VARIANCE-AWARE FINE-GRAINED GAP-DEPENDENT BOUNDS FOR ONLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study model-free online reinforcement learning (RL) for episodic tabular Markov decision processes, focusing on both gap-dependent regret and policy switching cost. While fine-grained gap-dependent analysis has been established for model-free RL algorithms using Hoeffding-type exploration bonuses, such results for model-free algorithms with variance-based exploration bonuses remain unknown, despite their superior worst-case and coarse-grained gap-dependent guarantees. In this paper, we resolve this open problem by establishing the first fine-grained gap-dependent regret upper bound for UCB-Bernstein+, a refined UCB-Bernstein algorithm, in variance-aware model-free online RL. Moreover, by integrating a stage-wise policy update design into our fine-grained framework and using refined variance-based bonuses, we achieve the best-known gap-dependent local switching cost to date. In addition, our analysis yields improved worst-case guarantees for both regret and local switching cost over the original UCB-Bernstein algorithm. Numerical experiments further demonstrate that UCB-Bernstein+ achieves favorable empirical performance in both regret and local switching cost.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 2018) provides a framework for sequential decision-making, where an agent learns to maximize cumulative rewards through interaction with the environment. RL algorithms are commonly divided into model-based and model-free methods. Model-free methods directly learn value functions and policies, and are widely used due to their simple implementation and low memory requirements (Jin et al., 2018). In contrast, model-based methods learn an empirical model of the transition dynamics and rewards, often incurring higher memory costs.

In this paper, we focus on model-free RL for episodic tabular Markov decision processes (MDPs) with inhomogeneous transition kernels. Specifically, we consider an episodic tabular MDP with S states, A actions, and H steps per episode. For this setting, the minimax regret lower bound over K episodes is $\Omega(\sqrt{H^2SAT})$, where $T = KH$ is the total number of steps (Jin et al., 2018).

Many model-free RL algorithms achieve worst-case $\sqrt{T}$-type regret bounds (Jin et al., 2018; Zhang et al., 2020; Li et al., 2023; Zhang et al., 2025b), among which Zhang et al. (2020); Li et al. (2023); Zhang et al. (2025b) attain the nearly minimax-optimal rate $\tilde{O}(\sqrt{H^2SAT})$. However, in practice, RL algorithms often perform better than their worst-case guarantees when a positive suboptimality gap exists, meaning that the optimal actions at each state-step pair are separated from suboptimal ones by a nontrivial margin. This observation has motivated the study of gap-dependent analysis.

In the model-free setting, Yang et al. (2021) established the first *coarse-grained* gap-dependent regret for UCB-Hoeffding (Jin et al., 2018), which uses Hoeffding-type exploration bonuses, showing a regret of $\tilde{O}\big(H^6SA/\Delta_{\min}\big)$, where $\Delta_{\min}$ denotes the minimum positive suboptimality gap across all state-action-step triples (s, a, h). Here, "coarse-grained" means that the bound depends only on the global minimum gap $\Delta_{\min}$ and does not account for individual suboptimality gaps $\Delta_h(s, a)$.

Beyond Hoeffding-type bonuses, variance-based exploration bonuses have been shown to improve regret performance in both worst-case studies and coarse-grained gap-dependent studies. In particular, several algorithms that leverage variance-based bonuses, including UCB-Bernstein (Jin et al., 2018), UCB-Advantage (Zhang et al., 2020), Q-EarlySettled-Advantage (Li et al., 2023), and Q-EarlySettled-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

LowCost (Zhang et al., 2025b), achieve better worst-case bounds or coarse-grained gap-dependent bounds (Zheng et al., 2025b; Zhang et al., 2025b) than UCB-Hoeffding (Jin et al., 2018; Yang et al., 2021), especially with the improved dependence on the horizon length H.

Meanwhile, Xu et al. (2021) propose the AMB algorithm, which still employs Hoeffding-type exploration bonuses but provides the first *fine-grained* gap-dependent regret bound. Here, "fine-grained" indicates that the bound incorporates the individual suboptimality gaps $\Delta_h(s,a)$ of each state-action-step triple, rather than relying solely on the global minimum gap $\Delta_{\min}$. Subsequently, Zhang et al. (2026a) reanalyze the UCB-Hoeffding algorithm and establish the first fine-grained gap-dependent regret bound for a UCB-based method. The fine-grained gap-dependent bounds for both algorithms take the form:

$$\tilde{O}\left(\left(\sum_{h=1}^{H}\sum_{\Delta_h(s,a)>0}\frac{1}{\Delta_h(s,a)}+\frac{|Z|}{\Delta_{\min}}+SA\right)\mathrm{poly}(H)\right),$$

where, under non-degenerate MDPs, $0 \le |Z| < HSA$ denotes the number of optimal state-action-step triples whose states have multiple optimal actions in Xu et al. (2021) and the number of optimal triples in Zhang et al. (2026a). While this bound significantly reduces reliance on the large factor $1/\Delta_{\min}$ compared to prior coarse-grained gap-dependent analyses, it suffers from the high polynomial dependence on the horizon due to the Hoeffding-type bonuses, similar to Yang et al. (2021).

These observations naturally lead to the following question:

Is it possible to establish improved fine-grained gap-dependent regret bounds for Q-learning algorithms that employ variance-based exploration bonuses?

Answering this question is challenging. Existing techniques for deriving fine-grained regret bounds (Xu et al., 2021; Zhang et al., 2026a) relate the regret to a weighted sum of exploration bonuses, which rely on the simple structure of Hoeffding-type bonuses: the bonus for each state-action-step triple depends only on its visit count. However, this property no longer holds for variance-based bonuses, which additionally depend on triple-specific variance estimates. As a result, no fine-grained regret bounds exist for model-free algorithms with variance-based bonuses, and it remains open whether UCB-Bernstein (Jin et al., 2018) admits even a coarse-grained gap-dependent regret bound.

Besides regret, policy switching cost is also an important evaluation criterion in online RL, particularly in applications where frequent policy changes are costly or restricted, such as hardware placement (Mirhoseini et al., 2017), and material discovery (Nguyen et al., 2019). Two types of switching costs are typically considered in the literature: local and global switching cost. Local switching cost counts the total number of policy changes across all state-step pairs between episodes, whereas global switching cost counts the total number of episodes in which the policy changes, regardless of the number of affected state–step pairs in a single switch.

In model-free RL, under worst-case MDPs, Bai et al. (2019) modify both UCB-Hoeffding and UCB-Bernstein of Jin et al. (2018) to achieve a local switching cost of $O(H^3SA\log T)$, while UCB-Advantage (Zhang et al., 2020) further reduces this cost to $O(H^2SA\log T)$. Both approaches improve upon the $O(T)$ local switching cost of standard Q-learning algorithms (e.g., Jin et al. (2018); Li et al. (2023)). For gap-dependent analyses, Zheng et al. (2025b) eliminate the dependence on the action space dimension A from the $\log T$ term when each state-step pair has a unique optimal action. All of these works also provide the same bound for the global switching cost. More recently, Zhang et al. (2025a;b) further improve these results by removing the dependence on the state dimension S in the $\log T$ term, resulting in the $H^2\log T$ dependence for the global switching cost.

In this paper, we resolve the open question above by introducing UCB-Bernstein+, a refined version of the UCB-Bernstein algorithm (Jin et al., 2018). We establish the **first fine-grained gap-dependent regret** upper bound for model-free RL using variance-based exploration bonuses, and the **best-known gap-dependent local switching cost**. Our main contributions are summarized as follows:

Variance-Aware Fine-Grained Gap-Dependent Regret (Theorem 3.1). We develop a novel generalization of the analytical framework in Zhang et al. (2026a), which is the first in the model-free RL literature to overcome the non-trivial technical challenges of integrating variance-based exploration bonuses into fine-grained regret analysis. By successfully incorporating these bonuses, our framework can yield a sharper dependence on the horizon length H compared to the Hoeffding-style analysis in Zhang et al. (2026a).

State-of-the-Art Gap-Dependent Local Switching Cost (Theorem 3.2). Integrating our fine-grained gap-dependent framework with variance-based bonuses, we achieve more precise control over the total number of visits to suboptimal state-action-step triples (Lemma I.1). This leads to the best-known gap-dependent local switching cost to date, featuring improved dependence on both the horizon length H and the state space dimension S.

Improved Worst-Case Guarantees. Our UCB-Bernstein+ introduces two refinements over the original UCB-Bernstein (Jin et al., 2018), targeting policy stability and exploration efficiency. First, we incorporate the stage design of Zhang et al. (2020) to enable lazy policy updates. This allows us not only to establish the best-known gap-dependent switching cost, but also to match the state-of-the-art worst-case local switching cost of $O(H^2SA\log T)$ (Theorem 3.4) for model-free RL (Zhang et al., 2020). Second, we employ refined variance-based bonuses that remove polynomial dependence on the state and action space dimensions, S and A. Crucially, although lazy updates may increase regret, our refined bonuses offset this overhead, ultimately improving the worst-case regret over the original UCB-Bernstein by removing its superlinear dependence on S and A (see Theorem 3.3).

Technical Novelty. Our analysis introduces three primary technical innovations:

(a) To establish the variance-aware fine-grained regret by controlling the variance-based bonuses, we introduce **pseudo value functions**, a key conceptual technical tool that does not introduce hyperparameters in the design of UCB-Bernstein+. This construction helps us approximate the non-uniform bonus that involves variance estimators with the conditional variances of the optimal value function, which reflect the variance structure of the MDP. It also plays a key role in the worst-case regret analysis. See Section 4 for the details.

(b) To analyze the gap-dependent local switching cost, our fine-grained framework yields a variance-aware bound on the total number of visits to suboptimal state-action-step triples (Lemma I.1). Since Q-values are updated only at stage ends, this visit bound also controls the number of policy switches. Relaxing its variance-dependent terms yields the guarantee stated in Theorem 3.2.

(c) In a significant refinement of the original UCB-Bernstein (Jin et al., 2018), we redesign the variance-based bonus through a refined concentration analysis to remove its explicit polynomial dependence on the state and action space dimensions, apart from logarithmic factors (see Lemma G.1). This refinement is a key ingredient in achieving the improved worst-case regret in Theorem 3.3.

2 PRELIMINARIES

For any $C \in \mathbb{N}$, we use $[C]$ to denote the set $\{1, 2, \ldots, C\}$. We use $\mathbb{I}[x]$ to denote the indicator function, which equals 1 if the event x holds and 0 otherwise. Throughout the paper, all $\arg\max$ operations use a fixed deterministic tie-breaking rule.

Tabular Episodic Markov Decision Process (MDP). A tabular episodic MDP is denoted as $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the set of states with $|\mathcal{S}| = S$, $\mathcal{A}$ is the set of actions with $|\mathcal{A}| = A$, H is the number of steps in each episode, $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ is the transition kernel so that $\mathbb{P}_h(\cdot \mid s, a)$ characterizes the distribution over the next state given the state-action pair (s, a) at step h, and $r := \{r_h\}_{h=1}^H$ is the collection of reward functions. We assume that $r_h(s, a) \in [0, 1]$ is a deterministic function of (s, a).

In each episode, an initial state s_1 is selected arbitrarily by an adversary. Then, at each step $h \in [H]$, an agent observes a state $s_h \in \mathcal{S}$, picks an action $a_h \in \mathcal{A}$, receives the reward $r_h = r_h(s_h, a_h)$ and then transitions to the next state s_{h+1}. The episode terminates after step H upon transitioning to the terminal state s_{H+1}.

For convenience, we denote $\mathbb{P}_{s,a,h} f = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} f(s')$, $\mathbb{1}_s f = f(s)$ and $\mathbb{V}_{s,a,h}(f) = \mathbb{P}_{s,a,h} f^2 - (\mathbb{P}_{s,a,h} f)^2$ for any function $f : \mathcal{S} \to \mathbb{R}$ and state-action-step triple (s, a, h).

Policies and Value Functions. A policy π is a collection of H functions $\{\pi_h : \mathcal{S} \to \Delta^{\mathcal{A}}\}_{h \in [H]}$, where $\Delta^{\mathcal{A}}$ is the set of probability distributions over $\mathcal{A}$. A policy is deterministic if for any $s \in \mathcal{S}$, $\pi_h(s)$ concentrates all the probability mass on an action $a \in \mathcal{A}$. In this case, we denote $\pi_h(s) = a$.

Denote state value functions by $V_h^\pi(s) := \sum_{h'=h}^H \mathbb{E}_{(s_{h'}, a_{h'}) \sim (\mathbb{P}, \pi)} [r_{h'}(s_{h'}, a_{h'}) \mid s_h = s]$ and action value functions by $Q_h^\pi(s, a) := r_h(s, a) + \sum_{h'=h+1}^H \mathbb{E}_{(s_{h'}, a_{h'}) \sim (\mathbb{P}, \pi)} [r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a]$.

Since the action space and the horizon are finite, there exists an optimal policy $\pi^\star$ that achieves the optimal value function $V_h^\star(s) = \sup_\pi V_h^\pi(s) = V_h^{\pi^\star}(s)$ for all $(s,h) \in \mathcal{S} \times [H]$ (Azar et al., 2017). Then for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the Bellman equation can be written as ($V_{H+1}^\pi \equiv V_{H+1}^\star \equiv 0$):

$$\begin{cases} V_h^\pi(s) = \mathbb{E}_{a' \sim \pi_h(s)}[Q_h^\pi(s,a')] \\ Q_h^\pi(s,a) := r_h(s,a) + \mathbb{P}_{s,a,h} V_{h+1}^\pi \end{cases} \text{and} \begin{cases} V_h^\star(s) = \max_{a' \in \mathcal{A}} Q_h^\star(s,a') \\ Q_h^\star(s,a) := r_h(s,a) + \mathbb{P}_{s,a,h} V_{h+1}^\star. \end{cases} \tag{1}$$

For any algorithm over K episodes, let π^k be the policy used in the k-th episode, and s_1^k be the corresponding initial state. The regret over $T = HK$ steps is $\text{Regret}(T) := \sum_{k=1}^K (V_1^\star - V_1^{\pi^k})(s_1^k)$.

Suboptimality Gap. We begin by formally defining the suboptimality gap for any given MDP.

Definition 2.1. *For any (s,a,h), the suboptimality gap is defined as $\Delta_h(s,a) := V_h^\star(s) - Q_h^\star(s,a)$.*

The Bellman optimality equation (1) implies that $\Delta_h(s,a) \geq 0$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Building on this, we introduce the minimum gap at each step.

Definition 2.2. *Define the **minimum gap at step** h as $\Delta_{\min,h} := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}, \Delta_h(s,a) > 0} \Delta_h(s,a)$ for any $h \in [H]$. If the set $\{\Delta_h(s,a) | \Delta_h(s,a) > 0, (s,a) \in \mathcal{S} \times \mathcal{A}\}$ is empty, we set $\Delta_{\min,h} = \infty$.*

We further define the global minimum gap as follows.

Definition 2.3. *The **minimum gap** is defined as $\Delta_{\min} := \min_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \Delta_h(s,a) > 0} \Delta_h(s,a)$.*

If $\{\Delta_h(s,a) \mid \Delta_h(s,a) > 0, (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\} = \emptyset$, then all policies are optimal, leading to a degenerate MDP. Therefore, we assume that the set is nonempty and $\Delta_{\min} > 0$. Definitions 2.1 and 2.3 and the assumption $\Delta_{\min} > 0$ are standard in the gap-dependent analysis (Simchowitz & Jamieson, 2019; Yang et al., 2021; Xu et al., 2021; Zheng et al., 2025b; Zhang et al., 2026a; 2025b).

Local Switching Cost. Following Bai et al. (2019); Zhang et al. (2020), we define the local switching cost of any algorithm run over $U > 1$ episodes as follows.

Definition 2.4. *The local switching cost for any algorithm with U episodes is defined as $N_{\text{switch}} := \sum_{u=1}^{U-1} \sum_{h=1}^H \sum_{s \in \mathcal{S}} \mathbb{I}\left[\pi_h^{u+1}(s) \neq \pi_h^u(s)\right]$, where π^u is the policy used in the u-th episode.*

3 Main Results

3.1 UCB-Bernstein+ Algorithm

In this section, we introduce UCB-Bernstein+ in Algorithm 1, which enhances UCB-Bernstein (Jin et al., 2018) by refining its exploration bonus and incorporating the stage design (Zhang et al., 2020).

We first introduce the stage design. For any triple (s,a,h), the samples collected for the triple are divided into consecutive *stages*. The length of each stage grows approximately exponentially with a factor of $(1 + 1/H)$. Specifically, we define $e_1 = H$ and $e_{i+1} = \left\lfloor (1 + \frac{1}{H}) e_i \right\rfloor$ for $i \geq 1$, and let e_i represent the length of stage i. We also define the set of stage endpoints as $\mathcal{L} := \{\sum_{i=1}^j e_i \mid j \in \mathbb{N}_+\}$.

Next, we discuss the algorithm design. UCB-Bernstein+ proceeds in episodes. At the beginning of each episode, in addition to maintaining the standard value function estimates $Q_h(s,a)$ and $V_h(s)$, the algorithm tracks the following quantities: $N_h(s,a)$, the total number of visits to (s,a,h) across all episodes, and $n_h(s,a)$, the number of visits to (s,a,h) in the current stage. It also tracks two additional stage-wise accumulators: $\mu_h(s,a)$ and $\sigma_h(s,a)$. They are reset to zero at the end of each stage and updated at every step as (with the convention $V_{H+1}(s) = 0$ for all $s \in \mathcal{S}$.)

$$(\mu_h, \sigma_h)(s_h, a_h) \xleftarrow{+} \left(V_{h+1}, (V_{h+1})^2\right)(s_{h+1}). \tag{2}$$

After each transition, for the visited triple (s_h, a_h, h), the algorithm updates $Q_h(s,a), V_h(s)$ and $\pi_h(s)$ only if $N_h(s,a)$ reaches a stage endpoint in $\mathcal{L}$. Only the samples for (s,a,h) collected during the current stage, tracked by $n_h(s,a)$, $\mu_h(s,a)$, and $\sigma_h(s,a)$, are used in updates. Specifically,

$$Q_h(s,a) \leftarrow \min\{r_h + \mu_h/n_h + b_h, Q_h\}(s,a), \quad V_h(s) \leftarrow \max_{a'} Q_h(s,a'), \tag{3}$$

and $\pi_h(s) \leftarrow \arg\max_{a'} Q_h(s,a')$, where the variance-based bonus $b_h(s,a)$ in (3) is defined as

$$\min\left\{ C_1 \sqrt{\frac{W_h(s,a)\iota}{n_h(s,a)}} + \frac{C_2 H \iota}{n_h(s,a)}, C_3 \sqrt{\frac{H^2 \iota}{n_h(s,a)}} \right\}. \tag{4}$$

Here, $C_1, C_2, C_3 > 0$ are constants, and $W_h(s,a)$ denotes the stage-wise empirical variance of the V-estimates collected for the triple (s,a,h):

$$W_h(s,a) = \sigma_h(s,a)/n_h(s,a) - (\mu_h(s,a)/n_h(s,a))^2 .$$

$\iota = \log(2SAT/\delta)$, where $\delta \in (0,1)$ is the failure probability, and we treat ι as a positive tuning parameter in practice. Here, we remark that the stages are unaligned across different triples (s,a,h), meaning that the stage renewal of one triple does not necessarily imply the renewal of another.

Compared to the original UCB-Bernstein, the stage design leads to infrequent policy switching, and the refined bonus removes the polynomial dependence on S and A.

Algorithm 1 UCB-Bernstein+

Initialize: For all (s,a,h), let $V_h(s) \leftarrow H$, $Q_h(s,a) \leftarrow H$, and $(n_h, N_h, \mu_h, \sigma_h)(s,a) \leftarrow 0$.
for episodes $k \leftarrow 1, 2, \ldots, K$ **do**
  Observe s_1.
  for $h \leftarrow 1, 2, \ldots, H$ **do**
    Take action $a_h = \pi_h(s_h) \leftarrow \arg\max_{a'} Q_h(s_h, a')$, and observe the next state s_{h+1}.
    Update the accumulators via $N_h(s_h,a_h) \xleftarrow{+} 1$, $n_h(s_h,a_h) \xleftarrow{+} 1$ and Equation (2).
    if $N_h(s_h,a_h) \in \mathcal{L}$ **then**
      Update the Q-estimates and V-estimates for triple (s_h,a_h,h) according to Equation (3).
      Set $(n_h, \mu_h, \sigma_h)(s_h,a_h) \leftarrow 0$ to reset stage-wise accumulators.
    end if
  end for
end for

3.2 Theoretical Guarantees

We now present our main result: the first fine-grained gap-dependent regret bound for model-free algorithms that employ variance-based bonuses. We first define the maximum conditional variance at step $h \in [H]$ as $\mathbb{Q}_h^\star := \max_{s,a}\{\mathbb{V}_{s,a,h}(V_{h+1}^\star)\} \in [0, H^2]$. Then we have Theorem 3.1.

Theorem 3.1. *Let $\iota_1 = \log(SAT)$. For UCB-Bernstein+, set the failure probability $\delta = 1/(8T)$. Then there exist numerical constants $C_1, C_2, C_3 > 0$ such that $\mathbb{E}\,(\mathrm{Regret}(T))$ can be bounded by*

$$O\left(\inf_{\beta\in(0,H]}\left\{\sum_{h=1}^{H}\sum_{\Delta_h(s,a)>0}\frac{(\mathbb{Q}_h^\star+\beta^2)H^3\iota_1}{\Delta_h(s,a)}+\sum_{h=1}^{H}\frac{H(\sum_{t=h+1}^{H}\sqrt{(\mathbb{Q}_t^\star+\beta^2)|Z_{\mathrm{opt},t}|})^2\iota_1}{\Delta_{\min,h}}+\frac{H^6SA\iota_1^2}{\beta}\right\}\right)$$

where $Z_{\mathrm{opt},h}$ collects optimal state-action pairs at step h: $Z_{\mathrm{opt},h} = \{(s,a) \in \mathcal{S}\times\mathcal{A} | \Delta_h(s,a) = 0\}$.

In Theorem 3.1, $\beta \in (0, H]$ is introduced by our theoretical analysis and is not a tuning parameter; details can be found in Section 4. The full proof of Theorem 3.1 can be found in Appendix H.

Compared with the fine-grained gap-dependent regret upper bounds for the UCB-Hoeffding (Zhang et al., 2026a) and AMB (Xu et al., 2021) algorithms, our result is variance-aware and can yield sharper bounds in low-variance regimes. Using the worst-case bound $\mathbb{Q}_h^\star \le H^2$ for all $h \in [H]$ and setting $\beta = H$, the leading gap-dependent terms in our bound recover the same horizon dependence as those in Zhang et al. (2026a), up to logarithmic factors. In contrast, in the ideal scenario of deterministic MDPs with $\mathbb{Q}_h^\star = 0$ for any $h \in [H]$, the AM–GM inequality implies that, for appropriate choices of β, our bound depends sublinearly on the inverse of the suboptimality gaps, improving upon the linear dependence established by Zhang et al. (2026a); Xu et al. (2021). This sublinear improvement can be made explicit in the coarse-grained setting by substituting all $\Delta_h(s,a)$ and $\Delta_{\min,h}$ with $\Delta_{\min}$. In this case, the regret bound in Theorem 3.1 becomes $O(H^{16/3}SA\iota_1^{5/3}/\Delta_{\min}^{1/3})$ for $\beta = \Theta((H^2\Delta_{\min}\iota_1)^{1/3})$ with $\Delta_{\min} \le O(H/\iota_1)$, exhibiting a sublinear dependence on $\Delta_{\min}^{-1}$.

With the stage design, we also establish a gap-dependent upper bound on the local switching cost.

Theorem 3.2. *For any $\delta \in (0,1)$, run UCB-Bernstein+ with failure-probability parameter $\delta/8$, so that $\iota = \log(16SAT/\delta)$. Then there exist numerical constants $C_1, C_2, C_3 > 0$ such that, with probability at least $1-\delta$, the local switching cost N_{switch} is upper bounded by*

$$O\left(H|Z_{\mathrm{opt}}|\log\left(\frac{K}{|Z_{\mathrm{opt}}|}+1\right)+H|Z_{\mathrm{opt}}^c|\log\left(\frac{H^{\frac{5}{2}}S^{\frac{1}{2}}A^{\frac{1}{2}}\log(\frac{SAT}{\delta})}{|Z_{\mathrm{opt}}^c|^{\frac{1}{2}}\Delta_{\min}}\right)\right).$$

Here, $Z_{\text{opt}} = \{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H] \mid a \in \mathcal{A}_h^\star(s)\}$*, where* $\mathcal{A}_h^\star(s) = \{a \mid a \text{ maximizes } Q_h^\star(s,\cdot)\}$*, and* $Z_{\text{opt}}^c = (\mathcal{S} \times \mathcal{A} \times [H]) \backslash Z_{\text{opt}}$.

The proof of Theorem 3.2 is provided in Appendix I. Compared with Zheng et al. (2025b), our result improves the dependence on H and S in the gap-dependent logarithmic factors. Our fine-grained analysis yields a variance-aware bound on visits to suboptimal triples (Lemma I.1), which gives the stated switching-cost bound after a worst-case relaxation. This improvement is particularly pronounced if each state-step pair has a unique optimal action, in which case the bound simplifies to

$$O(H^2 S \log(K/HS + 1) + H^2 SA \log(H^2 \log(SAT/\delta)/\Delta_{\min})).$$

In this scenario, our result removes the $S \log S$ dependence and improves the dependence on H compared to the corresponding result for UCB-Advantage in Zheng et al. (2025b).

To complete the theoretical analysis of UCB-Bernstein+, we also provide worst-case guarantees for both regret and local switching cost in the following two theorems.

Theorem 3.3. *For any* $\delta \in (0,1)$*, run UCB-Bernstein+ with failure-probability parameter* $\delta/8$*, so that* $\iota = \log(16SAT/\delta)$*. Then there exist constants* $C_1, C_2, C_3 > 0$ *such that, with probability at least* $1-\delta$*, we have* $\text{Regret}(T) \le O(\sqrt{H^3 SAT\iota} + H^{\frac{9}{2}} SA\iota^2)$.

The proof is deferred to Appendix J. UCB-Bernstein (Jin et al., 2018) achieves a worst-case regret bound of $O(\sqrt{H^3 SAT\iota} + \sqrt{H^9 S^3 A^3 \iota^4})$. Our bound removes their superlinear dependence on the state and action space dimensions S and A, thanks to the refined bonus design in Equation (4).

UCB2-Bernstein (Bai et al., 2019) achieves a logarithmic local switching cost of $O(H^3 SA \log(K/A))$ while maintaining the same worst-case regret as UCB-Bernstein. Our analysis further shows that UCB-Bernstein+, with the stage design, improves upon the local switching cost of UCB2-Bernstein. The worst-case local switching cost bound is provided in the following theorem.

Theorem 3.4. *The local switching cost of UCB-Bernstein+ is bounded by* $O(H^2 SA \log(\frac{K}{SAH} + 1))$.

The proof of Theorem 3.4 is given in Appendix K. Compared to UCB2-Bernstein, our result improves the dependence on the horizon length by a factor of H. Moreover, it matches the best-known worst-case local switching cost achieved by UCB-Advantage (Zhang et al., 2020).

4 Our Technical Tool: Pseudo Value Functions

We introduce a key technical tool for the proofs of Theorems 3.1 and 3.3: the pseudo value functions. To facilitate the subsequent analysis, we first introduce some notation.

Let $Q_h^k(s,a)$ and $V_h^k(s)$ denote the values of $Q_h(s,a)$ and $V_h(s)$ at the beginning of episode k, respectively. Let $W_h^k(s,a)$ and $b_h^k(s,a)$ be the corresponding bonus terms used to update $Q_h^k(s,a)$, and let $n_h^k(s,a)$ denote the number of visits to (s,a,h) during the previous stage of the triple in episode k. For $i \in [n_h^k(s,a)]$, let $k_{l_i(s,a,h,k)}$ be the episode index of the i-th visit to (s,a,h) within the most recently completed stage. When (s,a,h,k) is clear from the context, we use k_{l_i} for brevity.

To upper bound the regret, a key step is to control the sum of value estimation errors $(Q_h^k - Q_h^\star)(s_h^k, a_h^k)$. The worst-case analysis considers the total sum over all episodes, whereas both coarse-grained and fine-grained gap-dependent analyses use a weighted sum. After the triple (s_h^k, a_h^k, h) has completed its first stage, by the update rule in Equation (3) and the Bellman Equation (1), the estimation error admits the following decomposition:

$$(Q_h^k - Q_h^\star)(s_h^k, a_h^k) \le \frac{\sum_{i=1}^{n_h^k} (V_{h+1}^{k_{l_i}} - V_{h+1}^\star)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k, a_h^k)} + \frac{\sum_{i=1}^{n_h^k} (\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s_h^k, a_h^k, h}) V_{h+1}^\star}{n_h^k(s_h^k, a_h^k)} + b_h^k(s_h^k, a_h^k).$$

Here, the first RHS term establishes a recursive relationship between estimation errors at steps h and $h+1$, while the second can be controlled using standard concentration inequalities. The main challenge lies in controlling the variance-based bonus in the last term. Lemma C.3 of Jin et al. (2018) directly approximates this bonus by the conditional variance of the optimal value function $V_{h+1}^\star$. However, summing the approximation errors introduces superlinear dependence on S and A, which

leads to the superlinear dependence in the regret bound of UCB-Bernstein. Moreover, it also remains unclear whether this direct approximation is suitable for fine-grained gap-dependent analysis.

To address this, we introduce the *pseudo value functions* $\hat{V}_h^k(s)$ for any $(s,h,k) \in \mathcal{S} \times [H] \times [K]$:

$$\hat{V}_h^k(s) := \max\left\{V_h^\star(s), \min\{V_h^\star(s) + \beta, V_h^k(s)\}\right\}, \tag{5}$$

where $\beta \in (0, H]$ is an arbitrary constant. The optimism of UCB-Bernstein+ (Lemma G.1 in Appendix G), which ensures that $V_h^k(s) \geq V_h^\star(s)$ for all $s \in \mathcal{S}$ with high probability, implies that

$$V_h^\star(s) \leq \hat{V}_h^k(s) = \min\left\{V_h^\star(s) + \beta, V_h^k(s)\right\}. \tag{6}$$

Next, we discuss how the pseudo value function helps bound the variance-based bonus, which satisfies

$$b_h^k(s,a) \leq C_b\left(\sqrt{\frac{W_h^k(s,a)\iota}{n_h^k(s,a)}} + \frac{H\iota}{n_h^k(s,a)}\right)$$

for some universal constant $C_b > 0$. Since $H\iota/n_h^k$ does not involve variance estimations, its cumulative sum can be controlled directly using Lemma E.2, and the corresponding weighted sum can be handled similarly when the weights are uniformly bounded. Therefore, we focus on the first term with the variance estimation $W_h^k(s,a)$. For a sequence $(f^j)_{j=1}^K$ given by (V_{h+1}^j), $(\widehat{V}_{h+1}^j)$, or the constant sequence $V_{h+1}^\star$, define, when $n_h^k(s,a) > 0$, the empirical mean and variance over the $n_h^k(s,a)$ visits to (s,a,h) as $\hat{\mathbb{E}}[f] = \sum_{i=1}^{n_h^k(s,a)} f^{k_{l_i}}(s_{h+1}^{k_{l_i}})/n_h^k(s,a)$, $\hat{\mathbb{V}}[f] = \hat{\mathbb{E}}[f^2] - (\hat{\mathbb{E}}[f])^2$. Since $\hat{V}_h^j(s) \leq V_h^j(s)$ for any $(s,j) \in \mathcal{S} \times [K]$ by Lemma G.1, $W_h^k(s,a)$ can be decomposed as:

$$\begin{aligned} W_h^k(s,a) &= \hat{\mathbb{V}}[V_{h+1}] \leq \hat{\mathbb{E}}\big[(V_{h+1})^2\big] - \big(\hat{\mathbb{E}}[\hat{V}_{h+1}]\big)^2 \\ &= (\hat{\mathbb{E}}[(V_{h+1})^2] - \hat{\mathbb{E}}[(\hat{V}_{h+1})^2]) + (\hat{\mathbb{V}}[\hat{V}_{h+1}] - 2\hat{\mathbb{V}}[V_{h+1}^\star]) + 2(\hat{\mathbb{V}}[V_{h+1}^\star] - \mathbb{V}[V_{h+1}^\star]) + 2\mathbb{V}[V_{h+1}^\star]. \end{aligned} \tag{7}$$

Next, we control the terms on the right-hand side of Equation (7) separately. The first term satisfies

$$\begin{aligned} \hat{\mathbb{E}}\left[(V_{h+1})^2\right] - \hat{\mathbb{E}}\big[(\hat{V}_{h+1})^2\big] &\leq 2H\hat{\mathbb{E}}[V_{h+1} - \hat{V}_{h+1}] \\ &\leq 2H\hat{\mathbb{E}}\left[\left(V_{h+1} - V_{h+1}^\star\right)\mathbb{I}\left[V_{h+1} - V_{h+1}^\star \geq \beta\right]\right]. \end{aligned} \tag{8}$$

Here, the first inequality follows from factoring the difference of two squares and the fact that $V_{h+1}^j, \hat{V}_{h+1}^j \leq H$ for any $j \in [K]$. The last inequality holds because for any $j \in [K]$, $V_{h+1}^j - \hat{V}_{h+1}^j > 0$ can occur only when $V_{h+1}^j - V_{h+1}^\star \geq \beta$ due to Equation (6), and $V_{h+1}^j - \hat{V}_{h+1}^j \leq V_{h+1}^j - V_{h+1}^\star$. The RHS of Equation (8) characterizes the value estimation error during the early exploration phase and can be properly controlled (see Lemma G.3 in Appendix G for details). In addition to the refined bonus design, this step is another key to avoiding the superlinear dependence on the state and action space dimensions in the worst-case regret of Jin et al. (2018). We also remark that this step resembles the control of the reference settling error for algorithms using variance reduction (Zhang et al., 2020; Li et al., 2023; Zhang et al., 2025b). However, it differs in that the target value gap β is conceptual and does not appear in the UCB-Bernstein+ design.

The second term in Equation (7), $\hat{\mathbb{V}}[\hat{V}_{h+1}] - 2\hat{\mathbb{V}}[V_{h+1}^\star]$, can be bounded by $2\hat{\mathbb{V}}[\hat{V}_{h+1} - V_{h+1}^\star]$ with the Cauchy–Schwarz inequality. It can be further bounded by $O(\beta^2)$ since $|\hat{V}_{h+1}^j - V_{h+1}^\star| \leq \beta$ for any $j \in [K]$ by Equation (6). The third term in Equation (7), $2(\hat{\mathbb{V}}[V_{h+1}^\star] - \mathbb{V}[V_{h+1}^\star])$, focuses on the empirical estimation error of the variance and can be bounded with the Azuma-Hoeffding inequality. See Equation (58) in Appendix H for details.

Finally, bounding the last term, $\mathbb{V}[V_{h+1}^\star]$, highlights the advantage of variance-based exploration bonuses. In the worst-case analysis, we approximate it with $\mathbb{V}[V_{h+1}^\pi]$ as in Jin et al. (2018), achieving improved H-dependence compared to algorithms using Hoeffding-type bonuses. For the gap-dependent analysis, we bound it with $\mathbb{Q}_h^\star$, which captures the MDP's variance structure. This completes the explanation of how the variance-aware bonus is controlled.

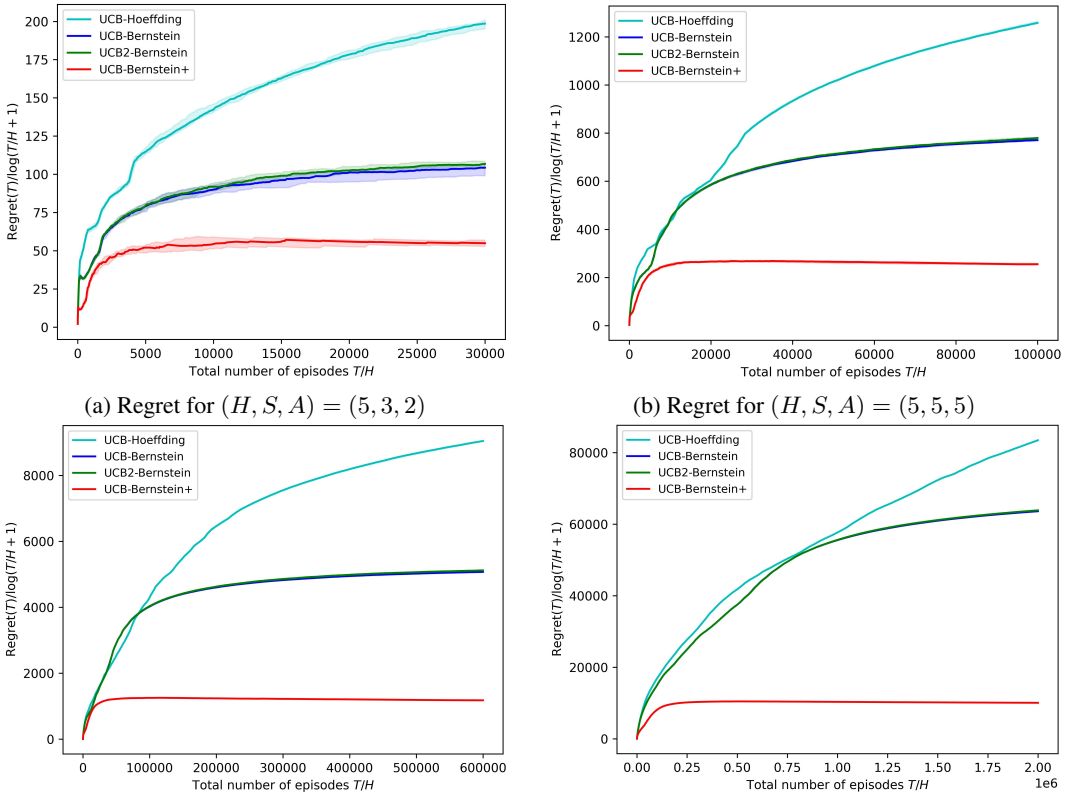
Using the pseudo value function and the bounding procedure for variance-based exploration bonuses described above, we derive the fine-grained gap-dependent regret bound by separately analyzing optimal and suboptimal state-action-step triples and recursively summing the weighted value estimation errors across steps (Lemmas H.1 and H.2), following the framework in Zhang et al. (2026a). Full details are provided in Appendix H. Similarly, the worst-case regret can be obtained by controlling the full sum of value estimation errors using the recursive procedure in Jin et al. (2018); Zhang et al. (2020), with full details in Appendix J.

5 NUMERICAL EXPERIMENTS

In this section, we conduct numerical experiments. We first evaluate UCB-Bernstein+ against representative model-free algorithms using Hoeffding-type or Bernstein-type variance-based bonuses, including UCB-Hoeffding (Jin et al., 2018), UCB-Bernstein (Jin et al., 2018), and UCB2-Bernstein (Bai et al., 2019), in a synthetic environment. We assess their performance in terms of both regret and local switching cost, with particular emphasis on the comparison with UCB2-Bernstein, the low-switching-cost variant of UCB-Bernstein. For evaluation, $V^\star$ and each episode's greedy-policy value are computed exactly by backward dynamic programming using the true MDP, which is unavailable to the learning algorithms. The local switching cost is computed according to Definition 2.4 by comparing consecutive greedy policies over all state–step pairs.

We further provide additional comparisons of the same four algorithms in the GridWorld environment (Domingues et al., 2021) in Appendix B.1. Finally, in Appendix B.2, we compare UCB-Bernstein+ with two state-of-the-art baselines: Q-EarlySettled-LowCost (Zhang et al., 2025b), which achieves the best-known regret guarantee, and UCB-Advantage (Zhang et al., 2020), which achieves the best-known local switching cost guarantee. These additional comparisons further demonstrate the empirical effectiveness of UCB-Bernstein+.

We now describe the setup for comparing the four algorithms in the synthetic environment. We consider four **experiment scales** with $(H, S, A, K) = (5, 3, 2, 3 \times 10^4)$, $(5, 5, 5, 10^5)$, $(7, 8, 6, 6 \times 10^5)$, and $(10, 15, 10, 2 \times 10^6)$. For each (s, a, h), the reward $r_h(s, a)$ is sampled independently from the uniform distribution over $[0, 1]$, and the transition kernel $\mathbb{P}_h(\cdot \mid s, a)$ is drawn uniformly from the probability simplex over S states. The initial state of each episode is chosen uniformly at random from the state space. We set $\iota = 1$ and use the bonus coefficient $\sqrt{2}$ for the leading term that scales as $\sqrt{1/n}$ in all four algorithms. Here, n represents the number of visits. For the lower-order term that scales as $1/n$ in the three algorithms with Bernstein-type bonuses, we use the coefficient 1. Accordingly, for the UCB-Bernstein+ algorithm, we set $C_1 = C_3 = \sqrt{2}$ and $C_2 = 1$ in Equation (4).



(a) Regret for $(H, S, A) = (5, 3, 2)$

(b) Regret for $(H, S, A) = (5, 5, 5)$

(c) Regret for $(H, S, A) = (7, 8, 6)$

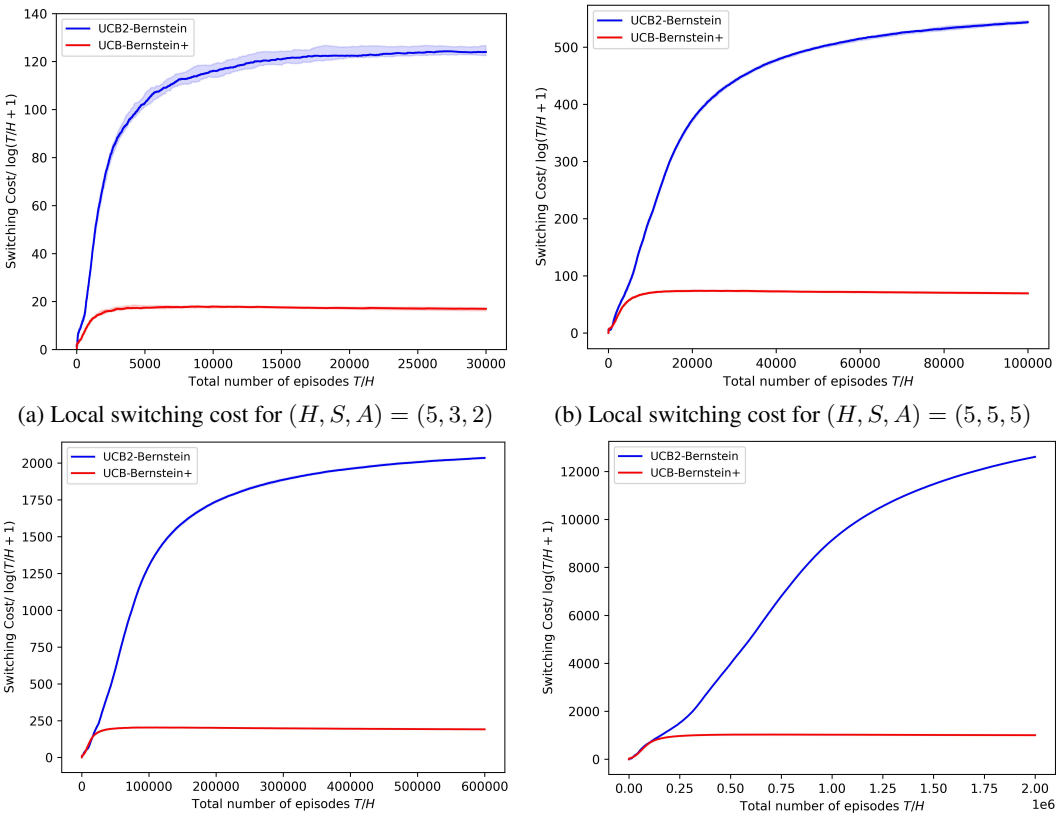
(d) Regret for $(H, S, A) = (10, 15, 10)$

Figure 1: Regret comparison of different algorithms.

To quantify uncertainty, for each experiment scale, we generate one MDP and run 10 independent trials for each algorithm on this fixed MDP. In the following figures, solid lines represent the median over the 10 trials, and shaded regions indicate the 10th–90th percentile intervals.

Figure 1 presents the regret comparison results, where we plot $\text{Regret}(T)/\log(K+1)$ against the number of episodes K. We observe that UCB-Bernstein+ achieves the best regret among the compared four algorithms. In addition, across all experiment scales, its curve flattens as K increases, suggesting approximately logarithmic regret growth, qualitatively consistent with Theorem 3.1.

We further compare the local switching costs of UCB2-Bernstein and UCB-Bernstein+, with the results shown in Figure 2. We plot Local Switching Cost$/\log(K+1)$ against the number of episodes K. The results show that UCB-Bernstein+ consistently incurs a substantially lower local switching cost than UCB2-Bernstein across the experiment scales. Moreover, the local switching-cost curve of UCB-Bernstein+ gradually stabilizes as K increases, suggesting approximately logarithmic growth of its local switching cost with the number of episodes. These empirical observations are consistent with both the gap-dependent and worst-case local switching cost guarantees in Theorems 3.2 and 3.4.



(a) Local switching cost for $(H, S, A) = (5, 3, 2)$

(b) Local switching cost for $(H, S, A) = (5, 5, 5)$

(c) Local switching cost for $(H, S, A) = (7, 8, 6)$

(d) Local switching cost for $(H, S, A) = (10, 15, 10)$

Figure 2: Local switching cost comparison of different algorithms.

6 Conclusion

In this work, we study model-free online RL for episodic tabular MDPs, focusing on gap-dependent regret and policy switching cost. We propose UCB-Bernstein+, a refined UCB-Bernstein algorithm with variance-based exploration bonuses. Our analysis establishes the first fine-grained gap-dependent regret upper bound for variance-aware model-free online RL, closing the gap left by prior Hoeffding-based analyses. By incorporating a stage-wise policy update design, UCB-Bernstein+ further achieves the best-known gap-dependent local switching cost to date. We also establish improved worst-case guarantees for both regret and local switching cost over the original UCB-Bernstein algorithm. Numerical experiments further demonstrate the empirical effectiveness of UCB-Bernstein+.

AI USE STATEMENT

AI tools were used for language polishing, presentation refinement, mathematics proofreading, assistance with method implementation, and code debugging. All scientific ideas, problem formulation, mathematical claims, proof strategies and key arguments, experimental design, and interpretation of results were independently developed and verified by the authors.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or real-world experiments requiring ethical approval. The empirical evaluation is conducted entirely in synthetic tabular MDPs and simulated GridWorld environments. We have followed the guidelines outlined in the ICLR Code of Ethics and have aimed to ensure transparency, integrity, and fairness throughout the research process. We do not foresee any direct ethical concerns or potential harms arising from this study.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the complete specification of UCB-Bernstein+, including its stage design, update rules, and exploration bonuses, in the main text. Detailed proofs of all theoretical results, together with the auxiliary lemmas and high-probability events used in the analysis, are provided in the appendix. The experimental sections describe the synthetic tabular MDPs, GridWorld environments, baseline algorithms, hyperparameter settings, computational resources, and evaluation metrics used in our empirical comparisons. The implementation code for the proposed method is included in the supplementary materials. All the experiments are run on a server with Intel Xeon E5-2650v4 (2.2GHz) and 100 cores. Each replication is limited to a single core and 16GB of RAM. The total execution time is about 10 hours.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 49–56. MIT Press, 2007.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32, 2019.

Shulun Chen, Runlong Zhou, Zihan Zhang, Maryam Fazel, and Simon Du. Sharp gap-dependent variance-aware regret bounds for tabular mdps. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 112169–112205. Curran Associates, Inc., 2025. doi: 10.52202/085713-3747. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a2fb8d64706276cc6c49a52f143e0e5c-Paper-Conference.pdf.

Omar Darwiche Domingues, Yannis Flet-Berliac, Edouard Leurent, Pierre Ménard, Xuedong Shang, and Michal Valko. rlberry - A Reinforcement Learning Library for Research and Education, October 2021. URL https://github.com/rlberry-py/rlberry.

Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in Neural Information Processing Systems*, 31, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Safwan Labbi, Daniil Tiapkin, Lorenzo Mancini, Paul Mangold, and Eric Moulines. Federated ucbvi: Communication-efficient federated regret minimization with heterogeneous agents. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1315–1323. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/labbi25a.html.

Gen Li, Laixi Shi, Yuxin Chen, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Information and Inference: A Journal of the IMA*, 12 (2):969–1043, 2023.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024.

Zihang Liang, Haochen Zhang, and Lingzhou Xue. Provably efficient federated reinforcement learning with linear function approximation and logarithmic communication cost. *arXiv preprint arXiv:2609.00193*, 2026.

Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pp. 7609–7618. PMLR, 2021.

Azalia Mirhoseini, Hieu Pham, Quoc V Le, Benoit Steiner, Rasmus Larsen, Yuefeng Zhou, Naveen Kumar, Mohammad Norouzi, Samy Bengio, and Jeff Dean. Device placement optimization with reinforcement learning. In *International Conference on Machine Learning*, pp. 2430–2439. PMLR, 2017.

Phuoc Nguyen, Truyen Tran, Sunil Gupta, Santu Rana, Matthew Barnett, and Svetha Venkatesh. Incomplete conditional density estimation for fast materials discovery. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 549–557. SIAM, 2019.

Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, and Raman Arora. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9310–9318, 2023.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34:16371–16383, 2021.

Dan Qiao, Ming Yin, Ming Min, and Yu-Xiang Wang. Sample-efficient reinforcement learning with loglog (t) switching cost. In *International Conference on Machine Learning*, pp. 18031–18061. PMLR, 2022.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Ambuj Tewari and Peter Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems*, pp. 1505–1512, 2008.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pp. 4438–4472. PMLR, 2021.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

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

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Gap-dependent bounds for federated q-learning. In *Forty-second International Conference on Machine Learning*, 2025a.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Regret-optimal q-learning with low cost for single-agent and federated reinforcement learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 128711–128799. Curran Associates, Inc., 2025b. doi: 10.52202/085713-4288. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bb203e938836544655996d1bb94a0fd7-Paper-Conference.pdf.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Q-learning with fine-grained gap-dependent regret. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 48118–48162, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4f8d1e02a057beb7920618ac73788078-Paper-Conference.pdf.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Gap-dependent bounds for nearly minimax optimal reinforcement learning with linear function approximation. *arXiv preprint arXiv:2602.20297*, 2026b.

Weitong Zhang, Zhiyuan Fan, Jiafan He, and Quanquan Gu. Achieving constant regret in linear markov decision processes. *Advances in Neural Information Processing Systems*, 37:130694–130738, 2024a.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021.

Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. In *Conference on Learning Theory*, pp. 5213–5219. PMLR, 2024b.

Zhong Zheng, Fengyu Gao, Lingzhou Xue, and Jing Yang. Federated q-learning: Linear regret speedup with low communication cost. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhong Zheng, Haochen Zhang, and Lingzhou Xue. Federated q-learning with reference-advantage decomposition: Almost optimal regret and logarithmic communication cost. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zhong Zheng, Haochen Zhang, and Lingzhou Xue. Gap-dependent bounds for q-learning using reference-advantage decomposition. In *The Thirteenth International Conference on Learning Representations*, 2025b.

In the appendix, Appendix A reviews related work. Appendix B presents additional numerical experiments. Appendix C summarizes the notation used throughout the paper. Appendix D provides several auxiliary lemmas that support our analysis. Appendix E summarizes basic properties of UCB-Bernstein+, while Appendix F introduces the high-probability events used in the proofs. Appendix G establishes several properties of the value function estimates. The proofs of Theorem 3.1 and Theorem 3.2 are provided in Appendix H and Appendix I, respectively, whereas Appendix J and Appendix K contain the proofs of Theorem 3.3 and Theorem 3.4, respectively.

A Related Work

Online RL for Tabular Episodic MDPs. For tabular episodic MDPs in the worst-case regret setting, Auer et al. (2008), Agrawal & Jia (2017), Azar et al. (2017), Agarwal et al. (2020), Zanette & Brunskill (2019), Zhang et al. (2021), and Zhang et al. (2024b) study model-based algorithms. Notably, Zhang et al. (2024b) propose the MVP algorithm that achieves a regret of $\tilde{O}(\min\{\sqrt{SAH^2T}, T\})$, matching the information-theoretic lower bound up to logarithmic factors. Model-free algorithms have been studied by Jin et al. (2018), Zhang et al. (2020), Li et al. (2023), Ménard et al. (2021), and Zhang et al. (2025b), with the latter four achieving a near-optimal regret of $\tilde{O}(\sqrt{SAH^2T})$. Several works also consider online federated RL, including Liang et al. (2026), Zheng et al. (2024), Labbi et al. (2025), Zheng et al. (2025a), and Zhang et al. (2025b), with the last three attaining near-optimal regret bounds up to logarithmic factors.

Gap-Dependent RL. When a strictly positive suboptimality gap exists, logarithmic regret becomes achievable. Early studies establish asymptotic logarithmic regret bounds (Auer & Ortner, 2007; Tewari & Bartlett, 2008), while more recent work develops non-asymptotic bounds (Jaksch et al., 2010; Ok et al., 2018; Simchowitz & Jamieson, 2019; He et al., 2021). Specifically, Jaksch et al. (2010) design a model-based algorithm whose regret depends on the policy gap rather than the action gap considered in this paper. Ok et al. (2018) derive problem-specific logarithmic-type lower bounds for both structured and unstructured MDPs, and Simchowitz & Jamieson (2019) extend the model-based algorithm of Zanette & Brunskill (2019) to obtain logarithmic regret bounds. More recently, Chen et al. (2025) further improve model-based gap-dependent results. Logarithmic regret bounds have also been established under linear function approximation (He et al., 2021), and Nguyen-Tang et al. (2023) provide gap-dependent guarantees for offline RL with linear function approximation.

For model-free algorithms, Yang et al. (2021) show that UCB-Hoeffding (Jin et al., 2018) achieves a gap-dependent regret bound of $\tilde{O}(H^6SA/\Delta_{\min})$. This bound is subsequently refined by Xu et al. (2021) with the Adaptive Multi-step Bootstrap (AMB) algorithm. Zheng et al. (2025b) provide gap-dependent analyses for algorithms with reference-advantage decomposition (Zhang et al., 2020; Li et al., 2023), and Zhang et al. (2025a) and Zhang et al. (2025b) extend gap-dependent analysis to federated Q-learning. More recently, Zhang et al. (2026a) provide the first fine-grained gap-dependent regret upper bound for UCB-Hoeffding.

For RL with linear function approximation, gap-dependent regret bounds have also been studied. For online RL, He et al. (2021) provided the first gap-dependent regret bound, showing that LSVI-UCB (Jin et al., 2020) achieves $\tilde{O}(d^3H^5/\Delta_{\min})$ in linear MDPs, and UCRL-VTR (Ayoub et al., 2020) achieves $\tilde{O}(d^2H^5/\Delta_{\min})$ in linear mixture MDPs. Papini et al. (2021) established constant-regret guarantees for LSVI-UCB under the UNISOFT condition. Subsequently, Zhang et al. (2024a) showed that Cert-LSVI-UCB can also achieve constant regret with high probability under their corresponding structural conditions, independent of the total number of steps T. More recently, Zhang et al. (2026b) established a gap-dependent analysis for the nearly minimax-optimal LSVI-UCB++ algorithm.

RL with Low Switching Cost. Research on low switching cost aims to minimize the number of policy switches while maintaining regret comparable to fully adaptive algorithms. Bai et al. (2019) introduce this problem and propose two Q-learning algorithms with delayed updates and achieve a local switching cost of $\tilde{O}(SAH^3 \log T)$. This line of work is further advanced by Zhang et al. (2020), which improves both the regret upper bound and the local switching cost. More recently, Qiao et al. (2022) propose a model-based algorithm with the global switching cost $\tilde{O}(\log \log T)$, and Zhang et al. (2025b) introduce Q-EarlySettled-LowCost, achieving the best-known regret among model-free algorithms while maintaining logarithmic global switching cost. Gao et al. (2021) also provide a logarithmic global switching cost bound in RL with linear function approximation.

B Additional Numerical Experiments

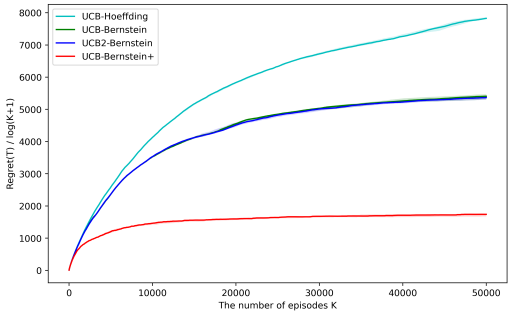
B.1 GridWorld Experiments

We also compare the four algorithms based on Hoeffding-type or Bernstein-type variance-based bonuses in the GridWorld environment, including UCB-Hoeffding, UCB-Bernstein, UCB2-Bernstein, and UCB-Bernstein+. In this environment, the agent navigates a grid to reach a target state.

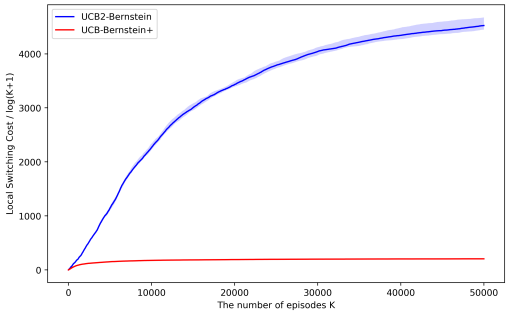
Following a similar experimental setup to Labbi et al. (2025), we consider two grid sizes, 3×3 and 5×5, each with a wall at coordinate $(1, 1)$. This gives $S = 8$ and $S = 24$ valid states, respectively. The horizons are set to $H = 10$ and $H = 15$, and the numbers of episodes range up to $K = 5 \times 10^4$ and $K = 10^5$, respectively. The agent starts from coordinate $(0, 0)$ and, at each step, chooses one of four actions: up, down, left, or right. Under the transition kernel, the agent moves to the intended neighboring cell with probability 0.8 and to a uniformly random other valid neighboring cell with the remaining probability. If the intended action would move the agent to an invalid cell, the agent remains in its current cell with probability 1. The target states are $(2, 2)$ for the 3×3 grid and $(4, 4)$ for the 5×5 grid, and each target state is absorbing. The per-step reward is 1 if the realized next state is the target state and 0 otherwise.

In these experiments, the algorithms observe both the realized next state and the realized reward. UCB-Hoeffding, UCB-Bernstein, and UCB2-Bernstein use the observed reward in each Q-update. For UCB-Bernstein+, the observed rewards are accumulated within each stage, and their empirical mean replaces $r_h(s, a)$ in the stage update in Equation (3); the next-state value samples are averaged separately. The empirical variance used in the Bernstein-type bonuses is computed from next-state value samples only, without a separate reward-variance or reward–value covariance term. For these four algorithms, we adopt the same parameter choices as specified in Section 5. Because the reward can be stochastic conditional on (s, a), this experiment lies outside the deterministic-reward setting of our theoretical guarantees and serves as an empirical robustness check.

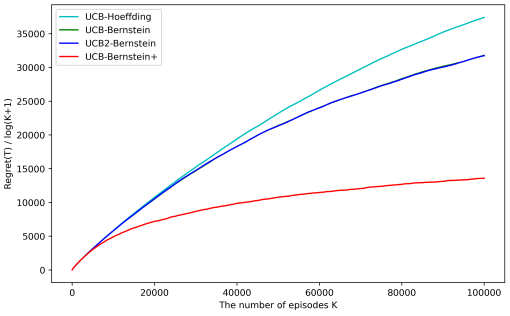
To quantify uncertainty, for each experiment scale, we run 10 independent trials for each algorithm on the same grid. In the figures, solid lines represent the median over the 10 trials, and shaded regions indicate the 10th–90th percentile intervals.



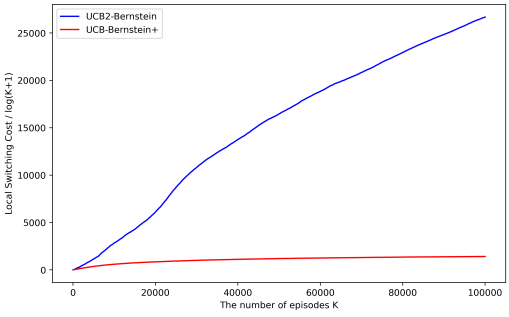
(a) Regret on the 3×3 GridWorld



(b) Local switching cost on the 3×3 GridWorld



(c) Regret on the 5×5 GridWorld



(d) Local switching cost on the 5×5 GridWorld

Figure 3: GridWorld results.

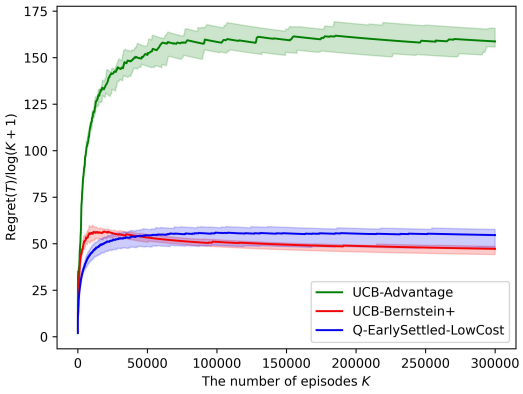
Figure 3 presents the GridWorld comparison results for the two grid sizes. Plots (a) and (c) show the normalized regret $\text{Regret}(T)/\log(K+1)$ as a function of the number of episodes K for each algorithm, while plots (b) and (d) show the normalized local switching cost $\text{Local Switching Cost}/\log(K+1)$ as a function of K.

The empirical findings are consistent with the synthetic-environment results in Section 5. From the regret plots (a) and (c), UCB-Bernstein+ achieves the lowest regret among the four compared algorithms. Moreover, across the two experiment scales, its normalized regret curve becomes flatter as K increases, suggesting approximately logarithmic cumulative regret growth. From the local switching-cost plots (b) and (d), UCB-Bernstein+ incurs a substantially lower normalized local switching cost than UCB2-Bernstein. Its normalized switching-cost curve also stabilizes as K increases, suggesting approximately logarithmic growth of the cumulative local switching cost.

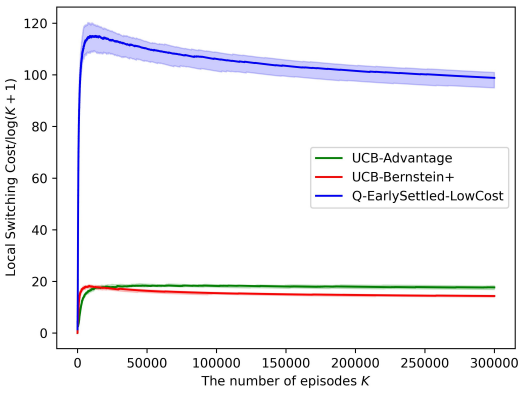
B.2 Experimental Comparison with State-of-the-Art Baselines

In this subsection, we further evaluate the empirical performance of UCB-Bernstein+ by comparing it with two state-of-the-art model-free RL baselines in a synthetic environment: Q-EarlySettled-LowCost (Zhang et al., 2025b), which achieves the best-known regret guarantee, and UCB-Advantage (Zhang et al., 2020), which achieves the best-known local switching cost guarantee.

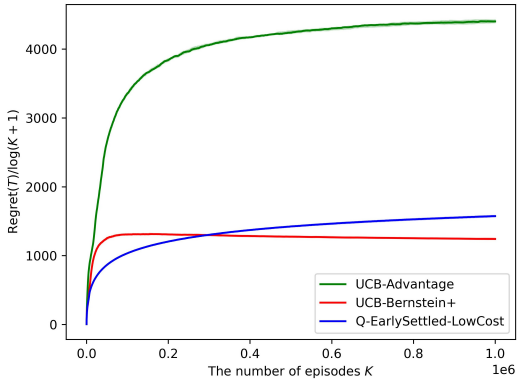
Similar to the synthetic environment introduced in Appendix B.1 of Zhang et al. (2025b), we consider two experiment scales with $(H, S, A, K) = (5, 3, 2, 3 \times 10^5)$, and $(7, 10, 5, 10^6)$. For each (s, a, h), the reward $r_h(s, a)$ is sampled independently from the uniform distribution on $[0, 1]$, and the transition kernel $\mathbb{P}_h(\cdot \mid s, a)$ is drawn uniformly from the probability simplex over S states. The initial state of each episode is sampled uniformly from the state space. We set $\iota = 1$ and use the bonus coefficients specified in Appendix B.1 of Zhang et al. (2025b) for Q-EarlySettled-LowCost and UCB-Advantage. For UCB-Bernstein+, we still set $C_1 = C_3 = \sqrt{2}$ and $C_2 = 1$ in Equation (4).



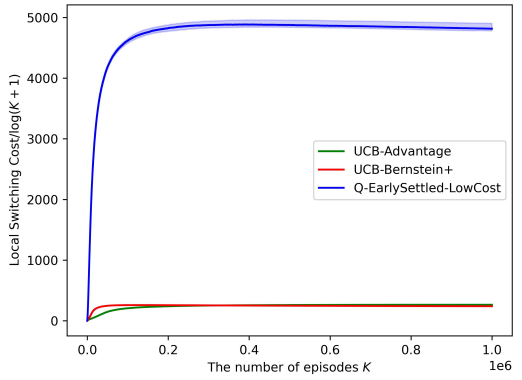
(a) Regret for $(H, S, A) = (5, 3, 2)$



(b) Local switching cost for $(H, S, A) = (5, 3, 2)$



(c) Regret for $(H, S, A) = (7, 10, 5)$



(d) Local switching cost for $(H, S, A) = (7, 10, 5)$

Figure 4: Experimental comparison with state-of-the-art baselines.

To quantify uncertainty, for each experiment scale, we generate one MDP and run 10 independent trials for each algorithm on this fixed MDP. In the following figures, solid lines represent the median over the 10 trials, and shaded regions indicate the 10th–90th percentile intervals.

The results for the two experiment scales are presented in Figure 4. We observe that UCB-Bernstein+ achieves better long-term regret performance than Q-EarlySettled-LowCost while incurring a significantly lower local switching cost. Compared with UCB-Advantage, UCB-Bernstein+ attains substantially better long-term regret performance while maintaining a comparable local switching cost. Overall, these results suggest that UCB-Bernstein+ offers a more favorable empirical trade-off between regret and local switching cost among the compared algorithms.

C Notation Tables

For readability, we summarize below the notation used throughout the algorithm and proofs.

Table 1: Basic MDP, value-function, and gap notation.

Notation	Meaning
$\mathcal{S}, \mathcal{A}$	State and action spaces, with $\lvert\mathcal{S}\rvert = S$ and $\lvert\mathcal{A}\rvert = A$.
H, K, T	Horizon length, number of episodes, and total number of steps, respectively, where $T = KH$.
$P_h(\cdot\lvert s, a)$	Transition distribution at step h from state-action pair (s, a).
$\mathbb{P}_{s,a,h} f$	Conditional expectation $\mathbb{P}_{s,a,h} f := \mathbb{E}_{s' \sim P_h(\cdot\lvert s,a)}[f(s')]$.
$\mathbb{V}_{s,a,h}(f)$	Conditional variance $\mathbb{V}_{s,a,h}(f) := \mathbb{P}_{s,a,h} f^2 - (\mathbb{P}_{s,a,h} f)^2$.
$r_h(s, a)$	Reward associated with the state-action-step triple (s, a, h).
s_h^k, a_h^k	State and action encountered at step h of episode k.
π^k	Policy executed in episode k.
V_h^π, Q_h^π	State-value and action-value functions under policy π.
$V_h^\star, Q_h^\star$	Optimal state-value and action-value functions.
$\Delta_h(s, a)$	Suboptimality gap $\Delta_h(s, a) := V_h^\star(s) - Q_h^\star(s, a)$.
$\Delta_{\min,h}$	Minimum positive suboptimality gap among state-action pairs at step h.
$\Delta_{\min}$	Minimum positive suboptimality gap among all state-action-step triples.
$Z_{\text{opt},h}$	Set of optimal state-action pairs at step h: $Z_{\text{opt},h} := \{(s, a) \in \mathcal{S} \times \mathcal{A} : \Delta_h(s, a) = 0\}$.
Z_{opt}	Set of optimal state-action-step triples: $Z_{\text{opt}} := \{(s, a, h) : \Delta_h(s, a) = 0\}$.
Z_{opt}^c	Set of suboptimal state-action-step triples.
$\mathbb{Q}_h^\star$	Maximum conditional variance of the optimal next-step value: $\mathbb{Q}_h^\star := \max_{s,a} \mathbb{V}_{s,a,h}(V_{h+1}^\star)$.

Table 2: Algorithmic quantities and stage-related notation.

Notation	Meaning
$Q_h^k(s, a), V_h^k(s)$	Values of the algorithmic estimates $Q_h(s, a)$ and $V_h(s)$ at the beginning of episode k.
$\pi_h^k(s)$	Greedy action induced by Q_h^k: $\pi_h^k(s) \in \arg\max_{a \in \mathcal{A}} Q_h^k(s, a)$.
$\mu_h(s, a)$	Stage-wise accumulator of the next-step value estimates.
$\sigma_h(s, a)$	Stage-wise accumulator of the squared next-step value estimates.
$t_h^k(s, a)$	Index of the stage associated with triple (s, a, h) at episode k.
$N_h^k(s, a)$	Total number of visits to (s, a, h) in the first $t_h^k(s, a) - 1$ completed stages for $k \in [K]$.
$N_h^{K+1}(s, a)$	Total number of visits to (s, a, h) over the first K episodes.
$n_h^k(s, a)$	Number of visits to (s, a, h) in the stage $t_h^k(s, a) - 1$.
$W_h^k(s, a)$	Stage-wise empirical variance used to construct $Q_h^k(s, a)$.
$b_h^k(s, a)$	Exploration bonus associated with the update producing $Q_h^k(s, a)$.
ι	Logarithmic confidence factor, typically $\iota = \log(2SAT/\delta)$.
e_j	Length of stage j, defined recursively by $e_1 = H, e_{j+1} = \left\lfloor \left(1 + \frac{1}{H}\right) e_j \right\rfloor$.
$\mathcal{L}$	Set of stage endpoints: $\mathcal{L} := \left\{ \sum_{j=1}^m e_j : m \in \mathbb{N}_+ \right\}$.
$e_{t_h^k(s,a)}$	Length of the stage indexed by $t_h^k(s, a)$.
$l_i(s, a, h, k)$	The index of the i-th visit to the state-action-step triple (s, a, h) within stage $t_h^k(s, a) - 1$.
$k_{l_i(s,a,h,k)}$	Episode in which the sample indexed by $l_i(s, a, h, k)$ was collected.

Table 3: Pseudo value functions and empirical operators used in the variance analysis.

Notation	Meaning
β	An analysis parameter in $(0, H]$ used to truncate the value-estimation error. It is not an algorithmic tuning parameter.
$\hat{V}_h^k(s)$	Pseudo value function $\hat{V}_h^k(s) := \max\{V_h^\star(s), \min\{V_h^\star(s)+\beta, V_h^k(s)\}\}$.
$\hat{\mathbb{E}}_{h,s,a,k}[f]$	Empirical average of a function f over the $n_h^k(s,a)$ samples in the relevant stage: $\hat{\mathbb{E}}_{h,s,a,k}[f] := \frac{1}{n_h^k(s,a)} \sum_{i=1}^{n_h^k(s,a)} f^{k_{l_i(s,a,h,k)}}\left(s_{h+1}^{k_{l_i(s,a,h,k)}}\right)$.
$\hat{\mathbb{V}}_{h,s,a,k}[f]$	Corresponding stage-wise empirical variance: $\hat{\mathbb{V}}_{h,s,a,k}[f] := \hat{\mathbb{E}}_{h,s,a,k}[f^2] - \left(\hat{\mathbb{E}}_{h,s,a,k}[f]\right)^2$.
$\hat{\mathbb{E}}[f]$, $\hat{\mathbb{V}}[f]$	Shorthand for $\hat{\mathbb{E}}_{h,s,a,k}[f]$ and $\hat{\mathbb{V}}_{h,s,a,k}[f]$ when (h,s,a,k) is clear from the context.

Table 4: Recursive weights for propagating cumulative weighted estimation errors across steps.

Notation	Meaning
$\{\omega_h^k\}_{k=1}^K$	A nonnegative initial weight sequence assigned to the estimation errors at the starting step h.
$\omega_h(k', h)$	Initial propagated weight at step h, defined by $\omega_h(k', h) := \omega_h^{k'}$.
$\omega_h(k', h'+1)$	Weight assigned to the estimation error at episode k' and step $h'+1$ after recursively propagating the weights from step h'. It captures the contribution of the episode-k', step-$(h'+1)$ estimation error when recursively bounding the cumulative weighted estimation error from step h' to step $h'+1$.
$\omega_h(k', h', s, a)$	State-action-specific propagated weight: $\omega_h(k', h', s, a) := \omega_h(k', h')\mathbb{I}[(s_{h'}^{k'}, a_{h'}^{k'}) = (s, a)]$. It isolates the contribution associated with state-action pair (s, a) at step h'.
$\tilde{\omega}_h(k', h'+1, s, a)$	The state-action-specific propagated weight assigned to the episode-k', step-$(h'+1)$ estimation error when bounding the cumulative weighted estimation error associated with (s, a) at step h': $\tilde{\omega}_h(k', h'+1, s, a) := \omega_h(k', h'+1)\mathbb{I}[(s_{h'}^{k'}, a_{h'}^{k'}) = (s, a)]$.
$\|\omega_h(\cdot, h')\|_1$	The ℓ_1-norm of the propagated weights at step h': $\|\omega_h(\cdot, h')\|_1 := \sum_{k'=1}^K \omega_h(k', h')$.
$\|\omega_h(\cdot, h')\|_\infty$	The ℓ_∞-norm of the propagated weights at step h': $\|\omega_h(\cdot, h')\|_\infty := \max_{k' \in [K]} \omega_h(k', h')$.
$\|\omega_h(\cdot, h', s, a)\|_1$	The state-action-specific ℓ_1-norm of the propagated weights at step h': $\|\omega_h(\cdot, h', s, a)\|_1 := \sum_{k'=1}^K \omega_h(k', h', s, a)$.
$\|\omega_h(\cdot, h', s, a)\|_\infty$	The state-action-specific ℓ_∞-norm of the propagated weights at step h': $\|\omega_h(\cdot, h', s, a)\|_\infty := \max_{k' \in [K]} \omega_h(k', h', s, a)$.

D Auxiliary Lemmas

Lemma D.1. (Azuma-Hoeffding inequality) *Suppose $\{X_k\}_{k=0}^\infty$ is a martingale and $|X_k - X_{k-1}| \le c_k$, $\forall k \in \mathbb{N}_+$ almost surely. Then for any positive integer N and any $\epsilon > 0$, it holds that:*

$$\mathbb{P}\left(|X_N - X_0| \ge \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^N c_k^2}\right).$$

Lemma D.2. (Freedman's inequality, Theorem EC.1 in Li et al. (2024).) *Consider a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, and let $\mathbb{E}_k$ stand for the expectation conditioned on $\mathcal{F}_k$. Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence adapted to $\{\mathcal{F}_k\}$ and obeying*

$$|X_k| \le R \quad \text{and} \quad \mathbb{E}_{k-1}[X_k] = 0 \quad \text{for all } k \ge 1$$

for some quantity $R < \infty$. We also define $W_n := \sum_{k=1}^n \mathbb{E}_{k-1}[X_k^2]$. In addition, suppose that $W_n \le \sigma^2$ holds deterministically for some given quantity $\sigma^2 < \infty$. Then for any positive integer

$m \geq 1$, *with probability at least* $1-\delta$, *one has*

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

E BASIC PROPERTIES OF UCB-BERNSTEIN+

Let $Q_h^k(s,a)$ and $V_h^k(s)$ denote the values of $Q_h(s,a)$ and $V_h(s)$ at the beginning of the k-th episode, respectively. We note that the definition of stages is with respect to each triple (s,a,h). For every (s,a,h,k), define $t_h^k(s,a)$ as the unique stage index $j \geq 1$ such that the number of visits to (s,a,h) before episode k lies in the interval $[\sum_{i=1}^{j-1} e_i, \sum_{i=1}^{j} e_i)$. Thus, if $(s_h^k, a_h^k) = (s,a)$, the sample collected at step h of episode k belongs to stage $t_h^k(s,a)$. When there is no ambiguity, we write t_h^k for short.

Let $N_h^k(s,a)$ denote the total number of visits to (s,a,h) in the first $t_h^k(s,a)-1$ completed stages, and let $n_h^k(s,a)$ denote the total number of visits to (s,a,h) in stage $t_h^k(s,a)-1$. Separately, for notational convenience, $N_h^{K+1}(s,a)$ denotes the total number of visits to (s,a,h) over the first K episodes. We also let $l_i(s,a,h,k)$, $i \in [n_h^k(s,a)]$, denote the i-th visit to (s,a,h) during stage $t_h^k(s,a)-1$, and let $k_{l_i(s,a,h,k)}$ be the corresponding episode index of the visit $l_i(s,a,h,k)$. When (s,a,h,k) is clear from the context, we use l_i and k_{l_i} for simplicity.

Throughout the appendices, whenever a summand contains the indicator $\mathbb{I}[t_h^k(s,a) > 1]$ together with quantities associated with the preceding stage, such as $n_h^k(s,a)$ or k_{l_i}, the summand is understood to be evaluated only when $t_h^k(s,a) > 1$ and to be zero otherwise. Equivalently, such sums may be restricted to indices satisfying $t_h^k(s,a) > 1$.

For any given step h and non-negative weight sequence $\{\omega_h^k\}_{k=1}^K$, we define the following weights for any $k' \in [K],\ h \leq h' < H$:

$$\begin{aligned}
&\omega_h(k',h) := \omega_h^{k'};\\
&\omega_h(k',h'+1) := \sum_{k=1}^{K} \omega_h(k,h') \frac{\sum_{i=1}^{n_{h'}^k(s_{h'}^k,a_{h'}^k)} \mathbb{I}\left[k_{l_i(s_{h'}^k,a_{h'}^k,h',k)} = k', t_{h'}^k(s_{h'}^k,a_{h'}^k) > 1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}
\end{aligned} \tag{9}$$

with the norms

$$\|\omega_h(\cdot,h')\|_\infty := \max_{k' \in [K]} \omega_h(k',h'),\ \|\omega_h(\cdot,h')\|_1 := \sum_{k'=1}^{K} \omega_h(k',h'). \tag{10}$$

The weight $\omega_h(k',h'+1)$ captures the contribution of the term $(Q_{h'+1}^{k'} - Q_{h'+1}^\star)(s_{h'+1}^{k'}, a_{h'+1}^{k'})$ when recursively bounding the cumulative weighted estimation error from step h' to $h'+1$ as shown in the second conclusion of Lemma H.1 later.

For each state-action pair (s,a), we define the state-action specific weight at any step $h \leq h' \leq H$ as $\omega_h(k',h',s,a) := \omega_h(k',h') \cdot \mathbb{I}[(s_{h'}^{k'}, a_{h'}^{k'}) = (s,a)]$ with the corresponding norms given by

$$\|\omega_h(\cdot,h',s,a)\|_\infty := \max_{k' \in [K]} \omega_h(k',h',s,a),\ \|\omega_h(\cdot,h',s,a)\|_1 := \sum\nolimits_{k'=1}^{K} \omega_h(k',h',s,a).$$

Additionally, for any state-action pair (s,a), we also define

$$\tilde{\omega}_h(k',h'+1,s,a) = \omega_h(k',h'+1)\mathbb{I}[(s_{h'}^{k'}, a_{h'}^{k'}) = (s,a)].$$

The weight $\tilde{\omega}_h(k',h'+1,s,a)$ characterizes the weight of the term $(Q_{h'+1}^{k'} - Q_{h'+1}^\star)(s_{h'+1}^{k'}, a_{h'+1}^{k'})$ when bounding the cumulative weighted estimation error for each state-action pair (s,a) at step h' as shown in the first conclusion of Lemma H.1 later. For convenience, we set $\omega_h(k,H+1) = 0$ and $\widetilde{\omega}_h(k,H+1,s,a) = 0$ for all $k \in [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Lemma E.1. *For any given non-negative weight sequence* $\{\omega_h^k\}_{k\in[K]}$ *at step* h, *the following relationships hold for any* $k' \in [K]$ *and* $h \leq h' < H$:

(a) $\sum_{s,a} \tilde{\omega}_h(k', h'+1, s, a) = \omega_h(k', h'+1)$.

(b) $\|\omega_h(\cdot, h', s, a)\|_\infty \le \|\omega_h(\cdot, h')\|_\infty$.

(c) $\|\omega_h(\cdot, h', s, a)\|_1 \le \|\omega_h(\cdot, h')\|_\infty N_{h'}^{K+1}(s, a)$.

(d) $\|\omega_h(\cdot, h')\|_1 = \sum_{s,a} \|\omega_h(\cdot, h', s, a)\|_1$.

(e) $\|\omega_h(\cdot, h'+1)\|_\infty \le \left(1 + \frac{1}{H}\right) \|\omega_h(\cdot, h')\|_\infty$, $\|\omega_h(\cdot, h'+1)\|_1 \le \|\omega_h(\cdot, h')\|_1$.

Proof of Lemma E.1. (a) is because

$$\sum_{s,a} \tilde{\omega}_h(k', h'+1, s, a) = \sum_{s,a} \omega_h(k', h'+1) \mathbb{I}[(s_{h'}^{k'}, a_{h'}^{k'}) = (s, a)] = \omega_h(k', h'+1).$$

(b) is because for any $k' \in [K]$

$$\omega_h(k', h', s, a) = \omega_h(k', h') \cdot \mathbb{I}\left[(s_{h'}^{k'}, a_{h'}^{k'}) = (s, a)\right] \le \omega_h(k', h') \le \|\omega_h(\cdot, h')\|_\infty.$$

(c) is because

$$\|\omega_h(\cdot, h', s, a)\|_1 \le \|\omega_h(\cdot, h')\|_\infty \sum_{k'=1}^{K} \mathbb{I}\left[(s_{h'}^{k'}, a_{h'}^{k'}) = (s, a)\right] = \|\omega_h(\cdot, h')\|_\infty N_{h'}^{K+1}(s, a).$$

(d) is because

$$\sum_{s,a} \|\omega_h(\cdot, h', s, a)\|_1 = \sum_{s,a} \sum_{k'=1}^{K} \omega_h(k', h') \cdot \mathbb{I}[(s_{h'}^{k'}, a_{h'}^{k'}) = (s, a)] = \sum_{k'=1}^{K} \omega_h(k', h') = \|\omega_h(\cdot, h')\|_1.$$

For (e), note that $\sum_{i=1}^{n_{h'}^k} \mathbb{I}\left[k_{l_i} = k', t_{h'}^k > 1\right] = 1$ if and only if $(s_{h'}^k, a_{h'}^k) = (s_{h'}^{k'}, a_{h'}^{k'})$ and $1 < t_{h'}^k(s_{h'}^k, a_{h'}^k) = t_{h'}^{k'}(s_{h'}^k, a_{h'}^k) + 1$. In this case, we have $t_{h'}^{k'}(s_{h'}^k, a_{h'}^k) > 0$ and $n_{h'}^k(s_{h'}^k, a_{h'}^k) = e_{t_{h'}^{k'}}$. Then for any given k', it holds that:

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{i=1}^{n_{h'}^k} \mathbb{I}\left[k_{l_i} = k', t_{h'}^k > 1\right] \\
&= \sum_{k=1}^{K} \mathbb{I}\left[(s_{h'}^k, a_{h'}^k) = (s_{h'}^{k'}, a_{h'}^{k'}), t_{h'}^k(s_{h'}^k, a_{h'}^k) = t_{h'}^{k'}(s_{h'}^k, a_{h'}^k) + 1\right] \le e_{t_{h'}^{k'}+1}.
\end{aligned}$$

Therefore, we have:

$$\sum_{k=1}^{K} \frac{\sum_{i=1}^{n_{h'}^k} \mathbb{I}[k_{l_i} = k', t_{h'}^k > 1]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)} \le \frac{e_{t_{h'}^{k'}+1}}{e_{t_{h'}^{k'}}} \le 1 + \frac{1}{H}. \tag{11}$$

Then according to the definition of $\|\omega_h(\cdot, h'+1)\|_\infty$, we have:

$$\|\omega_h(\cdot, h'+1)\|_\infty \le \|\omega_h(\cdot, h')\|_\infty \max_{k' \in [K]} \sum_{k=1}^{K} \frac{\sum_{i=1}^{n_{h'}^k} \mathbb{I}[k_{l_i} = k', t_{h'}^k > 1]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)} \le \left(1 + \frac{1}{H}\right) \|\omega_h(\cdot, h')\|_\infty.$$

We also have:

$$\begin{aligned}
\|\omega_h(\cdot, h'+1)\|_1 &= \sum_{k'=1}^{K} \sum_{k=1}^{K} \omega_h(k, h') \sum_{i=1}^{n_{h'}^k} \frac{\mathbb{I}[k_{l_i} = k', t_{h'}^k > 1]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)} \\
&= \sum_{k=1}^{K} \omega_h(k, h') \sum_{i=1}^{n_{h'}^k} \sum_{k'=1}^{K} \frac{\mathbb{I}[k_{l_i} = k', t_{h'}^k > 1]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)} \\
&\le \|\omega_h(\cdot, h')\|_1.
\end{aligned}$$

□

Lemma E.2. *For any non-negative weight sequence $\{\omega_h^k\}_k$ at step $h \in [H]$, any subsequent step $h' \in [h, H]$, any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, and any $\alpha \in (0, 1)$, it holds that:*

$$\sum_{k=1}^{K} \frac{\omega_h(k, h', s, a)\mathbb{I}\left[t_{h'}^k(s_{h'}^k, a_{h'}^k) > 1\right]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)^\alpha} \leq \frac{(4H)^\alpha}{1-\alpha}\|\omega_h(\cdot, h')\|_\infty^\alpha \|\omega_h(\cdot, h', s, a)\|_1^{1-\alpha},$$

and

$$\sum_{k=1}^{K} \frac{\omega_h(k, h')\mathbb{I}\left[t_{h'}^k(s_{h'}^k, a_{h'}^k) > 1\right]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)^\alpha} \leq \frac{(4H)^\alpha}{1-\alpha}(SA\|\omega_h(\cdot, h')\|_\infty)^\alpha \|\omega_h(\cdot, h')\|_1^{1-\alpha},$$

In particular, for $\alpha = 1$ and any $h \in [H]$, we have the following conclusion:

$$\sum_{k=1}^{K} \frac{\mathbb{I}\left[t_h^k(s_h^k, a_h^k) > 1\right]}{n_h^k(s_h^k, a_h^k)} \leq 10HSA\iota.$$

Proof of Lemma E.2. We first note that

$$\sum_{k=1}^{K} \frac{\omega_h(k, h', s, a)\mathbb{I}\left[t_{h'}^k(s_{h'}^k, a_{h'}^k) > 1\right]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)^\alpha} = \sum_{k=1}^{K} \frac{\omega_h(k, h')\mathbb{I}[t_{h'}^k(s, a) > 1, (s_{h'}^k, a_{h'}^k) = (s, a)]}{n_{h'}^k(s, a)^\alpha}. \quad (12)$$

Then we have

$$\Omega_h(s, a) := \sum_{k=1}^{K} \omega_h(k, h')\mathbb{I}\left[t_{h'}^k(s, a) > 1\right]\mathbb{I}[(s_{h'}^k, a_{h'}^k) = (s, a)] \leq \|\omega_h(\cdot, h', s, a)\|_1.$$

If $\Omega_h(s, a) = 0$, the left-hand side of Equation (12) is zero, and the desired bounds hold trivially. Hence, in the remainder of the proof, we assume $\Omega_h(s, a) > 0$.

Given the term on RHS of Equation (12), when the weights $\omega_h(k, h')\mathbb{I}\left[t_{h'}^k(s, a) > 1\right]\mathbb{I}[(s_{h'}^k, a_{h'}^k) = (s, a)]$ concentrate on the earlier terms with smaller index $k \geq 1$, we can obtain the largest value. Let

$$c_{s,a,h'} = \max\left\{c : \|\omega_h(\cdot, h')\|_\infty \sum_{i=2}^{c} e_i \leq \Omega_h(s, a)\right\},\ d_{s,a,h'} = \Omega_h(s, a) - \|\omega_h(\cdot, h')\|_\infty \sum_{i=2}^{c_{s,a,h'}} e_i.$$

Then we have:

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{\omega_h(k, h', s, a)\mathbb{I}\left[t_{h'}^k(s_{h'}^k, a_{h'}^k) > 1\right]}{n_{h'}^k(s_{h'}^k, a_{h'}^k)^\alpha} \\
&\leq \sum_{i=1}^{c_{s,a,h'}-1} \frac{e_{i+1}\|\omega_h(\cdot, h')\|_\infty}{e_i^\alpha} + \frac{d_{s,a,h'}}{e_{c_{s,a,h'}}^\alpha} \\
&\leq (4H)^\alpha \left(\sum_{i=1}^{c_{s,a,h'}-1} \frac{e_{i+1}\|\omega_h(\cdot, h')\|_\infty}{\left(\sum_{j=2}^{i+1} e_j\right)^\alpha} + \frac{d_{s,a,h'}}{\left(\sum_{i=2}^{c_{s,a,h'}+1} e_i\right)^\alpha}\right) \quad (13) \\
&\leq (4H)^\alpha \left(\|\omega_h(\cdot, h')\|_\infty \sum_{i=1}^{c_{s,a,h'}-1} \frac{(\sum_{j=2}^{i+1} e_j)^{1-\alpha} - (\sum_{j=2}^{i} e_j)^{1-\alpha}}{1-\alpha} + \frac{d_{s,a,h'}}{\left(\sum_{i=2}^{c_{s,a,h'}+1} e_i\right)^\alpha}\right) \quad (14) \\
&= (4H)^\alpha \left(\frac{\|\omega_h(\cdot, h')\|_\infty \left(\sum_{i=2}^{c_{s,a,h'}} e_i\right)^{1-\alpha}}{1-\alpha} + \frac{d_{s,a,h'}}{\left(\sum_{i=2}^{c_{s,a,h'}+1} e_i\right)^\alpha}\right) \\
&= (4H)^\alpha \|\omega_h(\cdot, h')\|_\infty^\alpha \left(\frac{\left(\|\omega_h(\cdot, h')\|_\infty \sum_{i=2}^{c_{s,a,h'}} e_i\right)^{1-\alpha}}{1-\alpha} + \frac{d_{s,a,h'}}{\left(\|\omega_h(\cdot, h')\|_\infty \sum_{i=2}^{c_{s,a,h'}+1} e_i\right)^\alpha}\right) \\
&\leq (4H)^\alpha \|\omega_h(\cdot, h')\|_\infty^\alpha \left(\frac{\left(\|\omega_h(\cdot, h')\|_\infty \sum_{i=2}^{c_{s,a,h'}} e_i\right)^{1-\alpha}}{1-\alpha} + \frac{d_{s,a,h'}}{\Omega_h(s, a)^\alpha}\right). \quad (15)
\end{aligned}$$

Here the inequality in Equation (13) is because for any $i \in \mathbb{N}_+$, $\sum_{j=2}^{i+1} e_j \leq 4He_i$. The inequality in Equation (14) is because for any $0 < y < x$ and $\alpha \in (0,1)$, we have:

$$\frac{x-y}{x^\alpha} \leq \frac{1}{1-\alpha}(x^{1-\alpha} - y^{1-\alpha}). \tag{16}$$

Then, let $x = \sum_{j=2}^{i+1} e_j$ and $y = \sum_{j=2}^{i} e_j$, it holds that:

$$\frac{e_{i+1}}{\left(\sum_{j=2}^{i+1} e_j\right)^\alpha} \leq \frac{1}{1-\alpha}\left(\left(\sum_{j=2}^{i+1} e_j\right)^{1-\alpha} - \left(\sum_{j=2}^{i} e_j\right)^{1-\alpha}\right).$$

The last inequality in Equation (15) is because $\|\omega_h(\cdot,h')\|_\infty \sum_{i=2}^{c_{s,a,h'}+1} e_i \geq \Omega_h(s,a)$.

For Equation (15), let $x = \Omega_h(s,a)$ and $y = \|\omega_h(\cdot,h')\|_\infty \sum_{i=2}^{c_{s,a,h'}} e_i$ in Equation (16), we have:

$$\frac{d_{s,a,h'}}{\Omega_h(s,a)^\alpha} + \frac{\left(\|\omega_h(\cdot,h')\|_\infty \sum_{i=2}^{c_{s,a,h'}} e_i\right)^{1-\alpha}}{1-\alpha} \leq \frac{\Omega_h(s,a)^{1-\alpha}}{1-\alpha} \leq \frac{\|\omega_h(\cdot,h',s,a)\|_1^{1-\alpha}}{1-\alpha}.$$

Applying this inequality to Equation (15), we have:

$$\sum_{k=1}^{K} \frac{\omega_h(k,h',s,a)\mathbb{I}\left[t_{h'}^k(s_{h'}^k,a_{h'}^k) > 1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)^\alpha} \leq \frac{(4H)^\alpha}{1-\alpha}\|\omega_h(\cdot,h')\|_\infty^\alpha \|\omega_h(\cdot,h',s,a)\|_1^{1-\alpha}.$$

Therefore, we have proved the first conclusion. By summing this conclusion for all state-action pairs (s,a), we reach:

$$\begin{aligned}\sum_{k=1}^{K} \frac{\omega_h(k,h')\mathbb{I}\left[t_{h'}^k(s_{h'}^k,a_{h'}^k) > 1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)^\alpha} &\leq \sum_{s,a} \frac{(4H)^\alpha}{1-\alpha}\|\omega_h(\cdot,h')\|_\infty^\alpha \|\omega_h(\cdot,h',s,a)\|_1^{1-\alpha} \\ &\leq \frac{(4H)^\alpha}{1-\alpha}(SA\|\omega_h(\cdot,h')\|_\infty)^\alpha \|\omega_h(\cdot,h')\|_1^{1-\alpha}.\end{aligned}$$

The last inequality is by Hölder's inequality, as

$$\sum_{s,a} \|\omega_h(\cdot,h',s,a)\|_1^{1-\alpha} \leq (SA)^\alpha \|\omega_h(\cdot,h')\|_1^{1-\alpha}.$$

For $\alpha = 1$, it holds that:

$$\begin{aligned}\sum_{k=1}^{K} \frac{\mathbb{I}\left[t_h^k(s_h^k,a_h^k) > 1\right]}{n_h^k(s_h^k,a_h^k)} &= \sum_{s,a}\sum_{k=1}^{K} \frac{\mathbb{I}\left[t_h^k(s,a) > 1, (s_h^k,a_h^k) = (s,a)\right]}{n_h^k(s,a)} \\ &\leq \sum_{s,a}\sum_{i=1}^{t_h^K - 1} \frac{e_{i+1}}{e_i} \leq 2\sum_{s,a}(t_h^K(s,a) - 1).\end{aligned} \tag{17}$$

The last inequality is because $e_{i+1} \leq 2e_i$ by the definition of the sequence $\{e_i\}_{i\in\mathbb{N}_+}$. Since at each stage of any given state-action pair (s,a), the number of visits increases by at least $(1+1/2H)$ times, for any $t_h^K(s,a) \geq 1$, the total number of visits to (s,a) is at least

$$\left[H\left(1+\frac{1}{2H}\right)^{t_h^K(s,a)-1} - H\right]$$

Therefore, for any step $h \in [H]$, we have:

$$\sum_{s,a}\left[H\left(1+\frac{1}{2H}\right)^{t_h^K(s,a)-1} - H\right] \leq K$$

By Jensen's inequality, we have

$$K + HSA \geq H \sum_{s,a} \left(1 + \frac{1}{2H}\right)^{t_h^K(s,a)-1} \geq HSA \left(1 + \frac{1}{2H}\right)^{\sum_{s,a} t_h^K(s,a)/SA - 1}$$

Solving the inequality yields

$$\sum_{s,a} t_h^K(s,a) \leq SA \frac{\log(K/(HSA)+1)}{\log\left(1+\frac{1}{2H}\right)} + SA \leq 4HSA \log\left(\frac{K}{HSA}+1\right) + SA \leq 5HSA\iota. \quad (18)$$

The second inequality is because

$$\log\left(1+\frac{1}{2H}\right) \geq \frac{1}{4H}.$$

Plugging Equation (18) into Equation (17), we finish the proof for the conclusion when $\alpha = 1$. □

F Probability Events

Let $\mathcal{X} = (\mathcal{S}, \mathcal{A}, H, T, \iota)$. The notation $f(\mathcal{X}) \lesssim g(\mathcal{X})$ means that there exists a universal constant $C > 0$ such that $f(\mathcal{X}) \leq Cg(\mathcal{X})$. We first provide a lemma of concentration inequalities.

Lemma F.1. *Let $\iota = \log(2SAT/\delta)$ with $\delta \in (0,1)$. We use $\forall(s,a,h,k)$ as the simplified notation for $\forall(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$. Whenever $n_h^k(s,a)$ appears in a denominator, the corresponding statement is restricted to tuples satisfying $t_h^k(s,a) > 1$, for which $n_h^k(s,a) > 0$. For any function $f : \mathcal{S} \to \mathbb{R}$, we denote $\mathbb{V}_{s,a,h}(f) = \mathbb{P}_{s,a,h} f^2 - (\mathbb{P}_{s,a,h} f)^2$. Then we have the following conclusions.*

(a) With probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_1 = \left\{ \left| \frac{1}{n_h^k} \sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^{k_{l_i}} \right| \leq \frac{5}{n_h^k} \sqrt{\sum_{i=1}^{n_h^k} \mathbb{V}_{s,a,h}(V_{h+1}^{k_{l_i}})\iota} + \frac{9H\iota}{n_h^k}, \forall(s,a,h,k) \right\}.$$

(b) With probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_2 = \left\{ \left| \frac{1}{n_h^k} \sum_{i=1}^{n_h^k} (\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h})(V_{h+1}^{k_{l_i}})^2 \right| \leq \frac{5}{n_h^k} \sqrt{\sum_{i=1}^{n_h^k} \mathbb{V}_{s,a,h}((V_{h+1}^{k_{l_i}})^2)\iota} + \frac{9H^2\iota}{n_h^k}, \forall(s,a,h,k) \right\}.$$

(c) With probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_3 = \left\{ \left| \frac{\sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^\star}{n_h^k(s,a)} \right| \leq H \sqrt{\frac{2\iota}{n_h^k(s,a)}},\ \forall(s,a,h,k) \right\}.$$

(d) With probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_4 = \left\{ \sum_{h'=h}^{H} \left(1+\frac{1}{H}\right)^{h'-h} \sum_{k=1}^{K} \left(\mathbb{P}_{s_{h'}^k, a_{h'}^k, h'} - \mathbb{1}_{s_{h'+1}^k}\right)(V_{h'+1}^\star - V_{h'+1}^{\pi^k}) \leq 3\sqrt{2H^2T\iota},\ \forall h \in [H] \right\}.$$

(e) With probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_5 = \left\{ \left| \frac{1}{n_h^k} \sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^\star \right| \leq 5\sqrt{\frac{\iota}{n_h^k} \mathbb{V}_{s,a,h}(V_{h+1}^\star)} + \frac{9H\iota}{n_h^k}, \forall(s,a,h,k) \right\}.$$

(f) With probability at least $1-\delta$, the following event holds:

$$\mathcal{E}_6 = \left\{ \frac{\sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) (V_{h+1}^\star)^2}{n_h^k(s,a)} \leq H^2 \sqrt{\frac{2\iota}{n_h^k(s,a)}},\ \forall(s,a,h,k) \right\}.$$

(g) (Lemma C.5 in Jin et al. (2018)) *With probability at least* $1-\delta$*, the following event holds:*

$$\mathcal{E}_7 = \left\{ \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{V}_{s_h^k, a_h^k, h}(V_{h+1}^{\pi^k}) \lesssim HT + H^3\iota \right\}.$$

(h) With probability at least $1-\delta$*, the following event holds:*

$$\mathcal{E}_8 = \left\{ \sum_{h=1}^{H} \sum_{k=1}^{K} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k} \right) \left(V_{h+1}^{\star} - V_{h+1}^{\pi^k} \right) \leq \sqrt{2H^2T\iota} \right\}.$$

Proof of Lemma F.1. (a) By Freedman's inequality in Lemma D.2, for any $\delta \in (0,1)$, with probability at least $1-\delta/SAT^2$, setting $\sigma^2 = TH$, $R = H$, and $m = \max\{1, \lceil \log_2(TH) \rceil\}$, it holds for a given $n_h^k(s,a) = n \in \mathbb{N}_+$ that

$$\left| \sum_{i=1}^{n} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^{k_{l_i}} \right| \leq \sqrt{8 \max\left\{ \sum_{i=1}^{n} \mathbb{V}_{s,a,h}(V_{h+1}^{k_{l_i}}), \frac{TH}{2^m} \right\} \log\left(\frac{2mSAT^2}{\delta} \right)} + \frac{4}{3} H \log\left(\frac{2mSAT^2}{\delta} \right).$$

Since $H \leq T$, we have $m \leq 1 + 2\log_2 T \leq 2^{3/2}\sqrt{T}$. Together with $\iota = \log(2SAT/\delta)$, this implies

$$\log\left(\frac{2mSAT^2}{\delta} \right) = \iota + \log(mT) \leq \frac{5}{2}\iota.$$

Moreover, by the definition of m, $TH/2^m \leq 1$. Therefore, using $H \geq 1$ and $\iota \geq \log 2$, the preceding inequality yields

$$\frac{1}{n} \left| \sum_{i=1}^{n} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^{k_{l_i}} \right| \leq \frac{5}{n} \sqrt{\sum_{i=1}^{n} \mathbb{V}_{s,a,h}(V_{h+1}^{k_{l_i}})\iota} + \frac{9H\iota}{n}.$$

For any $k \in [K]$, we have $n_h^k(s,a) \in [\frac{T}{H}]$. The total number of possible combinations $(s,a,h,k,n) \in \mathcal{S} \times \mathcal{A} \times [H] \times [\frac{T}{H}] \times [\frac{T}{H}]$ is at most SAT^2. Hence, taking a union bound, with probability at least $1-\delta$, it holds simultaneously for all $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ that:

$$\left| \frac{1}{n_h^k} \sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^{k_{l_i}} \right| \leq \frac{5}{n_h^k} \sqrt{\sum_{i=1}^{n_h^k} \mathbb{V}_{s,a,h}(V_{h+1}^{k_{l_i}})\iota} + \frac{9H\iota}{n_h^k}.$$

(b) The proof is similar to (a).

(c) The sequence

$$\left\{ \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^{\star} \right\}_{i \in \mathbb{N}^+}$$

is a martingale-difference sequence bounded by H. Then according to Azuma-Hoeffding inequality, for any $\delta \in (0,1)$, with probability at least $1-\delta/SAT$, it holds for a given $n_h^k(s,a) = n \in \mathbb{N}_+$ that:

$$\frac{1}{n} \left| \sum_{i=1}^{n} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^{\star} \right| \leq H\sqrt{\frac{2\iota}{n}}.$$

For any $k \in [K]$, we have $n_h^k(s,a) \in [\frac{T}{H}]$. Considering all the possible combinations $(s,a,h,n) \in \mathcal{S} \times \mathcal{A} \times [H] \times [\frac{T}{H}]$, with probability at least $1-\delta$, it holds simultaneously for all $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ that:

$$\frac{1}{n_h^k} \left| \sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h} \right) V_{h+1}^{\star} \right| \leq H\sqrt{\frac{2\iota}{n_h^k(s,a)}}.$$

(d) For $\mathcal{E}_4$, fix $h \in [H]$ and order the pairs (k, h'), $k \in [K]$ and $h \le h' \le H$, chronologically, first by episode and then by step. The corresponding martingale-difference sequence

$$\left\{ \left(1+\frac{1}{H}\right)^{h'-h} \left(\mathbb{P}_{s^k_{h'},a^k_{h'},h'} - \mathbb{1}_{s^k_{h'+1}}\right)\left(V^\star_{h'+1} - V^{\pi^k}_{h'+1}\right) \right\}_{(k,h')}$$

is bounded by $3H$. According to Azuma-Hoeffding inequality, for any $\delta \in (0,1)$ and given $h \in [H]$, with probability at least $1-\delta/H$, it holds that:

$$\sum_{h'=h}^{H} \left(1+\frac{1}{H}\right)^{h'-h} \sum_{k=1}^{K} \left(\mathbb{P}_{s^k_{h'},a^k_{h'},h'} - \mathbb{1}_{s^k_{h'+1}}\right)\left(V^\star_{h'+1} - V^{\pi^k}_{h'+1}\right) \le 3\sqrt{2H^2T\iota}.$$

Considering all the possible values of h, we finish the proof.

(e) The proof is similar to (a).

(f) The proof is similar to (c) with

$$\left\{ \left(\mathbb{1}_{s^{k_{l_i}}_{h+1}} - \mathbb{P}_{s,a,h}\right)(V^\star_{h+1})^2 \right\}_{i\in\mathbb{N}^+}$$

being a martingale-difference sequence bounded by H^2.

(h) The conclusion is proved directly by Azuma-Hoeffding inequality.

□

G Properties of Value Function Estimates

In this section, we provide some properties of value function estimates. Throughout the theoretical analysis, we take $C_1 = 8$, $C_2 = 119$, and $C_3 = 2$ in Algorithm 1. Before proceeding, we first introduce some notation.

For any (s,a,h,k) satisfying $t^k_h(s,a) > 1$, let $\tilde{\mu}^k_h(s,a)$ and $\tilde{\sigma}^k_h(s,a)$ denote the values of $\mu_h(s,a)$ and $\sigma_h(s,a)$ at the end of stage $t^k_h(s,a)-1$, respectively. More specifically,

$$\tilde{\mu}^k_h(s,a) = \sum_{i=1}^{n^k_h} V^{k_{l_i}}_{h+1}(s^{k_{l_i}}_{h+1}),\ \tilde{\sigma}^k_h(s,a) = \sum_{i=1}^{n^k_h} \left(V^{k_{l_i}}_{h+1}(s^{k_{l_i}}_{h+1})\right)^2.$$

We also denote

$$\tilde{b}^{k+1}_h(s,a) = \min\left\{ 8\sqrt{\frac{\tilde{W}^{k+1}_h(s,a)\iota}{n^{k+1}_h(s,a)}} + \frac{119H\iota}{n^{k+1}_h(s,a)}, 2\sqrt{\frac{H^2\iota}{n^{k+1}_h(s,a)}} \right\}, \tag{19}$$

with

$$\tilde{W}^k_h(s,a) = \frac{\tilde{\sigma}^k_h(s,a)}{n^k_h(s,a)} - \left(\frac{\tilde{\mu}^k_h(s,a)}{n^k_h(s,a)}\right)^2. \tag{20}$$

For $t^k_h(s,a) > 1$, define

$$\widetilde{Q}^k_h(s,a) := r_h(s,a) + \frac{\widetilde{\mu}^k_h(s,a)}{n^k_h(s,a)} + \widetilde{b}^k_h(s,a).$$

Then by induction and the update rule in Equation (3), it holds that

$$\begin{aligned} Q^{k+1}_h(s,a) &= \mathbb{I}\left[t^{k+1}_h(s,a)=1\right]H \\ &\quad + \mathbb{I}\left[t^{k+1}_h(s,a)>1\right]\min\left\{ r_h(s,a) + \frac{\tilde{\mu}^{k+1}_h(s,a)}{n^{k+1}_h(s,a)} + \tilde{b}^{k+1}_h(s,a), Q^k_h(s,a)\right\} \\ &= \mathbb{I}[t^{k+1}_h = 1]H + \mathbb{I}[t^{k+1}_h > 1]\min\{\widetilde{Q}^{k+1}_h(s,a), Q^k_h(s,a)\}. \end{aligned} \tag{21}$$

We then prove the optimism properties of the Q-estimates.

Lemma G.1. *Under the event $\bigcap_{i=1}^{3}\mathcal{E}_i$ in Lemma F.1, it holds that for any $(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$:*

$$Q_h^k(s,a)\geq Q_h^\star(s,a).$$

Then we have $V_h^k(s)\geq V_h^\star(s)$ for any $(s,h,k)\in\mathcal{S}\times[H]\times[K]$.

Proof of Lemma G.1. We use mathematical induction on k to prove $Q_h^k(s,a)\geq Q_h^\star(s,a)$ for any $(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$. For $k=1$, $Q_h^1(s,a)=H\geq Q_h^\star(s,a)$ for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$. For $k\geq 2$, assume we already have $Q_h^{k'}(s,a)\geq Q_h^\star(s,a)$ for any $(s,a,h,k')\in\mathcal{S}\times\mathcal{A}\times[H]\times[k-1]$, then we will prove for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, $Q_h^k(s,a)\geq Q_h^\star(s,a)$. According to Equation (21), the following relationship holds:

$$Q_h^k(s,a)=\mathbb{I}\left[t_h^k=1\right]H+\mathbb{I}\left[t_h^k>1\right]\min\{\tilde{Q}_h^k(s,a),Q_h^{k-1}(s,a)\}.$$

Then for any given $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, we have the following three cases:

(a) If $t_h^k(s,a)=1$, then $Q_h^k(s,a)=H\geq Q_h^\star(s,a)$.

(b) If $t_h^k(s,a)>1$ and $Q_h^k(s,a)=Q_h^{k-1}(s,a)$, then the conclusion holds.

(c) If $t_h^k(s,a)>1$ and $Q_h^k(s,a)=r_h(s,a)+\tilde{\mu}_h^k(s,a)/n_h^k(s,a)+\tilde{b}_h^k(s,a)$, then by Equation (1), we have the following equality:

$$Q_h^\star(s,a)=r_h(s,a)+\mathbb{P}_{s,a,h}V_{h+1}^\star.$$

Then we have:

$$\begin{aligned}Q_h^k(s,a)-Q_h^\star(s,a)&=\tilde{\mu}_h^k(s,a)/n_h^k(s,a)+\tilde{b}_h^k(s,a)-\mathbb{P}_{s,a,h}V_{h+1}^\star\\&=\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(V_{h+1}^{k_{l_i}}(s_{h+1}^{k_{l_i}})-\mathbb{P}_{s,a,h}V_{h+1}^\star\right)+\tilde{b}_h^k(s,a).\end{aligned}\tag{22}$$

According to the definition of $l_i(s,a,h,k)$, we know $k_{l_i}<k$ for $i\in[n_h^k(s,a)]$. Then $Q_h^{k_{l_i}}(s,a)\geq Q_h^\star(s,a)$ based on the induction. Therefore, according to the update rule Equation (3) and Equation (1), for any $(s,h)\in\mathcal{S}\times[H]$ and any $i\in[n_h^k(s,a)]$, we have:

$$V_{h+1}^{k_{l_i}}(s)=\max_{a\in\mathcal{A}}Q_{h+1}^{k_{l_i}}(s,a)\geq\max_{a\in\mathcal{A}}Q_{h+1}^\star(s,a)=V_{h+1}^\star(s),$$

and thus for any $i\in[n_h^k(s,a)]$ it holds:

$$\mathbb{P}_{s,a,h}V_{h+1}^{k_{l_i}}\geq\mathbb{P}_{s,a,h}V_{h+1}^\star.$$

If the second branch in Equation (19) attains the minimum, then by the pointwise optimism above and the event $\mathcal{E}_3$ in Lemma F.1,

$$\begin{aligned}Q_h^k(s,a)-Q_h^\star(s,a)&\geq\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(\mathbb{1}_{s_{h+1}^{k_{l_i}}}-\mathbb{P}_{s,a,h}\right)V_{h+1}^\star+2H\sqrt{\frac{\iota}{n_h^k}}\\&\geq-H\sqrt{\frac{2\iota}{n_h^k}}+2H\sqrt{\frac{\iota}{n_h^k}}\geq 0.\end{aligned}$$

If the first branch in Equation (19) attains the minimum, applying the inequality above to Equation (22) and using the event $\mathcal{E}_1$ in Lemma F.1, we have

$$\begin{aligned}Q_h^k(s,a)-Q_h^\star(s,a)&\geq\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(V_{h+1}^{k_{l_i}}(s_{h+1}^{k_{l_i}})-\mathbb{P}_{s,a,h}V_{h+1}^{k_{l_i}}\right)+\tilde{b}_h^k(s,a)\\&\geq 8\sqrt{\frac{\tilde{W}_h^k(s,a)\iota}{n_h^k}}-\frac{5}{n_h^k}\sqrt{\sum_{i=1}^{n_h^k}\mathbb{V}_{s,a,h}(V_{h+1}^{k_{l_i}})\iota}+\frac{110H\iota}{n_h^k}.\end{aligned}\tag{23}$$

Next we will bound the difference

$$I_1 = \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}_{s,a,h}(V_{h+1}^{k_{l_i}}) - \tilde{W}_h^k(s,a) \triangleq X - \tilde{W}_h^k(s,a),$$

where

$$X = \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}_{s,a,h}(V_{h+1}^{k_{l_i}}).$$

According to the definition of $\mathbb{V}_{s,a,h}$, we also have

$$\mathbb{V}_{s,a,h}V_{h+1}^{k_{l_i}} = \mathbb{P}_{s,a,h}(V_{h+1}^{k_{l_i}})^2 - \left(\mathbb{P}_{s,a,h}V_{h+1}^{k_{l_i}}\right)^2. \tag{24}$$

Combining the results of Equation (20) and Equation (24), we can decompose the difference I_1 as:

$$I_1 = \sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\left(\mathbb{P}_{s,a,h} - \mathbb{1}_{s_{h+1}^{k_{l_i}}}\right)(V_{h+1}^{k_{l_i}})^2 + \left[\left(\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}V_{h+1}^{k_{l_i}}\left(s_{h+1}^{k_{l_i}}\right)\right)^2 - \sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\left(\mathbb{P}_{s,a,h}V_{h+1}^{k_{l_i}}\right)^2\right]. \tag{25}$$

For the first term, by the event $\mathcal{E}_2$ in Lemma F.1 we have

$$\begin{aligned}
&\left|\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\left(\mathbb{P}_{s,a,h} - \mathbb{1}_{s_{h+1}^{k_{l_i}}}\right)(V_{h+1}^{k_{l_i}})^2\right| && (26)\\
&\le 5\sqrt{\frac{\iota}{n_h^k}\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\mathbb{V}_{s,a,h}((V_{h+1}^{k_{l_i}})^2)} + \frac{9H^2\iota}{n_h^k}\\
&\le 10\sqrt{\frac{H^2\iota}{n_h^k}\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\mathbb{V}_{s,a,h}V_{h+1}^{k_{l_i}}} + \frac{9H^2\iota}{n_h^k}\\
&= 10\sqrt{\frac{H^2\iota}{n_h^k}X} + \frac{9H^2\iota}{n_h^k}. && (27)
\end{aligned}$$

The last inequality is by $\mathbb{V}_{s,a,h}(X^2) \le 4C^2\mathbb{V}_{s,a,h}(X)$ for $|X| \le C$. For the second term of Equation (25), by the Cauchy–Schwarz inequality, we reach

$$\begin{aligned}
&\left(\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}V_{h+1}^{k_{l_i}}\left(s_{h+1}^{k_{l_i}}\right)\right)^2 - \sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\left(\mathbb{P}_{s,a,h}V_{h+1}^{k_{l_i}}\right)^2\\
&\le \left(\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}V_{h+1}^{k_{l_i}}\left(s_{h+1}^{k_{l_i}}\right)\right)^2 - \left(\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\mathbb{P}_{s,a,h}V_{h+1}^{k_{l_i}}\right)^2\\
&\le 2H\left|\sum_{i=1}^{n_h^k}\frac{1}{n_h^k}\left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s,a,h}\right)V_{h+1}^{k_{l_i}}\right|\\
&\le 10\sqrt{\frac{H^2\iota}{n_h^k}X} + \frac{18H^2\iota}{n_h^k}. && (28)
\end{aligned}$$

The last inequality is by the event $\mathcal{E}_1$. Combining Equation (26) and Equation (28), back to Equation (25), we know

$$I_1 = X - \tilde{W}_h^k(s,a) \le 20\sqrt{\frac{H^2\iota}{n_h^k}X} + \frac{27H^2\iota}{n_h^k}.$$

Solving the inequality, we have

$$X \leq 2\tilde{W}_h^k(s,a) + \frac{454H^2\iota}{n_h^k}.$$

Applying this inequality to Equation (23) proves optimism for the first branch. Together with the second-branch argument above, this establishes $Q_h^k(s,a) \geq Q_h^\star(s,a)$. Finally, since

$$V_{h+1}^k(s) = \max_{a\in\mathcal{A}} Q_{h+1}^k(s,a) \geq \max_{a\in\mathcal{A}} Q_{h+1}^\star(s,a) = V_{h+1}^\star(s),$$

the proof is complete. □

The following lemma upper bounds the weighted cumulative estimation errors of $Q_h^k(s,a) - Q_h^\star(s,a)$.

Lemma G.2. *For UCB-Bernstein+, under the event $\bigcap_{i=1}^3 \mathcal{E}_i$ in Lemma F.1, for any given $h \in [H]$ and non-negative weight sequence $\{\omega_h^k = \omega_h(k,h)\}_{k=1}^K$ at step h, we have the following conclusion:*

$$\begin{aligned}
&\sum_{k=1}^K \omega_h(k,h)\left(Q_h^k - Q_h^\star\right)(s_h^k, a_h^k) \\
&\lesssim \sum_{h_1=h}^H \sum_{k=1}^K \omega_h(k,h_1)\mathbb{I}\left[t_{h_1}^k = 1\right] H + \sqrt{H^5SA\|\omega_h(\cdot,h)\|_\infty \|\omega_h(\cdot,h)\|_1 \iota}.
\end{aligned}$$

Proof of Lemma G.2. By Equation (21), we have the following relationship

$$Q_h^k(s_h^k,a_h^k) \leq \mathbb{I}\left[t_h^k(s_h^k,a_h^k)=1\right]H + \mathbb{I}\left[t_h^k(s_h^k,a_h^k)>1\right]\left(r_h(s_h^k,a_h^k) + \frac{\tilde{\mu}_h^k(s_h^k,a_h^k)}{n_h^k(s_h^k,a_h^k)} + \tilde{b}_h^k(s_h^k,a_h^k)\right)$$

and according to the Bellman equality Equation (1), we have:

$$Q_h^\star(s_h^k,a_h^k) = r_h(s_h^k,a_h^k) + \mathbb{P}_{s_h^k,a_h^k,h}V_{h+1}^\star.$$

Combining these two inequalities, it holds for any $h \in [H]$ that:

$$\begin{aligned}
&(Q_h^k - Q_h^\star)(s_h^k,a_h^k) \leq \mathbb{I}[t_h^k=1]H + \\
&\mathbb{I}\left[t_h^k>1\right]\left(\frac{\sum_{i=1}^{n_h^k}(V_{h+1}^{k_{l_i}} - V_{h+1}^\star)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k,a_h^k)} + \frac{\sum_{i=1}^{n_h^k}\left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s_h^k,a_h^k,h}\right)V_{h+1}^\star}{n_h^k(s_h^k,a_h^k)} + \tilde{b}_h^k(s_h^k,a_h^k)\right) \quad (29)\\
&\leq \mathbb{I}\left[t_h^k=1\right]H + \mathbb{I}\left[t_h^k>1\right]\frac{\sum_{i=1}^{n_h^k}\left(V_{h+1}^{k_{l_i}} - V_{h+1}^\star\right)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k,a_h^k)} + 4\mathbb{I}\left[t_h^k>1\right]\sqrt{\frac{H^2\iota}{n_h^k(s_h^k,a_h^k)}}. \quad (30)
\end{aligned}$$

The last inequality is by the definition of $\tilde{b}_h^k$ in Equation (19) and the event $\mathcal{E}_3$ in Lemma F.1. Then for any $h \in [H]$, we have

$$\begin{aligned}
&\sum_{k=1}^K \omega_h(k,h)(Q_h^k - Q_h^\star)(s_h^k,a_h^k) \\
&\leq \sum_{k=1}^K \omega_h(k,h)\mathbb{I}\left[t_h^k=1\right]H + \sum_{k=1}^K \omega_h(k,h)\mathbb{I}\left[t_h^k>1\right]\frac{\sum_{i=1}^{n_h^k}\left(V_{h+1}^{k_{l_i}} - V_{h+1}^\star\right)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k,a_h^k)} \\
&\quad + 4\sum_{k=1}^K \omega_h(k,h)\mathbb{I}\left[t_h^k>1\right]\sqrt{\frac{H^2\iota}{n_h^k(s_h^k,a_h^k)}}. \quad (31)
\end{aligned}$$

For the second term in Equation (31), we have:

$$\begin{aligned}
&\sum_{k=1}^{K}\omega_h(k,h)\mathbb{I}[t_h^k>1]\frac{\sum_{i=1}^{n_h^k}\left(V_{h+1}^{k_{l_i}}-V_{h+1}^{\star}\right)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k,a_h^k)}\\
&=\sum_{k=1}^{K}\sum_{i=1}^{n_h^k}\omega_h(k,h)\frac{\left(V_{h+1}^{k_{l_i}}-V_{h+1}^{\star}\right)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k,a_h^k)}\mathbb{I}[t_h^k>1]\cdot\sum_{k'=1}^{K}\mathbb{I}\left[k_{l_i}=k'\right]\\
&=\sum_{k=1}^{K}\sum_{i=1}^{n_h^k}\sum_{k'=1}^{K}\omega_h(k,h)\frac{\left(V_{h+1}^{k'}-V_{h+1}^{\star}\right)(s_{h+1}^{k'})}{n_h^k(s_h^k,a_h^k)}\mathbb{I}[t_h^k>1]\cdot\mathbb{I}[k_{l_i}=k']\\
&=\sum_{k'=1}^{K}\left(\sum_{k=1}^{K}\omega_h(k,h)\frac{\sum_{i=1}^{n_h^k}\mathbb{I}\left[k_{l_i}=k',t_h^k>1\right]}{n_h^k(s_h^k,a_h^k)}\right)\left(V_{h+1}^{k'}-V_{h+1}^{\star}\right)(s_{h+1}^{k'})\\
&=\sum_{k'=1}^{K}\omega_h(k',h+1)\left(V_{h+1}^{k'}-V_{h+1}^{\star}\right)(s_{h+1}^{k'})\\
&\le\sum_{k'=1}^{K}\omega_h(k',h+1)(Q_{h+1}^{k'}-Q_{h+1}^{\star})(s_{h+1}^{k'},a_{h+1}^{k'}).
\end{aligned}\tag{32}$$

The inequality is by $Q_{h+1}^{k'}(s_{h+1}^{k'},a_{h+1}^{k'})=V_{h+1}^{k'}(s_{h+1}^{k'})$, $Q_{h+1}^{\star}(s_{h+1}^{k'},a_{h+1}^{k'})\le V_{h+1}^{\star}(s_{h+1}^{k'})$.

For the last term in Equation (31), by Lemma E.2 with $\alpha=1/2$, we have:

$$\sum_{k=1}^{K}\omega_h(k,h)\mathbb{I}\left[t_h^k>1\right]\sqrt{\frac{H^2\iota}{n_h^k(s_h^k,a_h^k)}}\le 4\sqrt{H^3SA\|\omega_h(\cdot,h)\|_\infty\|\omega_h(\cdot,h)\|_1\iota}.\tag{33}$$

Based on Equation (31), by combining the results of Equation (32) and Equation (33), we have

$$\begin{aligned}
\sum_{k=1}^{K}\omega_h(k,h)\left(Q_h^k-Q_h^{\star}\right)(s_h^k,a_h^k)\le&\sum_{k'=1}^{K}\omega_h(k',h+1)(Q_{h+1}^{k'}-Q_{h+1}^{\star})(s_{h+1}^{k'},a_{h+1}^{k'})\\
&+\sum_{k=1}^{K}\omega_h(k,h)\mathbb{I}\left[t_h^k=1\right]H+16\sqrt{H^3SA\|\omega_h(\cdot,h)\|_\infty\|\omega_h(\cdot,h)\|_1\iota}.
\end{aligned}\tag{34}$$

Due to the property (e) of Lemma E.1, it holds for any $h_1\in[h,H]$ that:

$$\|\omega_h(\cdot,h_1)\|_1\le\|\omega_h(\cdot,h)\|_1,\ \|\omega_h(\cdot,h_1)\|_\infty\le(1+1/H)^{h_1-h}\|\omega_h(\cdot,h)\|_\infty\le e\|\omega_h(\cdot,h)\|_\infty.$$

Therefore, by recursively applying Equation (34) over steps $h,h+1,\ldots,H$ and using $Q_{H+1}^k(s,a)=Q_{H+1}^{\star}(s,a)=0$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{K}\omega_h(k,h)\left(Q_h^k-Q_h^{\star}\right)(s_h^k,a_h^k)\\
&\le\sum_{h_1=h}^{H}\sum_{k=1}^{K}\omega_h(k,h_1)\mathbb{I}\left[t_{h_1}^k=1\right]H+16\sum_{h_1=h}^{H}\sqrt{H^3SA\|\omega_h(\cdot,h_1)\|_\infty\|\omega_h(\cdot,h_1)\|_1\iota}\\
&\le\sum_{h_1=h}^{H}\sum_{k=1}^{K}\omega_h(k,h_1)\mathbb{I}\left[t_{h_1}^k=1\right]H+16\sqrt{3H^5SA\|\omega_h(\cdot,h)\|_\infty\|\omega_h(\cdot,h)\|_1\iota},
\end{aligned}$$

which proves the lemma. □

With the lemma above, we can further establish the following two conclusions.

Lemma G.3. *Under the event $\bigcap_{i=1}^{3}\mathcal{E}_i$ in Lemma F.1, for all $\epsilon\in(0,H]$, we have the following two conclusions:*

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[Q_h^k(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)\geq\epsilon\right]\lesssim\frac{H^6SA\iota}{\epsilon^2},$$

and

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_h^k-Q_h^\star\right)(s_h^k,a_h^k)\mathbb{I}\left[\left(Q_h^k-Q_h^\star\right)(s_h^k,a_h^k)\geq\epsilon\right]\lesssim\frac{H^6SA\iota}{\epsilon}.$$

Proof of Lemma G.3. Let $N=\max\{1,\lceil\log_2(H/\epsilon)\rceil\}$. For any given $h\in[H]$, any $k\in[K]$ and $i\in[N-1]$, define:

$$\omega_h^i(k,h)=\mathbb{I}\left[(Q_h^k-Q_h^\star)(s_h^k,a_h^k)\in\left[2^{i-1}\epsilon,2^i\epsilon\right)\right]\leq 1,$$

and

$$\omega_h^N(k,h)=\mathbb{I}\left[(Q_h^k-Q_h^\star)(s_h^k,a_h^k)\in\left[2^{N-1}\epsilon,H\right]\right]\leq 1.$$

Then for any $k'\in[K]$ and $h\leq h'\leq H$, we recursively define $\omega_h^i(k',h')$ according to Equation (9):

$$\omega_h^i(k',h'+1):=\sum_{k=1}^{K}\omega_h^i(k,h')\frac{\sum_{i=1}^{n_{h'}^k(s_{h'}^k,a_{h'}^k)}\mathbb{I}\left[k_{l_i(s_{h'}^k,a_{h'}^k,h',k)}=k',t_{h'}^k(s_{h'}^k,a_{h'}^k)>1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}\tag{35}$$

with norms $\|\omega_h^i(\cdot,h')\|_1$ and $\|\omega_h^i(\cdot,h')\|_\infty$ defined accordingly by Equation (10) as:

$$\|\omega_h^i(\cdot,h')\|_\infty:=\max_{k'\in[K]}\omega_h^i(k',h'),\ \|\omega_h^i(\cdot,h')\|_1:=\sum_{k'=1}^{K}\omega_h^i(k',h').\tag{36}$$

Then we have the following relationship:

$$\sum_{k=1}^{K}\omega_h^i(k,h)\left(Q_h^k-Q_h^\star\right)(s_h^k,a_h^k)\geq 2^{i-1}\epsilon\|\omega_h^i(\cdot,h)\|_1.\tag{37}$$

By Lemma G.2, since $\omega_h^i(k,h)\leq 1$, we have

$$\sum_{k=1}^{K}\omega_h^i(k,h)\left(Q_h^k-Q_h^\star\right)(s_h^k,a_h^k)\lesssim\sum_{h_1=h}^{H}\sum_{k=1}^{K}\omega_h^i(k,h_1)\mathbb{I}\left[t_{h_1}^k=1\right]H+\sqrt{H^5SA\|\omega_h^i(\cdot,h)\|_1\iota}.\tag{38}$$

Combining the results of Equation (38) and Equation (37), we have the following inequality:

$$2^{i-1}\epsilon\|\omega_h^i(\cdot,h)\|_1\lesssim\sqrt{H^5SA\|\omega_h^i(\cdot,h)\|_1\iota}+\sum_{h_1=h}^{H}\sum_{k=1}^{K}\omega_h^i(k,h_1)\mathbb{I}\left[t_{h_1}^k=1\right]H.\tag{39}$$

Solving the inequality in Equation (39), we can derive the following relationship:

$$\|\omega_h^i(\cdot,h)\|_1\lesssim\frac{H^5SA\iota}{4^i\epsilon^2}+\frac{\sum_{h_1=h}^{H}\sum_{k=1}^{K}\omega_h^i(k,h_1)\mathbb{I}\left[t_{h_1}^k=1\right]H}{2^i\epsilon}.\tag{40}$$

Note that for any $k'\in[K]$ and $h\leq h'\leq H-1$, by Equation (35), we have

$$\sum_{i=1}^{N}\omega_h^i(k',h'+1)=\sum_{k=1}^{K}\left(\sum_{i=1}^{N}\omega_h^i(k,h')\right)\frac{\sum_{i=1}^{n_{h'}^k}\mathbb{I}\left[k_{l_i}=k',t_{h'}^k>1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}.$$

Then by mathematical induction on $h'\in[h,H]$, it is straightforward to prove that for any $k'\in[K]$,

$$\sum_{i=1}^{N}\omega_h^i(k',h')\leq\left(1+\frac{1}{H}\right)^{h'-h}<3,\tag{41}$$

given Equation (11) and the base case $\sum_{i=1}^{N} \omega_h^i(k', h) \leq 1$.

We now prove the first conclusion. By

$$\mathbb{I}\left[\left(Q_h^k - Q_h^\star\right)(s_h^k, a_h^k) \geq \epsilon\right] = \sum_{i=1}^{N} \omega_h^i(k, h),$$

we have

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{I}\left[Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \geq \epsilon\right] = \sum_{h=1}^{H}\sum_{i=1}^{N} \|\omega_h^i(\cdot, h)\|_1. \tag{42}$$

By Equation (40), it holds that

$$\begin{aligned}
\sum_{i=1}^{N} \|\omega_h^i(\cdot, h)\|_1 &\lesssim \sum_{i=1}^{N} \frac{H^5 SA\iota}{4^i\epsilon^2} + \sum_{i=1}^{N} \frac{\sum_{h_1=h}^{H}\sum_{k=1}^{K} \omega_h^i(k, h_1)\mathbb{I}\left[t_{h_1}^k = 1\right] H}{2^i\epsilon} \\
&\lesssim \frac{H^5 SA\iota}{\epsilon^2} + \sum_{i=1}^{N} \frac{\sum_{h_1=h}^{H}\sum_{k=1}^{K} \mathbb{I}\left[t_{h_1}^k = 1\right] H}{2^i\epsilon} \\
&\lesssim \frac{H^5 SA\iota}{\epsilon^2} + \frac{H^3 SA}{\epsilon}.
\end{aligned} \tag{43}$$

Here, the second inequality is because $0 \leq \omega_h^i(k, h_1) < 3$ by Equation (41). The last inequality is because

$$\begin{aligned}
\sum_{h_1=1}^{H}\sum_{k=1}^{K} \mathbb{I}\left[t_{h_1}^k = 1\right] H &= H \sum_{h_1=1}^{H}\sum_{k=1}^{K} \mathbb{I}\left[t_{h_1}^k(s_{h_1}^k, a_{h_1}^k) = 1\right] \left(\sum_{s,a} \mathbb{I}\left[(s_{h_1}^k, a_{h_1}^k) = (s, a)\right]\right) \\
&= H \sum_{h_1=1}^{H}\sum_{s,a}\sum_{k=1}^{K} \mathbb{I}\left[(s_{h_1}^k, a_{h_1}^k) = (s, a), t_{h_1}^k(s, a) = 1\right] \\
&\leq H \sum_{h_1=1}^{H}\sum_{s,a} e_1 = H^3 SA.
\end{aligned} \tag{44}$$

Combining the results of Equation (42) and Equation (43), we reach

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{I}\left[Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \geq \epsilon\right] \lesssim \frac{H^6 SA\iota}{\epsilon^2}.$$

Now we finish the proof of the first conclusion. Further, noting that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K} \left(Q_h^k - Q_h^\star\right)(s_h^k, a_h^k)\mathbb{I}\left[\left(Q_h^k - Q_h^\star\right)(s_h^k, a_h^k) \geq \epsilon\right] \\
&\leq \sum_{h=1}^{H}\sum_{i=1}^{N} 2^i\epsilon\|\omega_h^i(\cdot, h)\|_1 \\
&\lesssim \sum_{h=1}^{H}\sum_{i=1}^{N} \frac{H^5 SA\iota}{2^i\epsilon} + \sum_{h=1}^{H}\sum_{h_1=h}^{H}\sum_{k=1}^{K} \left(\sum_{i=1}^{N} \omega_h^i(k, h_1)\right)\mathbb{I}\left[t_{h_1}^k = 1\right] H \\
&\lesssim \frac{H^6 SA\iota}{\epsilon} + \sum_{h=1}^{H}\sum_{h_1=h}^{H}\sum_{k=1}^{K} \mathbb{I}\left[t_{h_1}^k = 1\right] H \\
&\lesssim \frac{H^6 SA\iota}{\epsilon}.
\end{aligned}$$

Here, the second inequality is by Equation (40). The second last inequality is by Equation (41), and the last inequality is because of Equation (44). □

The following lemma provides a preliminary upper bound on the regret using the Hoeffding-type bonus.

Lemma G.4. *For any $h \in [H]$, under the event $\bigcap_{i=1}^{4} \mathcal{E}_i$ in Lemma F.1, the following conclusion holds:*

$$\sum_{k=1}^{K} \left(V_h^\star - V_h^{\pi^k}\right)(s_h^k) \lesssim \sqrt{H^4SAT\iota} + H^3SA.$$

Proof of Lemma G.4. For $h \in [H+1]$, denote:

$$\delta_h^k = \left(V_h^k - V_h^\star\right)(s_h^k),\ \zeta_h^k = \left(V_h^k - V_h^{\pi^k}\right)(s_h^k).$$

Here, $\delta_{H+1}^k = \zeta_{H+1}^k = 0$. Because $V_h^\star(s) = \sup_\pi V_h^\pi(s)$, we have $\delta_h^k \le \zeta_h^k$ for any $h \in [H+1]$. In addition, as $V_h^k(s) \ge V_h^\star(s)$ for all $(s,h,k) \in \mathcal{S} \times [H] \times [K]$ according to Lemma G.1, we only need to bound $\sum_{k=1}^K \zeta_h^k$.

$$\begin{aligned}
\sum_{k=1}^{K} \zeta_h^k &= \sum_{k=1}^{K} (Q_h^k - Q_h^{\pi^k})(s_h^k, a_h^k) \\
&= \sum_{k=1}^{K} (Q_h^k - Q_h^\star)(s_h^k, a_h^k) + \sum_{k=1}^{K} (Q_h^\star - Q_h^{\pi^k})(s_h^k, a_h^k) \\
&= \sum_{k=1}^{K} (Q_h^k - Q_h^\star)(s_h^k, a_h^k) + \sum_{k=1}^{K} \mathbb{P}_{s_h^k, a_h^k, h}(V_{h+1}^\star - V_{h+1}^{\pi^k}).
\end{aligned} \tag{45}$$

The last equality holds due to Equation (1).

By Equation (30), we have

$$\begin{aligned}
\sum_{k=1}^{K} (Q_h^k - Q_h^\star)(s_h^k, a_h^k) \le \sum_{k=1}^{K} \mathbb{I}[t_h^k = 1]H + \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} (V_{h+1}^{k_{l_i}} - V_{h+1}^\star)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k, a_h^k)} \\
+ 4 \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{H^2\iota}{n_h^k}}.
\end{aligned} \tag{46}$$

For the first term of Equation (46), similar to Equation (44), we have:

$$\sum_{k=1}^{K} \mathbb{I}\left[t_h^k = 1\right] H \le H \sum_{s,a} e_1 = H^2SA. \tag{47}$$

For the last term of Equation (46), by Lemma E.2 with $\alpha = 1/2$, it holds that

$$\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{H^2\iota}{n_h^k}} \le 4\sqrt{H^2SAT\iota}. \tag{48}$$

For the second term of Equation (46), by setting $\omega_h(k,h) = 1$ and applying Equation (11) to Equation (32), it holds that

$$\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} \left(V_{h+1}^{k_{l_i}} - V_{h+1}^\star\right)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k, a_h^k)} \le \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \delta_{h+1}^k. \tag{49}$$

Moreover,

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{P}_{s_h^k, a_h^k, h}(V_{h+1}^\star - V_{h+1}^{\pi^k}) \\
&= \sum_{k=1}^{K} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right) + \sum_{k=1}^{K} \left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right)(s_{h+1}^k) \\
&= \sum_{k=1}^{K} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right) + \sum_{k=1}^{K} \left(\zeta_{h+1}^k - \delta_{h+1}^k\right).
\end{aligned}$$

Taking the above results Equation (47), Equation (48) and Equation (49) together with Equation (46) and Equation (45), we obtain

$$\begin{aligned}
\sum_{k=1}^{K}\zeta_h^k &\le \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\delta_{h+1}^k + \sum_{k=1}^{K}\left(\zeta_{h+1}^k-\delta_{h+1}^k\right) + H^2SA + 16\sqrt{H^2SAT\iota} \\
&\quad + \sum_{k=1}^{K}\left(\mathbb{P}_{s_h^k,a_h^k,h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right) \\
&= \sum_{k=1}^{K}\zeta_{h+1}^k + \frac{1}{H}\sum_{k=1}^{K}\delta_{h+1}^k + H^2SA + 16\sqrt{H^2SAT\iota} + \sum_{k=1}^{K}\left(\mathbb{P}_{s_h^k,a_h^k,h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right) \\
&\le \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\zeta_{h+1}^k + H^2SA + 16\sqrt{H^2SAT\iota} + \sum_{k=1}^{K}\left(\mathbb{P}_{s_h^k,a_h^k,h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right),
\end{aligned}$$

where the last inequality uses $\delta_{h+1}^k \le \zeta_{h+1}^k$.

By recursively applying the inequality above over $h, h+1, \ldots, H$, since $\zeta_{H+1}^k = 0$, we have

$$\begin{aligned}
\sum_{k=1}^{K}\zeta_h^k &\le \sum_{h'=h}^{H}\left(1+\frac{1}{H}\right)^{h'-h}\left(H^2SA + 16\sqrt{H^2SAT\iota}\right) \\
&\quad + \sum_{h'=h}^{H}\left(1+\frac{1}{H}\right)^{h'-h}\sum_{k=1}^{K}\left(\mathbb{P}_{s_{h'}^k,a_{h'}^k,h'} - \mathbb{1}_{s_{h'+1}^k}\right)\left(V_{h'+1}^\star - V_{h'+1}^{\pi^k}\right) \\
&\le eH^3SA + 16e\sqrt{H^4SAT\iota} + \sum_{h'=h}^{H}\left(1+\frac{1}{H}\right)^{h'-h}\sum_{k=1}^{K}\left(\mathbb{P}_{s_{h'}^k,a_{h'}^k,h'} - \mathbb{1}_{s_{h'+1}^k}\right)\left(V_{h'+1}^\star - V_{h'+1}^{\pi^k}\right) \\
&\lesssim \sqrt{H^4SAT\iota} + H^3SA.
\end{aligned}$$

The second inequality uses $(1+1/H)^H \le e$, and the last inequality follows from the event $\mathcal{E}_4$ in Lemma F.1. Then we finish the proof. □

H Proof of Theorem 3.1

Lemma H.1. *For the UCB-Bernstein+ algorithm, under the event $\bigcap_{i=1}^{6}\mathcal{E}_i$ in Lemma F.1, for any given step $h \in [H]$ and any non-negative weight sequence $\{\omega_h^k\}_{k=1}^K$ at step h, it holds simultaneously for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and subsequent step $h' \in [h,H]$ that:*

$$\begin{aligned}
\sum_{k=1}^{K}\omega_h(k,h',s,a)(Q_{h'}^k - Q_{h'}^\star)(s_{h'}^k,a_{h'}^k) &\le \sum_{k'=1}^{K}\tilde{\omega}_h(k',h'+1,s,a)(Q_{h'+1}^{k'} - Q_{h'+1}^\star)(s_{h'+1}^{k'},a_{h'+1}^{k'}) \\
&\quad + 252\sqrt{(\mathbb{Q}_{h'}^\star+\beta^2)H\|\omega_h(\cdot,h')\|_\infty\|\omega_h(\cdot,h',s,a)\|_1\iota} + 152\sum_{k=1}^{K}\omega_h(k,h',s,a)Z_{h'}^k \qquad (50)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{k=1}^{K}\omega_h(k,h')\left(Q_{h'}^k - Q_{h'}^\star\right)(s_{h'}^k,a_{h'}^k) &\le \sum_{k'=1}^{K}\omega_h(k',h'+1)(Q_{h'+1}^{k'} - Q_{h'+1}^\star)(s_{h'+1}^{k'},a_{h'+1}^{k'}) \\
&\quad + 252\sum_{s,a}\sqrt{(\mathbb{Q}_{h'}^\star+\beta^2)H\|\omega_h(\cdot,h')\|_\infty\|\omega_h(\cdot,h',s,a)\|_1\iota} + 152\sum_{k=1}^{K}\omega_h(k,h')Z_{h'}^k, \qquad (51)
\end{aligned}$$

where

$$Z_{h'}^k = \mathbb{I}\left[t_{h'}^k = 1\right]H + \mathbb{I}\left[t_{h'}^k > 1\right]\frac{H^2\iota/\beta + \sqrt{H\Psi_{h'}^k(s_{h'}^k,a_{h'}^k)\iota}}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}. \qquad (52)$$

with

$$\Psi_h^k(s_h^k, a_h^k) = \sum_{i=1}^{n_h^k} \left(V_{h+1}^{k_{l_i}} - \hat{V}_{h+1}^{k_{l_i}}\right)(s_{h+1}^{k_{l_i}}).$$

Proof of Lemma H.1. By Equation (29) and the definition of $\tilde{b}_{h'}^k$ in Equation (19), for any $h' \in [h, H]$, we have

$$\begin{aligned}
&(Q_{h'}^k - Q_{h'}^\star)(s_{h'}^k, a_{h'}^k) \\
&\leq \mathbb{I}\left[t_{h'}^k = 1\right] H + \mathbb{I}\left[t_{h'}^k > 1\right] \frac{\sum_{i=1}^{n_{h'}^k} (V_{h'+1}^{k_{l_i}} - V_{h'+1}^\star)(s_{h'+1}^{k_{l_i}})}{n_{h'}^k(s_{h'}^k, a_{h'}^k)} \\
&+ \mathbb{I}\left[t_{h'}^k > 1\right] \left(8\sqrt{\frac{\tilde{W}_{h'}^k(s_{h'}^k, a_{h'}^k)\iota}{n_{h'}^k}} + \frac{119H\iota}{n_{h'}^k}\right) + \mathbb{I}\left[t_{h'}^k > 1\right] \frac{\sum_{i=1}^{n_{h'}^k} \left(\mathbb{1}_{s_{h'+1}^{k_{l_i}}} - \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'}\right) V_{h'+1}^\star}{n_{h'}^k(s_{h'}^k, a_{h'}^k)}.
\end{aligned} \tag{53}$$

By the event $\mathcal{E}_5$ in Lemma F.1, we can bound the last term in Equation (53) as:

$$\frac{\sum_{i=1}^{n_{h'}^k} \left(\mathbb{1}_{s_{h'+1}^{k_{l_i}}} - \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'}\right) V_{h'+1}^\star}{n_{h'}^k(s_{h'}^k, a_{h'}^k)} \leq 5\sqrt{\frac{\mathbb{Q}_{h'}^\star \iota}{n_{h'}^k(s_{h'}^k, a_{h'}^k)}} + \frac{9H\iota}{n_{h'}^k(s_{h'}^k, a_{h'}^k)}. \tag{54}$$

For the variance-dependent bonus term in Equation (53), we will upper bound $\tilde{W}_h^k(s_h^k, a_h^k)$. Define the **pseudo value function** $\hat{V}_h^k(s)$ as

$$\hat{V}_h^k(s) = \max\left\{\min\left\{V_h^\star(s) + \beta, V_h^k(s)\right\}, V_h^\star(s)\right\}.$$

Due to the optimistic property in Lemma G.1, it holds that $V_h^k(s) \geq \hat{V}_h^k(s)$. Then according to the definition of $\tilde{W}_h^k(s_h^k, a_h^k)$ in Equation (20), we have

$$\sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)}{n_h^k(s_h^k, a_h^k)}} \leq \sqrt{\frac{J_{1,h}^k + J_{2,h}^k}{n_h^k(s_h^k, a_h^k)}}, \tag{55}$$

where:

$$J_{1,h}^k = \frac{\sum_{i=1}^{n_h^k} \left(V_{h+1}^{k_{l_i}}(s_{h+1}^{k_{l_i}})\right)^2}{n_h^k(s_h^k, a_h^k)} - \frac{\sum_{i=1}^{n_h^k} \left(\hat{V}_{h+1}^{k_{l_i}}(s_{h+1}^{k_{l_i}})\right)^2}{n_h^k(s_h^k, a_h^k)}$$

and

$$J_{2,h}^k = \frac{\sum_{i=1}^{n_h^k} \left(\hat{V}_{h+1}^{k_{l_i}}(s_{h+1}^{k_{l_i}})\right)^2}{n_h^k(s_h^k, a_h^k)} - \left(\frac{\sum_{i=1}^{n_h^k} \hat{V}_{h+1}^{k_{l_i}}(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k, a_h^k)}\right)^2.$$

Now we want to bound both $J_{1,h}^k$ and $J_{2,h}^k$. Note that

$$J_{1,h}^k = \frac{\sum_{i=1}^{n_h^k} \left(V_{h+1}^{k_{l_i}} + \hat{V}_{h+1}^{k_{l_i}}\right)\left(V_{h+1}^{k_{l_i}} - \hat{V}_{h+1}^{k_{l_i}}\right)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k, a_h^k)} \leq \frac{2H\Psi_h^k(s_h^k, a_h^k)}{n_h^k(s_h^k, a_h^k)}. \tag{56}$$

For the second term $J_{2,h}^k$, using the Cauchy–Schwarz inequality, we have:

$$J_{2,h}^k = \frac{\sum_{i=1}^{n_h^k} \left(\hat{V}_{h+1}^{k_{l_i}}(s_{h+1}^{k_{l_i}}) - \frac{\sum_{n=1}^{n_h^k} \hat{V}_{h+1}^{k_{l_n}}(s_{h+1}^{k_{l_n}})}{n_h^k(s_h^k, a_h^k)}\right)^2}{n_h^k(s_h^k, a_h^k)} \leq 2\left(J_{2,h,1}^k + J_{2,h,2}^k\right),$$

where:

$$J_{2,h,1}^k = \frac{\sum_{i=1}^{n_h^k} \left(\left(\hat{V}_{h+1}^{k_{l_i}} - V_{h+1}^\star\right)(s_{h+1}^{k_{l_i}}) + \frac{\sum_{n=1}^{n_h^k} \left(V_{h+1}^\star - \hat{V}_{h+1}^{k_{l_n}}\right)(s_{h+1}^{k_{l_n}})}{n_h^k(s_h^k, a_h^k)}\right)^2}{n_h^k(s_h^k, a_h^k)},$$

and

$$J^k_{2,h,2} = \frac{\sum_{i=1}^{n_h^k}\left(V^\star_{h+1}(s^{k_{l_i}}_{h+1}) - \frac{\sum_{n=1}^{n_h^k} V^\star_{h+1}(s^{k_{l_n}}_{h+1})}{n_h^k(s_h^k,a_h^k)}\right)^2}{n_h^k(s_h^k,a_h^k)}$$
$$= \frac{\sum_{i=1}^{n_h^k}\left(V^\star_{h+1}(s^{k_{l_i}}_{h+1})\right)^2}{n_h^k(s_h^k,a_h^k)} - \left(\frac{\sum_{i=1}^{n_h^k} V^\star_{h+1}(s^{k_{l_i}}_{h+1})}{n_h^k(s_h^k,a_h^k)}\right)^2.$$

Since $V^\star_{h+1}(s) \le \hat{V}^{k_{l_i}}_{h+1}(s) \le V^\star_{h+1}(s)+\beta$, it holds that:

$$\left|\left(\hat{V}^{k_{l_i}}_{h+1} - V^\star_{h+1}\right)(s^{k_{l_i}}_{h+1}) + \frac{\sum_{n=1}^{n_h^k}\left(V^\star_{h+1} - \hat{V}^{k_{l_n}}_{h+1}\right)(s^{k_{l_n}}_{h+1})}{n_h^k(s_h^k,a_h^k)}\right|$$
$$\le \left|\left(\hat{V}^{k_{l_i}}_{h+1} - V^\star_{h+1}\right)(s^{k_{l_i}}_{h+1})\right| + \left|\frac{\sum_{n=1}^{n_h^k}\left(V^\star_{h+1} - \hat{V}^{k_{l_n}}_{h+1}\right)(s^{k_{l_n}}_{h+1})}{n_h^k(s_h^k,a_h^k)}\right|$$
$$\le 2\beta.$$

Therefore, applying this inequality to $J^k_{2,h,1}$, we have $J^k_{2,h,1} \le 4\beta^2$.

Moreover, we claim that:

$$J^k_{2,h,2} \le \mathbb{V}_{s_h^k,a_h^k,h}(V^\star_{h+1}) + 3\sqrt{2}H^2\sqrt{\frac{\iota}{n_h^k(s_h^k,a_h^k)}}. \tag{57}$$

This is because

$$J^k_{2,h,2} - \mathbb{V}_{s_h^k,a_h^k,h}(V^\star_{h+1})$$
$$= \frac{\sum_{i=1}^{n_h^k}\left(V^\star_{h+1}(s^{k_{l_i}}_{h+1})\right)^2}{n_h^k(s_h^k,a_h^k)} - \left(\frac{\sum_{i=1}^{n_h^k} V^\star_{h+1}(s^{k_{l_i}}_{h+1})}{n_h^k(s_h^k,a_h^k)}\right)^2 - \mathbb{P}_{s_h^k,a_h^k,h}\left(V^\star_{h+1}\right)^2 + \left(\mathbb{P}_{s_h^k,a_h^k,h}V^\star_{h+1}\right)^2$$
$$= \frac{\sum_{i=1}^{n_h^k}\left(\mathbb{1}_{s^{k_{l_i}}_{h+1}} - \mathbb{P}_{s_h^k,a_h^k,h}\right)\left(V^\star_{h+1}\right)^2}{n_h^k(s_h^k,a_h^k)} + \left(\mathbb{P}_{s_h^k,a_h^k,h}V^\star_{h+1}\right)^2 - \left(\frac{\sum_{i=1}^{n_h^k} V^\star_{h+1}(s^{k_{l_i}}_{h+1})}{n_h^k(s_h^k,a_h^k)}\right)^2$$
$$\le \frac{\sum_{i=1}^{n_h^k}\left(\mathbb{1}_{s^{k_{l_i}}_{h+1}} - \mathbb{P}_{s_h^k,a_h^k,h}\right)\left(V^\star_{h+1}\right)^2}{n_h^k(s_h^k,a_h^k)} + 2H\left|\frac{\sum_{i=1}^{n_h^k}\left(\mathbb{1}_{s^{k_{l_i}}_{h+1}} - \mathbb{P}_{s_h^k,a_h^k,h}\right)V^\star_{h+1}}{n_h^k(s_h^k,a_h^k)}\right|$$
$$\le 3\sqrt{2}H^2\sqrt{\frac{\iota}{n_h^k(s_h^k,a_h^k)}}. \tag{58}$$

The last inequality is because of the events $\mathcal{E}_3$ and $\mathcal{E}_6$ in Lemma F.1.

Applying the results of Equation (56), Equation (57) and $J^k_{2,h,1} \le 4\beta^2$ to Equation (55), we reach

$$\sqrt{\frac{\tilde{W}_h^k(s_h^k,a_h^k)}{n_h^k(s_h^k,a_h^k)}} \le \frac{\sqrt{2H\Psi_h^k(s_h^k,a_h^k)}}{n_h^k(s_h^k,a_h^k)} + \sqrt{\frac{2\mathbb{V}_{s_h^k,a_h^k,h}(V^\star_{h+1})}{n_h^k(s_h^k,a_h^k)}} + \sqrt{\frac{8\beta^2}{n_h^k(s_h^k,a_h^k)}} + \frac{\sqrt{6\sqrt{2}}H\iota^{\frac{1}{4}}}{(n_h^k(s_h^k,a_h^k))^{\frac{3}{4}}}. \tag{59}$$

Using the definition of $\mathbb{Q}^\star_h$, we further derive the following upper bound for $\tilde{W}_h^k(s_h^k,a_h^k)$:

$$\sqrt{\frac{\tilde{W}_{h'}^k(s_{h'}^k,a_{h'}^k)\iota}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}} \le \frac{\sqrt{2H\Psi_{h'}^k(s_{h'}^k,a_{h'}^k)\iota}}{n_{h'}^k(s_{h'}^k,a_{h'}^k)} + \sqrt{\frac{2\mathbb{Q}^\star_{h'}\iota}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}} + \sqrt{\frac{8\beta^2\iota}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}} + \frac{\sqrt{6\sqrt{2}}H\iota^{\frac{3}{4}}}{n_{h'}^k(s_{h'}^k,a_{h'}^k)^{\frac{3}{4}}}. \tag{60}$$

Here, we use the AM-GM inequality

$$\frac{H\iota^{\frac{3}{4}}}{n_{h'}^k(s_{h'}^k,a_{h'}^k)^{\frac{3}{4}}} \le \sqrt{\frac{\beta^2\iota}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}} + \frac{H^2\iota}{\beta n_{h'}^k(s_{h'}^k,a_{h'}^k)}.$$

Combining Equation (53), Equation (54), and Equation (60), and using $\beta \le H$, we then reach

$$\begin{aligned}
&(Q_{h'}^k - Q_{h'}^\star)(s_{h'}^k,a_{h'}^k) \\
&\le \mathbb{I}\left[t_{h'}^k>1\right]\frac{\sum_{i=1}^{n_{h'}^k}(V_{h'+1}^{k_{l_i}} - V_{h'+1}^\star)(s_{h'+1}^{k_{l_i}})}{n_{h'}^k(s_{h'}^k,a_{h'}^k)} + 63\mathbb{I}\left[t_{h'}^k>1\right]\sqrt{\frac{(\mathbb{Q}_{h'}^\star+\beta^2)\iota}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}} + 152Z_{h'}^k,
\end{aligned}$$

where we use $\beta \le H$, together with

$$8\sqrt{2}+5+16\sqrt{2}+8\sqrt{6\sqrt{2}} < 63,\ 128+8\sqrt{6\sqrt{2}} < 152.$$

Therefore, it holds that

$$\begin{aligned}
&\sum_{k=1}^K \omega_h(k,h',s,a)(Q_{h'}^k - Q_{h'}^\star)(s_{h'}^k,a_{h'}^k) \\
&\le \sum_{k=1}^K \omega_h(k,h',s,a)\mathbb{I}\left[t_{h'}^k>1\right]\frac{\sum_{i=1}^{n_{h'}^k}\left(V_{h'+1}^{k_{l_i}} - V_{h'+1}^\star\right)(s_{h'+1}^{k_{l_i}})}{n_{h'}^k(s_{h'}^k,a_{h'}^k)} \\
&\quad + 63\sum_{k=1}^K \omega_h(k,h',s,a)\mathbb{I}\left[t_{h'}^k>1\right]\sqrt{\frac{(\mathbb{Q}_{h'}^\star+\beta^2)\iota}{n_{h'}^k}} + 152\sum_{k=1}^K \omega_h(k,h',s,a)Z_{h'}^k. \qquad (61)
\end{aligned}$$

For the first term of Equation (61), note that:

$$\begin{aligned}
&\sum_{k=1}^K \omega_h(k,h',s,a)\mathbb{I}[t_{h'}^k>1]\frac{\sum_{i=1}^{n_{h'}^k}\left(V_{h'+1}^{k_{l_i}} - V_{h'+1}^\star\right)(s_{h'+1}^{k_{l_i}})}{n_{h'}^k(s_{h'}^k,a_{h'}^k)} \\
&= \sum_{k=1}^K\sum_{i=1}^{n_{h'}^k} \omega_h(k,h',s,a)\frac{\left(V_{h'+1}^{k_{l_i}} - V_{h'+1}^\star\right)(s_{h'+1}^{k_{l_i}})}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}\mathbb{I}[t_{h'}^k>1]\cdot\sum_{k'=1}^K\mathbb{I}\left[k_{l_i}=k'\right] \\
&= \sum_{k=1}^K\sum_{i=1}^{n_{h'}^k}\sum_{k'=1}^K \omega_h(k,h',s,a)\frac{\left(V_{h'+1}^{k'} - V_{h'+1}^\star\right)(s_{h'+1}^{k'})}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}\mathbb{I}[t_{h'}^k>1]\cdot\mathbb{I}[k_{l_i}=k'] \\
&= \sum_{k'=1}^K\left(\sum_{k=1}^K \omega_h(k,h',s,a)\frac{\sum_{i=1}^{n_{h'}^k}\mathbb{I}\left[k_{l_i}=k',t_{h'}^k>1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)}\right)\left(V_{h'+1}^{k'} - V_{h'+1}^\star\right)(s_{h'+1}^{k'}) \\
&\le \sum_{k'=1}^K \tilde{\omega}_h(k',h'+1,s,a)(Q_{h'+1}^{k'} - Q_{h'+1}^\star)(s_{h'+1}^{k'},a_{h'+1}^{k'}). \qquad (62)
\end{aligned}$$

The inequality is because

$$Q_{h'+1}^{k'}(s_{h'+1}^{k'},a_{h'+1}^{k'}) = V_{h'+1}^{k'}(s_{h'+1}^{k'}),\ Q_{h'+1}^\star(s_{h'+1}^{k'},a_{h'+1}^{k'}) \le V_{h'+1}^\star(s_{h'+1}^{k'})$$

and

$$\begin{aligned}
&\sum_{k=1}^K \omega_h(k,h',s,a)\frac{\sum_{i=1}^{n_{h'}^k}\mathbb{I}\left[k_{l_i}=k',t_{h'}^k>1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)} \\
&= \mathbb{I}\left[(s_{h'}^{k'},a_{h'}^{k'})=(s,a)\right]\sum_{k=1}^K \omega_h(k,h')\frac{\sum_{i=1}^{n_{h'}^k}\mathbb{I}\left[k_{l_i}=k',t_{h'}^k>1\right]}{n_{h'}^k(s_{h'}^k,a_{h'}^k)} = \tilde{\omega}_h(k',h'+1,s,a)
\end{aligned}$$

because $\sum_{i=1}^{n_{h'}^k} \mathbb{I}\left[k_{l_i} = k', t_{h'}^k > 1\right] = 1$ if and only if $(s_{h'}^k, a_{h'}^k) = (s_{h'}^{k'}, a_{h'}^{k'})$ and $1 < t_{h'}^k(s_{h'}^k, a_{h'}^k) = t_{h'}^{k'}(s_{h'}^k, a_{h'}^k) + 1$.

For the second term of Equation (61), by Lemma E.2 with $\alpha = 1/2$, it holds that:

$$\sum_{k=1}^{K} \omega_h(k, h', s, a) \sqrt{\frac{(\mathbb{Q}_{h'}^\star + \beta^2)\iota}{n_{h'}^k}} \mathbb{I}\left[t_{h'}^k > 1\right] \leq 4\sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) H \|\omega_h(\cdot, h')\|_\infty \|\omega_h(\cdot, h', s, a)\|_1 \iota}. \tag{63}$$

Applying Equation (62) and Equation (63) to Equation (61), for any $h' \in [h, H]$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{K} \omega_h(k, h', s, a)(Q_{h'}^k - Q_{h'}^\star)(s_{h'}^k, a_{h'}^k) \\
&\leq \sum_{k'=1}^{K} \tilde{\omega}_h(k', h'+1, s, a)(Q_{h'+1}^{k'} - Q_{h'+1}^\star)(s_{h'+1}^{k'}, a_{h'+1}^{k'}) \\
&\quad + 252\sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) H \|\omega_h(\cdot, h')\|_\infty \|\omega_h(\cdot, h', s, a)\|_1 \iota} + 152 \sum_{k=1}^{K} \omega_h(k, h', s, a) Z_{h'}^k,
\end{aligned}$$

which proves the first conclusion Equation (50). The second conclusion follows by summing the inequality above over all state-action pairs and using property (a) of Lemma E.1. □

Lemma H.2. *For UCB-Bernstein+, under the event $\bigcap_{i=1}^{6} \mathcal{E}_i$ in Lemma F.1, for any given $h \in [H]$ and non-negative weight sequence $\{\omega_h^k\}_k$ at step h, it holds simultaneously for any subsequent step $h' \in [h, H]$ that:*

$$\begin{aligned}
&\sum_{k=1}^{K} \omega_h(k, h')(Q_{h'}^k - Q_{h'}^\star)(s_{h'}^k, a_{h'}^k) \\
&\leq 252 \sum_{h_1=h'}^{H} \sum_{s,a} \sqrt{(\mathbb{Q}_{h_1}^\star + \beta^2) H \|\omega_h(\cdot, h_1)\|_\infty \|\omega_h(\cdot, h_1, s, a)\|_1 \iota} + 152 \sum_{h_1=h'}^{H} \sum_{k=1}^{K} \omega_h(k, h_1) Z_{h_1}^k.
\end{aligned}$$

Proof of Lemma H.2. Repeatedly applying Equation (51) over steps $h', h'+1, \ldots, H$ and using the terminal condition at step $H+1$ completes the proof. □

Lemma H.3. *For UCB-Bernstein+, under the event $\bigcap_{i=1}^{6} \mathcal{E}_i$ in Lemma F.1, it holds simultaneously for any $h \in [H]$ that:*

$$\begin{aligned}
\sum_{s,a} \Delta_h(s,a) N_h^{K+1}(s,a) \lesssim \sum_{h'=h}^{H} \sum_{\Delta_{h'}(s,a)>0} &\frac{(\mathbb{Q}_{h'}^\star + \beta^2) H^2 \iota}{\Delta_{h'}(s,a)} + \frac{H\left(\sum_{t=h+1}^{H} \sqrt{(\mathbb{Q}_t^\star + \beta^2)|Z_{\text{opt},t}|}\right)^2 \iota}{\Delta_{\min,h}} \\
&+ \sum_{h'=h+1}^{H} \frac{\left(\sum_{t=h'+1}^{H} \sqrt{(\mathbb{Q}_t^\star + \beta^2)|Z_{\text{opt},t}|}\right)^2 \iota}{\Delta_{\min,h'}} + \sum_{h'=h}^{H} \sum_{k=1}^{K} Z_{h'}^k.
\end{aligned}$$

Proof of Lemma H.3. For any $h \in [H]$, define

$$A_h := \sum_{s,a} \Delta_h(s,a) N_h^{K+1}(s,a).$$

Let $Z_{\text{sub},h} := \{(s,a) \mid \Delta_h(s,a) > 0\}$ and define

$$\omega_h^k = \omega_h(k,h) := \mathbb{I}\left[(s_h^k, a_h^k) \in Z_{\text{sub},h}\right] \leq 1.$$

For any $(h,k) \in [H] \times [K]$, if $(s_h^k, a_h^k) \in Z_{\text{sub},h}$, then

$$Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) = V_h^k(s_h^k) - Q_h^\star(s_h^k, a_h^k) \geq V_h^\star(s_h^k) - Q_h^\star(s_h^k, a_h^k) = \Delta_h(s_h^k, a_h^k) > 0,$$

where the inequality follows from Lemma G.1. Hence, for any $(s,a) \in Z_{\text{sub},h}$,

$$\|\omega_h(\cdot,h,s,a)\|_1 = \sum_{k=1}^{K} \mathbb{I}\left[(s_h^k, a_h^k) = (s,a)\right] = N_h^{K+1}(s,a),$$

while $\|\omega_h(\cdot,h,s,a)\|_1 = 0$ for $(s,a) \in Z_{\text{opt},h}$.

By Equation (50) in Lemma H.1, for any $(s,a) \in Z_{\text{sub},h}$,

$$\begin{aligned}
&\sum_{k=1}^{K} \omega_h(k,h,s,a)\left(Q_h^k - Q_h^\star\right)(s_h^k, a_h^k) \\
&\le \sum_{k'=1}^{K} \tilde{\omega}_h(k',h+1,s,a)\left(Q_{h+1}^{k'} - Q_{h+1}^\star\right)(s_{h+1}^{k'}, a_{h+1}^{k'}) \\
&\quad + 252\sqrt{(\mathbb{Q}_h^\star + \beta^2) H N_h^{K+1}(s,a)\iota} + 152\sum_{k=1}^{K} \omega_h(k,h,s,a) Z_h^k. \qquad (64)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\sum_{k=1}^{K} \omega_h(k,h,s,a)\left(Q_h^k - Q_h^\star\right)(s_h^k, a_h^k) &\ge \Delta_h(s,a)\sum_{k=1}^{K} \mathbb{I}\left[(s_h^k, a_h^k) = (s,a)\right] \\
&= \Delta_h(s,a) N_h^{K+1}(s,a). \qquad (65)
\end{aligned}$$

Combining Equation (64) and Equation (65), and using the AM-GM inequality,

$$252\sqrt{(\mathbb{Q}_h^\star + \beta^2) H N_h^{K+1}(s,a)\iota} \le \frac{1}{2}\Delta_h(s,a) N_h^{K+1}(s,a) + 31752\frac{(\mathbb{Q}_h^\star + \beta^2) H\iota}{\Delta_h(s,a)},$$

we obtain

$$\begin{aligned}
\Delta_h(s,a) N_h^{K+1}(s,a) &\le 2\sum_{k'=1}^{K} \tilde{\omega}_h(k',h+1,s,a)\left(Q_{h+1}^{k'} - Q_{h+1}^\star\right)(s_{h+1}^{k'}, a_{h+1}^{k'}) \\
&\quad + 63504\frac{(\mathbb{Q}_h^\star + \beta^2) H\iota}{\Delta_h(s,a)} + 304\sum_{k=1}^{K} \omega_h(k,h,s,a) Z_h^k.
\end{aligned}$$

Since $Q_{h+1}^k(s,a) \ge Q_{h+1}^\star(s,a)$ by Lemma G.1, summing the inequality above over $(s,a) \in Z_{\text{sub},h}$ and using property (a) of Lemma E.1 gives

$$\begin{aligned}
A_h &\le 2\sum_{k'=1}^{K} \omega_h(k',h+1)\left(Q_{h+1}^{k'} - Q_{h+1}^\star\right)(s_{h+1}^{k'}, a_{h+1}^{k'}) \\
&\quad + 63504\sum_{\Delta_h(s,a)>0} \frac{(\mathbb{Q}_h^\star + \beta^2) H\iota}{\Delta_h(s,a)} + 304\sum_{k=1}^{K} Z_h^k. \qquad (66)
\end{aligned}$$

By Lemma H.2,

$$\begin{aligned}
&\sum_{k'=1}^{K} \omega_h(k',h+1)\left(Q_{h+1}^{k'} - Q_{h+1}^\star\right)(s_{h+1}^{k'}, a_{h+1}^{k'}) \\
&\le 252\sum_{h'=h+1}^{H}\sum_{s,a} \sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) H \|\omega_h(\cdot,h')\|_\infty \|\omega_h(\cdot,h',s,a)\|_1 \iota} + 152\sum_{h'=h+1}^{H}\sum_{k=1}^{K} \omega_h(k,h') Z_{h'}^k. \qquad (67)
\end{aligned}$$

By part (e) of Lemma E.1,

$$\|\omega_h(\cdot,h')\|_\infty \le \left(1 + \frac{1}{H}\right)^{h'-h} \|\omega_h(\cdot,h)\|_\infty \le e < 3, \qquad (68)$$

and

$$\|\omega_h(\cdot, h')\|_1 \le \|\omega_h(\cdot, h)\|_1 = \sum_{\Delta_h(s,a)>0} N_h^{K+1}(s,a). \tag{69}$$

Moreover, by part (c) of Lemma E.1,

$$\|\omega_h(\cdot, h', s, a)\|_1 \le \|\omega_h(\cdot, h')\|_\infty N_{h'}^{K+1}(s,a) \le 3N_{h'}^{K+1}(s,a). \tag{70}$$

We next divide the state-action pairs at each step h' into $Z_{\text{sub},h'}$ and $Z_{\text{opt},h'}$. For the suboptimal state-action pairs, by Equation (68), Equation (70), and the Cauchy–Schwarz inequality,

$$\begin{aligned}
&504 \sum_{h'=h+1}^{H} \sum_{\Delta_{h'}(s,a)>0} \sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) H \|\omega_h(\cdot, h')\|_\infty \|\omega_h(\cdot, h', s, a)\|_1 \iota} \\
&\le 1512 \sqrt{\left(\sum_{h'=h+1}^{H} \sum_{\Delta_{h'}(s,a)>0} \frac{(\mathbb{Q}_{h'}^\star + \beta^2) H \iota}{\Delta_{h'}(s,a)} \right) \left(\sum_{h'=h+1}^{H} A_{h'} \right)} \\
&\le \frac{1}{2H} \sum_{h'=h+1}^{H} A_{h'} + 1143072 \sum_{h'=h+1}^{H} \sum_{\Delta_{h'}(s,a)>0} \frac{(\mathbb{Q}_{h'}^\star + \beta^2) H^2 \iota}{\Delta_{h'}(s,a)},
\end{aligned} \tag{71}$$

where the last inequality follows from the AM-GM inequality.

For the optimal state-action pairs, by parts (d) and (e) of Lemma E.1,

$$\sum_{(s,a)\in Z_{\text{opt},h'}} \|\omega_h(\cdot, h', s, a)\|_1 \le \|\omega_h(\cdot, h')\|_1 \le \sum_{\Delta_h(s,a)>0} N_h^{K+1}(s,a).$$

Thus, again by Equation (68) and the Cauchy–Schwarz inequality,

$$\begin{aligned}
&504 \sum_{h'=h+1}^{H} \sum_{(s,a)\in Z_{\text{opt},h'}} \sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) H \|\omega_h(\cdot, h')\|_\infty \|\omega_h(\cdot, h', s, a)\|_1 \iota} \\
&\le 504\sqrt{3} \left(\sum_{h'=h+1}^{H} \sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) H |Z_{\text{opt},h'}| \iota} \right) \sqrt{\sum_{\Delta_h(s,a)>0} N_h^{K+1}(s,a)}.
\end{aligned} \tag{72}$$

If $|Z_{\text{sub},h}| = 0$, then $A_h = 0$ and the desired conclusion holds trivially. Otherwise, $\Delta_{\min,h} \in (0, +\infty)$ and

$$A_h = \sum_{\Delta_h(s,a)>0} \Delta_h(s,a) N_h^{K+1}(s,a) \ge \Delta_{\min,h} \sum_{\Delta_h(s,a)>0} N_h^{K+1}(s,a).$$

Hence, applying the AM-GM inequality to Equation (72), we obtain

$$\begin{aligned}
&504\sqrt{3} \left(\sum_{h'=h+1}^{H} \sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) H |Z_{\text{opt},h'}| \iota} \right) \sqrt{\sum_{\Delta_h(s,a)>0} N_h^{K+1}(s,a)} \\
&\le \frac{1}{2} A_h + 381024 \frac{H \left(\sum_{h'=h+1}^{H} \sqrt{(\mathbb{Q}_{h'}^\star + \beta^2) |Z_{\text{opt},h'}|} \right)^2 \iota}{\Delta_{\min,h}}.
\end{aligned} \tag{73}$$

Moreover, by Equation (68),

$$304 \sum_{h'=h+1}^{H} \sum_{k=1}^{K} \omega_h(k, h') Z_{h'}^k \le 912 \sum_{h'=h+1}^{H} \sum_{k=1}^{K} Z_{h'}^k.$$

Multiplying Equation (67) by 2, applying Equation (71) and Equation (73), and then substituting the resulting bound into Equation (66), we obtain

$$\begin{aligned}A_h \le \frac{1}{2}A_h + \frac{1}{2H}\sum_{h'=h+1}^{H} A_{h'} \\ + 1143072\sum_{h'=h+1}^{H}\sum_{\Delta_{h'}(s,a)>0}\frac{(\mathbb{Q}^\star_{h'}+\beta^2)H^2\iota}{\Delta_{h'}(s,a)} + 63504\sum_{\Delta_h(s,a)>0}\frac{(\mathbb{Q}^\star_h+\beta^2)H\iota}{\Delta_h(s,a)} \\ + 381024\frac{H\left(\sum_{h'=h+1}^{H}\sqrt{(\mathbb{Q}^\star_{h'}+\beta^2)|Z_{\text{opt},h'}|}\right)^2\iota}{\Delta_{\min,h}} + 304\sum_{k=1}^{K}Z_h^k + 912\sum_{h'=h+1}^{H}\sum_{k=1}^{K}Z_{h'}^k.\end{aligned}$$

Moving $A_h/2$ to the left-hand side, multiplying both sides by 2, and enlarging the numerical constants if necessary, we obtain the explicit recursion

$$A_h \le F_h + \frac{1}{H}\sum_{h'=h+1}^{H} A_{h'}, \tag{74}$$

where

$$\begin{aligned}F_h := 2286144\sum_{h'=h}^{H}\sum_{\Delta_{h'}(s,a)>0}\frac{(\mathbb{Q}^\star_{h'}+\beta^2)H^2\iota}{\Delta_{h'}(s,a)} \\ + 762048\frac{H\left(\sum_{h'=h+1}^{H}\sqrt{(\mathbb{Q}^\star_{h'}+\beta^2)|Z_{\text{opt},h'}|}\right)^2\iota}{\Delta_{\min,h}} + 1824\sum_{h'=h}^{H}\sum_{k=1}^{K}Z_{h'}^k.\end{aligned}$$

When $|Z_{\text{sub},h}| = 0$, $A_h = 0$, so Equation (74) also holds trivially.

It remains to solve the recursion Equation (74). We claim that

$$A_h \le F_h + \frac{1}{H}\sum_{h'=h+1}^{H}\left(1+\frac{1}{H}\right)^{h'-h-1}F_{h'}. \tag{75}$$

This follows by backward induction. The claim is immediate for $h = H$. Assuming that it holds for all $t \in \{h+1,\dots,H\}$, Equation (74) gives

$$\begin{aligned}A_h &\le F_h + \frac{1}{H}\sum_{t=h+1}^{H}\left(F_t + \frac{1}{H}\sum_{h'=t+1}^{H}\left(1+\frac{1}{H}\right)^{h'-t-1}F_{h'}\right) \\ &= F_h + \frac{1}{H}\sum_{h'=h+1}^{H}\left(1+\frac{1}{H}\right)^{h'-h-1}F_{h'},\end{aligned}$$

which proves Equation (75).

Finally, since

$$\left(1+\frac{1}{H}\right)^{H} \le e,$$

substituting the definition of $F_{h'}$ into Equation (75) and exchanging the order of summation yields

$$\begin{aligned}A_h \lesssim \sum_{h'=h}^{H}\sum_{\Delta_{h'}(s,a)>0}\frac{(\mathbb{Q}^\star_{h'}+\beta^2)H^2\iota}{\Delta_{h'}(s,a)} + \frac{H\left(\sum_{t=h+1}^{H}\sqrt{(\mathbb{Q}^\star_t+\beta^2)|Z_{\text{opt},t}|}\right)^2\iota}{\Delta_{\min,h}} \\ + \sum_{h'=h+1}^{H}\frac{\left(\sum_{t=h'+1}^{H}\sqrt{(\mathbb{Q}^\star_t+\beta^2)|Z_{\text{opt},t}|}\right)^2\iota}{\Delta_{\min,h'}} + \sum_{h'=h}^{H}\sum_{k=1}^{K}Z_{h'}^k.\end{aligned}$$

The hidden constant in the final inequality is universal and independent of h, H, S, A, and K. This proves the lemma. □

Now we continue to bound the gap-dependent regret. We claim that

$$\sum_{h'=1}^{H}\sum_{k=1}^{K} Z_{h'}^{k} \lesssim \frac{H^5SA\iota^2}{\beta}. \tag{76}$$

Proof of Equation (76): By definition of $Z_{h'}^{k}$ in Equation (52), we have the following relationship

$$Z_{h'}^{k} = \mathbb{I}\left[t_{h'}^{k}=1\right]H + \mathbb{I}\left[t_{h'}^{k}>1\right]\frac{H^2\iota/\beta + \sqrt{H\Psi_{h'}^{k}(s_{h'}^{k},a_{h'}^{k})\iota}}{n_{h'}^{k}(s_{h'}^{k},a_{h'}^{k})}. \tag{77}$$

For the first term of Equation (77), by Equation (44), we have

$$\sum_{h'=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_{h'}^{k}=1\right]H \lesssim H^3SA.$$

For the second term of Equation (77), by Lemma E.2 with $\alpha=1$, we know

$$\sum_{h'=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_{h'}^{k}>1\right]\frac{H^2\iota}{\beta n_{h'}^{k}(s_{h'}^{k},a_{h'}^{k})} \lesssim \frac{H^4SA\iota^2}{\beta}. \tag{78}$$

We also claim that

$$\sum_{h'=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_{h'}^{k}>1\right]\frac{\sqrt{H\Psi_{h'}^{k}(s_{h'}^{k},a_{h'}^{k})\iota}}{n_{h'}^{k}(s_{h'}^{k},a_{h'}^{k})} \lesssim \frac{H^5SA\iota^2}{\beta}, \tag{79}$$

which will be proved later. Then combining the results of Equation (78) and Equation (79), we can bound the second term of Equation (77) as

$$\sum_{h'=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_{h'}^{k}>1\right]\frac{H^2\iota/\beta + \sqrt{H\Psi_{h'}^{k}(s_{h'}^{k},a_{h'}^{k})\iota}}{n_{h'}^{k}(s_{h'}^{k},a_{h'}^{k})} \lesssim \frac{H^5SA\iota^2}{\beta}.$$

Combining the results above, we complete the proof of Equation (76).

Let $\delta=\frac{1}{8T}$, then $\mathcal{E}:=\bigcap_{i=1}^{8}\mathcal{E}_i$ holds with probability at least $1-\frac{1}{T}$ and $\iota\le O(\log(SAT))$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}(\text{Regret}(T))\\
&=\mathbb{E}\left(\sum_{h=1}^{H}\sum_{s,a}\Delta_h(s,a)N_h^{K+1}(s,a)\right)\\
&=\mathbb{E}\left(\sum_{h=1}^{H}\sum_{s,a}\Delta_h(s,a)N_h^{K+1}(s,a)\middle|\mathcal{E}\right)\mathbb{P}(\mathcal{E})+\mathbb{E}\left(\sum_{h=1}^{H}\sum_{s,a}\Delta_h(s,a)N_h^{K+1}(s,a)\middle|\mathcal{E}^c\right)\mathbb{P}(\mathcal{E}^c)\\
&\le O\left(\sum_{h=1}^{H}\sum_{\Delta_h(s,a)>0}\frac{(\mathbb{Q}_h^{\star}+\beta^2)H^3\iota_1}{\Delta_h(s,a)}+\sum_{h=1}^{H}\frac{H(\sum_{t=h+1}^{H}\sqrt{(\mathbb{Q}_t^{\star}+\beta^2)|Z_{\text{opt},t}|})^2\iota_1}{\Delta_{\min,h}}+\frac{H^6SA\iota_1^2}{\beta}\right)\\
&\quad+\frac{1}{T}\cdot HT\\
&\le O\left(\sum_{h=1}^{H}\sum_{\Delta_h(s,a)>0}\frac{(\mathbb{Q}_h^{\star}+\beta^2)H^3\iota_1}{\Delta_h(s,a)}+\sum_{h=1}^{H}\frac{H(\sum_{t=h+1}^{H}\sqrt{(\mathbb{Q}_t^{\star}+\beta^2)|Z_{\text{opt},t}|})^2\iota_1}{\Delta_{\min,h}}+\frac{H^6SA\iota_1^2}{\beta}\right).
\end{aligned}$$

The first inequality holds due to Lemma H.3 under the event $\mathcal{E}$, and under the event $\mathcal{E}^c$,

$$\sum_{h=1}^{H}\sum_{s,a}\Delta_h(s,a)N_h^{K+1}(s,a)\le HT.$$

Proof of Equation (79):

For each (s,a,h) and stage $t \geq 2$, let $\Phi_h^t(s,a)$ denote the common value of $\Psi_h^k(s,a)$ over episodes k satisfying $t_h^k(s,a) = t$. This value is well-defined because $\Psi_h^k(s,a)$ depends only on the samples collected in the preceding stage $t-1$, and hence remains unchanged throughout stage t.

Then we have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sqrt{H\Psi_h^k(s_h^k,a_h^k)}}{n_h^k(s_h^k,a_h^k)} \\
&= \sum_{h=1}^{H}\sum_{s,a}\sum_{k=1}^{K} \frac{\sqrt{H\Psi_h^k(s,a)}}{n_h^k(s,a)} \mathbb{I}\left[(s_h^k,a_h^k)=(s,a)\right] \left(\sum_{t=2}^{t_h^K(s,a)} \mathbb{I}\left[t_h^k(s,a)=t\right]\right) \\
&= \sum_{h=1}^{H}\sum_{s,a}\sum_{t=2}^{t_h^K(s,a)}\sum_{k=1}^{K} \frac{\sqrt{H\Psi_h^k(s,a)}}{e_{t-1}} \mathbb{I}\left[(s_h^k,a_h^k)=(s,a)\right] \mathbb{I}\left[t_h^k(s,a)=t\right] \\
&\leq \sum_{h=1}^{H}\sum_{s,a}\sum_{t=2}^{t_h^K(s,a)} \sqrt{H\Phi_h^t(s,a)}\frac{e_t}{e_{t-1}} \\
&\lesssim \sum_{h=1}^{H}\sum_{s,a}\sum_{t=2}^{t_h^K(s,a)} \sqrt{H\Phi_h^t(s,a)} \\
&\lesssim \sqrt{SAH^3\iota \sum_{h=1}^{H}\sum_{s,a}\sum_{t=2}^{t_h^K(s,a)} \Phi_h^t(s,a)}.
\end{aligned} \tag{80}$$

Here, the first inequality holds due to the definition of $\Phi_h^t(s,a)$. The second inequality is because $e_t \leq 2e_{t-1}$. The last inequality is by the Cauchy–Schwarz inequality with $\sum_{s,a,h} t_h^K(s,a) \lesssim SAH^2\iota$ by Equation (18). Furthermore, note that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{s,a}\sum_{t=2}^{t_h^K(s,a)} \Phi_h^t(s,a) \\
&\leq \sum_{h=1}^{H}\sum_{k=1}^{K} \left(V_{h+1}^k - \hat{V}_{h+1}^k\right)(s_{h+1}^k) \\
&\leq \sum_{h=1}^{H}\sum_{k=1}^{K} \left(V_{h+1}^k - V_{h+1}^\star\right)(s_{h+1}^k) \cdot \mathbb{I}\left[\left(V_{h+1}^k - V_{h+1}^\star\right)(s_{h+1}^k) > \beta\right] \\
&\overset{\text{(i)}}{\lesssim} \sum_{h=1}^{H}\sum_{k=1}^{K} \left(Q_{h+1}^k - Q_{h+1}^\star\right)(s_{h+1}^k,a_{h+1}^k) \cdot \mathbb{I}\left[\left(Q_{h+1}^k - Q_{h+1}^\star\right)(s_{h+1}^k,a_{h+1}^k) > \beta\right] \\
&\lesssim \frac{H^6SA\iota}{\beta}.
\end{aligned} \tag{81}$$

Here, (i) is because $Q_{h+1}^k(s_{h+1}^k,a_{h+1}^k) = V_{h+1}^k(s_{h+1}^k)$ and $Q_{h+1}^\star(s_{h+1}^k,a_{h+1}^k) \leq V_{h+1}^\star(s_{h+1}^k)$. The last inequality holds due to Lemma G.3. Then by applying Equation (81) to Equation (80), we complete the proof of Equation (79).

I Proof of Theorem 3.2

In this section, we provide the proof for Theorem 3.2.

Proof of Theorem 3.2. For any target failure probability $\delta \in (0,1)$, run Algorithm 1 with its failure-probability parameter set to $\delta/8$, so that $\iota = \log(16SAT/\delta)$. Applying Lemma F.1 with failure parameter $\delta/8$ and taking a union bound, $\bigcap_{i=1}^{8} \mathcal{E}_i$ holds with probability at least $1-\delta$.

Next, we will prove Theorem 3.2 under the event $\bigcap_{i=1}^{8} \mathcal{E}_i$.

From the proof of Theorem 2 in Zhang et al. (2020), we have:

$$N_{\text{switch}} \leq \sum_{s,a,h} 4H \log\left(\frac{N_h^{K+1}(s,a)}{2H} + 1\right).$$

Next for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we will bound the term $N_h^{K+1}(s,a)$. Let $\mathcal{A}_h^\star(s) = \{a \mid a \text{ maximizes } Q_h^\star(s,\cdot)\}$, which is the set of optimal actions for state-step pair (s,h). For $a \notin \mathcal{A}_h^\star(s)$, we have $\Delta_h(s,a) > 0$ and then $\Delta_h(s,a) \geq \Delta_{\text{min}}$. For any $h \in [H]$, let Z_h be the set of all triples of (s,a,h) such that $a \notin \mathcal{A}_h^\star(s)$, i.e., $Z_h = \{(s,a,h) | a \notin \mathcal{A}_h^\star(s)\}$.

We also let the set $Z = \bigcup_{h=1}^{H} Z_h$ and the set $Z_{\text{opt}} = \{(s,a,h) | a \in \mathcal{A}_h^\star(s)\}$. Then we have $|Z| + |Z_{\text{opt}}| = SAH$. Since every state-step pair (s,h) has at least one optimal action, we know $|Z_{\text{opt}}| \geq SH$ and then $0 < |Z| \leq SH(A-1)$.

If for given $(h,k) \in [H] \times [K]$, $(s_h^k, a_h^k, h) \in Z_h$, we have $\Delta_h(s_h^k, a_h^k) \geq \Delta_{\text{min}}$. Then it holds that:

$$Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) = V_h^k(s_h^k) - Q_h^\star(s_h^k, a_h^k) \geq \Delta_h(s_h^k, a_h^k) \geq \Delta_{\text{min}}.$$

The first inequality is because $V_h^k(s) \geq V_h^\star(s)$ by Lemma G.1. Therefore, we have

$$\sum_{(s,a,h)\in Z_h} \mathbb{I}[(s_h^k, a_h^k) = (s,a)] = \mathbb{I}[(s_h^k, a_h^k, h) \in Z_h] \leq \mathbb{I}\left[Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \geq \Delta_{\text{min}}\right]$$

and thus

$$\begin{aligned}
\sum_{(s,a,h)\in Z} N_h^{K+1}(s,a) &= \sum_{h=1}^{H} \sum_{(s,a,h)\in Z_h} \sum_{k=1}^{K} \mathbb{I}[(s_h^k, a_h^k) = (s,a)] \\
&\leq \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{I}\left[Q_h^k(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \geq \Delta_{\text{min}}\right] \\
&\leq O\left(\frac{\sum_{h=1}^{H} \left(\mathbb{Q}_h^\star + \beta^2\right) H^3 SA\iota}{\Delta_{\text{min}}^2} + \frac{H^6 SA\iota^2}{\beta \Delta_{\text{min}}}\right). \qquad (82)
\end{aligned}$$

The last inequality is by Lemma I.1. Therefore we have:

$$\begin{aligned}
N_{\text{switch}} &\leq \sum_{s,a,h} 4H \log\left(\frac{N_h^{K+1}(s,a)}{2H} + 1\right) \\
&= \sum_{(s,a,h)\in Z} 4H \log\left(\frac{N_h^{K+1}(s,a)}{2H} + 1\right) + \sum_{(s,a,h)\in Z_{\text{opt}}} 4H \log\left(\frac{N_h^{K+1}(s,a)}{2H} + 1\right) \\
&\leq O\Bigg(H(SAH - |Z_{\text{opt}}|) \log\left(\frac{\sum_{h=1}^{H} \left(\mathbb{Q}_h^\star + \beta^2\right) H^2 SA\iota}{(SAH - |Z_{\text{opt}}|)\Delta_{\text{min}}^2} + \frac{H^5 SA\iota^2}{(SAH - |Z_{\text{opt}}|)\beta\Delta_{\text{min}}}\right) \\
&\qquad + H|Z_{\text{opt}}| \log\left(\frac{K}{|Z_{\text{opt}}|} + 1\right)\Bigg) \\
&\leq O\left(H|Z_{\text{opt}}^c| \log\left(\frac{H^6 SA\iota^2}{|Z_{\text{opt}}^c|\beta\Delta_{\text{min}}^2}\right) + H|Z_{\text{opt}}| \log\left(\frac{K}{|Z_{\text{opt}}|} + 1\right)\right) \\
&= O\left(H|Z_{\text{opt}}| \log\left(\frac{K}{|Z_{\text{opt}}|} + 1\right) + H|Z_{\text{opt}}^c| \log\left(\frac{H^{\frac{5}{2}} S^{\frac{1}{2}} A^{\frac{1}{2}} \log(\frac{SAT}{\delta})}{|Z_{\text{opt}}^c|^{\frac{1}{2}} \Delta_{\text{min}}}\right)\right).
\end{aligned}$$

The second inequality is because of Jensen's inequality and Equation (82). The last inequality holds due to

$$\frac{\sum_{h=1}^{H} \left(\mathbb{Q}_h^\star + \beta^2\right) H^2 SA\iota}{(SAH - |Z_{\text{opt}}|)\,\Delta_{\text{min}}^2} \leq \frac{H^6 SA\iota^2}{(SAH - |Z_{\text{opt}}|)\,\beta\Delta_{\text{min}}^2},$$

and

$$\frac{H^5SA\iota^2}{(SAH-|Z_{\text{opt}}|)\beta\Delta_{\min}} \le \frac{H^6SA\iota^2}{(SAH-|Z_{\text{opt}}|)\beta\Delta_{\min}^2}$$

by $\mathbb{Q}_h^\star \le H^2$ for all h and $0 < \beta, \Delta_{\min} \le H$. The last equation holds by setting $\beta = H$. Then we finish the proof.

In particular, if each state-step pair has a unique optimal action, which means $|Z_{\text{opt}}| = SH$, then the policy switching cost is upper bounded by:

$$O\left(H^2SA\log\left(\frac{H^2\log(\frac{SAT}{\delta})}{\Delta_{\min}}\right) + H^2S\log\left(\frac{K}{HS}+1\right)\right).$$

□

Now we prove Lemma I.1.

Lemma I.1. *Under the event $\bigcap_{i=1}^6 \mathcal{E}_i$ in Lemma F.1, for all $\epsilon \in (0, H]$, we have the following conclusion:*

$$\sum_{h=1}^H\sum_{k=1}^K \mathbb{I}\left[Q_h^k(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \ge \epsilon\right] \lesssim \frac{\sum_{h=1}^H\left(\mathbb{Q}_h^\star+\beta^2\right)H^3SA\iota}{\epsilon^2} + \frac{H^6SA\iota^2}{\beta\epsilon}.$$

Proof of Lemma I.1. By Lemma H.2, for any given $h \in [H]$ and non-negative weight sequence $\{\omega_h^k\}_k$ at step h, it holds simultaneously for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and subsequent step $h' \in [h, H]$ that:

$$\begin{aligned}
&\sum_{k=1}^K \omega_h(k,h')(Q_{h'}^k - Q_{h'}^\star)(s_{h'}^k, a_{h'}^k) \\
&\lesssim \sum_{h_1=h'}^H\sum_{s,a}\sqrt{(\mathbb{Q}_{h_1}^\star+\beta^2)H\|\omega_h(\cdot,h_1)\|_\infty\|\omega_h(\cdot,h_1,s,a)\|_1\iota} + \sum_{h_1=h'}^H\sum_{k=1}^K\omega_h(k,h_1)Z_{h_1}^k \\
&\lesssim \sqrt{\sum_{h_1=h'}^H(\mathbb{Q}_{h_1}^\star+\beta^2)H^2SA\|\omega_h(\cdot,h_1)\|_\infty\|\omega_h(\cdot,h_1)\|_1\iota} + \sum_{h_1=h'}^H\sum_{k=1}^K\omega_h(k,h_1)Z_{h_1}^k \\
&\lesssim \sqrt{\sum_{h_1=h'}^H\left(\mathbb{Q}_{h_1}^\star+\beta^2\right)H^2SA\|\omega_h(\cdot,h)\|_\infty\|\omega_h(\cdot,h)\|_1\iota} + \sum_{h_1=h'}^H\sum_{k=1}^K\omega_h(k,h_1)Z_{h_1}^k. \qquad (83)
\end{aligned}$$

Here, the second inequality follows from the Cauchy–Schwarz inequality. The last inequality holds due to

$$\|\omega_h(\cdot,h_1)\|_\infty \le \left(1+\frac{1}{H}\right)^{h_1-h}\|\omega_h(\cdot,h)\|_\infty \lesssim \|\omega_h(\cdot,h)\|_\infty,\ \|\omega_h(\cdot,h_1)\|_1 \le \|\omega_h(\cdot,h)\|_1$$

for any $h_1 \in [h, H]$ by part (e) of Lemma E.1.

Then, following the same argument as in the proof of Lemma G.3, let $N = \max\{1, \lceil\log_2(H/\epsilon)\rceil\}$. For any fixed $h \in [H]$, $k \in [K]$, and $i \in [N-1]$, define

$$\omega_h^i(k,h) = \mathbb{I}\left[(Q_h^k - Q_h^\star)(s_h^k,a_h^k) \in \left[2^{i-1}\epsilon, 2^i\epsilon\right)\right] \le 1$$

and

$$\omega_h^N(k,h) = \mathbb{I}\left[(Q_h^k - Q_h^\star)(s_h^k,a_h^k) \in \left[2^{N-1}\epsilon, H\right]\right] \le 1.$$

Moreover, for any $k' \in [K]$ and $h \le h' \le H$, we define $\omega_h^i(k',h')$ analogously as in Equation (35), and denote the corresponding norms $\|\omega_h^i(\cdot,h')\|_1$ and $\|\omega_h^i(\cdot,h')\|_\infty$ in the same manner as in Equation (36). Then we have:

$$\sum_{k=1}^K \omega_h^i(k,h)\left(Q_h^k - Q_h^\star\right)(s_h^k,a_h^k) \ge 2^{i-1}\epsilon\|\omega_h^i(\cdot,h)\|_1. \qquad (84)$$

By Equation (83), since $\omega_h^i(k,h) \le 1$, we have

$$\sum_{k=1}^{K} \omega_h^i(k,h) \left(Q_h^k - Q_h^\star\right)(s_h^k, a_h^k)$$
$$\lesssim \sum_{h_1=h}^{H}\sum_{k=1}^{K} \omega_h^i(k,h_1) Z_{h_1}^k + \sqrt{\sum_{h_1=h}^{H} \left(\mathbb{Q}_{h_1}^\star + \beta^2\right) H^2 SA \|\omega_h^i(\cdot,h)\|_1 \iota}.$$

Combining the result above and Equation (84), we have the following inequality:

$$2^{i-1}\epsilon\|\omega_h^i(\cdot,h)\|_1 \lesssim \sqrt{\sum_{h_1=h}^{H} \left(\mathbb{Q}_{h_1}^\star + \beta^2\right) H^2 SA \|\omega_h^i(\cdot,h)\|_1 \iota} + \sum_{h_1=h}^{H}\sum_{k=1}^{K} \omega_h^i(k,h_1) Z_{h_1}^k. \tag{85}$$

Solving the inequality in Equation (85), we can derive the following relationship:

$$\|\omega_h^i(\cdot,h)\|_1 \lesssim \frac{\sum_{h_1=h}^{H} \left(\mathbb{Q}_{h_1}^\star + \beta^2\right) H^2 SA\iota}{4^i\epsilon^2} + \frac{\sum_{h_1=h}^{H}\sum_{k=1}^{K} \omega_h^i(k,h_1) Z_{h_1}^k}{2^i\epsilon}. \tag{86}$$

We now prove the conclusion. By Equation (42) and Equation (86), we have

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{I}\left[Q_h^k(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \ge \epsilon\right]$$
$$= \sum_{h=1}^{H}\sum_{i=1}^{N} \|\omega_h^i(\cdot,h)\|_1$$
$$\lesssim \sum_{h=1}^{H}\sum_{i=1}^{N} \frac{\sum_{h_1=h}^{H} \left(\mathbb{Q}_{h_1}^\star + \beta^2\right) H^2 SA\iota}{4^i\epsilon^2} + \sum_{h=1}^{H}\sum_{i=1}^{N} \frac{\sum_{h_1=h}^{H}\sum_{k=1}^{K} \omega_h^i(k,h_1) Z_{h_1}^k}{2^i\epsilon}$$
$$\lesssim \frac{\sum_{h=1}^{H} \left(\mathbb{Q}_h^\star + \beta^2\right) H^3 SA\iota}{\epsilon^2} + \sum_{h=1}^{H}\sum_{i=1}^{N} \frac{\sum_{h_1=h}^{H}\sum_{k=1}^{K} Z_{h_1}^k}{2^i\epsilon}$$
$$\lesssim \frac{\sum_{h=1}^{H} \left(\mathbb{Q}_h^\star + \beta^2\right) H^3 SA\iota}{\epsilon^2} + \frac{H^6 SA\iota^2}{\beta\epsilon}.$$

Here, the second inequality is because $0 \le \omega_h^i(k,h_1) < 3$ by Equation (41). The last inequality holds due to Equation (76). Then we finish the proof. □

J Proof of Theorem 3.3

In this section, we bound the worst-case regret under the event $\bigcap_{i=1}^{8} \mathcal{E}_i$ in Lemma F.1.

As $V_h^k(s) \ge V_h^\star(s)$ for all $(s,h,k) \in \mathcal{S} \times [H] \times [K]$ by Lemma G.1, we have:

$$\text{Regret}(T) = \sum_{k=1}^{K} \left(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \le \sum_{k=1}^{K} \left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right) = \sum_{k=1}^{K} \zeta_1^k.$$

Thus, we only need to bound $\sum_{k=1}^{K} \zeta_1^k$. By Equation (29) and the definition of $\tilde{b}_h^k$ in Equation (19), we have

$$\sum_{k=1}^{K} (Q_h^k - Q_h^\star)(s_h^k, a_h^k)$$
$$\le \sum_{k=1}^{K} \mathbb{I}\left[t_h^k = 1\right] H + 8\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\tilde{W}_h^k(s_h^k,a_h^k)\iota}{n_h^k}} + 119\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{H\iota}{n_h^k}$$
$$+ \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} (V_{h+1}^{k_{l_i}} - V_{h+1}^\star)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k,a_h^k)} + \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s_h^k,a_h^k,h}\right) V_{h+1}^\star}{n_h^k(s_h^k,a_h^k)}. \tag{87}$$

For the lower-order bonus term in Equation (87), by Lemma E.2, it holds that

$$\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{H\iota}{n_h^k} \leq 10H^2SA\iota^2. \tag{88}$$

Moreover, by Equation (47) and Equation (49),

$$\sum_{k=1}^{K} \mathbb{I}\left[t_h^k = 1\right] H \leq H^2SA,$$

$$\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} \left(V_{h+1}^{k_{l_i}} - V_{h+1}^{\star}\right)(s_{h+1}^{k_{l_i}})}{n_h^k(s_h^k, a_h^k)} \leq \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \delta_{h+1}^k.$$

Since $\iota = \log(2SAT/\delta) \geq \log 2$, we have $1 \leq 3\iota^2$. Therefore, combining the above inequalities with Equation (87) yields

$$\begin{aligned}\sum_{k=1}^{K}(Q_h^k - Q_h^\star)(s_h^k, a_h^k) \leq \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \delta_{h+1}^k + 1193H^2SA\iota^2 + 8\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)\iota}{n_h^k}} \\ + \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s_h^k, a_h^k, h}\right) V_{h+1}^\star}{n_h^k(s_h^k, a_h^k)}.\end{aligned}$$

Furthermore,

$$\sum_{k=1}^{K} \mathbb{P}_{s_h^k, a_h^k, h}(V_{h+1}^\star - V_{h+1}^{\pi^k}) = \sum_{k=1}^{K} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right) + \sum_{k=1}^{K} \left(\zeta_{h+1}^k - \delta_{h+1}^k\right).$$

Hence, together with Equation (45), we obtain

$$\begin{aligned}&\sum_{k=1}^{K} \zeta_h^k \leq \sum_{k=1}^{K} \zeta_{h+1}^k + \frac{1}{H}\sum_{k=1}^{K} \delta_{h+1}^k + 1193H^2SA\iota^2 + 8\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)\iota}{n_h^k}} \\ &+ \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s_h^k, a_h^k, h}\right) V_{h+1}^\star}{n_h^k(s_h^k, a_h^k)} + \sum_{k=1}^{K} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right) \\ &\leq \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \zeta_{h+1}^k + 1193H^2SA\iota^2 + 8\sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)\iota}{n_h^k}} \\ &\quad + \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s_h^k, a_h^k, h}\right) V_{h+1}^\star}{n_h^k(s_h^k, a_h^k)} + \sum_{k=1}^{K} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right),\end{aligned}$$

where the last inequality uses $\delta_{h+1}^k \leq \zeta_{h+1}^k$.

By recursively applying the inequality above over $h = 1, 2, \ldots, H$, since $\zeta_{H+1}^k = 0$ and $(1 + 1/H)^H \leq e$, we obtain

$$\text{Regret}(T) \leq \sum_{k=1}^{K} \zeta_1^k \leq 1193eH^3SA\iota^2 + 8R_1 + R_2 + R_3,$$

where

$$R_1 = \sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^{h-1} \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)\iota}{n_h^k}},$$

$$R_2 = \sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^{h-1} \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{\sum_{i=1}^{n_h^k} \left(\mathbb{1}_{s_{h+1}^{k_{l_i}}} - \mathbb{P}_{s_h^k, a_h^k, h}\right) V_{h+1}^\star}{n_h^k(s_h^k, a_h^k)},$$

and

$$R_3 = \sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^{h-1} \sum_{k=1}^{K} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right) \left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right).$$

Under the event $\mathcal{E}_4$ in Lemma F.1, we can bound R_3 by $3\sqrt{2H^2T\iota}$. Then, we reach

$$\text{Regret}(T) \lesssim \sqrt{H^2T\iota} + H^3SA\iota^2 + R_1 + R_2 \lesssim (1+\beta)\sqrt{H^3SAT\iota} + \left(1 + \frac{1}{\beta}\right) H^{\frac{9}{2}}SA\iota^2.$$

The last inequality holds due to Lemma J.1 and J.2.

For any target failure probability $\delta \in (0,1)$, run Algorithm 1 with its failure-probability parameter set to $\delta/8$, so that $\iota = \log(16SAT/\delta)$. Applying Lemma F.1 with failure parameter $\delta/8$ and taking a union bound, the event $\bigcap_{i=1}^{8} \mathcal{E}_i$ holds with probability at least $1-\delta$. On this event, setting $\beta = 1$, we have

$$\text{Regret}(T) \lesssim \sqrt{H^3SAT\iota} + H^{\frac{9}{2}}SA\iota^2.$$

J.1 Upper Bounds of R_1 and R_2

Lemma J.1. *Under the event $\bigcap_{i=1}^{8} \mathcal{E}_i$ in Lemma F.1, we have*

$$R_1 \lesssim (1+\beta)\sqrt{H^3SAT\iota} + \left(1 + \frac{1}{\beta}\right) H^{\frac{9}{2}}SA\iota^{\frac{3}{2}}.$$

Proof of Lemma J.1. Since $\tilde{W}_h^k(s_h^k, a_h^k) \geq 0$, we have

$$R_1 \lesssim \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)\iota}{n_h^k(s_h^k, a_h^k)}}.$$

By Equation (59), we have

$$\sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)}{n_h^k(s_h^k, a_h^k)}} \lesssim \frac{\sqrt{H\Psi_h^k(s_h^k, a_h^k)}}{n_h^k(s_h^k, a_h^k)} + \sqrt{\frac{\mathbb{V}_{s_h^k, a_h^k, h}(V_{h+1}^\star)}{n_h^k(s_h^k, a_h^k)}} + \sqrt{\frac{\beta^2}{n_h^k(s_h^k, a_h^k)}} + \frac{H\iota^{\frac{1}{4}}}{(n_h^k(s_h^k, a_h^k))^{\frac{3}{4}}}. \tag{89}$$

Note that by Lemma E.2 with $\alpha = \frac{1}{2}$ and $\frac{3}{4}$, we have

$$\sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\beta^2}{n_h^k(s_h^k, a_h^k)}} \lesssim \beta\sqrt{H^2SAT},$$

and

$$\sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \frac{H\iota^{\frac{1}{4}}}{(n_h^k(s_h^k, a_h^k))^{\frac{3}{4}}} \lesssim H^{\frac{5}{2}}(SA)^{\frac{3}{4}}(T\iota)^{\frac{1}{4}}.$$

Applying these two inequalities to Equation (89), it holds that

$$\begin{aligned}\sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\tilde{W}_h^k(s_h^k, a_h^k)}{n_h^k(s_h^k, a_h^k)}} &\lesssim \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{I}[t_h^k > 1] \frac{\sqrt{H\Psi_h^k(s_h^k, a_h^k)}}{n_h^k(s_h^k, a_h^k)} \\ &+ \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{I}\left[t_h^k > 1\right] \sqrt{\frac{\mathbb{V}_{s_h^k, a_h^k, h}(V_{h+1}^\star)}{n_h^k(s_h^k, a_h^k)}} + H^{\frac{5}{2}}(SA)^{\frac{3}{4}}(T\iota)^{\frac{1}{4}} + \beta\sqrt{H^2SAT}. \end{aligned} \tag{90}$$

By Equation (80) and Equation (81), we have

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\frac{\sqrt{H\Psi_h^k(s_h^k,a_h^k)}}{n_h^k(s_h^k,a_h^k)}\lesssim\sqrt{\frac{1}{\beta}}H^{\frac{9}{2}}SA\iota. \tag{91}$$

We further claim the following conclusion

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\sqrt{\frac{\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)}{n_h^k(s_h^k,a_h^k)}}\lesssim\sqrt{H^3SAT}+H^{\frac{9}{2}}SA\sqrt{\iota}, \tag{92}$$

which will be proved later. Then combining the results of Equation (90), Equation (92) and Equation (91), we have

$$\begin{aligned}R_1&\lesssim(1+\beta)\sqrt{H^3SAT\iota}+H^{\frac{5}{2}}(SA\iota)^{\frac{3}{4}}T^{\frac{1}{4}}+\left(1+\sqrt{\frac{1}{\beta}}\right)H^{\frac{9}{2}}SA\iota^{\frac{3}{2}}\\&\lesssim(1+\beta)\sqrt{H^3SAT\iota}+\left(1+\frac{1}{\beta}\right)H^{\frac{9}{2}}SA\iota^{\frac{3}{2}}.\end{aligned}$$

The last inequality follows from the AM–GM inequality:

$$\begin{aligned}H^{\frac{5}{2}}(SA\iota)^{\frac{3}{4}}T^{\frac{1}{4}}&=\sqrt{\sqrt{H^3SAT\iota}\,H^{\frac{7}{2}}SA\iota}\\&\le\frac{1}{2}\sqrt{H^3SAT\iota}+\frac{1}{2}H^{\frac{7}{2}}SA\iota\\&\lesssim\sqrt{H^3SAT\iota}+H^{\frac{9}{2}}SA\iota^{\frac{3}{2}}.\end{aligned}$$

Next, we will prove Equation (92).

Proof of Equation (92):

Because of Lemma E.2 with $\omega_h^k=1$, $\alpha=\frac{1}{2}$, we know

$$\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\frac{\mathbb{I}\left[(s_h^k,a_h^k)=(s,a)\right]}{\sqrt{n_h^k(s,a)}}\lesssim\sqrt{HN_h^{K+1}(s,a)}. \tag{93}$$

Then we can derive the following relationship

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\sqrt{\frac{\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)}{n_h^k(s_h^k,a_h^k)}}\\&=\sum_{h=1}^{H}\sum_{s,a}\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\sqrt{\frac{\mathbb{V}_{s,a,h}(V_{h+1}^\star)}{n_h^k(s,a)}}\mathbb{I}\left[(s_h^k,a_h^k)=(s,a)\right]\\&\lesssim\sum_{h=1}^{H}\sum_{s,a}\sqrt{HN_h^{K+1}(s,a)\mathbb{V}_{s,a,h}(V_{h+1}^\star)}\\&\le\sqrt{H^2SA}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)}.\end{aligned} \tag{94}$$

The first inequality holds due to Equation (93). The last inequality is by the Cauchy–Schwarz inequality. Next, we will bound the right-hand side of Equation (94). We first note that

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)&\le\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^{\pi^k})+\sum_{h=1}^{H}\sum_{k=1}^{K}\left|\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)-\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^{\pi^k})\right|\\&\lesssim HT+H^3\iota+\sum_{h=1}^{H}\sum_{k=1}^{K}\left|\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)-\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^{\pi^k})\right|.\end{aligned} \tag{95}$$

The last inequality follows directly from $\mathcal{E}_7$ in Lemma F.1. The last term on the right-hand side of Equation (95) can be bounded as follows

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\left|\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)-\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^{\pi^k})\right| \\
&=\sum_{h=1}^{H}\sum_{k=1}^{K}\left|\mathbb{P}_{s_h^k,a_h^k,h}(V_{h+1}^\star)^2-\mathbb{P}_{s_h^k,a_h^k,h}(V_{h+1}^{\pi^k})^2-(\mathbb{P}_{s_h^k,a_h^k,h}V_{h+1}^\star)^2+(\mathbb{P}_{s_h^k,a_h^k,h}V_{h+1}^{\pi^k})^2\right| \\
&\le\sum_{h=1}^{H}\sum_{k=1}^{K}\left\{\mathbb{P}_{s_h^k,a_h^k,h}\left((V_{h+1}^\star-V_{h+1}^{\pi^k})(V_{h+1}^\star+V_{h+1}^{\pi^k})\right)+\left|(\mathbb{P}_{s_h^k,a_h^k,h}V_{h+1}^\star)^2-(\mathbb{P}_{s_h^k,a_h^k,h}V_{h+1}^{\pi^k})^2\right|\right\} \\
&\le 4H\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{P}_{s_h^k,a_h^k,h}\left(V_{h+1}^\star-V_{h+1}^{\pi^k}\right) \\
&=4H\sum_{h=1}^{H}\sum_{k=1}^{K}\left\{V_{h+1}^\star(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k)+\left(\mathbb{P}_{s_h^k,a_h^k,h}-\mathbb{1}_{s_{h+1}^k}\right)(V_{h+1}^\star-V_{h+1}^{\pi^k})\right\} \\
&\overset{\text{(i)}}{\lesssim} H\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_{h+1}^k-V_{h+1}^{\pi^k}\right)(s_{h+1}^k)+H^2\sqrt{T\iota} \\
&\overset{\text{(ii)}}{\lesssim} H^4\sqrt{SAT\iota}+H^5SA.
\end{aligned}
\tag{96}
$$

Here, (i) holds due to the event $\mathcal{E}_8$ in Lemma F.1. (ii) follows from the bound on $\sum_{k=1}^{K}\zeta_{h+1}^k$ established by the proof of Lemma G.4. Therefore, applying Equation (96) to Equation (95), we reach

$$
\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)\lesssim HT+H^7SA\iota. \tag{97}
$$

Here we use $H^4\sqrt{SAT\iota}\le HT+H^7SA\iota$ by the AM-GM inequality. Applying Equation (97) to Equation (94), we finish the proof of Equation (92):

$$
\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\sqrt{\frac{\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)}{n_h^k(s_h^k,a_h^k)}}\lesssim\sqrt{H^2SA}\sqrt{HT+H^7SA\iota}\le\sqrt{H^3SAT}+H^{\frac{9}{2}}SA\sqrt{\iota}.
$$

□

Lemma J.2. *Under the event $\bigcap_{i=1}^{8}\mathcal{E}_i$ in Lemma F.1, we have*

$$
R_2\lesssim\sqrt{H^3SAT\iota}+H^{\frac{9}{2}}SA\iota^2.
$$

Proof of Lemma J.2. By the event $\mathcal{E}_5$ in Lemma F.1, we have

$$
\begin{aligned}
R_2&=\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\frac{\sum_{i=1}^{n_h^k}\left(\mathbb{1}_{s_{h+1}^{k_{l_i}}}-\mathbb{P}_{s_h^k,a_h^k,h}\right)V_{h+1}^\star}{n_h^k(s_h^k,a_h^k)} \\
&\lesssim\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{I}\left[t_h^k>1\right]\left(\sqrt{\frac{\iota}{n_h^k}\mathbb{V}_{s_h^k,a_h^k,h}(V_{h+1}^\star)}+\frac{H\iota}{n_h^k}\right) \\
&\lesssim\sqrt{H^3SAT\iota}+H^{\frac{9}{2}}SA\iota^2.
\end{aligned}
$$

The last inequality is by Equation (92) and Lemma E.2 with $\alpha=1$. □

K PROOF OF THEOREM 3.4

Because $\pi_h^k(s) = \arg\max_a Q_h^k(s,a)$, we have

$$\mathbb{I}\left[\pi_h^k(s) \neq \pi_h^{k+1}(s)\right] \leq \mathbb{I}\left[\exists a,\ Q_h^{k+1}(s,a) \neq Q_h^k(s,a)\right].$$

Therefore,

$$\begin{aligned}\sum_{k=1}^{K-1} \mathbb{I}\left[\pi_h^k(s) \neq \pi_h^{k+1}(s)\right] &\leq \sum_{k=1}^{K-1} \mathbb{I}\left[\exists a,\ Q_h^{k+1}(s,a) \neq Q_h^k(s,a)\right] \\ &\leq \sum_{a\in\mathcal{A}} \left(t_h^K(s,a) - 1\right).\end{aligned}$$

Hence, the total local switching cost of UCB-Bernstein+ is bounded by

$$\begin{aligned}N_{\text{switch}} &= \sum_{s\in\mathcal{S}} \sum_{h=1}^{H} \sum_{k=1}^{K-1} \mathbb{I}\left[\pi_h^k(s) \neq \pi_h^{k+1}(s)\right] \\ &\leq \sum_{h=1}^{H} \sum_{s,a} \left(t_h^K(s,a) - 1\right) \\ &\leq 4H^2SA \log\left(\frac{K}{HSA} + 1\right),\end{aligned}$$

where the last inequality follows from Equation (18).