}_{\mathsf{P}}(f) \le 2\,\mathrm{Var}_{\mathsf{P}}(g) + 2b\mathsf{P}|f-g| \quad \text{and} \tag{61}$$

$$\mathrm{Var}_{\mathsf{Q}}(f) \le \mathrm{Var}_{\mathsf{P}}(f) + 3b^2\|\mathsf{P}-\mathsf{Q}\|_1 \tag{62}$$

where we denote the absolute operator by $|f|(s) = |f(s)|$ for all $s\in\mathcal{S}$.

Lemma F.7 (Theorem 4 by Maurer and Pontil 2009)**.** . *Consider any $\delta > 0$ and any integer $n \geq 2$. Let $Y, Y_1, \ldots, Y_n$ be a collection of i.i.d. random variables falling within $[0,1]$. Define the empirical mean $\bar{Y} := \frac{1}{n}\sum_{i=1}^n Y_i$ and empirical variance $\widehat{Y}_n := \frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i\right| > \sqrt{\frac{2\widehat{Y}_n \log(2/\delta)}{n-1}} + \frac{7\log(2/\delta)}{3(n-1)}\right] \leq \delta$$

Below, we state the self-normalized Bernstein-type inequality by Domingues et al. (2021c). Let $(Y_t)_{t\in\mathbb{N}^\star}$, $(w_t)_{t\in\mathbb{N}^\star}$ be two sequences of random variables adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. We assume that the weights are in the unit interval $w_t \in [0,1]$ and predictable, i.e. $\mathcal{F}_{t-1}$ measurable. We also assume that the random variables Y_t are bounded $|Y_t| \leq b$ and centered $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$. Consider the following quantities

$$S_t := \sum_{s=1}^t w_s Y_s\,, \qquad V_t := \sum_{s=1}^t w_s^2 \cdot \mathbb{E}[Y_s^2|\mathcal{F}_{s-1}]\,,$$

and let $h(x) \triangleq (x+1)\log(x+1) - x$ be the Cramér transform of a Poisson distribution of parameter 1.

Theorem F.1 (Anytime Bernstein-type concentration inequality for martingales)**.** *For all $\delta > 0$,*

$$\mathbb{P}\left[\exists t \geq 1, (V_t/b^2 + 1)h\left(\frac{b|S_t|}{V_t + b^2}\right) \geq \log(1/\delta) + \log\left(4e(2t+1)\right)\right] \leq \delta.$$

The previous inequality can be weakened to obtain a more explicit bound: if $b \geq 1$ with probability at least $1-\delta$, for all $t \geq 1$,

$$|S_t| \leq \sqrt{2V_t \log\left(4\mathrm{e}(2t+1)/\delta\right)} + 3b\log\left(4\mathrm{e}(2t+1)/\delta\right)\,.$$

Next, we apply this Bernstein inequality to a particular distribution. Let $\mathcal{F}_t$ for $t \in \mathbb{N}$ be a filtration and $(X_t)_{t\in\mathbb{N}^\star}$ be a sequence of Bernoulli random variables with $\mathbb{P}(X_t = 1|\mathcal{F}_{t-1}) = P_t$ with P_t being $\mathcal{F}_{t-1}$-measurable and X_t being $\mathcal{F}_t$-measurable.

Corollary F.1. *For all $\delta > 0$,*

$$\mathbb{P}\left(\exists n : \left|\sum_{t=1}^n X_t - P_t\right| > \frac{1}{8}\sum_{t=1}^n P_t + 11\log\left(\frac{4\mathrm{e}(2n+1)}{\delta}\right)\right) \leq \delta.$$

Proof. Given a simplified version, we have with probability at least $1-\delta$ by applying inequality $2ab \leq a^2 + b^2$ for $a, b \geq 0$

$$\left|\sum_{t=1}^n X_t - P_t\right| \leq \sqrt{2 \cdot \frac{V_n}{8} \cdot 8\log\left(4\mathrm{e}(2n+1)/\delta\right)} + 3\log\left(4\mathrm{e}(2n+1)/\delta\right) \leq \frac{1}{8}\sum_{t=1}^n P_t + 11\log\left(4\mathrm{e}(2n+1)/\delta\right)\,.$$

□

F.3 Performance-difference Lemma

Lemma F.8 (Lemma 3 of Russo 2019)**.** *Let us consider two MDPs $\mathcal{M}_1 = (\mathcal{S}, \mathcal{A}, H, \mathsf{r}^{(1)}, \mathsf{P}^{(1)})$ and $\mathcal{M}_2 = (\mathcal{S}, \mathcal{A}, H, \mathsf{r}^{(2)}, \mathsf{P}^{(2)}$. Let $\mathcal{V}_1^{(1),\pi}(s)$ and $\mathcal{V}_1^{(2),\pi}(s)$ are values of a fixed policy π in MDP $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. Then it holds*

$$\mathcal{V}_1^{(1),\pi}(s) - \mathcal{V}_1^{(2),\pi}(s) = \mathbb{E}_{\pi,\mathcal{M}_1}\left[\sum_{h=1}^H \left(\mathsf{r}_h^{(1)} - \mathsf{r}_h^{(2)}\right)(s_h, a_h) + \left(\mathsf{P}_h^{(1)} - \mathsf{P}_h^{(2)}\right)\mathcal{V}_{h+1}^{(2),\pi}(s_h, a_h)\right],$$

where expectation is taken over the trajectories $(s_1, a_1, \ldots, s_H, a_H)$ generated by policy π in an MDP $\mathcal{M}_1$.

Corollary F.2. *Let us consider two MDPs* $\mathcal{M}_1 = (\mathcal{S}, \mathcal{A}, H, \mathsf{r}^{(1)}, \mathsf{P}^{(1)})$ *and* $\mathcal{M}_2 = (\mathcal{S}, \mathcal{A}, H, \mathsf{r}^{(2)}, \mathsf{P}^{(2)}$, *such that* $\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : |\mathsf{r}_h^{(1)}(s,a) - \mathsf{r}^{(2)}(s,a)| \le \varepsilon_{\mathsf{r}}$, $|\mathcal{V}_h^{(1),\pi}(s)| \le c$, $|\mathcal{V}_h^{(2),\pi}(s)| \le c$, *and* $\|\mathsf{P}_h^{(1)}(s,a) - \mathsf{P}_h^{(2)}(s,a)\|_1 \le \varepsilon_{\mathsf{P}}$ *where* $c > 0$ *is a positive constant and* $\mathcal{V}_1^{(1),\pi}(s)$ *and* $\mathcal{V}_1^{(2),\pi}(s)$ *are values of a fixed policy* π *in MDP* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *respectively. Then it holds*

$$\mathcal{V}_1^{(1),\pi}(s_1) - \mathcal{V}_1^{(2),\pi}(s_1) \le \varepsilon_{\mathsf{P}}\, cH + \varepsilon_{\mathsf{r}} H\,.$$

Proof. Follows directly from combination of Lemma F.8, Holder's inequality and a fact that $\|\mathcal{V}_h^{(2),\pi}\|_1 \le c$. □

Inspired by a construction of Ross and Bagnell (2010), we can show that dependence H^2 in terms of ℓ_1-distance between two models in non-improvable.

Lemma F.9. *There exist two MDPs* $\mathcal{M}_1 = (\mathcal{S}, \mathcal{A}, H, \mathsf{r}, \mathsf{P}^1)$ *and* $\mathcal{M}_2 = (\mathcal{S}, \mathcal{A}, H, \mathsf{r}, \mathsf{P}^2)$ *with the same reward function and different kernels,* $H \ge 2$ *such that* $\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : \|\mathsf{P}_h^1(s,a) - \mathsf{P}_h^2(s,a)\|_1 \le \varepsilon_{\mathsf{P}}$ *for* $0 < \varepsilon_{\mathsf{p}} < 2/H$. *Then there is a policy* π *such that values* $\mathcal{V}_1^{1,\pi}(s)$ *and* $\mathcal{V}_1^{2,\pi}(s)$ *in MDPs* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *satisfy*

$$\mathcal{V}_1^{1,\pi}(s_1) - \mathcal{V}_1^{2,\pi}(s_1) = \Omega(\varepsilon_{\mathsf{p}} H^2)\,.$$

Proof. Consider the problem with 2 states $\{s_1, s_2\}$ and 1 action $\{a\}$, the agent always starts at s_1. The reward function satisfies $\mathsf{r}_h(s_1, a) = 1, \mathsf{r}_h(s_2, a) = 0$ for all $h \in H$. Finally, the transition kernels are the same for all h and are defined as

$$\mathsf{P}_h^i(s_1|s_1,a) = 1 - p_i, \quad \mathsf{P}_h^i(s_2|s_1,a) = p_i, \quad \mathsf{P}_h^i(s_1|s_2,a) = 0, \quad \mathsf{P}_h^i(s_2|s_2,a) = 1\,,$$

for $i \in \{1,2\}$. In other words, the state s_2 is a sink with zero reward. Since there is only one action, the value is the same for any policy π. Let us take $p_1 = 0$ and $p_2 = \varepsilon_{\mathsf{p}}$, then under the kernel P^1 the value $\mathcal{V}_1^{1,\pi}(s_1)$ is equal to H, whereas under the kernel P^2, the value function $\mathcal{V}_1^{1,\pi}(s_1)$ it is equal to

$$\mathcal{V}_1^{2,\pi}(s_1) = 1 + (1-\varepsilon_{\mathsf{p}}) + (1-\varepsilon_{\mathsf{p}})^2 + \ldots + (1-\varepsilon_{\mathsf{p}})^{H-1} = \frac{1-(1-\varepsilon_{\mathsf{p}})^H}{\varepsilon_{\mathsf{p}}}.$$

Then we have

$$\mathcal{V}_1^{1,\pi}(s_1) - \mathcal{V}_1^{2,\pi}(s_1) = \frac{H\varepsilon_{\mathsf{p}} - 1 + (1-\varepsilon_{\mathsf{p}})^H}{\varepsilon_{\mathsf{p}}}\,.$$

Now as $0 < \varepsilon_{\mathsf{p}} < 2/H$, Bernoulli's inequality yields

$$\mathcal{V}_1^{1,\pi}(s_1) - \mathcal{V}_1^{2,\pi}(s_1) = \frac{H\varepsilon_{\mathsf{p}} - 1 + (1-\varepsilon_{\mathsf{p}})^{H/2}(1-\varepsilon_{\mathsf{p}})^{H/2}}{\varepsilon_{\mathsf{p}}} \ge \frac{H\varepsilon_{\mathsf{p}} - 1 + (1 - H\varepsilon_{\mathsf{p}}/2)(1 - H\varepsilon_{\mathsf{p}}/2)}{\varepsilon_{\mathsf{p}}} = \frac{\varepsilon_{\mathsf{p}} H^2}{4}\,,$$

where the first inequality comes from $(1-x)^r \ge 1 - rx$ for $0 \le x \le 1$ and $r > 1$. □