

A RELATED WORKS

Single-agent episodic MDPs. Significant contributions have been made in both model-based and model-free frameworks. In the model-based category, a series of algorithms have been proposed by Auer et al. (2008), Agrawal & Jia (2017), Azar et al. (2017), Kakade et al. (2018), Agarwal et al. (2020), Dann et al. (2019), Zanette & Brunskill (2019), and Zhang et al. (2021), with more recent contributions from Zhou et al. (2023) and Zhang et al. (2023). Notably, Zhang et al. (2023) proved that a modified version of MVP (proposed by Zhang et al. (2021)) achieves a regret of $\tilde{O}\left(\min\{\sqrt{SAH^2T}, T\}\right)$ which matches the minimax lower bound. Within the model-free framework, Jin et al. (2018) proposed a Q-learning with UCB exploration algorithm, achieving regret of $\tilde{O}\left(\sqrt{SAH^3T}\right)$, which has been advanced further by Yang et al. (2021), Zhang et al. (2020), Li et al. (2021) and Ménard et al. (2021). The latter three have introduced algorithms that achieve minimax regret of $\tilde{O}\left(\sqrt{SAH^2T}\right)$.

Federated and distributed RL. Existing literature on federated and distributed RL algorithms sheds light on different aspects. Guo & Brunskill (2015) showed that applying concurrent RL to identical MDPs can linearly speed up sample complexity. Agarwal et al. (2021) proposed a parallel RL algorithm with low communication cost. Jin et al. (2022), Khodadadian et al. (2022), Fan et al. (2023) and Woo et al. (2023) investigated federated Q-learning algorithms in different settings. Fan et al. (2021), Wu et al. (2021) and Chen et al. (2023) focused on robustness. Particularly, Chen et al. (2023) proposed algorithms in both offline and online settings, obtaining near-optimal sample complexities and achieving a superior robustness guarantee. Doan et al. (2019), Doan et al. (2021), Chen et al. (2021b), Sun et al. (2020), Wai (2020), Wang et al. (2020a), Zeng et al. (2021) and Liu & Olshevsky (2023) analyzed the convergence of decentralized temporal difference algorithms. Fan et al. (2021) and Chen et al. (2021a) studied communication-efficient policy gradient algorithms. Shen et al. (2023b), Shen et al. (2023a) and Chen et al. (2022) have analyzed the convergence of distributed actor-critic algorithms. Assran et al. (2019), Espeholt et al. (2018) and Mnih et al. (2016) explored federated actor-learner architectures.

RL with low switching cost and batched RL. Research in RL with low-switching cost aims to minimize the number of policy switching while maintaining comparable regret bounds to its fully adaptive counterparts and can be applied to federated RL. In batched RL (e.g., Perchet et al. (2016), Gao et al. (2019)), the agent sets the number of batches and length of each batch upfront, aiming for fewer batches and lower regret. Bai et al. (2019) first introduced the problem of RL with low-switching cost and proposed a Q -learning algorithm with lazy update, achieving $\tilde{O}(SAH^3 \log T)$ switching costs. This work was advanced by Zhang et al. (2020), which improved the regret upper bound. Besides, Wang et al. (2021) studied the problem of RL under the adaptivity constraint. Recently, Qiao et al. (2022) proposed a model-based algorithm with $\tilde{O}(\log \log T)$ switching costs. Zhang et al. (2022) proposed a batched RL algorithm that is well-suited for the federated setting.

Federated/distributed bandits. Federated bandits with low communication costs have been studied extensively recently in the literature Wang et al. (2020b); Li & Wang (2022); Shi & Shen (2021); Shi et al. (2021); Huang et al. (2021); Wang et al. (2022); He et al. (2022); Li et al. (2022b). Shi & Shen (2021) and Shi et al. (2021) investigated efficient client-server communication and coordination protocols for federated MAB without and with personalization, respectively. Wang et al. (2020b) investigated communication-efficient distributed linear bandits, while Huang et al. (2021) studied federated linear contextual bandits. Li & Wang (2022) focused on the asynchronous communication protocol.

When data privacy is explicitly considered, Li et al. (2020); Zhu et al. (2021) studied federated bandits with item-level differential privacy (DP) guarantee. Dubey & Pentland (2022) considered private and byzantine-proof cooperative decision making in multi-armed bandits. Dubey & Pentland (2020); Zhou & Chowdhury (2023) considered the linear contextual bandit model with joint DP guarantee. Huang et al. (2023) recently investigated linear contextual bandits under user-level DP constraints. Private distributed bandits with partial feedback was also studied in Li et al. (2022a).

B AUXILIARY LEMMAS

In this section, we introduce some useful lemmas which will be used in the proofs. Before starting, we describe the global indexing mechanism mentioned in Section 4. Global visiting indices $i = 1, 2, \dots$ are assigned, based on the chronological order, to the visits of any given $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. With this, we can establish a map between the global visiting index i , and k, m, j , where k is the round index, m is the agent index and j is the episode index for a given round and a given agent. For (x, a, h) , we define functions that recover k, m, j from i as $k_h(i; x, a), m_h(i; x, a), j_h(i; x, a)$. When there is no ambiguity, we will use the simplified notations k^i, m^i, j^i . The visiting indices are utilized to construct a sequence, ensuring that quantities with smaller indices are observed prior to those with larger indices. Under the synchronization and zero-latency assumption, we have the following formulas for m^i, k^i, j^i .

$$\begin{aligned} k_h(i; x, a) &= \sup \{k \in \mathbb{N}_+ : N_h^k(x, a) < i\}, \\ j_h(i; x, a) &= \sup \left\{ j \in \mathbb{N}_+ : \sum_{j'=1}^{j-1} \sum_{m=1}^M \mathbb{I} \left[(x, a) = (x_h^{m, k^i, j'}, a_h^{m, k^i, j'}) \right] < i - N_h^{k^i}(x, a) \right\}, \\ m_h(i; x, a) &= \sup \left\{ m \in \mathbb{N}_+ : \sum_{m'=1}^{m-1} \mathbb{I} \left[(x, a) = (x_h^{m', k^i, j^i}, a_h^{m', k^i, j^i}) \right] \right. \\ &\quad \left. < i - N_h^{k^i}(x, a) - \sum_{j'=1}^{j^i-1} \sum_{m=1}^M \mathbb{I} \left[(x, a) = (x_h^{m, k^i, j'}, a_h^{m, k^i, j'}) \right] \right\}. \end{aligned}$$

We also introduce a new notation $\hat{T} = MT$ that represents the total number of samples generated by all the agents.

Next, we begin to introduce the lemmas. First, Lemma B.1 establishes some relationships between some quantities used in Algorithms 1 and 2.

Lemma B.1. *Denote $\tilde{C} = 1/(H(H+1))$. The following relationships hold for both algorithms.*

- (a) $K \leq K_0$.
- (b) $N_h^K(x, a) \leq T_0/H$.
- (c) For any $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we have

$$n_h^{m, k}(x, a) \leq \max \left\{ 1, \frac{\tilde{C} N_h^k(x, a)}{M} \right\}, \forall m \in [M]. \quad (11)$$

and

$$n_h^k(x, a) \leq \max\{M, \tilde{C} N_h^k(x, a)\}. \quad (12)$$

If $N_h^k(x, a) \geq i_0$,

$$n_h^k(x, a) \leq \tilde{C} N_h^k(x, a).$$

- (d) For any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $N_h^{K+1}(x, a) \leq (1 + \tilde{C})T_0/H + M$.
- (e) $\hat{T} \leq (1 + \tilde{C})T_0 + HM$.

Proof of Lemma B.1. (a)-(c) are obvious given Algorithms 1 and 4. (d) and (e) can be directly obtained from (b) and (c). \square

Next, Lemma B.2 provides some properties about θ_t^i 's.

Lemma B.2. *(Lemma 4.1 in Jin et al. (2018) and beyond) The following properties hold for all $t \in \mathbb{N}_+$ for both algorithms.*

- (a) $1/\sqrt{t} \leq \sum_{i=1}^t \theta_t^i / \sqrt{i} \leq 2/\sqrt{t}$, which implies that $\beta_t \in [2c\sqrt{H^3\iota/t}, 4c\sqrt{H^3\iota/t}]$, $\forall t \in \mathbb{N}_+$.
- (b) $\max_{i \in [t]} \theta_t^i \leq 2H/t$.
- (c) $\sum_{i=1}^t (\theta_t^i)^2 \leq 2H/t$.
- (d) $\sum_{t=i}^\infty \theta_t^i = 1 + 1/H$.
- (e) For any $t \in \mathbb{N}_+$ and $i \in [t] - \{t\}$, $\theta_t^{i+1}/\theta_t^i = 1 + H/i > 1$.
- (f) For both algorithms, for any $t \in \mathbb{N}_+$ and $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, if $i_1, i_2 \in [t]$, $k_h(i_1, x, a) = k_h(i_2, x, a)$ and $N_h^{k_h(i_1, x, a)}(x, a) \geq i_0$ we have that $\theta_t^{i_1}/\theta_t^{i_2} \leq \exp(1/H)$.

Proof of Lemma B.2. (a)-(e) are obvious based on θ_t^i 's definition and Lemma 4.1 in Jin et al. (2018). For (f), denoting $t_0 = N_h^{k_h(i_1, x, a)}(x, a) + 1$ and $t_1 = N_h^{k_h(i_1, x, a)+1}(x, a)$, based on (e), we have

$$\theta_t^{i_1}/\theta_t^{i_2} \leq \theta_t^{t_1}/\theta_t^{t_0} = \prod_{t'=t_0}^{t_1-1} (1 + H/t').$$

Based on (c) in Lemma B.1, we further have that

$$\prod_{t'=t_0}^{t_1-1} (1 + H/t') \leq (1 + H/t_0)^{t_1-t_0} \leq \exp(H(t_1 - t_0)/t_0) \leq \exp(1/H).$$

□

Next, we rigorously define the weights $\tilde{\theta}_t^i$ mentioned in Section 4. For any $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we let $t = N_h^{K'}(x, a)$ and $i \in [t] \cup \{0\}$. Letting $t' = N_h^{k^i}(x, a)$ and $t'' = N_h^{k^i+1}(x, a)$, we denote

$$\tilde{\theta}_t^i(x, a, h) = \theta_t^i \mathbb{I}[t' < i_0] + \frac{1 - \alpha^c(t' + 1, t'')}{t'' - t'} \alpha^c(t'' + 1, t) \mathbb{I}[t' \geq i_0],$$

and we will use the simplified notation $\tilde{\theta}_t^i$ when there is no ambiguity. Lemma B.3 provides properties of $\tilde{\theta}_t^i$ and its relationship with θ_t^i .

Lemma B.3. *The following relationships hold for any $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ with $t = N_h^{K'}(x, a)$ for both algorithms.*

- (a) $\tilde{\theta}_t^i(x, a, h) = \tilde{\theta}_{t'}^i(x, a, h) \alpha^c(t' + 1, t)$ with $t' = N_h^{k_h(i, x, a)+1}(x, a)$.
- (b) For any $i_1, i_2 \in [t]$, if $k_h(i_1, x, a) = k_h(i_2, x, a)$ and $N_h^{k_h(i_1, x, a)}(x, a) \geq i_0$, we have that $\tilde{\theta}_t^{i_1}(x, a, h) = \tilde{\theta}_t^{i_2}(x, a, h)$.
- (c) For any $k' \leq K'$, we have that

$$\sum_{i'=N_h^{k'}(x, a)+1}^{N_h^{k'+1}(x, a)} \tilde{\theta}_t^{i'}(x, a, h) = \sum_{i'=N_h^{k'}(x, a)+1}^{N_h^{k'+1}(x, a)} \theta_t^{i'},$$

which indicates that

$$\sum_{i=1}^t \tilde{\theta}_t^i = \mathbb{I}[t > 0].$$

- (d) For any $i \in [t]$, when $N_h^{k_i(x, a, h)}(x, a) \geq i_0$, we have that

$$(1 + H/(1 + N_i))^{1-n_i} \leq \tilde{\theta}_t^i/\theta_t^i \leq (1 + H/(1 + N_i))^{n_i-1},$$

in which $N_i = N_h^{k_h(i, x, a)}(x, a)$ and $n_i = n_h^{k_h(i, x, a)}(x, a)$.

(e) For any $i \in [t]$, when $N_h^{k_i(x,a,h)}(x,a) \geq i_0$, we have that

$$\tilde{\theta}_t^i / \theta_t^i \leq \exp(1/H).$$

Proof of Lemma B.3. (a)-(c) can be obtained directly through the definition of $\tilde{\theta}_t^i$. Next, we prove (d) and (e). Denote $t_0 = N_h^{k_h(i,x,a)}(x,a) + 1$ and $t_1 = N_h^{k_h(i,x,a)+1}(x,a)$. By (c) and (e) in Lemma B.2, we have that $\theta_{t_1}^{t_0} / \theta_{t_1}^{t_1} \leq \tilde{\theta}_t^i / \theta_t^i \leq \theta_{t_1}^{t_1} / \theta_{t_1}^{t_0}$. Then, (d) can be proved by noticing that $\theta_{t_1}^{t_1} / \theta_{t_1}^{t_0} \leq (1 + H/(1 + N_i))^{n_i-1}$. This implies that (e) holds because of (c) in Lemma B.1. \square

C PROOF OF THEOREM 4.1

C.1 ROBUSTNESS AGAINST ASYNCHRONIZATION

In this subsection, we discuss a more general situation for Algorithms 1 and 2, where agent m generates $n^{m,k}$ episodes during round k . We no longer assume that $n^{m,k}$ has the same value n^k for different clients. The difference can be caused by latency (the time gap between an agent sending an abortion signal and other agents receiving the signal) and asynchronization (the heterogeneity among clients on the computation speed and process of collecting trajectories). In this case, for K rounds, the total number of samples generated by all the clients is

$$\hat{T} = H \sum_{k=1}^K \sum_{m=1}^M n^{m,k}.$$

Thus, we generalize the notation $T = \hat{T}/M$, which characterizes the mean number of samples generated by an agent. Accordingly, the definition of $\text{Regret}(T)$ can be generalized as

$$\text{Regret}(T) = \sum_{k=1}^K \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} V_1^*(x_1^{m,k,j}) - V_1^{\pi^k}(x_1^{m,k,j}),$$

Similarly, the definitions of $n_h^{m,k}(x,a)$, $N_h^{m,k}(x,a)$, $v_h^{m,k}(x,a)$ are also generalized by replacing $\sum_{j=1}^{n^k}$ with $\sum_{j=1}^{n^{m,k}}$.

We note that Algorithms 1 and 2 naturally accommodate such asynchronicity. Therefore, in the following analysis of the regret, we adopt the general notation $n^{m,k}$. However, for the proof of Theorem 4.2 pertaining to communication, we will maintain the synchronization assumption that $n^{m,k} = n^k, \forall m \in [M]$.

C.2 BOUNDS ON $Q_h^k - Q_h^*$

Lemma C.1. *For Algorithms 1 and 2, there exists a positive constant $c > 0$ such that, for any $p \in (0, 1)$, the following relationship holds for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ with probability at least $1 - p$:*

$$0 \leq Q_h^{K'}(x, a) - Q_h^*(x, a) \leq \theta_t^0 H + \sum_{i=1}^t \tilde{\theta}_t^i (V_{h+1}^{k^i} - V_{h+1}^*)(x_{h+1}^{m^i, k^i, j^i}) + \beta_t, \quad (13)$$

in which $t = N_h^{K'}(x, a)$.

We first provide Lemma C.2 to formally state Equation (14) and Equation (15), which establish the relationship between Q_h^k and Q_h^* . The proof is the same as the proof of Equation (4.3) in Jin et al. (2018).

Lemma C.2. For the Hoeffding-type Algorithms 1 and 2, for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, denoting $t = N_h^{K'}(x, a)$, we have

$$Q_h^{K'}(x, a) = \theta_t^0 H + \sum_{i=1}^t \tilde{\theta}_t^i \left(r_h(x, a) + V_{h+1}^{k^i}(x_{h+1}^{m^i, k^i, j^i}) \right) + \sum_{i=1}^t \theta_t^i b_i, \quad (14)$$

$$Q_h^*(x, a) = \tilde{\theta}_t^0 Q_h^* + \sum_{i=1}^t \tilde{\theta}_t^i \left(r_h(x, a) + \left([\mathbb{P}_h V_{h+1}^*](x, a) - \tilde{\mathbb{E}}_{x, a, h, i} V_{h+1}^* \right) + \tilde{\mathbb{E}}_{x, a, h, i} V_{h+1}^* \right).$$

Furthermore, we have

$$\begin{aligned} (Q_h^{K'} - Q_h^*)(x, a) &= \tilde{\theta}_t^0 (H - Q_h^*(x, a)) + \sum_{i=1}^t \tilde{\theta}_t^i (\tilde{\mathbb{E}}_{x, a, h, i} - \mathbb{E}_{x, a, h}) V_{h+1}^* \\ &\quad + \sum_{i=1}^t \tilde{\theta}_t^i (V_{h+1}^{k^i} - V_{h+1}^*)(x_{h+1}^{m^i, k^i, j^i}) + \sum_{i=1}^t \theta_t^i b_i, \end{aligned} \quad (15)$$

$$(Q_h^{K'} - Q_h^*)(x, a) = \tilde{\theta}_t^0 (H - Q_h^*(x, a)) + \sum_{i=1}^t \tilde{\theta}_t^i (\tilde{\mathbb{E}}_{x, a, h, i} - \mathbb{E}_{x, a, h}) V_{h+1}^*$$

in which

$$\begin{aligned} \mathbb{E}_{x, a, h} V_{h+1}^* &= \mathbb{E}_{x, a, h} V_{h+1}^*(x_{h+1}) = \mathbb{E} [V_{h+1}^*(x_{h+1}) | (x_h, a_h) = (x, a)], \\ \tilde{\mathbb{E}}_{x, a, h, i} V_{h+1}^* &= \tilde{\mathbb{E}}_{x, a, h, i} V_{h+1}^*(x_{h+1}) = V_{h+1}^*(x_{h+1}^{m^i, k^i, j^i}). \end{aligned}$$

With this lemma, we derive a probabilistic upper bound for $|\sum_{i=1}^t \tilde{\theta}_t^i X_i|$ with $X_i = (\tilde{\mathbb{E}}_{x, a, h, i} - \mathbb{E}_{x, a, h}) V_{h+1}^*$ in Lemma C.3.

Lemma C.3. There exists $c_0 > 0$ such that, for any $p \in (0, 1)$, with probability at least $1 - p$, the following relationship holds for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ with $t = N_h^{K'}(x, a)$:

$$\left| \sum_{i=1}^t \tilde{\theta}_t^i (\tilde{\mathbb{E}}_{x, a, h, i} - \mathbb{E}_{x, a, h}) V_{h+1}^*(x_{h+1}) \right| \leq c_0 \sqrt{H^3 t / t}. \quad (16)$$

Proof. For a given $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, denote $X_i(x, a, h) = (\tilde{\mathbb{E}}_{x, a, h, i} - \mathbb{E}_{x, a, h}) V_{h+1}^*(x_{h+1})$. When there is no ambiguity, we use the simplified notation $X_i = X_i(x, a, h)$. We know that $\{X_i\}_{i=1}^\infty$ is a sequence of martingale differences with $|X_i| \leq H$. We decompose the summation as follows:

$$\sum_{i=1}^t \tilde{\theta}_t^i X_i = \sum_{i=1}^t \theta_t^i X_i + \sum_{i=1}^t (\tilde{\theta}_t^i - \theta_t^i) X_i.$$

Note that $t \leq T_0/H$.

First, we focus on the first term. By Azuma-Hoeffding Inequality, for any given $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and a given $t' \in \mathbb{N}_+$, for any $p \in (0, 1)$, with probability $1 - p$, there exists a numerical constant $c_1 > 0$ such that

$$\left| \sum_{i=1}^{t'} \theta_{t'}^i X_i \right| \leq c_1 H \sqrt{\left(\sum_{i=1}^{t'} (\theta_{t'}^i)^2 \right) \log \frac{2}{p}},$$

which indicates that $\left| \sum_{i=1}^{t'} \theta_{t'}^i X_i \right| \leq \frac{c_1}{\sqrt{2}} \sqrt{(H^3/t') \log \frac{2}{p}}$ based on (c) in Lemma B.2.

By considering all the possible combinations $(x, a, h, t') \in \mathcal{S} \times \mathcal{A} \times [H] \times [T_0/H]$, with a union bound and the realization of $t = t'$, we have, for any $p \in (0, 1)$, with at probability at least $1 - p$, the following relationship holds simultaneously for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$:

$$\left| \sum_{i=1}^t \theta_t^i X_i \right| \leq \frac{c_1}{\sqrt{2}} \sqrt{(H^3/t) \log \frac{2SAT_0}{p}} \leq \frac{c_1}{\sqrt{2}} \sqrt{\iota_0 H^3/t}.$$

We then focus on the second term $\sum_{i=1}^t (\tilde{\theta}_t^i - \theta_t^i) X_i$. For any given $(x, a, h, k_s) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we consider the part with samples generated by the k_s -th round, which is

$$\sum_{i=t_2}^{t_3} (\tilde{\theta}_{t_3}^i - \theta_{t_3}^i) X_i,$$

in which $t_2 = N_h^{k_s}(x, a) + 1$, $t_3 = N_h^{k_s+1}(x, a)$. We can control the second term by controlling $|\sum_{i=t_2}^{t_3} (\tilde{\theta}_{t_3}^i - \theta_{t_3}^i) X_i|$ for all $k_s \in [K]$.

We have

$$\sum_{i=1}^t (\tilde{\theta}_t^i - \theta_t^i) X_i = \sum_{k_s=1}^{K'-1} \left[\prod_{t'=t_3+1}^t (1 - \alpha_{t'}) \right] \sum_{i=t_2}^{t_3} (\tilde{\theta}_{t_3}^i - \theta_{t_3}^i) X_i. \quad (17)$$

To begin with, we prove that there exists a numerical constant $c_3 > 0$ such that

$$\sqrt{\sum_{i=t_2}^{t_3} (\tilde{\theta}_{t_3}^i - \theta_{t_3}^i)^2} \leq c_3 \sum_{t'=t_2}^{t_3} \theta_{t_3}^{t'} / \sqrt{t'}. \quad (18)$$

This relationship obviously holds when $N_h^{k_s}(x, a) < i_0$ as $LHS = 0$. When $N_h^{k_s}(x, a) \geq i_0$, we have

$$\sqrt{\sum_{i=t_2}^{t_3} (\tilde{\theta}_{t_3}^i - \theta_{t_3}^i)^2} \leq O \left(\sqrt{\sum_{i=t_2}^{t_3} \frac{H^2(t_3 - t_2)^2}{t_2^2} (\theta_{t_3}^i)^2} \right) \leq O \left(\frac{H(t_3 - t_2)^{3/2}}{t_2} \theta_{t_3}^{t_2} \right),$$

where the first inequality comes from (d) in Lemma B.3 and the second one comes from (f) in Lemma B.2.

We also have that

$$O \left(\frac{H(t_3 - t_2)^{3/2}}{t_2} \theta_{t_3}^{t_2} \right) = O \left(\frac{H(t_3 - t_2)^{1/2}}{\sqrt{t_2}} (t_3 - t_2) \theta_{t_3}^{t_2} / \sqrt{t_2} \right) = O \left(\sum_{t'=t_2}^{t_3} \theta_{t_3}^{t'} / \sqrt{t'} \right),$$

where the second relationship comes from (f) in Lemma B.2. This completes the proof of Equation (18).

Next, we proceed with discussions conditioning on all the information before starting the k_s -th round, which means that $N_h^{k_s}(x, a)$ and t_2 can be treated as constants. If $N_h^{k_s}(x, a) < i_0$, this quantity is equal to 0. Otherwise, given any $t'_3 \geq t_2$ and $i \in [t_2, t'_3]$, we denote

$$\hat{\theta}_{t'_3} = \left[1 - \prod_{t'=t_2}^{t'_3} (1 - \alpha_{t'}) \right] / (t'_3 - t_2 + 1),$$

Therefore, in the expression $\sum_{i=t_2}^{t'_3} (\hat{\theta}_{t'_3} - \theta_{t'_3}^i) X_i$, we can treat $\{X_{t_2}, X_{t_2+1} \dots X_{t'_3}\}$ as martingale differences and $\theta_{t'_3}^i$ s and $\hat{\theta}_{t'_3}$ as constants. Hence, by Azuma-Hoeffding Inequality, there exists a positive numerical constant c_2 such that, for any $p \in (0, 1)$, with probability at least $1 - p$,

$$\left| \sum_{i=t_2}^{t'_3} (\hat{\theta}_{t'_3} - \theta_{t'_3}^i) X_i \right| \leq c_2 H \sqrt{\log \frac{2}{p} \sum_{i=t_2}^{t'_3} (\hat{\theta}_{t'_3} - \theta_{t'_3}^i)^2}.$$

By considering all possible values of t'_3 with a union bound, we have that with probability at least $1 - p$, the following relationship holds simultaneously for any $t_2 \leq t_3 \leq t_2 + (1 + \tilde{C})T_0/H + M - 1$.

$$\left| \sum_{i=t_2}^{t'_3} (\hat{\theta}_{t'_3} - \theta_{t'_3}^i) X_i \right| \leq c_2 H \sqrt{\log \frac{2(T_0/H + M)(1 + \tilde{C})}{p} \sum_{i=t_2}^{t'_3} (\hat{\theta}_{t'_3} - \theta_{t'_3}^i)^2}.$$

So, noticing that $\hat{\theta}_{t'_3} = \tilde{\theta}_{t'_3}^i$ when $t'_3 = t_3$ and $i \in [t_2, t_3]$ and applying Equation (18), we have that, for any $k_s \in \mathbb{N}_+$ and any $p \in (0, 1)$, with probability at least $1 - p$,

$$\left| \sum_{i=t_2}^{t_3} (\tilde{\theta}_{t_3}^i - \theta_{t_3}^i) X_i \right| \leq c_2 c_3 H \sqrt{\log \frac{2(T_0/H + M)(1 + \tilde{C})}{p}} \sum_{i=t_2}^{t_3} \theta_{t_3}^i / \sqrt{i}.$$

We apply the union bound and claim that for any $p \in (0, 1)$, the following relationship holds with probability at least $1 - p$ for all $(x, a, h, k_s) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K_0]$.

$$\begin{aligned} \left| \sum_{i=t_2}^{t_3} (\tilde{\theta}_{t_3}^i - \theta_{t_3}^i) X_i \right| &\leq c_2 c_3 H \sqrt{\log \frac{2SAHK_0(T_0/H + M)(1 + \tilde{C})}{p}} \sum_{i=t_2}^{t_3} \theta_{t_3}^i / \sqrt{i} \\ &= c_2 c_3 H \sqrt{\iota_1} \sum_{i=t_2}^{t_3} \theta_{t_3}^i / \sqrt{i}. \end{aligned}$$

Under this event, with Equation (17), we have that

$$\left| \sum_{i=1}^t (\tilde{\theta}_t^i - \theta_t^i) X_i \right| \leq c_2 c_3 H \sum_{i=1}^t \sqrt{\iota_1} \theta_t^i / \sqrt{i}. \quad (19)$$

By (a) in Lemma B.2, we have $\sum_{i=1}^t \theta_t^i / \sqrt{i} \leq \sqrt{4/t}$. Combining the results for the two terms completes the proof. \square

Finally, we provide the proof for Lemma C.1.

Proof of Lemma C.1. We pick $c = c_0$ such that the event in Lemma C.3 holds. Under the event given in Lemma C.3 and noting Equation (15), we claim the conclusion by using the same proof as that for Lemma 4.3 in Jin et al. (2018). \square

C.3 PROOF OF THEOREM 4.1

Having proved Lemma C.1, we turn our attention to demonstrating the remaining parts of the proof. We use $n^{m,k}$ to denote the number of episodes by agent m in round k .

We first provide some additional notations. Define

$$\begin{aligned} \delta_h^k &= \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} \left(V_h^k - V_h^{\pi^k} \right) (x_h^{m,k,j}), \\ \phi_h^k &= \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} (V_h^k - V_h^*) (x_h^{m,k,j}), \forall h \in [H + 1], \end{aligned}$$

in which $\delta_{H+1}^k = \phi_{H+1}^k = 0$. We also define

$$\xi_{h+1}^k = \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} (\mathbb{P} - \hat{\mathbb{P}}) \left(V_{h+1}^* - V_{h+1}^{\pi^k} \right) (x_h^{m,k,j}, a_h^{m,k,j}), h \in [H]$$

with $\xi_{H+1}^k = 0$. Here,

$$(\mathbb{P}) \left(V_{h+1}^* - V_{h+1}^{\pi^k} \right) (x_h^{m,k,j}, a_h^{m,k,j}) = \mathbb{E} \left[\left(V_{h+1}^* - V_{h+1}^{\pi^k} \right) (x_{h+1}^{m,k,j}) | (\pi^k, x_h^{m,k,j}, a_h^{m,k,j}) \right],$$

and

$$(\hat{\mathbb{P}}) \left(V_{h+1}^* - V_{h+1}^{\pi^k} \right) (x_h^{m,k,j}, a_h^{m,k,j}) = \left(V_{h+1}^* - V_{h+1}^{\pi^k} \right) (x_{h+1}^{m,k,j}).$$

We first provide a Lemma related to ξ_{h+1}^k .

Lemma C.4. *There exists a numerical constant $c_5 > 0$ such that, for any $p \in (0, 1)$, with probability at least $1 - p$,*

$$\left| \sum_{k=1}^K C_h \sum_{h=1}^H \xi_{h+1}^k \right| \leq c_5 H \sqrt{\hat{T} \iota}, \quad (20)$$

where $C_h = \exp(3(h-1)/H)$.

Proof. Denote $V(m, k, j, h) = C_h(\mathbb{P} - \hat{\mathbb{P}}) \left(V_{h+1}^* - V_{h+1}^{\pi^k} \right) (x_h^{m,k,j}, a_h^{m,k,j})$ and use $\sum_{m,k,j,h}$ as a simplified notation for $\sum_{k=1}^K \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} \sum_{h=1}^{H-1}$. The quantity of interest can be rewritten as $\sum_{m,k,j,h} V(m, k, j, h)$, with $|V(m, k, j, h)| \leq O(H)$ as $C_h \leq \exp(3)$.

Let $\tilde{V}(\tilde{i})$ be the \tilde{i} -th term in the summation that contains $\hat{T}(H-1)/H$ terms, in which the order follows a ‘‘round first, episode second, step third, agent fourth’’ rule. Then the sequence $\{\tilde{V}(\tilde{i})\}$ is a martingale difference. By Azuma-Hoeffding Inequality, for any $p \in (0, 1)$ and $t \in \mathbb{N}_+$, with probability at least $1 - p$,

$$\left| \sum_{\tilde{i}=1}^t \tilde{V}(\tilde{i}) \right| \leq O \left(H \sqrt{t \log \frac{2}{p}} \right).$$

Then by applying a union bound over $t \in [(1 + \tilde{C})T_0 + HM]$ and knowing that $\hat{T}(H-1)/H \leq T_0(1 + \tilde{C}) + HM$ due to (e) in Lemma B.1, we have that, for any $p \in (0, 1)$, with probability at least $1 - p$,

$$\left| \sum_{k=1}^K C_h \sum_{h=1}^H \xi_{h+1}^k \right| = \left| \sum_{\tilde{i}=1}^{\hat{T}(H-1)/H} \tilde{V}(\tilde{i}) \right| \leq O(H \sqrt{\hat{T} \iota}).$$

This completes the proof. \square

Noticing that $\text{Regret}(T) \leq \sum_{k=1}^K \delta_1^k$ due to $\text{Regret}(T) = \sum_{k=1}^K \delta_1^k - \sum_{k=1}^K \phi_1^k$ and $\phi_1^k \geq 0$ shown in Equation (13), we attempt to establish a probability upper bound for $\sum_{k=1}^K \delta_1^k$. First, we have

$$\delta_h^k \leq \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} (Q_h^k - Q_h^*) (x_h^{m,k,j}, a_h^{m,k,j}) + \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} (Q_h^* - Q_h^{\pi^k}) (x_h^{m,k,j}, a_h^{m,k,j}), \quad (21)$$

which holds because $V_h^{\pi^k}(x_h^{m,k,j}) = Q_h^{\pi^k}(x_h^{m,k,j}, a_h^{m,k,j})$ and

$$V_h^k(x_h^{m,k,j}) \leq \max_{a' \in \mathcal{A}} Q_h^k(x_h^{m,k,j}, a') = Q_h^k(x_h^{m,k,j}, a_h^{m,k,j}).$$

Next, we attempt to bound the terms in RHS of Equation (21) separately. Our discussions are based on the events outlined in Lemma C.1. For any given h , denote $t_h^{m,k,j} = N_h^k(x_h^{m,k,j}, a_h^{m,k,j})$ and the corresponding k, m, j (round index, agent index, and episode index) for the i -th global visiting for $(x_h^{m,k,j}, a_h^{m,k,j}, h)$ are $k_{i,h}^{m,k,j}, m_{i,h}^{m,k,j}, j_{i,h}^{m,k,j}, i = 1, 2 \dots t_h^{m,k,j}$. For the first term, due to Equation (13), we have

$$\begin{aligned} & \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} (Q_h^k - Q_h^*) (x_h^{m,k,j}, a_h^{m,k,j}) \\ & \leq \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} \tilde{\theta}_{t_h^{m,k,j}}^0 H + \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}) \\ & \quad + \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} \beta_{t_h^{m,k,j}}. \end{aligned} \quad (22)$$

For the second term, due to Equation (1),

$$\sum_{m=1}^M \sum_{j=1}^{n^{m,k}} (Q_h^* - Q_h^{\pi^k})(x_h^{m,k,j}, a_h^{m,k,j}) = \delta_{h+1}^k - \phi_{h+1}^k + \xi_{h+1}^k. \quad (23)$$

Next, we try to find some bounds related to $\sum_{k=1}^K \delta_h^k$. For notation simplicity, we use $\sum_{m,k,j}$ to represent $\sum_{k=1}^K \sum_{m=1}^M \sum_{j=1}^{n^{m,k}}$. We can prove the following relationships with details referred to Appendix C.4:

$$\sum_{m,k,j} \tilde{\theta}_{t_h^{m,k,j}}^0 H \leq M H S A. \quad (24)$$

$$\sum_{m,k,j} \sum_{i=1}^{t_h^{m,k,j}} \theta_{t_h^{m,k,j}}^i (V_{h+1}^{k_i^{m,k,j}} - V_{h+1}^*)(x_{h+1}^{(m,k,j)_i^{m,k,j}}) \leq e^{3/H} \sum_{k=1}^K \phi_{h+1}^k + O(H^3 S A (M-1)). \quad (25)$$

$$\sum_{k=1}^K \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} \beta_{t_h^{m,k,j}} \leq O(\sqrt{H^2 \hat{\tau} S A} + S A (M-1) \sqrt{H^3 \hat{\tau}}). \quad (26)$$

Combining Equations (22) to (26), we have that for any $h \in [H]$,

$$\begin{aligned} \sum_{k=1}^K \delta_h^k &\leq \exp(3/H) \sum_{k=1}^K \phi_{h+1}^k + \sum_{k=1}^K \delta_{h+1}^k - \sum_{k=1}^K \phi_{h+1}^k + \sum_{k=1}^K \xi_{h+1}^k \\ &\quad + O\left(\sqrt{H^2 \hat{\tau} S A} + S A (M-1) \sqrt{H^3 \hat{\tau}} + M H S A + H^3 S A (M-1)\right). \end{aligned}$$

Noticing that $\delta_h^k \geq \phi_h^k, \forall (h, k) \in [H] \times [K]$ due to the optimality of π^* and $\exp(3/H)^H = O(1)$, by recursions on $1, 2 \dots H$, we have

$$\begin{aligned} \sum_{k=1}^K \delta_1^k &\leq \sum_{h=1}^{H-1} C_h \sum_{k=1}^K \xi_{h+1}^k \\ &\quad + O\left(\sqrt{H^4 \hat{\tau} S A} + H S A (M-1) \sqrt{H^3 \hat{\tau}} + M H^2 S A + H^4 S A (M-1)\right), \end{aligned}$$

in which $C_h = \exp(3(h-1)/H)$. With Lemma C.4, we can also show that, with high probability,

$$\left| \sum_{k=1}^K C_h \sum_{h=1}^H \xi_{h+1}^k \right| \leq O(H \sqrt{\hat{\tau} \hat{\tau}}). \quad (27)$$

This indicates that $\sum_{k=1}^K \delta_1^k = O(\sqrt{H^4 \hat{\tau} S A} + H S A (M-1) \sqrt{H^3 \hat{\tau}} + M H^2 S A + H^4 S A (M-1))$. With these discussions, we have already shown that, under the intersection of events given in Lemma C.1 and Lemma C.4,

$$\begin{aligned} \text{Regret}(T) &\leq \sum_{k=1}^K \delta_1^k \\ &\leq O\left(\sqrt{H^4 \hat{\tau} S A} + H S A (M-1) \sqrt{H^3 \hat{\tau}} + M H^2 S A + H^4 S A (M-1)\right). \end{aligned}$$

By replacing p for the events in Lemma C.1 and Lemma C.4 with $p/2$, we finish the proof.

C.4 PROOFS OF EQUATIONS (24) TO (26)

In this subsection, we try to give bounds the terms in RHS of Equation (21) separately. We make discussions based on the intersection of events given in Lemma C.1 and Lemma C.4. Under these events, we have already shown that Equation (22), Equation (23) and Equation (27) hold. So, we will provide the proof by establishing Equations (24) to (26).

Proof of Equation (24). First, we note that

$$\sum_{m,k,j} \tilde{\theta}_{t_h^{m,k,j}}^0 H \leq \sum_{m,k,j} H \cdot \mathbb{I}[t_h^{m,k,j} = 0].$$

For each $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we consider all the rounds indexed as $0 < k_1 < k_2 < \dots$ satisfying the condition $n_h^{k_1}(x, a) > 0$. Here, k_s s are simplified notations for functions of (x, a, h) , and we use the simplified notations when there is no ambiguity and the stated meaning of these notations is applicable only to the proof of Equation (24). So,

$$\sum_{m,k,j} \mathbb{I}[t_h^{m,k,j} = 0] \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)] \leq n_h^{k_1}(x, a).$$

As $N_h^{k_1}(x, a) = 0$, due to Equation (12), we have $n_h^{k_1}(x, a) \leq M$. Therefore,

$$\sum_{m,k,j} \tilde{\theta}_{t_h^{m,k,j}}^0 H = \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{m,k,j} H \mathbb{I}[t_h^{m,k,j} = 0] \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)] \leq M H S A.$$

This completes the proof for Equation (24). \square

Proof of Equation (25). We denote $i_1 = (M-1)H(H+1)$ and split the summation into two parts:

$$\begin{aligned} & \sum_{k,m,j} \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}) \\ &= \sum_{m,k,j} \mathbb{I}[t_h^{m,k,j} \leq i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}) \\ &+ \sum_{m,k,j} \mathbb{I}[t_h^{m,k,j} > i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}). \end{aligned}$$

To bound the first term, we first notice that

$$\sum_{m,k,j} \mathbb{I}[t_h^{m,k,j} \leq i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}) \leq H \cdot \sum_{m,k,j} \mathbb{I}[0 < t_h^{m,k,j} \leq i_1]$$

due to the fact that $(V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}) \leq H$ and $\sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i = \mathbb{I}[t_h^{m,k,j} > 0]$ given in (c) in Lemma B.3. For every $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, suppose that k' is the round index such that $N_h^{k'}(x, a) \leq i_1$ and $N_h^{k'+1}(x, a) > i_1$, and k'' is the round index such that $N_h^{k''}(x, a) = 0$ and $N_h^{k''+1}(x, a) > 0$. Here, k' and k'' are simplified notations for functions of (x, a, h) , and we use the simplified notations when there is no ambiguity and the stated meaning is only valid in the proof of Equation (25).

We have

$$\sum_{m,k,j} \mathbb{I}[0 < t_h^{m,k,j} \leq i_1] \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)] \leq i_1 + n_h^{k'}(x, a) - n_h^{k''}(x, a).$$

As $n_h^{k''}(x, a) \geq 1$ and $n_h^{k'}(x, a) \leq M$ due to $N_h^{k'}(x, a) < i_0$ and Equation (12),

$$\sum_{m,k,j} \mathbb{I}[0 < t_h^{m,k,j} \leq i_1] \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)] \leq i_1 + (M-1) = O(H^2(M-1)).$$

So,

$$\begin{aligned} & \sum_{m,k,j} \mathbb{I}[t_h^{m,k,j} \leq i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}) \\ & \leq H \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{m,k,j} \mathbb{I}[0 < t_h^{m,k,j} \leq i_1] \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)] \\ & = O(H^3 S A (M-1)). \end{aligned} \tag{28}$$

To bound the second term, we first notice that $V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^* \geq 0$ due to Equation (13). Then we regroup the summations in a different way. For every (m', k', j') , the term $(V_{h+1}^{k'} - V_{h+1}^*)(x_{h+1}^{m',k',j'})$ appears in the term $\mathbb{I}[t_h^{m,k,j} > i_1] \sum_{i=1}^{t_h^{m,k,j}} \theta_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*)(x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}})$ for (k, m, j) if and only if $k > k'$ and $(x_h^{m,k,j}, a_h^{m,k,j}) = (x_h^{m',k',j'}, a_h^{m',k',j'})$. Thus, for each (m', k', j') , we denote $(x, a) = (x_h^{m',k',j'}, a_h^{m',k',j'})$. We consider all the later round indices $k' = k_0 < k_1 < k_2 < \dots$ that satisfy $n_s = n_h^{k_s}(x, a) > 0, s \in \mathbb{N}$. Here, k_s 's are simplified notations for functions of (m', k', j', h) , and we use the simplified notations when there is no ambiguity and the stated meaning is only valid in proof of Equation (25). Then, the coefficient of summation related to (m', k', j') can be upper bounded by

$$\sum_{s=1}^{\infty} \mathbb{I}[N_h^{k_s}(x, a) > i_1] n_s \tilde{\theta}_{N_h^{k_s}(x, a)}^{i'},$$

in which i' is the global visiting number for (x, a, h) at (m', k', j') , which means that $(m', k', j') = m_h(i'; x, a), k_h(i'; x, a), j_h(i'; x, a)$.

First, based on (e) in Lemma B.3, we have that

$$\sum_{s=1}^{\infty} \mathbb{I}[N_h^{k_s}(x, a) > i_1] n_s \tilde{\theta}_{N_h^{k_s}(x, a)}^{i'} \leq \exp(1/H) \sum_{s=1}^{\infty} \mathbb{I}[N_h^{k_s}(x, a) > i_1] n_s \theta_{N_h^{k_s}(x, a)}^{i'}.$$

We know that, $N_h^{k_s}(x, a) + n_s = N_h^{k_{s+1}}(x, a)$, $i' \leq N_h^{k_1}(x, a)$, and by (d) in Lemma B.2, $\sum_{i=i'}^{\infty} \theta_i^{i'} = (1 + 1/H)$. Therefore, if we can find $C' > 1$ such that

$$C' \geq \max_{(i', i'') \in \tilde{A}} \frac{\theta_{N_h^{k_s}(x, a)}^{i'}}{\theta_{N_h^{k_s}(x, a) + i''}^{i'}}, \forall (x, a, h, s) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathbb{N},$$

where $\tilde{A} = \{(i', i'') \in \mathbb{N}^2 : N_h^{k_s}(x, a) > i_1, 0 < i'' < n_s\}$, we can have

$$\sum_{s=1}^{\infty} \mathbb{I}[N_h^{k_s}(x, a) > i_1] n_s \tilde{\theta}_{N_h^{k_s}(x, a)}^{i'} \leq C' \exp(2/H).$$

Next, we prove that, for any $(i', i'') \in \tilde{A}$, if $N_h^{k_s}(x, a) > i_1$,

$$\begin{aligned} \frac{\theta_{N_h^{k_s}(x, a)}^{i'}}{\theta_{N_h^{k_s}(x, a) + i''}^{i'}} &\leq \frac{\theta_{N_h^{k_s}(x, a)}^{i'}}{\theta_{N_h^{k_s}(x, a) + n_s - 1}^{i'}} \\ &= \prod_{d=N_h^{k_s}(x, a) + 1}^{N_h^{k_s}(x, a) + n_s - 1} (1 - \alpha_d)^{-1} \\ &\leq (1 - \alpha_{d_0})^{1 - n_s} \\ &\leq \exp(1/H), \end{aligned}$$

in which $d_0 = N_h^{k_s}(x, a) + 1$ so that we can let $C' = \exp(1/H)$. The first inequality holds because $\theta_t^i = \alpha_i \prod_{i'=i+1}^t (1 - \alpha_{i'})$ is a decreasing function with respect to t . The equality follows from the definition of θ_t^i . The second inequality holds because $\alpha_t = \frac{H+1}{H+t}$ is a decreasing function with respect to t . Then we focus on the last inequality. According to the definition of α_t , we have

$$(1 - \alpha_{d_0})^{1 - n_s} = \left(1 - \frac{H+1}{H + N_h^{k_s}(x, a) + 1}\right)^{1 - n_s} = \left(1 + \frac{H+1}{N_h^{k_s}(x, a)}\right)^{n_s - 1}.$$

If $N_h^{k_s}(x, a) > MH(H+1)$, according to Equation (12), we have that $n_s \leq \frac{N_h^{k_s}(x, a)}{H(H+1)}$. Then we have

$$\left(1 + \frac{H+1}{N_h^{k_s}(x, a)}\right)^{n_s - 1} \leq \left(1 + \frac{H+1}{N_h^{k_s}(x, a)}\right)^{\frac{N_h^{k_s}(x, a)}{H(H+1)}} \leq \exp(1/H).$$

If $i_1 < N_h^{k_s}(x, a) \leq MH(H+1)$, we can prove that

$$\begin{aligned} \left(1 + \frac{H+1}{N_h^{k_s}(x, a)}\right)^{n_s-1} &\stackrel{(a)}{\leq} \left(1 + \frac{H+1}{N_h^{k_s}(x, a)}\right)^{M-1} \\ &\stackrel{(b)}{<} \left(1 + \frac{1}{H(M-1)}\right)^{M-1} \\ &\leq \exp(1/H) \end{aligned}$$

where (a) holds because according to Equation (12) we have $n_s \leq M$ and (b) holds because $N_h^{k_s}(x, a) > (M-1)H(H+1)$.

Putting the two cases together, we have

$$\sum_{s=1}^{\infty} \mathbb{I}[N_h^{k_s}(x, a) > i_1] n_s \tilde{\theta}_{N_h^{k_s}(x, a)}^{j'} \leq \exp(1/H) \exp(2/H) \leq \exp(3/H).$$

Then we conclude that

$$\begin{aligned} \sum_{m,k,j} \mathbb{I}[t_h^{m,k,j} > i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i (V_{h+1}^{k_{i,h}^{m,k,j}} - V_{h+1}^*) (x_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}) \\ \leq \exp\left(\frac{3}{H}\right) \sum_{m',k',j'} (V_{h+1}^{k'} - V_{h+1}^*) (x_{h+1}^{m',k',j'}) \\ = \exp\left(\frac{3}{H}\right) \sum_{k=1}^K \phi_{h+1}^k. \end{aligned}$$

Combining with Equation (28), we complete the proof for Equation (25). \square

Proof of Equation (26). We split the summation into two parts:

$$\sum_{m,k,j} \beta_{t_h^{m,k,j}} = \sum_{m,k,j} \beta_{t_h^{m,k,j}} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1] + \sum_{m,k,j} \beta_{t_h^{m,k,j}} \mathbb{I}[t_h^{m,k,j} \geq M].$$

For every pair (x, a, h) , we consider all the rounds indexed as $0 < k_1 < k_2 < \dots$ satisfying the condition $n_s = N_h^{k_s}(x, a) > 0$. Suppose that k_p is the round index such that $N_h^{k_p}(x, a) \leq M-1$ and $N_h^{k_{p+1}}(x, a) > M-1$. Here, k_s s and p are simplified notations for functions of (x, a, h) , and we use the simplified notations when there is no ambiguity and the stated meaning is only valid in the proof of Equation (26).

To bound the first term, we use the fact that $\beta_{t_h^{m,k,j}} \leq O(1)\sqrt{H^3\iota}$. Then we have

$$\sum_{m,k,j} \beta_{t_h^{m,k,j}} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1] \leq O(1)\sqrt{H^3\iota} \sum_{m,k,j} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1]$$

and

$$\sum_{m,k,j} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1] \leq \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \left(M-1 + n_h^{k_p}(x, a) - n_h^{k_1}(x, a) \right).$$

It is straightforward that $n_h^{k_1}(x, a) \geq 1$. Additionally, due to Equation (12), we can establish that $n_h^{k_p}(x, a) \leq M$. Therefore

$$\sum_{m,k,j} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1] \leq 2SA(M-1)$$

and

$$\sum_{m,k,j} \beta_{t_h^{m,k,j}} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1] = O(SA(M-1)\sqrt{H^3\iota}). \quad (29)$$

To establish bounds for the second term, we define some notions first. For every pair $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we consider all the rounds indexed as $0 < \tilde{k}_1 < \tilde{k}_2 < \dots < \tilde{k}_g \leq K$ satisfying $n_h^{\tilde{k}_s}(x, a) > 0$, $N_h^{\tilde{k}_1}(x, a) \geq M$ and $N_h^{\tilde{k}_1-1}(x, a) < M$. Here, \tilde{k}_s s and g are simplified notations for functions of (x, a, h) , and we use the simplified notations when there is no ambiguity and the stated meaning is only valid in proof of Equation (26). Then we have

$$\sum_{m,k,j} \beta_{t_h^{m,k,j}} \mathbb{I}[t_h^{m,k,j} \geq M] = O(1) \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s=1}^g n_h^{\tilde{k}_s}(x, a) \sqrt{\frac{H^3 \iota}{N_h^{\tilde{k}_s}(x, a)}}.$$

Firstly, we prove that

$$\begin{aligned} \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s=1}^g \sum_{j''=1}^{n_h^{\tilde{k}_s}(x,a)} \sqrt{\frac{H^3 \iota}{N_h^{\tilde{k}_s}(x, a) + j'' - 1}} &= O(1) \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sqrt{H^3 \iota (N_h^{\tilde{k}_g}(x, a) + n_h^{\tilde{k}_g}(x, a) - 1)} \\ &= O(1) \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sqrt{H^3 \iota N_h^{K+1}(x, a)} \\ &\stackrel{(a)}{\leq} O(\sqrt{H^2 \iota \hat{T} S A}) \end{aligned}$$

where (a) holds because of the concavity of $f(x) = \sqrt{H^3 \iota x}$ and the fact that $\sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} N_h^{K+1}(x, a) \leq \hat{T}/H$.

Then we bound $\frac{1/\sqrt{N_h^{\tilde{k}_s}(x, a)}}{1/\sqrt{N_h^{\tilde{k}_s}(x, a) + d}}$. If we can find some numerical constant $C'' > 1$ such that

$$C'' \geq \max_{(j,d) \in \tilde{B}} \frac{1/\sqrt{N_h^{\tilde{k}_s}(x, a)}}{1/\sqrt{N_h^{\tilde{k}_s}(x, a) + d}}, \forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H],$$

in which $\tilde{B} = \{(s, d) \in \mathbb{N}^2 : 1 \leq s \leq g, 1 \leq d \leq n_h^{\tilde{k}_s}(x, a) - 1\}$, then we can have

$$\begin{aligned} \sum_{m,k,j} \beta_{t_h^{m,k,j}} \mathbb{I}[t_h^{m,k,j} \geq M] &= O(1) \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s=1}^g n_h^{\tilde{k}_s}(x, a) \sqrt{\frac{H^3 \iota}{N_h^{\tilde{k}_s}(x, a)}} \\ &\leq O(C'') \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s=1}^g \sum_{j''=1}^{n_h^{\tilde{k}_s}(x,a)} \sqrt{\frac{H^3 \iota}{N_h^{\tilde{k}_s}(x, a) + j'' - 1}} \\ &= O(\sqrt{H^2 \iota \hat{T} S A}). \end{aligned} \tag{30}$$

Next, we will prove that we can choose $C'' = \sqrt{2}$. We notice that

$$\max_{d \in [n_h^{\tilde{k}_s}(x, a) - 1]} \frac{1/\sqrt{N_h^{\tilde{k}_s}(x, a)}}{1/\sqrt{N_h^{\tilde{k}_s}(x, a) + d}} = \sqrt{\frac{N_h^{\tilde{k}_s}(x, a) + n_h^{\tilde{k}_s}(x, a) - 1}{N_h^{\tilde{k}_s}(x, a)}}.$$

If $M \leq N_h^{\tilde{k}_s}(x, a) < i_0$, according to Equation (12), $\tilde{n}_j(x, a) \leq M$, which indicates that

$$\sqrt{\frac{N_h^{\tilde{k}_s}(x, a) + n_h^{\tilde{k}_s}(x, a) - 1}{N_h^{\tilde{k}_s}(x, a)}} \leq \sqrt{2}.$$

If $N_h^{\tilde{k}_j(x, a, h)}(x, a) \geq i_0$, according to Equation (12), $n_h^{\tilde{k}_s}(x, a) \leq \tilde{C} N_h^{\tilde{k}_s}(x, a)$

$$\sqrt{\frac{N_h^{\tilde{k}_s}(x, a) + n_h^{\tilde{k}_s}(x, a) - 1}{N_h^{\tilde{k}_s}(x, a)}} \leq \sqrt{1 + \tilde{C}} \leq \sqrt{2}.$$

So, we can choose $C'' = \sqrt{2}$.

Combining Equation (29) and Equation (30), we obtain Equation (26). \square

D PROOF OF THEOREM 4.2

Proof of Theorem 4.2. This theorem is proved under the synchronization assumption, i.e., $n^{m,k} = n^k, \forall m \in [M]$. We only need to prove that when $k \geq H^2(H+1)SAM$,

$$\left[\left(1 + \frac{1}{2H(H+1)M} \right)^{\lceil K/(HSA) \rceil - H(H+1)M} \right] H^2(H+1)M^2 \leq \hat{T}.$$

For each $k \in [K]$, there exists at least one $(x, m, h) \in \mathcal{S} \times [M] \times [H]$ with $a = \pi_h^k(x)$ such that equality in Equation (11) holds. Thus, there exist at least K different tuples of $(x, a, h, m, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [M] \times [K]$ such that equality in Equation (11) holds. Define set \mathcal{K} to have all the different k 's satisfying that there exists $m \in [M]$ such that the equality in Equation (11) holds. Then, by Pigeonhole principle, there must exist a triple $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ such that $|\mathcal{K}| \geq \lceil K/(HSA) \rceil$.

We order the elements of \mathcal{K} as $0 < k_1 < k_2 \dots < k_g \leq K$, where $g \geq \lceil K/(HSA) \rceil$. We also denote $N_s = N_h^{k_s+1}(x, a)$, m_s as the first agent index such that equality in Equation (11) holds, and $n_s = n_h^{m_s, k_s}(x, a)$. Due to the synchronization assumption, we have

$$\hat{T} \geq HM \sum_{s=1}^g n_s. \quad (31)$$

When $N_s \geq H(H+1)M$, we have that

$$N_s \geq \sum_{s'=1}^s n_{s'}, n_{s+1} \geq \tilde{C}N_s/(2M)$$

due to Equation (11) and $\left\lfloor \frac{s'}{H(H+1)M} \right\rfloor \geq \frac{s'}{2H(H+1)M}, \forall s' \geq M(H+1)H$.

Thus, we have that $\sum_{s=1}^{H(H+1)M} n_s \geq H(H+1)M$ and

$$\sum_{s'=1}^{s+1} n_{s'} \geq \sum_{s'=1}^s n_{s'} + \tilde{C}N_s/(2M) \geq (1 + \tilde{C}/(2M)) \sum_{s'=1}^s n_{s'}, s \geq H(H+1)M.$$

Therefore,

$$\begin{aligned} \sum_{s'=1}^g n_{s'} &\geq \left[\left(1 + \tilde{C}/(2M) \right)^{g-H(H+1)M} \right] H(H+1)M \\ &\geq \left[\left(1 + \tilde{C}/(2M) \right)^{\lceil K/(HSA) \rceil - H(H+1)M} \right] H(H+1)M. \end{aligned}$$

Combining with Equation (31), we have

$$\left[\left(1 + \frac{1}{2H(H+1)M} \right)^{\lceil K/(HSA) \rceil - H(H+1)M} \right] H^2(H+1)M^2 \leq \hat{T},$$

which directly leads to the conclusion. \square

E THE BERNSTEIN-TYPE ALGORITHM

E.1 ALGORITHM DESIGN

The Bernstein-type algorithm differs from the Hoeffding-type algorithm Algorithms 1 and 2, in that it selects the upper confidence bound based on a variance estimator of X_i , akin to the approach used in the Bernstein-type algorithm in Jin et al. (2018). This is done to determine a probability upper bound of $|\sum_{i=1}^t \theta_i^i X_i|$. In this subsection, we first introduce the algorithm design.

To facilitate understanding, we introduce additional notations exclusive to Bernstein-type algorithms, supplementing the already provided notations for the Hoeffding-type algorithm.

$$\mu_h^{m,k}(x, a) = \frac{1}{n_h^{m,k}(x, a)} \sum_{j=1}^{n_h^{m,k}} \left[V_{h+1}^k \left(x_{h+1}^{m,k,j} \right) \right]^2 \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)].$$

$$\mu_h^k(x, a) = \frac{1}{N_h^{k+1}(x, a) - N_h^k(x, a)} \sum_{m=1}^M \mu_h^{m,k}(x, a) n_h^{m,k}(x, a).$$

Here, $\mu_h^{m,k}(x, a)$ is the sample mean of $\left[V_{h+1}^k(x_{h+1}^{m,k,j}) \right]^2$ for all the visits of (x, a, h) for the m -th agent during the k -th round and $\mu_h^k(x, a)$ corresponds to the mean for all the visits during the k -th round. We emphasize here that we adopt the general notation $n_h^{m,k}$ in the definition of $\mu_h^{m,k}$. We define $W_k(x, a, h)$ to denote the sample variance of all the visits before the k -th round, calculated using $V_{h+1}^{k^i}(x_{h+1}^{m^i,k^i,j^i})$, i.e.

$$W_k(x, a, h) = \frac{1}{N_h^k(x, a)} \sum_{i=1}^{N_h^k(x, a)} \left(V_{h+1}^{k^i}(x_{h+1}^{m^i,k^i,j^i}) - \frac{1}{N_h^k(x, a)} \sum_{i'=1}^{N_h^k(x, a)} V_{h+1}^{k^{i'}}(x_{h+1}^{m^{i'},k^{i'},j^{i'}}) \right)^2.$$

We can find that

$$W_k(x, a, h) = \frac{1}{N_h^k(x, a)} \sum_{k'=1}^{k-1} \mu_h^{k'}(x, a) n_h^{k'}(x, a) - \left[\frac{1}{N_h^k(x, a)} \sum_{k'=1}^{k-1} v_{h+1}^{k'}(x, a) n_h^{k'}(x, a) \right]^2,$$

which means that this quantity can be calculated efficiently in practice in the following way. Define

$$W_{1,k}(x, a, h) = \sum_{k'=1}^{k-1} \mu_h^{k'}(x, a) n_h^{k'}(x, a), W_{2,k}(x, a, h) = \sum_{k'=1}^{k-1} v_{h+1}^{k'}(x, a) n_h^{k'}(x, a), \quad (32)$$

we have that

$$W_{1,k+1}(x, a, h) = W_{1,k}(x, a, h) + \mu_h^k(x, a) n_h^k(x, a), \quad (33)$$

$$W_{2,k+1}(x, a, h) = W_{2,k}(x, a, h) + v_{h+1}^k(x, a) n_h^k(x, a) \quad (34)$$

and

$$W_{k+1}(x, a, h) = \frac{W_{1,k+1}(x, a, h)}{N_h^{k+1}(x, a)} - \left[\frac{W_{2,k+1}(x, a, h)}{N_h^{k+1}(x, a)} \right]^2. \quad (35)$$

This indicates that the central server, by actively maintaining and updating the quantities $W_{1,k}$ and $W_{2,k}$ and systematically collecting $n_h^{m,k}$ s, $\mu_h^{m,k}$ s and $v_{h+1}^{m,k}$ s, is able to compute W_{k+1} .

Next, we define

$$\beta_t(x, a, h) = c' \left(\min \left\{ \sqrt{\frac{H\iota}{t}} (W_{k^t+1}(x, a, h) + H) + \iota \frac{\sqrt{H^7 S A} + \sqrt{M S A H^6}}{t}, \sqrt{\frac{H^3 \iota}{t}} \right\} \right),$$

in which $c' > 0$ is a positive constant. Here, $W_{k^t+1}(x, a, h) = W^t(x, a, h)$ which is mentioned in Section 5. With this, the upper confidence bound $b_t(x, a, h)$ for a single visit is determined by

$$\beta_t(x, a, h) = 2 \sum_{i=1}^t \theta_t^i b_t(x, a, h),$$

which can be calculated as follows:

$$b_1(x, a, h) := \frac{\beta_1(x, a, h)}{2},$$

$$b_t(x, a, h) := \frac{\beta_t(x, a, h) - (1 - \alpha_t) \beta_{t-1}(x, a, h)}{2\alpha_t}.$$

When there is no ambiguity, we adopt the simplified notation $\tilde{b}_t = b_t(x, a, h)$ and $\tilde{\beta}_t = \beta_t(x, a, h)$. In the Bernstein-type algorithm, we let $\tilde{\beta} = \beta_{t^k}(x, a, h) - \alpha^c(t^{k-1} + 1, t^k)\beta_{t^{k-1}}(x, a, h)$ in replace of β^k in Equation (3) and Equation (4). We know that $\tilde{\beta}_t \leq \beta_t$ when $c = c'$, indicating that the Bernstein-type algorithm operates with a smaller upper confidence bound.

Next, we will delve into certain components of the algorithm in round k . We remark that we discuss our algorithm based on the general situation where there is no necessity for zero latency and synchronization assumptions. In this general scenario, agent m generates $n^{m,k}$ episodes in round k .

Coordinated Exploration for Agents. At the beginning of round k , the server decides a deterministic policy $\pi^k = \{\pi_h^k\}_{h=1}^H$, and then broadcasts it along with $\{N_h^k(x, \pi_h^k(x))\}_{x,h}$ and $\{V_h^k(x)\}_{x,h}$ to all of the agents. When $k = 1$, $N_h^1(x, a) = 0$, $Q_h^1(x, a) = V_h^1(x) = H$, $\forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and π^1 is an arbitrary deterministic policy.

Once receiving such information, the agents will execute policy π^k and start collecting trajectories.

Event-Triggered Termination of Exploration. During exploration, every agent m will monitor $n_h^{m,k}(x, a)$, i.e., the total number of visits for each (x, a, h) triple within the current round. For any agent m , at the end of each episode, it sequentially conducts two procedures. First, if any (x, a, h) has been visited by $\max\left\{1, \lfloor \frac{\tilde{C}}{M} N_h^k(x, a) \rfloor\right\}$ times by agent m , it will abort its own exploration and send an abortion signal to the server and other clients. Second, it checks whether it has received an abortion signal. If so, it will abort its exploration. We remark that Equation (2) still holds, and for any $k \in [K]$, there exists a tuple (x, a, h) such that the equality is met.

Local Updating of the Estimated Expected Return. Each agent updates the local estimate of the expected return $v_{h+1}^{m,k}(x, a)$ at the end of round k . Next, each agent m sends $\{r_h(x, \pi_h^k(x))\}_{x,h}$, $\{n_h^{m,k}(x, \pi_h^k(x))\}_{x,h}$, $\{v_{h+1}^{m,k}(x, \pi_h^k(x))\}_{x,h}$ and $\{\mu_h^{m,k}(x, \pi_h^k(x))\}_{x,h}$ to the central server for aggregation.

Server-side Information Aggregation. After receiving the information sent by the agents, for each (x, a, h) tuple visited by the agents, the server first calculates $W_{1,k+1}(x, a, h)$, $W_{2,k+1}(x, a, h)$ and $W_{k+1}(x, a, h)$ based on Equation (32), Equation (33), Equation (34) and Equation (35) for each pair (x, h) with $a = \pi_h^k(x)$. Then it sets $t^{k-1} = N_h^k(x, a)$, $t^k = N_h^{k+1}(x, a)$, $\alpha_{agg} = 1 - \alpha^c(t^{k-1} + 1, t^k)$ and $\tilde{\beta} = \beta_{t^k}(x, a, h) - \alpha^c(t^{k-1} + 1, t^k)\beta_{t^{k-1}}(x, a, h)$, and updates the global estimate of the value functions according to one of the following two cases.

- **Case 1:** $N_h^k(x, a) < i_0$. Due to Equation (2), this case implies that each client can visit each (x, a) pair at step h at most once. Then, we denote $1 \leq m_1 < m_2 \dots < m_{t^k - t^{k-1}} \leq M$ as the agent indices with $n_h^{m,k}(x, a) > 0$. The server then updates the global estimate of action values as follows:

$$Q_h^{k+1}(x, a) = (1 - \alpha_{agg})Q_h^k(x, a) + \alpha_{agg}r_h(x, a) + \sum_{t=1}^{t^k - t^{k-1}} \theta_{t^k}^{t^{k-1} + t} v_{h+1}^{m_t, k}(x, a) + \tilde{\beta}/2. \quad (36)$$

- **Case 2:** $N_h^k(x, a) \geq i_0$. In this case, the central server calculates $v_{h+1}^k(x, a)$ as and updates the Q -estimate as

$$Q_h^{k+1}(x, a) = (1 - \alpha_{agg})Q_h^k(x, a) + \alpha_{agg}(r_h(x, a) + v_{h+1}^k(x, a)) + \tilde{\beta}/2. \quad (37)$$

After finishing updating the estimated Q function, the central server updates the estimated value function and the policy based on Equations (5) and (6). The algorithm then proceeds to round $k + 1$.

Algorithms 3 and 4 formally present the Bernstein-type design. Inputs K_0 and T_0 in Algorithms 3 are termination conditions, where K_0 limits the total number of rounds and T_0 limits the total number of samples generated by all the agents before the last round.

We provide some remarks. First, the coordinated exploration for agents is designed based on the general situation where $n^{m,k}$ might be different across different agents, and clients share $\mu_h^{m,k}$ s in addition to the information shared during the coordinated exploration in the Hoeffding-type Algorithm 2. Second, the information aggregation at the central server differs from that in the Hoeffding-type Algorithm 1, in terms of specifying $\tilde{\beta}$ to set the upper confidence bound and in maintaining $W_{1,k}$, $W_{2,k}$ and W_k .

Algorithm 3 FedQ-Bernstein (Central Server)

```

1: Input:  $T_0, K_0 \in \mathbb{N}_+$ .
2: Initialization:  $k = 1, N_h^1(x, a) = W_{1,k}(x, a, h) = W_{2,k}(x, a, h) = 0, Q_h^1(x, a) = V_h^1(x) = H, \forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$  and  $\pi^1 = \{\pi_h^1 : \mathcal{S} \rightarrow \mathcal{A}\}_{h \in [H]}$  is an arbitrary deterministic policy.

3: while  $H \sum_{k'=1}^{k-1} Mn^{k'} < T_0$  &  $k \leq K_0$  do
4:   Broadcast  $\pi^k, \{N_h^k(x, \pi_h^k(x))\}_{x,h}$  and  $\{V_h^k(x)\}_{x,h}$  to all clients.
5:   Wait until receiving an abortion signal and send the signal to all agents.
6:   Receive  $\{r_h(x, \pi_h^k(x))\}_{x,h}, \{n_h^{m,k}(x, \pi_h^k(x))\}_{x,h,m}, \{v_{h+1}^{m,k}(x, \pi_h^k(x))\}_{x,h,m}$  and  $\{\mu_h^{m,k}(x, \pi_h^k(x))\}_{x,h,m}$  from clients.
7:   Calculate  $N_h^{k+1}(x, a), n_h^k(x, a), v_{h+1}^k(x, a), \forall (x, h) \in \mathcal{S} \times [H]$  with  $a = \pi_h^k(x)$ .
8:   Calculate  $W_k(x, a, h), W_{k+1}(x, a, h), W_{1,k+1}(x, a, h), W_{2,k+1}(x, a, h), \forall (x, h) \in \mathcal{S} \times [H]$  with  $a = \pi_h^k(x)$  based on Equation (32), Equation (33), Equation (34) and Equation (35).
9:   for  $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$  do
10:    if  $a \neq \pi_h^k(x)$  or  $n_h^k(x, a) = 0$  then
11:       $Q_h^{k+1}(x, a) \leftarrow Q_h^k(x, a)$ .
12:    else if  $N_h^k(x, a) < i_0$  then
13:      Update  $Q_h^{k+1}(x, a)$  according to Equation (36).
14:    else
15:      Update  $Q_h^{k+1}(x, a)$  according to Equation (37).
16:    end if
17:  end for
18:  Update  $V_h^{k+1}$  and  $\pi^{k+1}$  according to Equation (5) and Equation (6).
19:   $k \leftarrow k + 1$ .
20: end while

```

Algorithm 4 FedQ-Bernstein (Agent m in round k)

```

1:  $n_h^m(x, a) = v_{h+1}^m(x, a) = r_h(x, a) = \mu_h^m(x, a) = 0, \forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ .
2: Receive  $\pi^k, \{N_h^k(x, \pi_h^k(x))\}_{x,h}$  and  $\{V_h^k(x)\}_{x,h}$  from the central server.
3: while no abortion signal from the central server do
4:   while  $n_h^m(x_h, a_h) < \max\left\{1, \lfloor \frac{\tilde{C}}{M} N_h^k(x_h, a_h) \rfloor\right\}, \forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$  do
5:     Collect a new trajectory  $\{(x_h, a_h, r_h)\}_{h=1}^H$  with  $a_h = \pi_h^k(x_h)$ .
6:      $n_h^m(x_h, a_h) \leftarrow n_h^m(x_h, a_h) + 1, v_{h+1}^m(x_h, a_h) \leftarrow v_{h+1}^m(x_h, a_h) + V_{h+1}^k(x_{h+1}),$ 
        $\mu_h^m(x_h, a_h) \leftarrow \mu_h^m(x_h, a_h) + [V_{h+1}^k(x_{h+1})]^2$ , and  $r_h(x_h, a_h) \leftarrow r_h, \forall h \in [H]$ .
7:   end while
8:   Send an abortion signal to the central server.
9: end while
10:  $n_h^{m,k}(x, a) \leftarrow n_h^m(x, a), v_{h+1}^{m,k}(x, a) \leftarrow v_{h+1}^m(x, a)/n_h^m(x, a)$  and  $\mu_h^{m,k}(x, a) \leftarrow \mu_h^m(x, a)/n_h^m(x, a), \forall (x, h) \in \mathcal{S} \times [H]$  with  $a = \pi_h^k(x)$ .
11: Send  $\{r_h(x, \pi_h^k(x))\}_{x,h}, \{n_h^{m,k}(x, \pi_h^k(x))\}_{x,h}, \{\mu_h^{m,k}(x, \pi_h^k(x))\}_{x,h}$  and  $\{v_{h+1}^{m,k}(x, \pi_h^k(x))\}_{x,h}$  to the central server.

```

E.2 PROOF OF THEOREM 5.1

In this subsection, we provide proof for Theorem 5.1 which provides the regret of Algorithms 3 and 4.

E.2.1 BOUNDS ON $Q_h^k - Q_h^*$

We first try to provide a Lemma that has stronger results than Lemma C.1.

Lemma E.1. *Using Algorithms 3 and 4, there exists a positive constant $c' > 0$ such that, for any $p \in (0, 1)$, the following relationship holds simultaneously for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K']$*

with probability at least $1 - p$.

$$0 \leq Q_h^{K'}(x, a) - Q_h^*(x, a) \leq \theta_t^0 H + \sum_{i=1}^t \tilde{\theta}_t^i (V_{h+1}^{k^i} - V_{h+1}^*) (x_{h+1}^{m^i, k^i, j^i}) + \beta_t(x, a, h), \quad (38)$$

in which $t = N_h^{K'}(x, a)$.

The remaining content of Appendix E.2.1 is dedicated to proving this Lemma. First, we can easily find that Lemma C.2 still holds with b_t, β_t replaced by $\tilde{b}_t, \tilde{\beta}_t$, and Lemma C.3 still holds. Next, due to Equation (16) and Equation (15) with b_t, β_t replaced, we can easily obtain a similar one-sided result summarized in the following Lemma.

Lemma E.2. *Using the Bernstein-type algorithm, there exists a positive constant $c'_0 > 0$ such that, for any $p \in (0, 1)$, the following relationship holds simultaneously for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ with probability at least $1 - p$.*

$$Q_h^{K'}(x, a) - Q_h^*(x, a) \leq \theta_t^0 H + \sum_{i=1}^t \tilde{\theta}_t^i (V_{h+1}^{k^i} - V_{h+1}^*) (x_{h+1}^{m^i, k^i, j^i}) + c'_0 \sqrt{H^3 \iota / t}, \quad (39)$$

in which $t = N_h^{K'}(x, a)$.

Proof. This relationship can be directly obtained from Equation (16) and Equation (15) with b_t, β_t replaced. \square

With this, we can introduce the following technical Lemma.

Lemma E.3. *Suppose that Equation (39) holds. For any given $K' \in \mathbb{N}$, denote $\sum_{m,k,j}^{K'} = \sum_{k=1}^{K'} \sum_{m=1}^M \sum_{j=1}^{n_h^{m,k}}$ and $w = \text{vec}(\{w_{mkj}\})$ with $m \in [M], k \in [K'], j \in [n^{m,k}]$ be a non-negative vector. Then there exists a numerical constant $c'_1 > 0$ such that, for all $(h, K') \in [H] \times [K]$,*

$$\begin{aligned} & \sum_{m,k,j}^{K'} w_{mkj} \left(V_h^k(x_h^{m,k,j}) - V_h^*(x_h^{m,k,j}) \right) \\ & \leq c'_1 \left(\|w\|_\infty M S A \sqrt{H^5 \iota} + \sqrt{S A} \|w\|_\infty \|w\|_1 H^5 \iota + H^4 S A (M - 1) \|w\|_\infty \right). \end{aligned} \quad (40)$$

Proof. We denote

$$\tilde{V}_h^{m,k,j} = V_h^k(x_h^{m,k,j}) - V_h^*(x_h^{m,k,j}).$$

Noticing that $Q_h^k(x_h^{m,k,j}, a_h^{m,k,j}) \geq V_h^k(x_h^{m,k,j})$ and $Q_h^*(x_h^{m,k,j}, a_h^{m,k,j}) = \max_{a \in \mathcal{A}} Q_h^*(x_h^{m,k,j}) \leq V_h^*(x_h^{m,k,j})$, letting $k = K'$ and $(x, a) = (x_h^{m,k,j}, a_h^{m,k,j})$, we have that

$$\tilde{V}_h^{m,k,j} \leq \theta_{t_h^{m,k,j}}^0 H + \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}} + c'_0 \sqrt{H^3 \iota / t_h^{m,k,j}}.$$

Taking the summation with regard to k from 1 to K' and noticing that $\theta_{t_h^{m,k,j}}^0 = \mathbb{I}[t_h^{m,k,j} > 0]$, we have

$$\begin{aligned} \sum_{m,k,j}^{K'} w_{mkj} \tilde{V}_h^{m,k,j} & \leq H \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} = 0] + \sum_{m,k,j} w_{mkj} \sum_{i=1}^{t_h^{m,k,j}} \theta_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}} \\ & \quad + \sum_{m,k,j}^{K'} w_{mkj} \Omega \left(\sqrt{\frac{H^3 \iota}{t_h^{m,k,j}}} \right). \end{aligned}$$

Next, we find upper bounds with regard to the three terms.

Step 1: finding an upper bound for $H \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} = 0]$. Noticing that $w_{mkj} \leq \|w\|_\infty$, using the same way as Proof of Equation (24) in Appendix C.4, we can find that

$$H \sum_{m,k,j} w_{mkj} \mathbb{I}[t_h^{m,k,j} = 0] \leq MHS A \|w\|_\infty.$$

Step 2: finding an upper bound for $\sum_{m,k,j}^{K'} w_{mkj} \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}$. Similar to Proof of Equation (25) in Appendix C.4, with $i_1 = (M-1)H(H+1)$, we still split it into two parts as follows:

$$\begin{aligned} \sum_{m,k,j}^{K'} w_{mkj} \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}} &= \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} \leq i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}} \\ &\quad + \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} > i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}. \end{aligned}$$

For the first part, applying $w_{mkj} \leq \|w\|_\infty$, using the same way as Proof of Equation (25) in Appendix C.4, we can find that

$$\sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} \leq i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}} \leq O(H^3 SA(M-1) \|w\|_\infty).$$

For the second part, we regroup the summations in a different way. For every (m', k', j') , the term $\tilde{V}_{h+1}^{m',k',j'}$ appears in the term $w_{mkj} \mathbb{I}[t_h^{m,k,j} > i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}}$ for (k, m, j) if and only if $K' \geq k > k'$ and $(x_h^{m,k,j}, a_h^{m,k,j}) = (x_h^{m',k',j'}, a_h^{m',k',j'})$. So, for each (m', k', j') , we denote $(x, a) = (x_h^{m',k',j'}, a_h^{m',k',j'})$. We consider all the later round indices $k' = k_0 < k_1 < k_2 < \dots < k_g \leq K'$ that satisfy $n_s = n_h^{k_s}(x, a) > 0$, $s \in \mathbb{N}$. Here, k_s s and g are simplified notations for functions of (m', k', j', h) , and we use the simplified notations when there is no ambiguity and the stated meaning is only valid in proof of step 2. So, the summation of coefficients related to (m', k', j') equals to

$$\tilde{w}_{m'k'j'} = \sum_{s=1}^g \left(\sum_{i=N_h^{k_s}(x,a)+1}^{N_h^{k_s}(x,a)+n_s} w_{(mkj)_{i,h}^{m,k,j}} \right) \mathbb{I}[N_h^{k_s}(x,a) > i_1] \tilde{\theta}_{N_h^{k_s}(x,a)}^{i'},$$

in which i' is the global visiting number for (x, a, h) at (m', k', j') , which means that $(m', k', j') = m_h(i'; x, a), k_h(i'; x, a), j_h(i'; x, a)$. This means that

$$\sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} > i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}} = \sum_{m',k',j'}^{K'} \tilde{w}_{m'k'j'} \tilde{V}_{h+1}^{m',k',j'}.$$

Denote $\tilde{w} = \text{vec}(\{\tilde{w}_{m'k'j'}\})$, we have that

$$\|\tilde{w}\|_1 = \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} > i_1] \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \leq \sum_{m,k,j}^{K'} w_{mkj} \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \leq \|w\|_1$$

due to (c) in Lemma B.3. We can also find that

$$\tilde{w}_{m'k'j'} \leq \sum_{s=1}^g n_s \|w\|_\infty \mathbb{I}[N_h^{k_s}(x,a) > i_1] \tilde{\theta}_{N_h^{k_s}(x,a)}^{i'} \leq \exp(3/H) \|w\|_\infty,$$

where the proof of the last inequality is the same as the Proof of Equation (25). Combining the two parts, we have that

$$\sum_{m,k,j}^{K'} w_{mkj} \sum_{i=1}^{t_h^{m,k,j}} \tilde{\theta}_{t_h^{m,k,j}}^i \tilde{V}_{h+1}^{(m,k,j)_{i,h}^{m,k,j}} \leq \sum_{m',k',j'}^{K'} \tilde{w}_{m'k'j'} \tilde{V}_{h+1}^{m',k',j'} + O(H^3 SA(M-1) \|w\|_\infty).$$

Step 3: finding an upper bound for $\sum_{m,k,j}^{K'} w_{mkj} \Omega \left(\sqrt{\frac{H^3 \iota}{t_h^{m,k,j}}} \right)$. We split it into two parts as follows.

$$\begin{aligned} \sum_{m,k,j}^{K'} w_{mkj} \Omega \left(\sqrt{\frac{H^3 \iota}{t_h^{m,k,j}}} \right) &= \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1] \Omega \left(\sqrt{\frac{H^3 \iota}{t_h^{m,k,j}}} \right) \\ &\quad + \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} \geq M] \Omega \left(\sqrt{\frac{H^3 \iota}{t_h^{m,k,j}}} \right). \end{aligned}$$

For the first part, applying that $w_{mkj} \leq \|w\|_\infty$, similar to Proof of Equation (26) in Appendix C.4, we have that

$$\sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[0 < t_h^{m,k,j} \leq M-1] \Omega \left(\sqrt{\frac{H^3 \iota}{t_h^{m,k,j}}} \right) = \Omega \left(\|w\|_\infty S A (M-1) \sqrt{H^3 \iota} \right).$$

For the second part, we denote $w'_k(x, a, h) = \sum_{m=1}^M \sum_{j=1}^{n_h^{m,k}} w_{mkj} \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)]$. We also introduce the following notation. For every pair $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we consider all the rounds indexed as $0 < \tilde{k}_1 < \tilde{k}_2 < \dots < \tilde{k}_g \leq K$ satisfying $n_h^{\tilde{k}_s}(x, a) > 0$, $N_h^{\tilde{k}_1}(x, a) \geq M$ and $N_h^{\tilde{k}_1-1}(x, a) < M$. Here, \tilde{k}_s s and g are simplified notations for functions of (x, a, h) , and we use the simplified notations when there is no ambiguity and the stated meaning is only valid in proof of step 3. Then we have

$$\sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} \geq M] \Omega \left(\sqrt{\frac{H^3 \iota}{t_h^{m,k,j}}} \right) = \Omega(1) \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s=1}^g w'_{\tilde{k}_s}(x, a, h) \sqrt{\frac{H^3 \iota}{N_h^{\tilde{k}_s}(x, a)}}.$$

We also define that

$$w'(i, x, a, h) = w'_{\tilde{k}_s}(x, a, h) / n_h^{\tilde{k}_s}(x, a), \forall i \in \mathbb{N}_+, j \in [N_h^{\tilde{k}_s}(x, a), N_h^{\tilde{k}_s}(x, a) + n_h^{\tilde{k}_s}(x, a) - 1],$$

which indicates that

$$w'(j, x, a, h) \leq \|w\|_\infty.$$

Similar to Proof of Equation (26), we have that

$$\sqrt{2} \geq \max_{(j,d) \in \tilde{B}} \frac{1 / \sqrt{N_h^{\tilde{k}_s}(x, a)}}{1 / \sqrt{N_h^{\tilde{k}_s}(x, a) + d}}, \forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H],$$

in which $\tilde{B} = \{(s, d) \in \mathbb{N}^2 : 1 \leq s \leq g, 1 \leq d \leq n_h^{\tilde{k}_s}(x, a) - 1\}$. So we have

$$\sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{s=1}^g w'_{\tilde{k}_s}(x, a, h) \sqrt{\frac{H^3 \iota}{N_h^{\tilde{k}_s}(x, a)}} = O(1) \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{i=N_h^{\tilde{k}_1}(x, a)}^{N_h^{\tilde{k}_g+1}(x, a)-1} w'(i, x, a, h) \sqrt{H^3 \iota / i}$$

with

$$\sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{i=N_h^{\tilde{k}_1}(x, a)}^{N_h^{\tilde{k}_g+1}(x, a)-1} w'(i, x, a, h) \leq \|w\|_1.$$

Denote $w''(x, a, h) = \|w\|_\infty \left[\sum_{i=N_h^{\tilde{k}_1}(x, a)}^{N_h^{\tilde{k}_g+1}(x, a)-1} w'(i, x, a, h) / \|w\|_\infty \right]$, which indicates that

$$w''(x, a, h) \leq \sum_{i=N_h^{\tilde{k}_1}(x, a)}^{N_h^{\tilde{k}_g+1}(x, a)-1} w'(i, x, a, h) + \|w\|_\infty$$

so that

$$\sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} w''(x, a, h) \leq \|w\|_1 + SA\|w\|_\infty.$$

Then by letting the mass related to $\{w'(i, x, a, h)\}_i$ concentrate at large values for $\{\sqrt{H^3\iota/i}\}_i$ as much as possible, we have

$$\begin{aligned} \sum_{i=N_h^{k_1}(x,a)}^{N_h^{k_g+1}(x,a)-1} w'(i, x, a, h) \sqrt{H^3\iota/i} &\leq \|w\|_\infty \sum_{i=N_h^{k_1}(x,a)}^{N_h^{k_1}(x,a)+w''(x,a,h)/\|w\|_\infty-1} \sqrt{H^3\iota/i} \\ &= O\left(\sqrt{H^3\iota w''(x, a, h)\|w\|_\infty}\right). \end{aligned}$$

By the concavity of $f(x) = \sqrt{H^3\iota x}$, we have

$$\begin{aligned} \sum_{m,k,j}^{K'} w_{mkj} \mathbb{I}[t_h^{m,k,j} \geq M] \Omega\left(\sqrt{\frac{H^3\iota}{t_h^{m,k,j}}}\right) &\leq \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} O\left(\sqrt{H^3\iota w''(x, a, h)\|w\|_\infty}\right) \\ &\leq \left(\sqrt{H^3SA\iota(\|w\|_1 + SA\|w\|_\infty)\|w\|_\infty}\right) \\ &= \left(\sqrt{H^3SA\iota\|w\|_1\|w\|_\infty} + SA\|w\|_\infty\sqrt{H^3\iota}\right). \end{aligned}$$

To conclude, for step 3, we have that

$$\sum_{m,k,j}^{K'} w_{mkj} \Omega\left(\sqrt{\frac{H^3\iota}{t_h^{m,k,j}}}\right) \leq O\left(\sqrt{H^3SA\iota\|w\|_1\|w\|_\infty} + MSA\|w\|_\infty\sqrt{H^3\iota}\right).$$

Combining the results for the three different steps, we have

$$\begin{aligned} \sum_{m,k,j} w_{mkj} \tilde{V}_h^{m,k,j} &\leq \sum_{m,k,j} \tilde{w}_{mkj} \tilde{V}_{h+1}^{m,k,j} + O\left(\sqrt{H^3SA\iota\|w\|_1\|w\|_\infty}\right. \\ &\quad \left.+ MSA\|w\|_\infty\sqrt{H^3\iota} + H^3SA(M-1)\|w\|_\infty\right), \end{aligned}$$

with $\|w\|_1 \leq \exp(3/H)\|\tilde{w}\|_1$ and $\|w\|_\infty \leq \|\tilde{w}\|_\infty$. So, by recursions with regard to $h, h+1 \dots H$, we can get the result. \square

Next, we will establish relationships between $W_k(x, a, h)$ and $[\mathbb{V}_h V_{h+1}^*](x, a)$, in which $[\mathbb{V}_h V_{h+1}^*](x, a)$ is a variance operator define below. We also need these definitions for any $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K+1]$ with $t = N_h^{K'}(x, a)$.

$$[\mathbb{P}_h V_{h+1}^*](x, a) = \mathbb{E}[V_{h+1}^*(x_{h+1}) | (x_h, a_h) = (x, a)].$$

$$[\mathbb{V}_h V_{h+1}^*](x, a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot | x, a)} [V_{h+1}^*(x') - [\mathbb{P}_h V_{h+1}^*](x, a)]^2 =: P_1$$

Here, P_1 depends on (x, a, h) and we will use the simplified notation when there is no ambiguity.

$$\frac{1}{t} \sum_{i=1}^t [V_{h+1}^*(x_{h+1}^{(m,k,j)_h(i;x,a)}) - [\mathbb{P}_h V_{h+1}^*](x, a)]^2 =: P_2.$$

$$\frac{1}{t} \sum_{i=1}^t \left[V_{h+1}^*(x_{h+1}^{(m,k,j)_h(i;x,a)}) - \frac{1}{t} \sum_{i'=1}^t V_{h+1}^*(x_{h+1}^{(m,k,j)_h(i';x,a)}) \right]^2 =: P_3$$

$$W_{K'}(x, a, h) = \frac{1}{t} \sum_{i=1}^t \left[V_{h+1}^{k_h(i;x,a)}(x_{h+1}^{(m,k,j)_h(i;x,a)}) - \frac{1}{t} \sum_{i'=1}^t V_{h+1}^{k_h(i';x,a)}(x_{h+1}^{(m,k,j)_h(i';x,a)}) \right]^2 =: P_4.$$

Here, P_2, P_3, P_4 depend on (x, a, h, k) and we use the simplified notations when there is no ambiguity. The following Lemmas establish the closeness of these quantities to illustrate the closeness between $W_{K'}(x, a, h)$ and $[\mathbb{V}_h V_{h+1}^*](x, a)$.

Lemma E.4. For any $p \in (0, 1)$ with probability at least $1 - p$, the following holds simultaneously for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K + 1]$ with $t = N_h^{K'}(x, a)$.

$$|P_1 - P_2| \leq O\left(H^2 \sqrt{\iota/t}\right).$$

Proof. We have that $\left\{ \left[V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) - [\mathbb{P}_h V_{h+1}^*](x, a) \right]^2 - P_1 \right\}_{i=1}^\infty$ is a martingale difference bounded by $O(H^2)$, and the random variable $t \leq T_0/H(1 + \tilde{C})$. By Azuma-Hoeffding Inequality, for any given $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and a given $t' \in \mathbb{N}_+$, for any $p \in (0, 1)$, with probability $1 - p$,

$$\frac{1}{t'} \left| \sum_{i=1}^{t'} \left(\left[V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) - [\mathbb{P}_h V_{h+1}^*](x, a) \right]^2 - P_1 \right) \right| \leq O\left(H^2 \sqrt{\frac{1}{t'} \log \frac{2}{p}}\right).$$

By considering all the possible combinations $(x, a, h, t') \in \mathcal{S} \times \mathcal{A} \times [H] \times [T_0(1 + \tilde{C})/H + M]$, with a union bound and the realization of $t = t'$, we can claim the conclusion. \square

Lemma E.5. For any $p \in (0, 1)$ with least $1 - p$ probability, the following holds simultaneously for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K + 1]$ with $t = N_h^{K'}(x, a)$:

$$|P_2 - P_3| \leq O\left(H^2 \sqrt{\iota/t}\right).$$

Proof. We can find that

$$|P_2 - P_3| \leq O\left(H \left| \frac{1}{t} \sum_{i'=1}^t V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i';x,a)} \right) - [\mathbb{P}_h V_{h+1}^*](x, a) \right| \right).$$

Knowing that $\left\{ V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i';x,a)} \right) - [\mathbb{P}_h V_{h+1}^*](x, a) \right\}_{i'=1}^\infty$ is a martingale difference bounded by $O(H)$, using the same procedure as proof for Lemma E.4, we can claim the result. \square

For $|P_3 - P_4|$, similar to the proof of Lemma C.3 in Jin et al. (2018), we have

$$|P_3 - P_4| \leq O\left(\frac{H}{t} \sum_{i=1}^t \left| V_{h+1}^{k_h(i;x,a)} \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) - V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) \right| \right).$$

We mark an event Equation (41) here, which means that the difference is always non-negative.

$$\begin{aligned} \text{Event}(K') &= \left\{ \sum_{i=1}^t \left| V_{h+1}^{k_h(i;x,a)} \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) - V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) \right| \right. \\ &\quad \left. = \sum_{i=1}^t \left(V_{h+1}^{k_h(i;x,a)} \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) - V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) \right), \forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H] \right\}. \end{aligned} \quad (41)$$

We do not need a new statistical lemma to prove that it holds with high probability. It will be shown to hold automatically based on some other statistical events that hold with high probability later. Under this event, we need to find an upper bound for

$$\frac{1}{t} \sum_{i=1}^t \left(V_{h+1}^{k_h(i;x,a)} \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) - V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) \right).$$

Under the event of Equation (39), based on Lemma E.3, letting $w_{mkj} = \frac{1}{t} \mathbb{I}[(x_h^{m,k,j}, a_h^{m,k,j}) = (x, a)]$, we have that

$$\begin{aligned} &\frac{1}{t} \sum_{i=1}^t \left(V_{h+1}^{k_h(i;x,a)} \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) - V_{h+1}^* \left(x_{h+1}^{(m,k,j)_h(i;x,a)} \right) \right) \\ &\leq O\left(\frac{MSA}{t} \sqrt{H^5 \iota} + \sqrt{\frac{SA}{t} H^5 \iota} + H^4 SA(M-1) \frac{1}{t}\right), \end{aligned}$$

which indicates that, under the intersections of the events of Equation (39), and $\bigcap_{k=1}^{K'} \text{Event}(k)$, for any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we have

$$|P_3 - P_4| \leq O \left(\frac{MSA}{t} \sqrt{H^7 \iota} + \frac{\sqrt{SAH^7 \iota}}{\sqrt{t}} + \frac{(M-1)SAH^5}{t} \right).$$

To conclude about the relationship between $W_k(x, a, h)$ and $[\mathbb{V}_h V_{h+1}^*](x, a)$, we have that, under the intersection of the events of Equation (39), $\bigcap_{k=1}^{K'} \text{Event}(k)$, Lemma E.4 and Lemma E.5, we have $\forall (x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K']$,

$$\begin{aligned} & |W_k(x, a, h) - [\mathbb{V}_h V_{h+1}^*](x, a)| \\ & \leq O \left(\frac{MSA}{t} \sqrt{H^7 \iota} + \frac{\sqrt{SAH^7 \iota}}{\sqrt{t}} + \frac{(M-1)SAH^5}{t} \right). \end{aligned} \quad (42)$$

With this relationship, we can provide the new concentration results. Similar to the proof of Lemma C.3, for a given $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we decompose the summation $\sum_{i=1}^t \tilde{\theta}_t^i X_i$ as follows.

$$\sum_{i=1}^t \tilde{\theta}_t^i X_i = \sum_{i=1}^t \theta_t^i X_i + \sum_{i=1}^t (\tilde{\theta}_t^i - \theta_t^i) X_i.$$

Equation (19) has already provided an upper bound for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ for the second summation. Next, we focus on $\left| \sum_{i=1}^t \theta_t^i X_i \right|$. By Azuma-Bernstein Inequality, for any fixed $t' \in \mathbb{N}_+$ and fixed $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, for any $p \in (0, 1)$, with probability at least $1 - p$, we have that

$$\left| \sum_{i=1}^{t'} \theta_{t'}^i X_i \right| \leq O \left(\sqrt{\frac{1}{t'} H [\mathbb{V}_h V_{h+1}^*](x, a) \log \frac{2}{p}} + \frac{1}{t'} H^2 \log \frac{2}{p} \right).$$

After considering the union bound with regard to $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $t' \leq T_0/H$, we can claim the following conclusion: for any $p \in (0, 1)$, with probability at least $1 - p$, the following relationship holds simultaneously for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$,

$$\left| \sum_{i=1}^t \theta_t^i X_i \right| \leq O \left(\sqrt{\frac{\iota}{t'} H [\mathbb{V}_h V_{h+1}^*](x, a)} + \frac{\iota}{t} H^2 \right), t = N_h^{K'}(x, a). \quad (43)$$

The intersection of events of Equation (43) and Equation (19) indicates that the following relationship holds simultaneously for all $(x, a, h, K') \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ with $t = N_h^{K'}(x, a)$:

$$\left| \sum_{i=1}^t \tilde{\theta}_t^i (\tilde{\mathbb{E}}_{x,a,h,i} - \mathbb{E}_{x,a,h}) V_{h+1}^*(x_{h+1}) \right| \leq O \left(\sqrt{\frac{\iota}{t} H [\mathbb{V}_h V_{h+1}^*](x, a)} + \frac{\iota}{t} H^2 + \sqrt{H \iota / t} \right). \quad (44)$$

Combining with the event of Equation (42) for K' replaced by $K' + 1$, we have that

$$\begin{aligned} & \left| \sum_{i=1}^t \tilde{\theta}_t^i \left((\tilde{\mathbb{E}}_{x,a,h,i} - \mathbb{E}_{x,a,h}) V_{h+1}^*(x_{h+1}) \right) \right| \\ & \leq O \left(\sqrt{\frac{\iota}{t} H \left(W_{K'+1}(x, a, h) + \frac{MSA}{t} \sqrt{H^7 \iota} + \frac{\sqrt{SAH^7 \iota}}{\sqrt{t}} + \frac{(M-1)SAH^5}{t} \right)} \right. \\ & \quad \left. + \frac{\iota}{t} H^2 + \sqrt{H \iota / t} \right). \end{aligned}$$

Due to $2\sqrt{\frac{H^7 SA \iota}{t}} \leq H + \frac{H^6 SA \iota}{t}$, we have that

$$\begin{aligned} & O \left(\sqrt{\frac{\iota}{t} H \left(W_{K'+1}(x, a, h) + \frac{MSA}{t} \sqrt{H^7 \iota} + \frac{\sqrt{SAH^7 \iota}}{\sqrt{t}} + \frac{(M-1)SAH^5}{t} \right)} \right) \\ & \leq O \left(\sqrt{\frac{\iota}{t} H \left(W_{K'+1}(x, a, h) + \frac{MSA}{t} \sqrt{H^7 \iota} + H + \frac{SAH^6 \iota}{t} + \frac{(M-1)SAH^5}{t} \right)} \right) \\ & \leq O \left(\sqrt{\frac{\iota}{t} H \left(W_{K'+1}(x, a, h) + \frac{MSA}{t} \sqrt{H^7 \iota} + H + \frac{SAH^6 \iota}{t} + \frac{(M-1)SAH^5}{t} \right)} \right). \end{aligned}$$

Noticing that

$$\frac{MSA}{t} \sqrt{H^7 \iota} + \frac{(M-1)SAH^5}{t} = O \left(\frac{MSA}{t} H^5 \iota \right),$$

we have

$$\begin{aligned} & \left| \sum_{i=1}^t \tilde{\theta}_t^i \left((\tilde{\mathbb{E}}_{x,a,h,i} - \mathbb{E}_{x,a,h}) V_{h+1}^*(x_{h+1}) + r_h^{(m,k,j)_h(i;x,a)} - r_h(x, a) \right) \right| \\ & \leq O \left(\sqrt{\frac{H \iota}{t} (W_{K'+1}(x, a, h) + H)} + \iota \frac{\sqrt{H^7 SA} + \sqrt{MSAH^6}}{t} \right), \end{aligned} \quad (45)$$

which indicates that

$$\left| \sum_{i=1}^t \tilde{\theta}_t^i \left((\tilde{\mathbb{E}}_{x,a,h,i} - \mathbb{E}_{x,a,h}) V_{h+1}^*(x_{h+1}) \right) \right| \leq \beta_t(x, a, h)/2$$

when combining with Equation (39) and c' is large enough.

Finally, we are ready to provide proof for Lemma E.1. We let c' to be large enough and Will provide discussion under the intersections of events for Equation (19), Equation (39), Lemma E.4, Lemma E.5 and Equation (43). We know that these events hold simultaneously with probability $1 - c_p p$ for some $c_p > 0$. Next, we will prove Equation (38) by induction. It obviously holds that for all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ when $K' = 1$. We suppose that it holds for every $K' \leq K'_0$. When $K' = K'_0 + 1$, LHS of Equation (38) indicates that $\bigcap_{k=1}^{K'_0+1} \text{Event}(k)$ holds. By the discussion above, this indicates that Equation (45) holds for $K'_0 + 1$, by recursions on $H, H-1, \dots, 1$ (similar to the proof of Lemma 4.3 in Jin et al. (2018)), we can prove that Equation (38) holds for all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ for $K'_0 + 1$. This finishes the induction. After we replace p with p/c_p , we finish the proof.

E.2.2 REMAINING PARTS FOR PROVING THEOREM 5.1

Next, we begin to discuss the overall complexity. Similar to Lemma C.5 in Jin et al. (2018), we will provide the following Lemma.

Lemma E.6. *For any $p \in (0, 1)$, with probability at least $1 - p$,*

$$\sum_{m,k,j} \left[\mathbb{V}_h V_{h+1}^k \right] (x_h^{m,k,j}, a_h^{m,k,j}) \leq O(H\hat{T} + H^3 \iota).$$

Proof. We assign an order for all the episodes based on the “round first, episode second, agent third” rule and suppose $m(i), k(i), j(i)$ recovers the agent index, round index and within round episode index for the i -th episode. Denote $R_i = \sum_{h=1}^H \mathbb{V}_h V_{h+1}^{\pi^k} (x_h^{(m,k,j)(i)}, a_h^{(m,k,j)(i)})$ and \mathcal{F}_{i-1} be the σ -field generated by the information before the i -th episode. Similar to the proof of Lemma C.5 in Jin et al. (2018), we have

$$\begin{aligned} \mathbb{E}[R_i | \mathcal{F}_{i-1}] & \leq H^2, \\ 0 & \leq R_i \leq H^3, \end{aligned}$$

$$\text{Var}[R_i|F_{i-1}] \leq H^5.$$

So, by Azuma-Hoeffding Inequality based on $\sum_{i=1}^t R_i$ with regard to the filtration $\{\mathcal{F}_i\}_{i=1}^\infty$ and a union bound for $t \leq T_0(1 + \tilde{C})/H + M$, we conclude that

$$\sum_{m,k,j} \left[\mathbb{V}_h V_{h+1}^{\pi^k} \right] (x_h^{m,k,j}, a_h^{m,k,j}) = \sum_{i=1}^{\hat{T}/H} R_i \leq O(H\hat{T} + H^3\iota).$$

□

We also provide a Lemma that focuses on the concentration of ξ_h^k .

Lemma E.7. *For any $p \in (0, 1)$, with probability at least $1 - p$, the following relationships holds simultaneously:*

$$\left| \sum_{k=1}^K C_h \sum_{h=h'}^H \xi_{h+1}^k \right| \leq O(H\sqrt{\hat{T}\iota}), \forall h' \in [H], \quad (46)$$

$$\left| \sum_{k=1}^K \sum_{h=1}^H \xi_{h+1}^k \right| \leq O(H\sqrt{\hat{T}\iota}), \quad (47)$$

in which $C_h = \exp(3(h-1)/H)$.

Proof. We first focus on the first event. Denote $V(m, k, j, h) = C_h(\mathbb{P} - \hat{\mathbb{P}})(V_{h+1}^{\pi^k} - V_{h+1}^{\pi^k})(x_h^{m,k,j}, a_h^{m,k,j})$ and a simplified notation $\sum_{m,k,j,h:h'} = \sum_{k=1}^K \sum_{m=1}^M \sum_{j=1}^{n^{m,k}} \sum_{h=h'}^{H-1}$. The quantity we focus on can be rewritten as

$$\sum_{m,k,j,h:h'} V(m, k, j, h),$$

with $|V(m, k, j, h)| \leq O(H)$ as $C_h \leq \exp(3)$. Let $\tilde{V}(\tilde{i})$ be the \tilde{i} -th term in the summation that contains $\hat{T}(H - h')/H$ terms, in which the order follows a “round first, episode second, step third, agent fourth” rule. Then the sequence $\{\tilde{V}(\tilde{i})\}$ is a martingale difference. By Azuma-Hoeffding Inequality, for any $p \in (0, 1)$ and $t \in N_+$, with probability at least $1 - p$,

$$\left| \sum_{\tilde{i}=1}^t \tilde{V}(\tilde{i}) \right| \leq O(H\sqrt{t}).$$

Then by applying a union bound with regard to $h' \in [H-1]$ and all possible t which is divisible by $H - h'$ and knowing that $\hat{T}(H - h')/H \leq T_0(1 + \tilde{C}) + HM$ due to (e) in Lemma B.1, we can claim that, for any $p \in (0, 1)$, with probability at least $1 - p$, the following relationship holds simultaneously:

$$\left| \sum_{k=1}^K C_h \sum_{h=h'}^H \xi_{h+1}^k \right| = \left| \sum_{\tilde{i}=1}^{\hat{T}(H-h')/H} \tilde{V}(\tilde{i}) \right| \leq O(H\sqrt{\hat{T}\iota}), \forall h' \in [H].$$

The second event can be analyzed similarly for the same conclusion. By combining these two events and re-scaling p , we can claim the result. □

Next, we try to find the upper bound for the regret. We pick c' to be large enough and discuss based on the intersection of events of Equation (19), Equation (39), Lemma E.4, Lemma E.5, Equation (43), Lemma E.7, Lemma E.1 and Lemma E.6. They hold simultaneously with probability at least $1 - c'_p p$ where $c'_p > 0$ is a numerical constant and indicates Equation (42). Similar to the discussions in Proof of Theorem 4.1, we can claim that for $\forall h \in [H]$,

$$\sum_{k=1}^K \delta_h^k \leq O\left(\sqrt{H^4 \iota \hat{T} S A} + H S A (M-1) \sqrt{H^3 \iota} + M H^2 S A + H^4 S A (M-1)\right), \quad (48)$$

due to $\beta_t(x, a, h) = O(\sqrt{H^3 \iota / t})$. In addition, due to the relationship

$$\begin{aligned} \sum_{k=1}^K \delta_h^k &\leq \exp(3/H) \sum_{k=1}^K \delta_{h+1}^k + \sum_{k=1}^K \xi_{h+1}^k + O(1) \sum_{k,m,j} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \\ &\quad + O(MHSA + H^3 SA(M-1)), \end{aligned}$$

which can be obtained similar to the situation in Appendix C.3 for Proof of Theorem 4.1, denoting $\sum_{k,m,j,h} = \sum_{k,m,j} \sum_{h=1}^H$, we have

$$\sum_{k=1}^K \delta_1^k \leq O(MH^2 SA + H^4 SA(M-1) + \sqrt{H^2 \hat{T} \iota}) + O(1) \sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \quad (49)$$

We will bound the last term by splitting it into two parts.

$$\begin{aligned} \sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) &= \sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \mathbb{I}[t_h^{m,k,j} \leq M-1] \\ &\quad + \sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \mathbb{I}[t_h^{m,k,j} \geq M]. \end{aligned}$$

For the first part, knowing that $\beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \leq O(\sqrt{H^3 \iota})$, using the similar technique as Proof of Equation (26), we have that

$$\sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \mathbb{I}[t_h^{m,k,j} \leq M-1] \leq O(HSA(M-1)\sqrt{H^3 \iota}).$$

For the second part, we have that

$$\begin{aligned} &\sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \mathbb{I}[t_h^{m,k,j} \geq M] \\ &\leq \sum_{k,m,j,h} O\left(\sqrt{\frac{H\iota}{t_h^{m,k,j}}} (W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h) + H) + \iota \frac{\sqrt{H^7 SA} + \sqrt{MSAH^6}}{t_h^{m,k,j}}\right) \\ &\quad \cdot \mathbb{I}[t_h^{m,k,j} \geq M]. \end{aligned}$$

Later on, we use another simplified notation $\sum_{k,m,j,h:M} = \sum_{k,m,j,h} \mathbb{I}[t_h^{m,k,j} \geq M]$. Using the same technique of finding C'' in Equation (26), we can find that

$$\sum_{k,m,j,h:M} 1/t_h^{m,k,j} \leq O(1) \sum_{(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]} \sum_{i=M}^{N_h^{K+1}(x,a)-1} 1/i \leq HSA\iota \quad (50)$$

and

$$\sum_{k,m,j,h:M} 1/\sqrt{t_h^{m,k,j}} \leq O(1) \sum_{(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]} \sum_{i=M}^{N_h^{K+1}(x,a)-1} 1/i \leq \sqrt{HSA\hat{T}}. \quad (51)$$

So, we have

$$\sum_{k,m,j,h:M} \iota \frac{\sqrt{H^7 SA} + \sqrt{MSAH^6}}{t_h^{m,k,j}} \leq \iota^2 HSA (\sqrt{H^7 SA} + \sqrt{MSAH^6}).$$

We also have

$$\begin{aligned} &\sum_{k,m,j,h:M} \left(\sqrt{\frac{H}{t_h^{m,k,j}}} (W_k(x_h^{m,k,j}, a_h^{m,k,j}, h) + H) \right) \\ &\leq O(1) \sqrt{\left(\sum_{m,k,j,h:M} (W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h) + H) \right) \left(\sum_{m,k,j,h:M} \frac{H}{t_h^{m,k,j}} \right)} \\ &\leq O(1) \sqrt{H^3 SA\hat{T}\iota} + O(1) \sqrt{H^2 SA\iota} \sqrt{\sum_{m,k,j,h:M} W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h)}, \end{aligned}$$

where the first inequality follows from Cauchy's inequality and the second inequality is due to Equation (50).

To conclude, we have

$$\begin{aligned} & \sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \\ &= O \left(HSA(M-1)\sqrt{H^3\iota} + \iota^2\sqrt{H^9S^3A^3} + \iota^2\sqrt{MS^3A^3H^8} + \right. \\ & \quad \left. + \sqrt{H^3SA\hat{T}\iota^2} + \sqrt{H^2SA\iota^2} \sqrt{\sum_{m,k,j,h:M} W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h)} \right). \end{aligned} \quad (52)$$

Next, we try to find an upper bound for

$$\sqrt{\sum_{m,k,j,h:M} W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h)}.$$

We know that

$$\begin{aligned} W_k(x, a, h) &\leq \mathbb{V}_h \left[V_{h+1}^{\pi^k} \right] (x, a) + \left| \left[\mathbb{V}_h V_{h+1}^* \right] (x, a) - W_k(x, a, h) \right| \\ &\quad + \left| \left[\mathbb{V}_h V_{h+1}^* \right] (x, a) - \left[\mathbb{V}_h V_{h+1}^{\pi^k} \right] (x, a) \right|. \end{aligned}$$

By Lemma E.6,

$$\sqrt{\sum_{m,k,j,h:M} \mathbb{V}_h \left[V_{h+1}^{\pi^k} \right] (x_h^{m,k,j}, a_h^{m,k,j})} \leq O \left(\sqrt{H\hat{T} + H^3\iota} \right).$$

By Equation (42), denoting $\tilde{t}_h^{m,k,j} = N_h^{k+1}(x_h^{m,k,j}, a_h^{m,k,j})$,

$$\begin{aligned} & \sqrt{\sum_{m,k,j,h:M} \left| \left[\mathbb{V}_h V_{h+1}^* \right] (x_h^{m,k,j}, a_h^{m,k,j}) - W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h) \right|} \\ & \leq O \left(\sqrt{\sum_{m,k,j,h:M} \left(\frac{MSA}{\tilde{t}_h^{m,k,j}} \sqrt{H^7\iota} + \frac{\sqrt{SAH^7\iota}}{\sqrt{\tilde{t}_h^{m,k,j}}} + \frac{(M-1)SAH^5}{\tilde{t}_h^{m,k,j}} \right)} \right). \end{aligned}$$

As $\tilde{t}_h^{m,k,j} \geq t_h^{m,j,k}$, we have that

$$\begin{aligned} & \sqrt{\sum_{m,k,j,h:M} \left| \left[\mathbb{V}_h V_{h+1}^* \right] (x_h^{m,k,j}, a_h^{m,k,j}) - W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h) \right|} \\ & \leq O \left(\sqrt{\sum_{m,k,j,h:M} \left(\frac{MSA}{t_h^{m,k,j}} \sqrt{H^7\iota} + \frac{\sqrt{SAH^7\iota}}{\sqrt{t_h^{m,k,j}}} + \frac{(M-1)SAH^5}{t_h^{m,k,j}} \right)} \right) \\ & = O \left(\sqrt{MH^{4.5}S^2A^2\iota^{1.5} + H^4SA\sqrt{\hat{T}\iota} + (M-1)H^6S^2A^2} \right), \end{aligned}$$

where the last inequality is due to Equation (50) and Equation (51). We also have

$$\begin{aligned} & \sqrt{\sum_{m,k,j,h:M} \left| \left[\mathbb{V}_h V_{h+1}^* \right] (x_h^{m,k,j}, a_h^{m,k,j}) - \left[\mathbb{V}_h V_{h+1}^{\pi^k} \right] (x_h^{m,k,j}, a_h^{m,k,j}) \right|} \\ & \leq \sqrt{\sum_{m,k,j,h} \left| \left[\mathbb{V}_h V_{h+1}^* \right] (x_h^{m,k,j}, a_h^{m,k,j}) - \left[\mathbb{V}_h V_{h+1}^{\pi^k} \right] (x_h^{m,k,j}, a_h^{m,k,j}) \right|}. \end{aligned}$$

Next, we will show that, for any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$\left| [\mathbb{V}_h V_{h+1}^*](x, a) - [\mathbb{V}_h V_{h+1}^{\pi^k}](x, a) \right| \leq O(H) \left([\mathbb{P}_h V_{h+1}^*](x, a) - [\mathbb{P}_h V_{h+1}^{\pi^k}](x, a) \right).$$

Suppose that u, v are random variables such that u follows the distribution of $V_{h+1}^*(x_{h+1})$ under π_* when $(x_h, a_h) = (x, a)$, and v follows the distribution of $V_{h+1}^{\pi^k}(x_{h+1})$ under π^k when $(x_h, a_h) = (x, a)$ and $u \geq v$. The third requirement is reasonable because the distribution of x_{h+1} only depends on (x, a) and $V_{h+1}^*(x_{h+1}) \geq V_{h+1}^{\pi^k}(x_{h+1})$. We have that $u, v \leq H$. So,

$$\begin{aligned} \left| [\mathbb{V}_h V_{h+1}^*](x, a) - [\mathbb{V}_h V_{h+1}^{\pi^k}](x, a) \right| &= |\text{Var}(u) - \text{Var}(v)| \\ &\leq |\mathbb{E}(u^2) - \mathbb{E}(v^2) + (\mathbb{E}v)^2 - (\mathbb{E}u)^2| \\ &\leq |\mathbb{E}(u - v)(u + v) + (\mathbb{E}v - \mathbb{E}u)(\mathbb{E}v + \mathbb{E}u)| \\ &\leq O(H)(|\mathbb{E}(u - v)| + \mathbb{E}|u - v|) \\ &= O(H)\mathbb{E}(u - v). \end{aligned}$$

This proves the conclusion. Using the conclusion, we can find that

$$\begin{aligned} &\sum_{m,k,j,h} \left| [\mathbb{V}_h V_{h+1}^*](x_h^{m,k,j}, a_h^{m,k,j}) - [\mathbb{V}_h V_{h+1}^{\pi^k}](x_h^{m,k,j}, a_h^{m,k,j}) \right| \\ &\leq O(H) \sum_{m,k,j,h} \left([\mathbb{P}_h V_{h+1}^*](x_h^{m,k,j}, a_h^{m,k,j}) - [\mathbb{P}_h V_{h+1}^{\pi^k}](x_h^{m,k,j}, a_h^{m,k,j}) \right) \\ &= O(H) \sum_{h=1}^H \sum_{k=1}^K (\delta_{h+1}^k - \phi_{h+1}^k + \xi_{h+1}^k) \\ &\leq O(H) \sum_{h=1}^H \sum_{k=1}^K (\delta_{h+1}^k + \xi_{h+1}^k), \end{aligned}$$

in which the last inequality is due to $\phi_h^k \geq 0$ based on Equation (38). By Equation (48) and Lemma E.7, we have

$$O(H) \sum_{h=1}^H \sum_{k=1}^K (\delta_{h+1}^k + \xi_{h+1}^k) = O\left(\sqrt{H^8 \hat{T} S A} + H^3 S A (M-1) \sqrt{H^3 \iota} + M H^4 S A + H^6 S A (M-1)\right).$$

So, we have

$$\begin{aligned} &\sum_{m,k,j,h:M} W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h) \\ &\leq O\left(H \hat{T} + H^3 \iota + M H^4 S A + \sqrt{H^8 \hat{T} S A \iota} + H^6 S A (M-1) + H^2 S A (M-1) \sqrt{H^5 \iota}\right) \\ &\quad + O\left(M S^2 A^2 \sqrt{H^9 \iota^3} + S A \sqrt{H^8 \hat{T} \iota} + S^2 A^2 H^6 (M-1)\right) \\ &= O\left(H \hat{T} + M H^{4.5} S^2 A^2 \iota^{1.5} + H^4 S A \sqrt{\hat{T} \iota} + H^6 S^2 A^2 (M-1)\right), \end{aligned}$$

where the last relationship is due to

$$(M-1)H^6 S^2 A^2 \geq (M-1)H^6 S A, H^4 S A \sqrt{\hat{T} \iota} \geq \sqrt{H^8 \hat{T} S A \iota}$$

and

$$M S^2 A^2 \sqrt{H^9 \iota^3} \geq M H^4 S A + H^3 \iota + H^2 S A (M-1) \sqrt{H^5 \iota}.$$

Inserting it into Equation (52), we have

$$\begin{aligned}
& \sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \\
&= O \left(HSA(M-1)\sqrt{H^3\iota} + \iota^2\sqrt{H^9S^3A^3} + \iota^2\sqrt{MS^3A^3H^8} + \right. \\
&\quad \left. + \sqrt{H^3SA\hat{T}\iota^2} + \sqrt{H^2SA\iota^2} \sqrt{\sum_{m,k,j,h:M} W_{k+1}(x_h^{m,k,j}, a_h^{m,k,j}, h)} \right) \\
&= O \left(HSA(M-1)\sqrt{H^3\iota} + \iota^2\sqrt{H^9S^3A^3} + \iota^2\sqrt{MS^3A^3H^8} + \sqrt{H^3SA\hat{T}\iota^2} \right. \\
&\quad \left. + \sqrt{MH^{6.5}S^3A^3\iota^{3.5}} + \sqrt{H^6S^2A^2\hat{T}^{0.5}\iota^{2.5}} + \sqrt{H^8S^3A^3(M-1)\iota^2} \right).
\end{aligned}$$

Due to

$$\begin{aligned}
\sqrt{MH^8S^3A^3\iota^4} &\geq \sqrt{MH^{6.5}S^3A^3\iota^{3.5}}, \\
\sqrt{H^8S^3A^3(M-1)\iota^2} &\leq \iota^2\sqrt{MS^3A^3H^8}
\end{aligned}$$

and

$$\sqrt{H^6S^2A^2\hat{T}^{0.5}\iota^{2.5}} \leq H^{4.5}S^{1.5}A^{1.5}\iota^{1.5} + \sqrt{\hat{T}SAH^3\iota^2} \leq H^{4.5}S^{1.5}A^{1.5}\iota^2 + \sqrt{\hat{T}SAH^3\iota^2},$$

we have

$$\begin{aligned}
& \sum_{k,m,j,h} \beta_{t_h^{m,k,j}}(x_h^{m,k,j}, a_h^{m,k,j}, h) \\
&\leq O \left(HSA(M-1)\sqrt{H^3\iota} + \iota^2\sqrt{H^9S^3A^3} + \iota^2\sqrt{MS^3A^3H^8} + \sqrt{H^3SA\hat{T}\iota^2} \right).
\end{aligned}$$

Inserting it into Equation (49), we have

$$\begin{aligned}
\text{Regret}(T) &\leq \sum_{k=1}^K \delta_1^k \\
&= O \left(MH^2SA + H^4SA(M-1) + HSA(M-1)\sqrt{H^3\iota} \right. \\
&\quad \left. + \iota^2\sqrt{H^9S^3A^3} + \iota^2\sqrt{MS^3A^3H^8} + \sqrt{H^3SA\hat{T}\iota^2} \right).
\end{aligned}$$

Finally, for the probability of the intersection of all the events, if we use p/c'_p to replace p , we complete the proof.

F NUMERICAL EXPERIMENTS

In this section, we conduct experiments in a synthetic environment to validate the theoretical performances of FedQ-Hoeffding, FedQ-Beinstein, and compare with their single-user counterparts UCB-H and UCB-B (Jin et al., 2018), respectively.

Synthetic Environment. We generate a synthetic environment to evaluate the proposed algorithms. We set the number of states S to be 3, the number of actions A for each state to be 2, and the episode length H to be 5. The reward $r_h(s, a)$ for each state-action pair and each step is generated independently and uniformly at random from $[0, 1]$. We also generate the transition kernel $P_h(\cdot | s, a)$ from an S -dimensional simplex independently and uniformly at random for each state-action pair and each step. Such procedure guarantees that the synthetic environment is a proper tabular MDP.

Under the given MDP, we set $M = 10$ and $T/H = 3 \times 10^4$ for FedQ-Hoeffding, FedQ-Beinstein, and $T/H = 3 \times 10^5$, $M = 1$ for UCB-H and UCB-B. Thus, the total number of episodes is 3×10^5 for all four algorithms. We choose $c = \iota = 1$ for all algorithms. For each episode, we randomly choose the initial state uniformly from S states. We collect 10 sample paths under all algorithms under the same MDP environment, and plot $\text{Regret}(T)/\sqrt{MT}$ versus MT/H in Figure 1. The

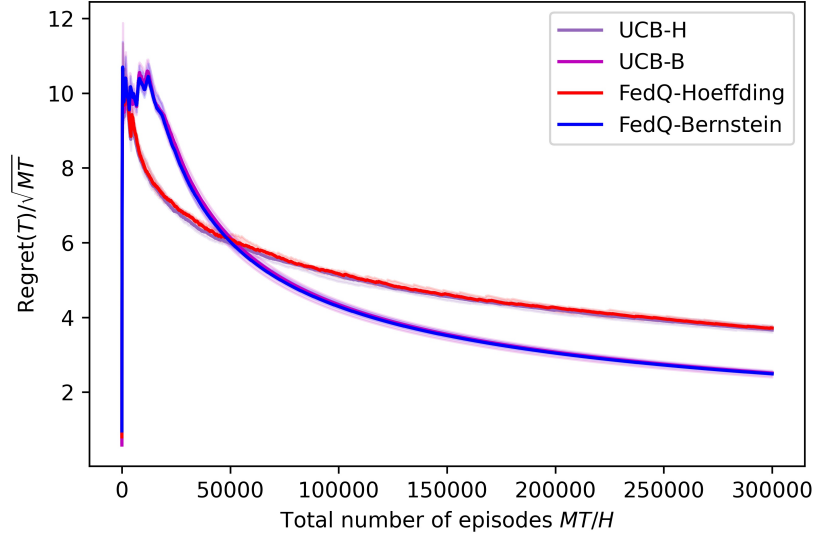


Figure 1: Regret comparison.

solid line represents the median of the 10 sample paths, while the shaded area shows the 10th and 90th percentiles. As we can see, both FedQ-Hoeffding and FedQ-Bernstein stay very close to their single-agent counterpart, indicating that FedQ-Hoeffding achieves linear speedup with respect to the number of clients M , as predicted by Theorem 4.1 and Theorem 5.1. Besides, as time progresses, FedQ-Bernstein achieves lower regret than FedQ-Hoeffding, which is consistent with the theoretical results as well.

We also track the number of communication rounds throughout the learning process under FedQ-Hoeffding and FedQ-Bernstein, and plot the median profiles as well as the 10th and 90th percentiles in Figure 2. Both curves exhibit sublinear growth, corroborating the theoretical result in Theorem 4.2. Besides, the total number of communication rounds under FedQ-Bernstein becomes lower than that under FedQ-Hoeffding as T becomes sufficiently large. This is because after the more active early-stage exploration of FedQ-Bernstein, it reaches a more stable policy, under which the synchronization triggered by (x, a, h) s that are less likely to be visited under the optimal policy rarely happens.

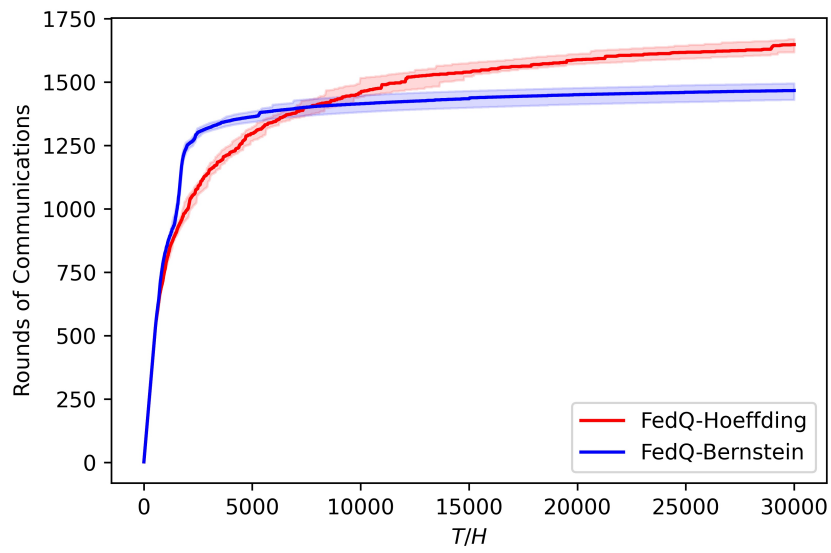


Figure 2: Total number of communication rounds as a function of T/H .