Provably Efficient and Agile Randomized Q-Learning

He Wang[*] Carnegie Mellon University

Xingyu Xu[*] Carnegie Mellon University

Yuejie Chi[†] Yale University

Abstract

While Bayesian-based exploration often demonstrates superior empirical performance compared to bonus-based methods in model-based reinforcement learning (RL), its theoretical understanding remains limited for model-free settings. Existing provable algorithms either suffer from computational intractability or rely on stage-wise policy updates which reduce responsiveness and slow down the learning process. In this paper, we propose a novel variant of Q-learning algorithm, referred to as RANDOMIZEDQ, which integrates *sampling-based exploration with agile, step-wise, policy updates*, for episodic tabular RL. We establish a sublinear regret bound $\widetilde{O}(\sqrt{H^5SAT})$, where S is the number of states, A is the number of actions, H is the episode length, and T is the total number of episodes. In addition, we present a logarithmic regret bound $O\left(\frac{H^6SA}{\Delta_{\min}}\log^5(SAHT)\right)$ when the optimal Q-function has a positive sub-optimality $\Delta_{\min}$. Empirically, RANDOMIZEDQ exhibits outstanding performance compared to existing Q-learning variants with both bonus-based and Bayesian-based exploration on standard benchmarks.

1 Introduction

In reinforcement learning (RL) [1], an agent aims to learn an optimal policy that maximizes its cumulative rewards through interactions with an unknown environment. Broadly speaking, RL algorithms can be categorized into two main approaches—model-based and model-free methods—depending on whether they first learn a model of the environment and plan within it, or directly learn the optimal policy from experience. While model-based approaches offer advantages in sample efficiency, model-free algorithms tend to be more computationally efficient and take lower space complexity, making them more attractive for deployment in many real-world applications, such as games [2], robotics control [3] and language model training [4].

As one of fundamental challenges in RL, the *exploitation-exploration dilemma* remains particularly difficult to address in the model-free paradigm, i.e., the learned policy needs to carefully balance between exploiting current observations and exploring unseen state-action pairs to maximize total rewards in the long term. To manage the trade-off, most provably efficient model-free algorithms adopt the principle of *optimism in the face of uncertainty*, incentivizing exploration by assigning bonuses to uncertain outcomes, constructed from their upper confidence bound (UCB) [5]. In particular, prior works [6–8] showed that Q-learning augmented with tailored bonus functions achieve comparable sample complexity to their model-based counterparts.

In contrast to bonus-based exploration methods aforementioned, Bayesian-based approaches have gained increasing attention for their superior empirical performance [11, 12]. These approaches enhance efficient exploration by leveraging the inherent randomness in sampling from posteriors that are updated based on prior observations. However, theoretical understandings have been limited, where the majority of prior work has focused on model-based RL [13–15].

[*]Department of Electrical and Computer Engineering, Carnegie Mellon University, PA, USA. Correspondence to: He Wang <hew2@andrew.cmu.edu>.

[†]Department of Statistics and Data Science, Yale University, CT, USA.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

Key Property	Conditional-PS [9]	Staged-RandQL [10]	RandQL [10]	RANDOMIZEDQ (This Work)
Computational tractability	✗	✓	✓	✓
Agile policy update	✓	✗	✓	✓
Gap-independent regret guarantee	✓	✓	✗	✓
Gap-dependent regret guarantee	✗	✗	✗	✓

Table 1: Comparison with the most relevant model-free RL methods with Bayesian-based exploration in tabular settings. A ✓ indicates the method possesses the corresponding property, while a ✗ denotes its absence. We identify and fix a technical gap in Tiapkin et al. [10], which preserves the gap-independent regret guarantee of Staged-RandQL. Notably, our method uniquely achieves *computational tractability*, *agile policy updates*, and *provable regret guarantees*, distinguishing it from prior work.

When it comes to model-free RL, research is even more limited in several aspects. Dann et al. [9] proposed a *sample-efficient* algorithm that draws Q-functions directly from the posterior distribution. Nevertheless, this approach suffers from *computational inefficiency*. More recently, Tiapkin et al. [10] introduced posterior sampling via randomized learning rates, but unfortunately they only provided theoretical guarantees[3] for *stage-wise policy updates*, which are known to be inefficient in practice as this staging approach does not allow agents to respond agilely to the environment. To this end, it is natural to ask:

Is it possible to design a model-free RL algorithm with Bayesian-based exploration, achieving ***sample efficiency***, ***computational efficiency***, *and* ***agile policy updates****?*

1.1 Main contribution

To answer this question, we focus on learning a near-optimal policy through sampling-based Q-learning, in a provably sample- and computation-efficient manner. As in Jin et al. [6], Dann et al. [9], Tiapkin et al. [10], throughout this paper, we consider tabular, finite-horizon Markov Decision Processes (MDPs) in the online setting. Below we summarize the highlights of this work:

- We propose RANDOMIZEDQ, a sampling-based Q-learning algorithm which leverages tailored randomized learning rates to enable both efficient exploration and agile policy updates.
- We establish a gap-independent regret bound on the order of $\widetilde{O}(\sqrt{H^5SAT})$, where S is the number of states, A is the number of actions, H is the episode length, and T is the number of episodes.
- Under a strictly positive sub-optimality gap $\Delta_{\min}$ of the optimal Q-function, we further prove a logarithmic regret bound of $O\left(H^6SA\log^5(SAHT)/\Delta_{\min}\right)$. To the best of our knowledge, this is the first result showing model-free algorithms can achieve logarithmic regret via sampling-based exploration.
- Empirically, RANDOMIZEDQ consistently outperforms existing bonus-based and sampling-based model-free algorithms on standard exploration benchmarks, validating its efficacy.

A detailed comparison with pertinent works is provided in Table 1.

1.2 Related works

In this section, we discuss closely-related prior works on optimistic Q-learning and online RL with Bayesian-based exploration, focusing on the tabular setting.

[3] A careful examination of their proof reveals a critical technical gap in their analysis. We provide a novel fix with substantial new analyses, which fortunately preserves their claimed theoretical guarantee. We discuss this in more detail in Section 4.1.

Q-learning with bonus-based exploration. Q-learning and its variants [16–18] are among the most widely studied model-free RL algorithms. To understand its theoretical guarantees, several works have equipped Q-learning with UCB bonuses derived from the principle of optimism in the face of uncertainty [6–8, 19–22]. Notably, Jin et al. [6] first introduced UCB-Q, which augments Q-learning with Hoeffding-type or Bernstein-type bonuses and established a nearly optimal regret bound. Building upon this, Zhang et al. [7] proposed a variance-reduced version of UCB-Q, achieving an optimal sample complexity, and Li et al. [8] further improved the performance by reducing the burn-in cost.

In addition to the worst-case regret bound, gap-dependent regret bounds often leverage benign properties of the environment and enjoy logarithmic regret bounds [21, 22] For instance, Yang et al. [21] showed that UCB-Q has a logarithmic regret bound under the positive sub-optimality gap assumption, and Zheng et al. [22] incorporated error decomposition to establish a gap-dependent bound for Q-learning with variance reduction techniques [7, 8].

Model-based RL with Bayesian-based exploration. Extensive works have investigated the theoretical and empirical performance of Bayesian-based exploration in model-based RL. One popular approach is posterior sampling for reinforcement learning [13, 14, 23–26], where the policy is iteratively learned by sampling a model from its posterior distribution over MDP models. The approach has been shown to achieve the optimal regret bound when UCB on Q-functions are also incorporated [15]. In addition, several works [27–30] have investigated posterior sampling with linear function approximation.

Model-free RL with Bayesian-based exploration. To overcome the computational inefficiency of model-based methods, several algorithms have been developed in model-free RL with Bayesian-based exploration, showing promising empirical results [11, 12, 31] but lacking theoretical guarantees. Recently, Dann et al. [9] sampled Q-functions directly from the posterior, but such an approach is computationally intractable. To address this, Tiapkin et al. [10] introduced RandQL, the first tractable model-free posterior sampling-based algorithm, which encourages exploration through using randomized learning rates and achieves a regret bound of $\widetilde{O}(\sqrt{SAH^5T})$ when RandQL is staged. However, the slow policy update empirically leads to a significantly degraded performance. This leaves a gap between the theoretical efficiency and practical performance in model-free RL with Bayesian-based exploration.

Notation. Throughout this paper, we define $\Delta(\mathcal{S})$ as the probability simplex over a set $\mathcal{S}$, and use $[H] := 1, \dots, H$ and $[T] := 1, \dots, T$ for positive integers $H, T > 0$. We denote $\mathbb{1}$ as the indicator function, which equals 1 if the specified condition holds and 0 otherwise. For any set $\mathcal{D}$, we write $|\mathcal{D}|$ to represent its cardinality (i.e., the number of elements in $\mathcal{D}$). The beta distribution with parameters α and β is denoted by $\mathrm{Beta}(\alpha, \beta)$. Finally, we use the notations $\widetilde{O}(\cdot)$ and $O(\cdot)$ to describe the order-wise non-asymptotic behavior, where the former omits logarithmic factors.

2 Problem Setup

Finite-horizon MDPs. Consider a tabular finite-horizon MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H, H)$, where $\mathcal{S}$ is the finite state space of cardinality S, $\mathcal{A}$ is the action space of cardinality A, $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function at time step $h \in [H]$, and H is the number of steps within each episode. In each episode, the agent starts from an initial state $s_1 \in \mathcal{S}$ and then interacts with the environment for H steps. In each step $h \in [H]$, the agent observes the current state $s_h \in \mathcal{S}$, selects an action $a_h \in \mathcal{A}$, receives a reward $r_h(s_h, a_h)$, and transitions to the next state $s_{h+1} \sim P_h(\cdot\,|s_h, a_h)$.

Policy, value function and Q-function. We denote $\pi = \{\pi_h\}_{h=1}^H$ as the *policy* of the agent within an episode of H steps, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the action selection probability over the action space $\mathcal{A}$ at the step $h \in [H]$. Given any finite-horizon MDP $\mathcal{M}$, we use the value function V_h^π (resp.Q-function) to denote the expected accumulative rewards starting from the state s (resp.the state-action pair (s, a)) at step h and following the policy π until the end of the episode: for any

$(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) | s_h = s \right],$$
$$Q_h^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a \right].$$

By convention, we set $V_{H+1}^\pi(s) = 0$ and $Q_{H+1}^\pi(s, a) = 0$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and policy π. In addition, we denote $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ as the *deterministic optimal policy*, which maximizes the value function (resp.Q-function) for all states (resp. state-action pairs) among all possible policies, i.e.

$$\begin{aligned} V_h^\star(s) &:= V_h^{\pi^\star}(s) = \max_\pi V_h^\pi(s), \\ Q_h^\star(s, a) &:= Q_h^{\pi^\star}(s, a) = \max_\pi Q_h^\pi(s, a), \end{aligned} \tag{1}$$

where the existence of the optimal policy is well-established [32].

Bellman equations. As the pivotal property of MDPs, the value function and Q-function satisfy the following Bellman consistency equations: for any policy π and any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$Q_h^\pi(s, a) = r_h(s, a) + P_{h,s,a} V_{h+1}^\pi, \tag{2}$$

where we use $P_{h,s,a} := P(\cdot | s, a) \in [0, 1]^{1 \times S}$ to represent the transition probability (row) vector for the state-action pair (s, a) at h-th step. Similarly, we also have the following Bellman optimality equation regarding the optimal policy $\pi^\star$:

$$Q_h^\star(s, a) = r_h(s, a) + P_{h,s,a} V_{h+1}^\star. \tag{3}$$

Learning goal. In this work, our goal is to learn a policy that minimizes the total regret during T episodes, defined as

$$\text{Regret}_T = \sum_{t=1}^{T} \left(V_1^\star(s_1) - V_1^{\pi^t}(s_1) \right), \tag{4}$$

in a computationally efficient and scalable fashion. Here, π^t denotes the learned policy executed in the t-th episode, for every $t \in [T]$.

3 Efficient and Agile Randomized Q-learning

In this section, we introduce a sampling-based variant of Q-learning, referred to as RANDOMIZEDQ, which ensures the policy to be updated in an agile manner and enhances efficient exploration based on random sampling rather than adding explicit bonuses.

3.1 Motivation

Before describing the proposed algorithm in detail, we first review the critical role of learning rates in Q-learning to achieve polynomial sample complexity guarantees, and why such choices cannot be directly extended to the context of Bayesian-based exploration.

The effect of learning rates in Q-learning with UCB bonuses. Upon observing a sample transition (s_h, a_h, s_{h+1}) at the h-th step, the celebrated UCB-Q algorithm [33] updates the corresponding entry of the Q-values as:

$$Q_h(s_h, a_h) \leftarrow (1 - w_m) Q_h(s_h, a_h) + w_m \left[r_h(s_h, a_h) + V_{h+1}(s_{h+1}) + b_m \right],$$

where m is the number of visits to the state-action pair (s_h, a_h) at the h-th step, $w_m = \frac{H+1}{H+m}$ is the learning rate and b_m is the UCB-style bonus term to drive efficient exploration. As detailed in [33], such a learning rate of $O(H/m)$ is essential to ensure that the first earlier observations have negligible influence on the most recent Q-value updates.

Challenges in randomized Q-Learning. In the absence of bonus terms, the exploration is guided by assigning higher weights to important states and leveraging the inherent randomness of sampling from the posterior. As directly sampling Q-functions is computationally intractable [9], recent work [10] encourages exploration by randomizing the learning rates according to Beta distribution with an expected order of $O(1/m)$. However, such a learning rate treats all episodes as equally informative, which can result in high bias and an exponential dependency of the sample complexity on the horizon H. To overcome the resulting exponential dependence on the horizon H, Tiapkin et al. [10] resort to split the learning process, update the policy at exponentially slower frequencies, and reset the temporary Q values to ensure enough optimism. While this strategy mitigates the sample complexity issue, it suffers from practical inefficiencies due to discarding valuable data across stages, and is unsuitable to deploy in time-varying environments. This inefficiency is empirically demonstrated in Tiapkin et al. [10, Appendix I] and Section 5.

Thus, it naturally raises the question: can we simply randomize learning rates with an expected order of $O(H/m)$, as used in UCB-Q [6]? Unfortunately, randomizing learning rates with an expected magnitude of $O(H/m)$ rapidly forgets the earlier episodes that includes the initialization and thus fails to maintain sufficient optimism.

3.2 Algorithm description

Motivated by these limitations, we propose an agile, bonus-free Q-learning algorithm for episodic tabular RL in the online setting, referred to as RANDOMIZEDQ and summarized in Algorithm 1.

The main idea behind RANDOMIZEDQ is to update the policy based on an *optimistic mixture of two Q-function ensembles*—each trained with a tailored distribution of learning rates—to balance between agile exploitation of recent observations and sufficiently optimistic exploration. Specifically, after initializing the counters, value and Q-functions (cf. Line 1 in Algorithm 1), if at the step h of episode t, the current state is $s_h \in \mathcal{S}$, the action a_h is selected greedily with respect to the *current* policy Q-function $Q_h^t(s_h, \cdot)$ (cf. Line 4 in Algorithm 1). Upon observing the next state s_{h+1}, the updates can be boiled down to the following key components.

Two Q-ensembles for adaptation and exploration. To ensure the mixed Q-function with the learning rate scaled as $O(H/m)$, we tailor the probability distribution of the randomized learning rate as:

$$w_m^j \sim \text{Beta}\left(\tfrac{H+1}{\kappa}, \tfrac{m+n_0}{\kappa}\right), \quad \forall j \in [J],$$

where m records the total number of visits to the state-action pair (s_h, a_h) just before current visit (cf. Line 5 in Algorithm 1), n_0 introduces pseudo-transitions to induce optimism, and $\kappa > 0$ controls the concentration level of the distribution and J is the size of temporary Q-ensembles. With these randomized learning rates in hand, the corresponding entry of temporary Q-ensembles is updated in parallel as:

$$\widetilde{Q}_h^j(s_h, a_h) \leftarrow (1 - w_m^j)\widetilde{Q}_h^j(s_h, a_h) + w_m^j\big(r_h(s_h, a_h) + \widetilde{V}_{h+1}(s_{h+1})\big), \tag{5}$$

for any $j \in [J]$, where $\widetilde{V}_{h+1}$ is the optimistic value estimate at the next step, computed *before* processing the current transition and held fixed within this visit. As discussed in Section 3.1, such a learning rate could guarantee a polynomial dependency on the horizon H, but lead to rapidly forgetting the earlier episodes and assigning exponentially decreasing weights on the optimistic initialization. To further emphasize the optimistic initialization, we also introduce another sequence of Q-ensembles as follows: for any $j \in [J]$,

$$\widetilde{Q}_h^{\flat,j}(s_h, a_h) \leftarrow (1 - w_m^{\flat,j})\widetilde{Q}_h^{\flat,j}(s_h, a_h) + w_m^{\flat,j}\big(r_h(s_h, a_h) + \widetilde{V}_{h+1}^{\flat}(s_{h+1})\big), \tag{6}$$

where the randomized learning rates are sampled from $w_m^{\flat,j} \sim \text{Beta}\left(\frac{1}{\kappa^\flat}, \frac{m^\flat + n_0^\flat}{\kappa^\flat}\right)$, $m^\flat$ represents the number of visits during the current stage, and $\widetilde{V}_{h+1}^{\flat}$ is the value estimate updated in a stage-wise manner and frozen for the current visit (cf. Line 21 in Algorithm 1).

Agile policy Q-function via optimistic mixing. Then, to promote optimism, the policy Q-function is computed via optimistic mixing (cf. Line 12 in Algorithm 1). For the current state–action pair (s_h, a_h),

$$Q_h^{t+1}(s_h, a_h) = \eta_{t,h} \max_{j \in [J]} \{\widetilde{Q}_h^{j,t+1}(s_h, a_h)\} + (1 - \eta_{t,h}) \cdot \widetilde{Q}_h^{\flat,t+1}(s_h, a_h), \tag{7}$$

Algorithm 1 RANDOMIZEDQ

Input: Initial state s_1, optimistically-initial value $\{V_h^0\}$, inflation coefficient $\kappa, \kappa^\flat > 0$, ensemble size J, the number of prior transitions $n_0, n_0^\flat > 0$, and mixing rates $\{\eta_{t,h}\}$.

1: **Initialize:** $n_h(s,a), n_h^\flat(s,a), q_h(s,a) \leftarrow 0$; $\widetilde{V}_h(s), \widetilde{V}_h^\flat(s) \leftarrow V_h^0$; $\widetilde{Q}_h^j(s,a), \widetilde{Q}_h^\flat(s,a), \widetilde{Q}_h^{\flat,j}(s,a) \leftarrow r_h(s,a) + V_{h+1}^0$, for any $(j,h,s,a) \in [J] \times [H] \times \mathcal{S} \times \mathcal{A}$.
2: **for** $t \in [T]$ **do**
3: **for** $h = 1, \ldots, H$ **do**
4: Play $a_h = \arg\max_{a \in \mathcal{A}} Q_h(s_h, a)$ and observe the next state $s_{h+1} \sim P_h(\cdot \,|s_h, a_h)$.
5: Set $m \leftarrow n_h(s_h, a_h)$ and $m^\flat \leftarrow n_h^\flat(s_h, a_h)$.
6: /* Update temporary Q-ensembles via randomized learning rates. */
7: **for** $j = 1, \ldots, J$ **do**
8: Sample $w_m^j \sim \text{Beta}\left(\frac{H+1}{\kappa}, \frac{m+n_0}{\kappa}\right)$ and $w_m^{\flat,j} \sim \text{Beta}\left(\frac{1}{\kappa^\flat}, \frac{m^\flat + n_0^\flat}{\kappa^\flat}\right)$.
9: Update $\widetilde{Q}_h^j$ and $\widetilde{Q}_h^{\flat,j}$ via (5) and (6).
10: **end for**
11: /* Update the agile policy Q-function by optimistic mixing. */
12: Update the policy Q-function Q_h via (7).
13: /*Update the policy with step-wise agility. */
14: Update policy $\pi_h(s_h) \leftarrow \arg\max_{a \in \mathcal{A}} Q_h(s_h, a)$.
15: /* Update $\widetilde{V}_h$ optimistically. */
16: Update $\widetilde{V}_h(s_h) \leftarrow \max_{j \in [J]} \widetilde{Q}_h^j(s_h, \pi_h(s_h))$.
17: /* Update visit counters. */
18: Update counter $n_h(s_h, a_h) \leftarrow n_h(s_h, a_h) + 1$ and $n_h^\flat(s_h, a_h) \leftarrow n_h^\flat(s_h, a_h) + 1$.
19: /* At the end of the stage: update $\widetilde{Q}_h^\flat$, $\pi_h^\flat$, $\widetilde{V}_h^\flat$ and reset $n_h^\flat$, $\{\widetilde{Q}_h^{\flat,j}\}$. */
20: **if** $n_h^\flat(s_h, a_h) = \lfloor (1 + 1/H)^q H \rfloor$ for the stage $q = q_h(s_h, a_h)$ **then**
21: Update $\widetilde{Q}_h^\flat(s_h, a_h) \leftarrow \max_{j \in [J]} \widetilde{Q}_h^{\flat,j}(s_h, a_h)$, $\pi_h^\flat(s_h) \leftarrow \arg\max_{a \in \mathcal{A}} \widetilde{Q}_h^\flat(s_h, a)$, and $\widetilde{V}_h^\flat(s_h) \leftarrow \widetilde{Q}_h^\flat(s_h, \pi_h^\flat(s_h))$.
22: Reset $\widetilde{Q}_h^{\flat,j}(s_h, a_h) \leftarrow r_h(s_h, a_h) + V_{h+1}^0$ for $j \in [J]$ and $n_h^\flat(s_h, a_h) \leftarrow 0$.
23: Move to the next stage: $q_h(s_h, a_h) \leftarrow q_h(s_h, a_h) + 1$.
24: **end if**
25: **end for**
26: **end for**

and for all $(s,a) \neq (s_h, a_h)$, we keep $Q_h^{t+1}(s,a) = Q_h^t(s,a)$. Here, $\widetilde{Q}_h^{j,t+1}(s_h, a_h)$ is the j-th temporary Q-value, updated at the current visit of episode t via (5), while the staged $\widetilde{Q}_h^{\flat,t+1}$ remains fixed throughout the current stage and is optimistically refreshed to the ensemble maximum only at the end of the stage (cf. Line 21 in Algorithm 1). Note that the first term—corresponding to the maximum over J temporary Q-values—is updated every step, which allows RANDOMIZEDQ to perform *agile policy updates* rather than the *exponentially slower* schedule that updates only at stage boundaries. Such optimistic mixing allows RANDOMIZEDQ to remain responsive to new data and adapt the policy efficiently without requiring periodic updates.

Reset for bias mitigation and optimism restoration. To mitigate outdated data and ensure optimism, we reset the temporary Q-ensembles $\widetilde{Q}_h^{\flat,j}$ according to the optimistic initialization V_{h+1}^0 and the visit counter (cf. Line 22 in Algorithm 1), when the number of visits in current stage exceeds a predefined threshold—specifically, when $n_h^\flat(s_h, a_h) = \lfloor (1 + 1/H)^q H \rfloor$ for the q-th stage. Meanwhile, the staged value estimate $\widetilde{V}_h^\flat$ is greedily updated with respect to $\widetilde{Q}_h^\flat$ at state s_h (cf. Line 21 in Algorithm 1), which will be reused to update the temporary Q-values in the subsequent visits within the new stage.

4 Theoretical Guarantee

In this section we provide both gap-independent and gap-dependent regret bounds for RANDOMIZEDQ, considering the worst-case scenario and favorable structural MDPs, respectively.

4.1 Gap-independent sublinear regret guarantee

To begin with, the following theorem shows that RANDOMIZEDQ has a $\sqrt{T}$-type regret bound, where the full proof is deferred to Appendix C.

Theorem 1. *Consider $\delta \in (0,1)$. Assume that $J = \lceil c \cdot \log(SAHT/\delta) \rceil$, $\kappa^\flat = c \cdot (\log(SAH/\delta) + \log(T))$, and $n_0^\flat = \lceil c \cdot \log(T) \cdot \kappa \rceil$, where c is some universal constant. Let the initialized value function $V_h^0 = 2(H-h+1)$ for any $h \in [H+1]$, and the mixing rate $\eta_{t,h} = \frac{1}{\sqrt{(1+1/H)^q H+1}}$ where $q = q_h^t(s_h^t, a_h^t)$ is the stage index for any $(t,h) \in [T] \times [H]$. Then, with probability at least $1-\delta$, Algorithm 1 guarantees that*

$$\mathrm{Regret}_T \le \widetilde{O}\left(\sqrt{H^5 SAT}\right).$$

Theorem 1 shows that RANDOMIZEDQ achieves a gap-independent regret bound of $\widetilde{O}(\sqrt{H^5SAT})$, matching the guarantees of UCB-Q with Hoeffding-type bonuses [6] in episodic tabular MDPs. This bound is minimax-optimal up to polynomial factors of H when compared to the known lower bound of $\Omega(\sqrt{H^3SAT})$ [6, 34].

Technical challenges. The primary challenge in analyzing RANDOMIZEDQ arises from several subtle requirements on the randomized learning rates. Specifically, these rates must:

- be sufficiently randomized to induce necessary optimism;
- avoid excessive randomness that could incur undesirable fluctuations;
- support efficient exploitation of the most recent observations to avoid introducing exponential dependence on the horizon H.

The subtle interplay among these conditions precludes the straightforward application of existing analytical techniques from the literature. For instance, the optimism may decay exponentially and be insufficient for sparse reward scenarios, so we re-inject the weighted optimistic values into the Q-ensembles at every stage to ensure necessary optimism at every step. In addition, to bound undesirable fluctuations of randomized learning rates, prior work [10] attempted to prove a concentration inequality based on Rosenthal's inequality [10, Theorem 6], which in turn requires a martingale property of the so-called *aggregated* learning rates. However, the martingale property in fact does not hold (detailed below), revealing a gap in their proof. We propose a new proof strategy to bridge this gap and to extend the concentration inequality to our setting.

These challenges jointly necessitate a carefully constructed mixing scheme that balances the efficient exploration and agile responses to latest observations, refined control of fluctuation and favorable properties of learning rates to ensure that RANDOMIZEDQ attains near-optimal sample complexity with agile updates.

Identifying and fixing a technical gap in the proof of [10]. While Tiapkin et al. [10] established a comparable regret bound for Staged-RandQL, it turns out that analysis has a crucial technical gap. Specifically, central to the analysis is to study the concentration of the weighted sum of the *aggregated randomized learning rates*, defined as

$$W_{j,m}^0 = \prod_{k=0}^{m-1} (1 - w_k^j)$$

$$W_{j,m}^i = w_{i-1}^j \prod_{k=i}^{m-1} (1 - w_k^j), \quad \forall i \in [m],$$

which involves bounding the sum

$$\left|\sum_{i=0}^{m} \lambda_i (W_{j,m}^i - \mathbb{E}[W_{j,m}^i])\right|, \tag{8}$$

for fixed real numbers $\lambda_i \in [-1, 1]$; see the proof of Lemma 4 in Tiapkin et al. [10]. To this end, Tiapkin et al. [10] asserted that the partial sums $S_i = \sum_{k=0}^{i} \lambda_k (W_{j,m}^k - \mathbb{E}[W_{j,m}^k])$ form a martingale with respect to some filtration $\mathcal{F}_i$ (cf. Proposition 7 there, which was invoked in the proof of Lemma 4 therein). The proof went on by a standard application of Rosenthal's inequality (i.e., Theorem 6 therein). To prove the martingale property, it was claimed that $W_{j,m}^i$ is adapted to $\mathcal{F}_i$ (cf. assumption of their Theorem 6), and $W_{j,m}^i$ is independent of $\mathcal{F}_{i-1}$ (cf. assumption of their Proposition 7). Unfortunately, it is *impossible* to achieve both adaptedness and independence except for trivial cases (e.g., deterministic learning rates), regardless of choice of $\mathcal{F}_i$.[4] Indeed, if $\{\mathcal{F}_i\}$ is such a filtration, then $W_{j,m}^i$ being adapted means that all the randomness of $W_{j,m}^0, W_{j,m}^1 \cdots, W_{j,m}^{i-1}$ is contained in $\mathcal{F}_{i-1}$. As $W_{j,m}^i$ is independent of $\mathcal{F}_{i-1}$, we see that $W_{j,m}^i$ is independent with all of $W_{j,m}^0, \cdots, W_{j,m}^{i-1}$. By induction, we readily see that $W_{j,m}^0, \cdots, W_{j,m}^m$ are jointly independent. However, since $\sum_{i=0}^{m} W_{j,m}^i = 1$ [10, Lemma 3], such independence is not possible unless all aggregated learning rates $W_{j,m}^i$, $0 \le i \le m$, are deterministic. Thus, Lemma 4 in Tiapkin et al. [10] does not hold, thereby leaving a gap in the analysis. We fix this gap by introducing a reverse filtration that is tailored to the form of the aggregated weights, and study (8) using a backward martingale construction in contrast with partial sums with substantial new analyses. With this approach, we established the correct concentration inequality not only for their setting but also for ours.

Memory and computation complexity. As the number of ensembles is $J = \widetilde{O}(1)$, the computational complexity is $O(H)$ per episode and the space complexity is $O(HSA)$, same as [10]. However, we note that due to the use of optimistic mixing, RandomizedQ requires maintaining two ensembles, which effectively doubles the memory and computational cost compared to Staged-RandQL [10].

4.2 Gap-dependent logarithmic regret guarantee

Note that such a $\sqrt{T}$-type regret bound holds for *any* episodic tabular MDPs, which might not be tight for environment with some benign structural properties. To this end, we further develop a gap-dependent regret bound, which improves the regret bound from sublinear to logarithmic under a strictly positive suboptimality gap condition, as follows.

Assumption 1 (Positive suboptimality gap)**.** *For any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we denote the sub-optimality gap as $\Delta_h(s, a) := V_h^\star(s) - Q_h^\star(s, a)$ and assume that the minimal gap*

$$\Delta_{\min} \triangleq \min_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]} \Delta_h(s, a) \mathbb{1}\{\Delta_h(s, a) \neq 0\} > 0.$$

Note that this assumption implies that there exist some strictly better actions (i.e., the optimal actions) outperform the others for every state. This assumption is mild, as the minimal suboptimality gap $\Delta_{\min} = 0$ only when the MDPs degenerates. Consequently, it is commonly fulfilled in environments with finite action spaces, such as Atari-games and control tasks, and it is also widely adopted in prior literature [21, 22].

Under this mild assumption, we have the following logarithmic gap-dependent regret bound, whose proof is deferred to Appendix D. To the best of our knowledge, this is the *first* guarantee that shows sampling-based Q-learning can also achieve the logarithmic regret in episodic tabular RL.

Theorem 2. *Consider $\delta \in (0, 1)$. Suppose all conditions in Theorem 1 and Assumption 1 hold. Let the mixing rate $\eta_{t,h} = \frac{1}{H(1+\frac{1}{H})^q}$, where $q = q_h^t(s_h^t, a_h^t)$ is the stage index for every $(t, h) \in [T] \times [H]$. Then, Algorithm 1 guarantees that*

$$\mathbb{E}[\mathrm{Regret}_T] \le O\left(\frac{H^6 SA}{\Delta_{\min}} \log^5(SAHT)\right).$$

[4] Tiapkin et al. [10] chose a filtration $\{\mathcal{F}_i\}$ to which $W_{j,m}^i$ is actually not adapted.

Note that the above logarithmic regret bound holds for RandomizedQ without assuming any prior knowledge on the minimal suboptimality gap $\Delta_{\min}$ during implementation. As shown in Simchowitz and Jamieson [35], any algorithm with an $\Omega(\sqrt{T})$ regret bound in the worst case, has a $\log T$-type lower bound of the expected gap-dependent regret. Also, our bound matches the expected regret for UCB-Q [6] under the same condition (i.e. Assumption 1) for episodic tabular MDPs [21], which is nearly tight in S, A, T up to the $\log(SAT)$ and H factors.

5 Experiments

In this section, we present the experimental results of RandomizedQ compared to baseline algorithms, using RLBERRY [36], in the following two environments. All the experiments are conducted on a machine equipped with 2 CPUs (Intel(R) Xeon(R) Gold 6244 CPU), running Red Hat Enterprise Linux 9.4, without GPU acceleration. The corresponding codes can be found at

https://github.com/IrisHeWANG/RandomizedQ

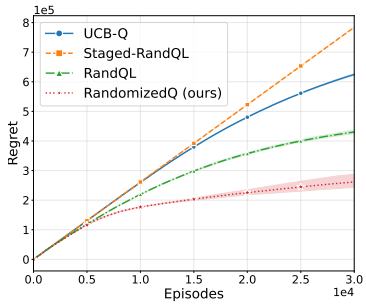
which is built upon the implementation of [10], using the RLBERRY library [36].

A grid-world environment. We first evaluate performance in a 10×10 grid-world environment as used in Tiapkin et al. [10], where each state is represented as a tuple $(i, j) \in [10] \times [10]$, and the agent can choose from four actions: left, right, up, and down. The episode horizon is set to $H = 50$. At each step, the agent moves in the planed direction with probability $1 - \epsilon$ and to a random neighboring state with probability $\epsilon = 0.2$. The agent starts at position $(1, 1)$, and the reward is 1 only at state $(10, 10)$, with all other states yielding zero reward. We also examine the performance in a larger 25×25 grid-world environment, where the agent receives a reward of 1 only at state $(25, 25)$, with the episode horizon set to $H = 200$. Compared to the 10×10 setting, the reward is significantly sparser. The corresponding results are shown in Figures 1a and 1b, respectively.

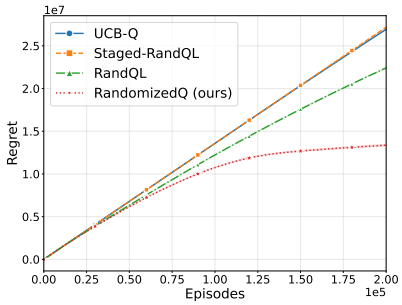
A chain MDP. We also consider a chain MDP environment as described in Osband et al. [11], which consists of $L = 20$ states (i.e., the length of the chain) and two actions: left and right. The episode horizon is set to $H = 50$. With each action, the agent transits in the intended direction with probability 0.9, and in the opposite direction with probability 0.1. The agent starts in the leftmost state, which provides a reward of 0.05, while the rightmost state yields the highest reward of 1. Additionally, we evaluate performance in a longer chain MDP with $L = 50$ states and a horizon of $H = 100$. The corresponding results are shown in Figures 1c and 1d, respectively.

Baselines and experiment setups. We compare RandomizedQ with (1) UCB-Q: model-free Q-learning with bonuses [6] (2) Staged-RandQL [10]: the staged version of RandQL with theoretical guarantees (3) RandQL [10]: the randomized version of UCB-Q, without provable guarantees. For all algorithms with randomized learning rates in both environments, we let the number of ensembles $J = 20$, the inflation coefficient $\kappa = 1$, and the number of pseudo-transitions $n_0 = 1/S$, where S corresponds to the size of the state space in different environments. For RandomizedQ, we set the mixing rate $\eta_{t,h} = (\sqrt{(1 + 1/H)^q H} + 1)^{-1}$, where $q = q_h^t(s_h^t, a_h^t)$, $\forall (t, h) \in [T] \times [H]$. For fair comparison, we run 4 trials per algorithm and show the average along with the 90% confidence interval in the figures above.

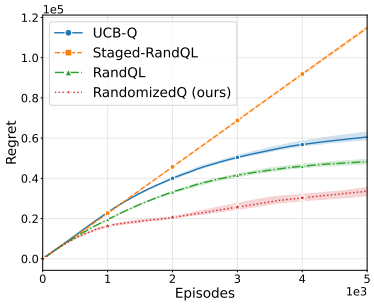
Results. From Figure 1, RandomizedQ exhibits significantly improved performance across all environment sizes, achieving substantially lower total regret. Unlike UCB-Q, which suffers from excessive over-exploration, and the Staged-RandQL that adapts the policy only at the end of each stage, RandomizedQ effectively balances exploration and exploitation through randomized learning rates and agile policy updates. We also observe that RandomizedQ performs even better than the empirical RandQL that lacks theoretical guarantees in prior work [10], especially for larger environments with more sparse rewards. These results validate the effectiveness of sampling-driven updates without explicit bonus terms and highlight the benefit of avoiding stage-wise policy updates in model-free reinforcement learning.



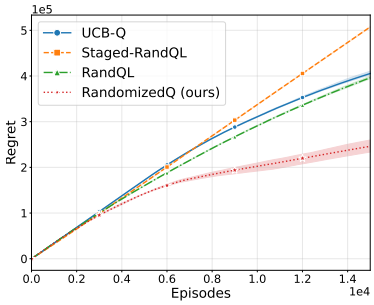
(a) A 10×10 grid-world environment.



(b) A 25×25 grid-world environment.



(c) A chain MDP with length $L = 20$.



(d) A chain MDP with length $L = 50$.

Figure 1: Comparison between RANDOMIZEDQ and baseline algorithms in the grid-world environment (cf. the first row) and the chain MDP (cf. the second row), where total regret is plotted against the number of episodes. RANDOMIZEDQ consistently achieves lower regret than UCB-Q, as well as both the standard randomized Q-learning (i.e., RandQL) and its stage-wise variant (i.e., Staged-RandQL), demonstrating superior sample efficiency and faster learning processes.

6 Conclusion

In this work, we study the performance of Q-learning without exploration bonuses for episodic tabular MDPs in the online setting. We identify two key challenges in existing approaches: the additional statistical dependency introduced by randomizing learning rates, and the inefficiency of slow, stage-wise policy updates, as the bottlenecks of theoretical analysis and algorithm design. To address these challenges, we develop a novel randomized Q-learning algorithm with agile updates called RANDOMIZEDQ, which efficiently adapts the policy to newly observed data. Theoretically, we establish a sublinear worst-case and a logarithmic gap-dependent regret bounds. Empirically, our experiments show that RANDOMIZEDQ significantly outperforms than baseline algorithms in terms of total regret, due to the effective exploration and agile updates.

There are several promising directions for future research. For example, extending our analysis to function approximation settings—such as linear or neural representations—would significantly broaden the applicability of RANDOMIZEDQ [37, 38]. In addition, incorporating variance reduction techniques could further improve the regret bounds and potentially match the theoretical lower bounds [7, 22].

Acknowledgement

This work is supported in part by grants ONR N00014-19-1-2404, NSF CCF-2106778, CNS-2148212, AFRL FA8750-20-2-0504, and by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program.

References

[1] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[3] C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone, "Deep reinforcement learning for robotics: A survey of real-world successes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, 2025, pp. 28 694–28 698.

[4] J. Hong, A. Dragan, and S. Levine, "Q-SFT: Q-learning for language models via supervised fine-tuning," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=v4MTnPiYXY

[5] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

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

[7] Z. Zhang, Y. Zhou, and X. Ji, "Almost optimal model-free reinforcement learning via reference-advantage decomposition," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[8] G. Li, L. Shi, Y. Chen, Y. Gu, and Y. Chi, "Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 762–17 776, 2021.

[9] C. Dann, M. Mohri, T. Zhang, and J. Zimmert, "A provably efficient model-free posterior sampling method for episodic reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 040–12 051, 2021.

[10] D. Tiapkin, D. Belomestny, D. Calandriello, E. Moulines, R. Munos, A. Naumov, P. Perrault, M. Valko, and P. Ménard, "Model-free posterior sampling via learning rate randomization," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[11] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[12] M. Fortunato, M. G. Azar, B. Piot, J. Menick, M. Hessel, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, and S. Legg, "Noisy networks for exploration," in *International Conference on Learning Representations*, 2018.

[13] I. Osband, D. Russo, and B. Van Roy, "(More) efficient reinforcement learning via posterior sampling," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[14] S. Agrawal and R. Jia, "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] D. Tiapkin, D. Belomestny, D. Calandriello, É. Moulines, R. Munos, A. Naumov, M. Rowland, M. Valko, and P. Ménard, "Optimistic posterior sampling for reinforcement learning with few samples and tight guarantees," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 737–10 751, 2022.

[16] C. J. C. H. Watkins, "Learning from delayed rewards," *PhD thesis, King's College, University of Cambridge*, 1989.

[17] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman, "PAC model-free reinforcement learning," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 881–888.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[19] Y. Bai, T. Xie, N. Jiang, and Y.-X. Wang, "Provably efficient Q-learning with low switching cost," *arXiv preprint arXiv:1905.12849*, 2019.

[20] C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain, "Model-free reinforcement learning in infinite-horizon average-reward markov decision processes," in *International conference on machine learning*. PMLR, 2020, pp. 10 170–10 180.

[21] K. Yang, L. Yang, and S. Du, "Q-learning with logarithmic regret," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1576–1584.

[22] Z. Zheng, H. Zhang, and L. Xue, "Gap-dependent bounds for q-learning using reference-advantage decomposition," in *The Thirteenth International Conference on Learning Representations*, 2025.

[23] M. Strens, "A bayesian framework for reinforcement learning," in *International Conference on Machine Learning*, 2000, pp. 943–950.

[24] T. Zhang, "Feel-good thompson sampling for contextual bandits and reinforcement learning," *SIAM Journal on Mathematics of Data Science*, vol. 4, no. 2, pp. 834–857, 2022.

[25] B. Hao and T. Lattimore, "Regret bounds for information-directed reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 575–28 587, 2022.

[26] A. Moradipari, M. Pedramfar, M. Shokrian Zini, and V. Aggarwal, "Improved bayesian regret bounds for thompson sampling in reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 23 557–23 569, 2023.

[27] I. Osband, B. Van Roy, and Z. Wen, "Generalization and exploration via randomized value functions," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2377–2386.

[28] P. Agrawal, J. Chen, and N. Jiang, "Improved worst-case regret bounds for randomized least-squares value iteration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6566–6573.

[29] A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric, "Frequentist regret bounds for randomized least-squares value iteration," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1954–1964.

[30] H. Ishfaq, Y. Tan, Y. Yang, Q. Lan, J. Lu, A. R. Mahmood, D. Precup, and P. Xu, "More efficient randomized exploration for reinforcement learning via approximate sampling," *arXiv preprint arXiv:2406.12241*, 2024.

[31] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," *Advances in neural information processing systems*, vol. 31, 2018.

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

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

[34] O. D. Domingues, P. Ménard, E. Kaufmann, and M. Valko, "Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited," in *Algorithmic Learning Theory*. PMLR, 2021, pp. 578–598.

[35] M. Simchowitz and K. G. Jamieson, "Non-asymptotic gap-dependent regret bounds for tabular mdps," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] O. D. Domingues, Y. Flet-Berliac, E. Leurent, P. Ménard, X. Shang, and M. Valko, "rlberry - A Reinforcement Learning Library for Research and Education," 10 2021. [Online]. Available: https://github.com/rlberry-py/rlberry

[37] F. S. Melo and M. I. Ribeiro, "Q-learning with linear function approximation," in *International Conference on Computational Learning Theory*. Springer, 2007, pp. 308–322.

[38] Z. Song and W. Sun, "Efficient model-free reinforcement learning in metric spaces," *arXiv preprint arXiv:1905.00475*, 2019.

[39] A. K. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications*. CRC press, 2004.

[40] I. Pinelis, "Optimum bounds for the distributions of martingales in banach spaces," *The Annals of Probability*, pp. 1679–1706, 1994.

[41] T.-T. Wong, "Generalized dirichlet distribution in bayesian analysis," *Applied Mathematics and Computation*, vol. 97, no. 2-3, pp. 165–181, 1998.

A Notation and Preliminaries

Before proceeding, we first introduce the following notation with the dependency on the episode index t and its short-hand notation whenever it is clear from the context.

- $n_h^t(s,a)$, or the shorthand n_h^t: the number of previous visits to (s,a) at step h before episode t.
- $n_h^{\flat,t}(s,a)$, or the shorthand $n_h^{\flat,t}$: the number of previous visits to (s,a) at step h during the current stage that the episode t belongs to.
- $q_h^t(s,a)$, or the shorthand q_h^t: the stage index of the i-th visit to (s,a) at step h and episode t.
- $\ell_h^i(s,a)$, or the shorthand ℓ^i: the episode index of the i-th visit to (s,a) at step h; by convention $\ell^0 = 0$.
- $\ell_{q,h}^{\flat,i}(s,a)$, or the shorthand ℓ_q^i: the episode index of the i-th visit to (s,a) at step h during the stage q; by convention, $\ell_0^0 = 0$ and $\ell_{q,h}^{\flat,0}(s,a)$ represents the episode when the q-th stage starts for (h,s,a).
- $e_q = \lceil (1+1/H)^q H \rceil$: the length of the q-th stage; by convention, $e_{-1} = 0$.
- J: number of ensemble heads (temporary Q-functions) per episode.
- $\widetilde{Q}_h^{j,t}(s,a)$: j-th temporary (ensemble) estimate of the optimal Q-value at the *beginning* of episode t, where the randomized learning rate follows $\mathrm{Beta}(\frac{H+1}{\kappa}, \frac{n_h^t + n_0 - 1}{\kappa})$.
- $\widetilde{Q}_h^{\flat,j,t}(s,a)$: j-th temporary (ensemble) estimate of the optimal Q-value at the *beginning* of episode t, where the randomized learning rate follows $\mathrm{Beta}(\frac{1}{\kappa^\flat}, \frac{n_h^{\flat,t} + n_0^\flat - 1}{\kappa^\flat})$.
- $\widetilde{Q}_h^{\flat,t}(s,a)$: the optimistic approximation of the optimal Q-function updated at the end of each stage.
- $Q_h^t(s,a)$: the policy Q-function at the start of episode t; its update at the visited pair (s_h^t, a_h^t) is
$$Q_h^t(s_h^t, a_h^t) = \eta_{t-1,h} \max_{j \in [J]} \widetilde{Q}_h^{j,t}(s_h^t, a_h^t) + (1 - \eta_{t-1,h}) \widetilde{Q}_h^{\flat,t}(s_h^t, a_h^t). \tag{9}$$
- $\widetilde{V}_h^{\ell^i}$, $\widetilde{V}_h^{\flat,\ell_q^i}$: optimistic value estimation of the optimal value function at episode ℓ^i and ℓ_q^i.
- π_h^t: the learned policy used at step h in episode t; the action a_h^t is drawn from the learned policy π_h^t at state s_h^t for any $(h,t) \in [H] \times [T]$.

For analysis, we also introduce the following notation.

- s_0: the optimistic pseudo-state s_0 with
$$r_h(s_0, a) = r_0 \ > \ 1, \quad p_h(s_0 \mid s, a) = \mathbb{1}\{s = s_0\}.$$
- $V_h^\star(s_0)$: the cumulative return obtained by always staying at the optimistic state s_0 from step h, i.e., $V_h^\star(s_0) = r_0\big(H - h + 1\big)$.

- $n_0, n_0^\flat$: prior pseudo-transition counts; thereby, each state-action pair (s,a) starts with n_0 prior pseudo-transitions, leading to

$$w_0^j \sim \mathrm{Beta}\big((H+1)/\kappa,\ n_0/\kappa\big), \quad w_0^{\flat,j} \sim \mathrm{Beta}\big(1/\kappa^\flat,\ n_0^\flat/\kappa^\flat\big), \qquad j \in [J].$$

- $\mathcal{K}_{\inf}(p,\mu)$: the information-theoretic distance between some measure $p \in \mathcal{P}[0,b]$ and $\mu \in [0,b]$, defined as

$$\mathcal{K}_{\inf}(p,\mu) = \inf\{\mathrm{KL}(p,q) : q \in \mathcal{P}[0,b], p \ll q, \mathbb{E}_{X\sim q}[X] \geq \mu\},$$

  where $\mathcal{P}[0,b]$ denotes all probability measures supported on $[0,b]$.
- δ_x: Dirac measure concentrated at a single point x.
- $[n]$: the indexing shorthand; for a positive integer n, we write $[n] \coloneqq \{1,2,\ldots,n\}$.
- $\|X\|_p$: the ℓ_p-norm of a vector $X \in \mathbb{R}^n$ where $p \geq 1$; formally defined as

$$\|X\|_p = \Big(\sum_{i=1}^n |X_i|^p\Big)^{1/p}.$$

- $(x)_k$: Pochhammer symbol, i.e., for $k \in \mathbb{N}$,

$$(x)_k = x(x+1)\cdots(x+k-1). \tag{10}$$

- $\mathbb{1}\{x \geq c\}$: an indicator function that equals 1 when $x \geq c$, and 0 otherwise.
- $|X|$: the cardinality of the set X.
- $A \lesssim B$: means $A \leq cB$ for some universal constant $c > 0$.
- $\mathbb{N}^\star$: the set consists of the positive integers, i.e., $\{1,2,3,\ldots\}$.

Beta distribution. We recall the definition and important properties of Beta distribution [39, Section 2], which is used in the follow-up analysis.

Definition 1 (Beta distribution)**.** *A continuous random variable X is said to follow a* Beta *distribution with shape parameters $\alpha > 0$ and $\beta > 0$, written as*

$$X \sim \mathrm{Beta}(\alpha,\beta),$$

if its probability density function is

$$f_X(x \mid \alpha,\beta) \ = \ \frac{1}{B(\alpha,\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}, \qquad 0 < x < 1,$$

and $f_X(x) = 0$ otherwise, where the Beta function $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ *serves as the normalizing constant.*

Lemma 1 (Moments of the Beta distribution)**.** *Let $X \sim \mathrm{Beta}(a,b)$ with $a,b > 0$, and recall the Pochhammer symbol $(x)_k$ defined in (10). Then, for any positive integer r,*

$$\mathbb{E}\left[X^r\right] = \frac{(a)_r}{(a+b)_r}. \tag{11}$$

In particular, the expectation and variance of X are

$$\mathbb{E}[X] = \frac{a}{a+b}, \qquad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{12}$$

We next collect a few auxiliary lemmas used in the proofs.

Lemma 2. *For any $a,b \geq 1, \kappa \geq 0$, we have*

$$\frac{a}{a+b} \leq \frac{a+\kappa}{a+b+\kappa} \leq (\kappa+1)\cdot\frac{a}{a+b}. \tag{13}$$

In addition, for any $r \in \mathbb{N}_+$,

$$\frac{a+r\kappa}{a+b+r\kappa} \leq r\cdot\frac{a+\kappa}{a+b+\kappa}. \tag{14}$$

Proof. The following proves (13) and (14), respectively. We start with the lower bound of (13). Define $f(t) = \dfrac{a+t}{a+b+t}$ for $t \geq 0$. Then $f'(t) = \dfrac{b}{(a+b+t)^2} > 0$, so f is increasing. Hence $f(0) = \dfrac{a}{a+b} \leq f(\kappa) = \dfrac{a+\kappa}{a+b+\kappa}$. Moving to the upper bound, it is equivalent to show

$$(a+\kappa)(a+b) \;\leq\; (\kappa+1)a(a+b+\kappa).$$

Expanding and rearranging leads to

$$(\kappa+1)a(a+b+\kappa) - (a+\kappa)(a+b) = \kappa\big(a^2 + ab + a\kappa - b\big) \geq 0,$$

which holds if $a, b \geq 1$ and $\kappa \geq 0$.

We now show (14). Let $r \in \mathbb{N}_+$ and

$$L = (a + r\kappa)(a+b+\kappa), \quad R = r(a+\kappa)(a+b+r\kappa).$$

We prove $L \leq R$ by first computing

$$\begin{aligned} R - L &= r(a+\kappa)(a+b+r\kappa) - (a+r\kappa)(a+b+\kappa) \\ &= (r-1)\Big[a(a+b) + \kappa\big(a(r+1) + \kappa r\big)\Big] \;\geq\; 0, \end{aligned}$$

because each bracketed term is non-negative and $r - 1 \geq 0$. Dividing by the common positive factor $(a+b+r\kappa)(a+b+\kappa)$ yields $\dfrac{a+r\kappa}{a+b+r\kappa} \leq r\dfrac{a+\kappa}{a+b+\kappa}$. All statements are therefore proved. □

Lemma 3 (Rosenthal inequality, Theorem 4.1 in [40])**.** *Let $X_1, \ldots, X_n$ be a martingale-difference sequence adapted to a filtration $\{\mathcal{F}_i\}_{i=1,\ldots,n}$:*

$$\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0.$$

Define $\mathcal{V}_i = \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}]$. Then there exist universal constants c_1 and c_2 such that for any $p \geq 2$ the following holds

$$\mathbb{E}^{1/p}\left[\left|\sum_{i=1}^n X_i\right|^p\right] \leq C_1 p^{1/2}\mathbb{E}^{1/p}\left[\left|\sum_{i=1}^n \mathcal{V}_i\right|^{p/2}\right] + 2C_2 p \cdot \mathbb{E}^{1/p}\left[\max_{i\in[n]} |X_i|^p\right].$$

Lemma 4 (Corrected version of Lemma 12 in Tiapkin et al. [10].[5])**.** *Let $\nu \in \mathcal{P}([0,b])$ be a probability measure over the segment $[0,b]$ and let $\bar{\nu} = (1-\alpha)\delta_{b_0} + \alpha \cdot \nu$ be a mixture between ν and a Dirac measure on $b_0 > b$, where $\alpha \in (0,1)$. Then for any $\mu \in (0,b)$,*

$$\mathcal{K}_{\inf}(\bar{\nu}, \mu) \leq \alpha \mathcal{K}_{\inf}(\nu, \mu). \tag{15}$$

Proof. From Tiapkin et al. [10, Lemma 9], for any probability measure $\nu \in \mathcal{P}[0,b]$ and $\mu \in (0,b)$,

$$\mathcal{K}_{\inf}(\bar{\nu}, \mu) = \max_{\lambda \in [0,\, 1/(b_0-\mu)]} \mathbb{E}_{X\sim\bar{\nu}}\big[\log\big(1 - \lambda\,(X-\mu)\big)\big].$$

The support of $\bar{\nu}$ is contained in $[0, b_0]$, so for any $\lambda \in [0, 1/(b_0-\mu)]$,

$$\mathbb{E}_{X\sim\bar{\nu}}\big[\log\big(1-\lambda\,(X-\mu)\big)\big] = (1-\alpha)\log\big(1-\lambda(b_0-\mu)\big) + \alpha\mathbb{E}_{X\sim\nu}\big[\log\big(1-\lambda(X-\mu)\big)\big].$$

For every admissible λ we have $0 \leq \lambda(b_0-\mu) \leq 1$, so $\log\big(1-\lambda(b_0-\mu)\big) \leq 0$. Hence

$$\mathcal{K}_{\inf}(\bar{\nu}, \mu) \leq \alpha\mathbb{E}_{X\sim\nu}\big[\log\big(1-\lambda(X-\mu)\big)\big].$$

Because $b_0 > b$, the interval $[0,\, 1/(b_0-\mu)]$ is a subset of $[0,\, 1/(b-\mu)]$. Taking the maximum over the smaller interval leads to

$$\mathcal{K}_{\inf}(\bar{\nu}, \mu) \leq \alpha \max_{0\leq\lambda\leq 1/(b-\mu)} \mathbb{E}_{X\sim\nu}\big[\log\big(1-\lambda(X-\mu)\big)\big] = \alpha\,\mathcal{K}_{\inf}(\nu, \mu).$$

□

[5] We clarify an earlier oversight in Tiapkin et al. [10, Lemma 12] by properly accounting for the Dirac measure's contribution, which was previously incorrectly separated after applying the variational formula—specifically, Lemma 9 in [10]. Consequently, the right-hand side of (15) is now scaled by α, instead of the $1-\alpha$ factor used in [10].

B Reformulation of the Update Equation and Aggregated Weights

In this section, we rewrite the update of each temporary Q-function for every trajectory $t \in [T]$ and $j \in [J]$ by recursively unrolling the update (5) and (6), for each $(h, s, a) \in [H] \times S \times A$.

For the ease of notation, we denote $m := n_h^t(s, a)$.

Unrolled update for $\widetilde{Q}_h^{j,t}$. For each $(j, t, h) \in [J] \times [T] \times [H]$, we can unroll (5) by

$$\begin{aligned}
\widetilde{Q}_h^{j,t}(s,a) = \widetilde{Q}_h^{j,\ell^m+1}(s,a) &= (1-w_{m-1}^j)\cdot \widetilde{Q}_h^{j,\ell^{m-1}+1}(s,a) + w_{m-1}^j \left[r_h(s,a) + \widetilde{V}_{h+1}^{\ell^m}(s_{h+1}^{\ell^m})\right], \\
\widetilde{Q}_h^{j,\ell^{m-1}+1}(s,a) &= (1-w_{m-2}^j)\cdot \widetilde{Q}_h^{j,\ell^{m-2}+1}(s,a) + w_{m-2}^j \left[r_h(s,a) + \widetilde{V}_{h+1}^{\ell^{m-1}}(s_{h+1}^{\ell^{m-1}})\right], \\
&\vdots \\
\widetilde{Q}_h^{j,\ell^1+1}(s,a) &= (1-w_0^j)\cdot \widetilde{Q}_h^{j,\ell^0+1}(s,a) + w_0^j \left[r_h(s,a) + \widetilde{V}_{h+1}^{\ell^1}(s_{h+1}^{\ell^1})\right], \\
\widetilde{Q}_h^{j,\ell^0+1}(s,a) &= r_h(s,a) + \widetilde{V}_{h+1}^{\ell^0}(s_{h+1}^{\ell^0}),
\end{aligned}$$

where we define $\widetilde{V}_{h+1}^{\ell^0}(s_{h+1}^{\ell^0}) = V_{h+1}^0$. For $m \geq 1$, let $W_{j,m} = \left(W_{j,m}^m, \ldots, W_{j,m}^1, W_{j,m}^0\right)$ be the aggregated weights defined as

$$W_{j,m}^0 = \prod_{k=0}^{m-1}(1-w_k^j) \quad \text{and} \quad W_{j,m}^i = w_{i-1}^j \prod_{k=i}^{m-1}(1-w_k^j), \quad \forall i \in [m], \tag{16}$$

where w_k^j is sampled from Beta $(\frac{H+1}{\kappa}, \frac{k+n_0}{\kappa})$. Then, we have

$$\widetilde{Q}_h^{j,t}(s,a) = r_h(s,a) + \sum_{i=0}^{m} W_{j,m}^i \left[\widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i})\right]. \tag{17}$$

Unrolled update for $\widetilde{Q}_h^{\flat,t}$. Suppose that q is the stage index of the episode t on (h, s, a). We let e_q be the length of the q-th stage.

Similar to the above unrolling steps for $\widetilde{Q}_h^{j,t}$, we define the corresponding aggregated weights as: For the q-th stage, let $W_{j,q}^\flat = \left(W_{j,e_q}^{\flat,0}, W_{j,e_q}^{\flat,1}, \ldots, W_{j,e_q}^{\flat,e_q}\right)$ be the aggregated weights at the defined as

$$W_{j,q}^{\flat,0} = \prod_{k=0}^{e_q-1}(1-w_{k,q}^{\flat,j}) \quad \text{and} \quad W_{j,q}^{\flat,i} = w_{i-1}^{\flat,j} \prod_{k=i}^{e_q-1}(1-w_{k,q}^{\flat,j}), \quad \forall i \in [e_q], \tag{18}$$

where $w_{k,q}^{\flat,j}$ is sampled from $\mathrm{Beta}(\frac{1}{\kappa^\flat}, \frac{k+n_0^\flat}{\kappa^\flat})$.

For each $(t, h) \in [T] \times [H]$, as shown in (6), $\widetilde{Q}_h^{\flat,t}(s,a)$ is updated via the most recent e_{q-1} samples before the q-th stage. Thus, for any $q \geq 1$, we have

$$\widetilde{Q}_h^{\flat,t}(s,a) = r_h(s,a) + \max_{j\in[J]}\left\{\sum_{i=0}^{e_{q-1}} W_{j,q-1}^{\flat,i}\left[\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}})\right]\right\}. \tag{19}$$

B.1 Properties of aggregated weights

In this section, we mainly discuss the properties of the aggregated weights $W_{j,m}$ for any $m \geq 1$.

To begin with, from (16), we can verify that the sum of the aggregated weights is equal to 1, i.e.,

$$\sum_{i=0}^{m} W_{j,m}^i = \prod_{k=0}^{m-1}(1-w_k^j) + \sum_{i=1}^{m} w_{i-1}^j \prod_{k=i}^{m-1}(1-w_k^j) = 1. \tag{20}$$

In the following proposition, we further show some desirable properties regarding the expectation and variance of the aggregated weights $W_{j,m}$.

Proposition 1. *The following properties hold for the aggregated weights $W_{j,m}$, $\forall j \in [J], m \geq 1$:*

(i) The moment of the aggregated weights is given by:

$$\mathbb{E}[(W_{j,m}^i)^d] = \left(\prod_{j=i+1}^{m} \frac{(\frac{n_0+j-1}{\kappa})_d}{(\frac{H+n_0+j}{\kappa})_d} \right) \cdot \frac{(\frac{H+1}{\kappa})_d}{(\frac{H+n_0+i}{\kappa})_d}; \tag{21}$$

(ii) The upper bound of expectations and the sum of variances:

$$\max_{i\in[m]} \mathbb{E}[W_{j,m}^i] \leq \frac{H+1}{H+n_0+m}, \qquad \sum_{i=1}^{m} \mathrm{Var}[W_{j,m}^i] \leq \frac{(H+1)\kappa}{H+n_0+m}; \tag{22}$$

(iii) For every $i \geq 1$, we have

$$\sum_{t=i}^{\infty} \mathbb{E}[W_{j,t}^i] \leq 1 + \frac{1}{H}. \tag{23}$$

Proof. We prove each item separately.

(i) Directly follows from Wong [41, Section 2].

(ii) Similar to Jin et al. [6, Lemma 4.1], we have that for $i \in [m]$

$$\begin{aligned}
\mathbb{E}[W_{j,m}^i] &= \frac{H+1}{H+n_0+i} \left(\frac{n_0+i}{H+n_0+i+1} \frac{n_0+i+1}{H+n_0+i+2} \cdots \frac{n_0+m-1}{H+n_0+m} \right) \\
&= \frac{H+1}{H+n_0+m} \left(\frac{n_0+i}{H+n_0+i} \frac{n_0+i+1}{H+n_0+i+1} \cdots \frac{n_0+m-1}{H+n_0+m-1} \right) \\
&\leq \frac{H+1}{H+n_0+m}.
\end{aligned}$$

Thus, $\max_{i\in[m]} \mathbb{E}[W_{j,m}^i] \leq \frac{H+1}{H+n_0+m}$ holds. From Lemma 2 and (21),

$$\begin{aligned}
\mathrm{Var}[W_{j,m}^i] &= \mathbb{E}[(W_{j,m}^i)^2] - \mathbb{E}[W_{j,m}^i]^2 \\
&= \mathbb{E}[W_{j,m}^i] \left(\left(\prod_{j=i+1}^{m} \frac{n_0+j-1+\kappa}{H+n_0+j+\kappa} \right) \cdot \frac{H+1+\kappa}{H+n_0+i+\kappa} - \mathbb{E}[W_{j,m}^i] \right) \\
&\leq \mathbb{E}[W_{j,m}^i] \left(\frac{H+1+\kappa}{H+n_0+m+\kappa} - \mathbb{E}[W_{j,m}^i] \right) \\
&\leq \kappa \mathbb{E}[W_{j,m}^i] \cdot \frac{H+1}{H+n_0+m},
\end{aligned}$$

and thus

$$\sum_{i=1}^{m} \mathrm{Var}[W_{j,m}^i] \leq \kappa \cdot \frac{H+1}{H+n_0+m} \cdot \left(\sum_{i=1}^{m} \mathbb{E}[W_{j,m}^i] \right) \leq \frac{(H+1)\kappa}{H+n_0+m},$$

where the last inequality used $\sum_{i=1}^{m} \mathbb{E}[W_{j,m}^i] \leq \sum_{i=0}^{m} \mathbb{E}[W_{j,m}^i] = 1$ (recalling (20)).

(iii) Following Jin et al. [6, equation (B.1)], it also holds for any positive integer n, k and $n \geq k$

$$\frac{n+n_0}{k} = 1 + \frac{n+n_0-k}{n+n_0+1} + \frac{(n+n_0-k)(n+n_0-k+1)}{(n+n_0+1)(n+n_0+2)} + \cdots \tag{24}$$

which can be verified by induction. Specifically, we let the terms of the right-hand side be $x_0 = 1, x_1 = \frac{n+n_0-k}{n+n_0+1}, \ldots$. Then, we will show $\frac{n+n_0}{k} - \sum_{j=0}^{i} x_j = \frac{n+n_0-k}{k} \prod_{j=1}^{i} \frac{n+n_0-k+j}{n+n_0+j}$ by induction.

- Base case when $i = 1$: It can easily verified that

$$\frac{n+n_0}{k} - 1 - \frac{n+n_0-k}{n+n_0+1} = \frac{(n+n_0-k)(n+n_0+1-k)}{k(n+n_0+1)}$$

- Suppose $i = r$, the claim holds, i.e., $\frac{n+n_0}{k} - \sum_{j=0}^{r} x_j = \frac{n+n_0-k}{k} \prod_{j=1}^{r} \frac{n+n_0-k+j}{n+n_0+j}$. Then, for $i = r+1$, we have

$$\begin{aligned}
&\frac{n+n_0}{k} - \sum_{j=0}^{r} x_j - x_{r+1} = \frac{n+n_0-k}{k} \prod_{j=1}^{r} \frac{n+n_0-k+j}{n+n_0+j} - \prod_{j=1}^{r+1} \frac{n+n_0-k+j-1}{n+n_0+j} \\
&= \frac{n+n_0-k}{k} \prod_{j=1}^{r} \frac{n+n_0-k+j}{n+n_0+j} \Big[1 - \frac{k}{n+n_0-k} \cdot \frac{n+n_0-k}{n+n_0+r+1}\Big] \\
&= \frac{n+n_0-k}{k} \prod_{j=1}^{r} \frac{n+n_0-k+j}{n+n_0+j} \frac{n+n_0-k+r+1}{n+n_0+r+1} \\
&= \frac{n+n_0-k}{k} \prod_{j=1}^{r+1} \frac{n+n_0-k+j}{n+n_0+j}.
\end{aligned}$$

By letting $i \to \infty$, we obtain (24). Thus, we have

$$\begin{aligned}
\sum_{t=i}^{\infty} \mathbb{E}[W_{j,t}^i] &= \frac{H+1}{H+n_0+i} \left(1 + \frac{i+n_0}{H+n_0+i+1} + \frac{i+n_0}{H+n_0+i+1} \frac{i+n_0+1}{H+n_0+i+2} + \dots\right) \\
&= \frac{H+1}{H+n_0+i} \frac{H+n_0+i}{H} = 1 + \frac{1}{H},
\end{aligned}$$

where the second equality uses (24) with $n = i + H$ and $k = H$.

□

B.2 Concentration of aggregated weights

For notational convenience and clearness, we slightly abuse the notation by ignoring the dependency on the ensemble index j, while the following concentration lemma of the aggregated weights holds for any $j \in [J]$.

Lemma 5. *Let $\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_m$ be nonnegative real numbers such that $\lambda_i \le 1$, $i = 1, 2, \cdots, m$. Then, with probability at least $1 - \frac{\delta}{T}$, we have*

$$\left| \sum_{i=0}^{m} \lambda_i \left(W_m^i - \mathbb{E}[W_m^i]\right) \right| \le c_1 \sqrt{\frac{(H+1)\kappa^2 \log^3(T/\delta)}{H+n_0+m}} + c_2 \frac{(H+1)\kappa \log^2(T/\delta)}{H+n_0+m},$$

where c_1 and c_2 are some positive universal constants.

Proof. By definition, we can find independent random variables $w_0, \cdots, w_{m-1}$ such that $w_i \sim \mathrm{Beta}(\frac{H+1}{\kappa}, \frac{n_0+i}{\kappa})$, and

$$W_m^0 = \prod_{k=1}^{m-1} (1 - w_k), \quad W_m^i = w_{i-1} \prod_{k=i}^{m-1} (1 - w_k), \quad 1 \le i \le m. \tag{25}$$

Let $\mathcal{F}_{-i}$ be the σ-algebra generated by $w_{m-1}, w_{m-2}, \cdots, w_{m-i}$, with the convention that $\mathcal{F}_{-0} = \mathcal{F}_0$ is the trivial σ-algebra. In the same spirit, denote

$$S_{-i} = \sum_{k=m-i+1}^{m} \lambda_k (W_m^k - \mathbb{E}[W_m^k]), \quad i = 0, 1, \cdots, m. \tag{26}$$

For conceptual reason, we will use the notation $S_{-\infty}$ to denote the sum $\sum_{k=0}^{m} \lambda_k (W_m^k - \mathbb{E}[W_m^k])$, which corresponds to $S_{-(m+1)}$ in the above notation.

We view $\mathcal{F}_0 \subset \mathcal{F}_{-1} \subset \cdots \subset \mathcal{F}_{-m}$ as a reverse filtration, and consider the backward martingale

$$M_{-i} \coloneqq \mathbb{E}[S_{-\infty}|\mathcal{F}_{-i}], \quad i = 0, 1, \cdots, m. \tag{27}$$

It is clear that $M_0 = \mathbb{E}[S_{-\infty}]$, while $M_{-m} = S_{-\infty}$. Therefore

$$\sum_{i=0}^{m} \lambda_i \left(W_m^i - \mathbb{E}[W_m^i]\right) = S_{-\infty} - \mathbb{E}[S_{-\infty}] = M_{-m} - M_0.$$

We may then apply the Rosenthal's inequality (i.e., Lemma 3) to obtain

$$\begin{aligned}
\left(\mathbb{E}[|M_{-m} - M_0|^p]\right)^{1/p} &\leq C\sqrt{p}\left(\mathbb{E}[\langle M\rangle_{-m}^{p/2}]\right)^{1/p} + Cp\left(\mathbb{E}[\max_{1\leq i\leq m}|M_{-i} - M_{-(i-1)}|^p]\right)^{1/p} \\
&\leq C\sqrt{p}\left(\mathbb{E}[\langle M\rangle_{-m}^{p/2}]\right)^{1/p} + Cp\left(\sum_{i=1}^{m}\mathbb{E}[|M_{-i} - M_{-(i-1)}|^p]\right)^{1/p}, \quad (28)
\end{aligned}$$

where

$$\langle M\rangle_{-m} \coloneqq \sum_{i=1}^{m}\mathbb{E}[(M_{-i} - M_{-(i-1)})^2|\mathcal{F}_{-(i-1)}].$$

To proceed, we calculate the martingale difference $M_{-i} - M_{-(i-1)}$ for any $i \in [m]$. To simplify the resulting expressions, we will denote

$$T_{-i} \coloneqq \lambda_0 \prod_{j=0}^{m-i-1}(1-w_j) + \sum_{k=1}^{m-i}\left(\lambda_k w_{k-1}\prod_{j=k}^{m-i-1}(1-w_j)\right).$$

With this notation in hand, we can decompose $S_{-\infty}$ by

$$\begin{aligned}
S_{-\infty} &= \lambda_{m-i+1}\left(W_m^{m-i+1} - \mathbb{E}[W_m^{m-i+1}]\right) + \sum_{k=0}^{m-i}\lambda_k(W_m^k - \mathbb{E}[W_m^k]) + \sum_{k=m-i+2}^{m}\lambda_k\left(W_m^k - \mathbb{E}[W_m^k]\right) \\
&= \lambda_{m-i+1}\left(w_{m-i}\prod_{k=m-(i-1)}^{m-1}(1-w_k) - \mathbb{E}\left[w_{m-i}\prod_{k=m-(i-1)}^{m-1}(1-w_k)\right]\right) \\
&\quad + \left(T_{-i}\prod_{k=m-i}^{m-1}(1-w_k) - \mathbb{E}\left[T_{-i}\prod_{k=m-i}^{m-1}(1-w_k)\right]\right) \\
&\quad + \sum_{k=m-i+2}^{m}\lambda_k\left(W_m^k - \mathbb{E}[W_m^k]\right),
\end{aligned}$$

where only the first two terms involve the observation $w_{m-i} = \mathcal{F}_{-i} \setminus \mathcal{F}_{-(i-1)}$. Then for $i \in [m]$,

$$\begin{aligned}
M_{-i} - M_{-(i-1)} &= \mathbb{E}[S_{-\infty}|\mathcal{F}_{-i}] - \mathbb{E}[S_{-\infty}|\mathcal{F}_{-(i-1)}] \quad (29) \\
&= \lambda_{m-i+1}\left(w_{m-i} - \mathbb{E}\left[w_{m-i}\right]\right)\prod_{k=m-(i-1)}^{m-1}(1-w_k) \\
&\quad + \mathbb{E}[T_{-i}]\prod_{k=m-(i-1)}^{m-1}(1-w_k)\left((1-w_{m-i}) - \mathbb{E}[1-w_{m-i}]\right) \\
&= (\lambda_{m-i+1} - \mathbb{E}[T_{-i}])\cdot(w_{m-i} - \mathbb{E}[w_{m-i}])\prod_{k=m-(i-1)}^{m-1}(1-w_k), \quad (30)
\end{aligned}$$

where the last line is from $\mathbb{E}[1 - w_{m-i}] = 1 - \mathbb{E}[w_{m-i}]$.

By our assumption $|\lambda_i| \leq 1$, $i = 1, \cdots, m$, it can be readily checked that

$$|T_{-i}| \leq \prod_{j=0}^{m-i-1}(1-w_j) + \sum_{k=1}^{m-i}\left(w_{k-1}\prod_{j=k}^{m-i-1}(1-w_j)\right) = 1, \quad \forall i = \{1,\ldots,m\}.$$

Consequently, the absolute value of the martingale difference can be bounded above by

$$|M_{-i} - M_{-(i-1)}| \leq 2|w_{m-i} - \mathbb{E}[w_{m-i}]| \prod_{k=m-(i-1)}^{m-1}(1-w_k). \tag{31}$$

Recall that $\langle M\rangle_{-m} \coloneqq \sum_{i=1}^{m}\mathbb{E}[(M_{-i} - M_{-(i-1)})^2|\mathcal{F}_{-(i-1)}]$. Together with (31), we have

$$\begin{aligned}
\langle M\rangle_{-m} &\leq 2\sum_{i=1}^{m}\mathbb{E}\left(|w_{m-i} - \mathbb{E}[w_{m-i}]| \prod_{k=m-(i-1)}^{m-1}(1-w_k)\,\middle|\,\mathcal{F}_{-(i-1)}\right)^2 \qquad (32)\\
&= 2\sum_{i=1}^{m}\mathrm{Var}[w_{m-i}]\prod_{k=m-(i-1)}^{m-1}(1-w_k)^2\\
&= 2\sum_{i=0}^{m-1}\mathrm{Var}[w_i]\prod_{k=i+1}^{m-1}(1-w_k)^2\\
&\leq 2\sum_{i=0}^{m-1}\frac{\kappa(H+1)(n_0+i)}{(H+1+n_0+i)^2(H+1+n_0+i+\kappa)}\prod_{k=i+1}^{m-1}(1-w_k)^2, \qquad (33)
\end{aligned}$$

where the second equality is due to the change of index, and the last inequality follows from (12). We may then apply triangle inequality to obtain

$$\begin{aligned}
\mathbb{E}^{2/p}[\langle M\rangle_{-m}^{p/2}] &\leq \sum_{i=0}^{m-1} 2\frac{\kappa(H+1)(n_0+i)}{(H+1+n_0+i)^2(H+1+n_0+i+\kappa)}\mathbb{E}^{2/p}\left[\left(\prod_{k=i+1}^{m-1}(1-w_k)^2\right)^{p/2}\right]\\
&= 2\sum_{i=0}^{m-1}\frac{\kappa(H+1)(n_0+i)}{(H+1+n_0+i)^2(H+1+n_0+i+\kappa)}\left(\prod_{k=i+1}^{m-1}\mathbb{E}\left[(1-w_k)^p\right]\right)^{2/p},
\end{aligned}$$

for $p \geq 2$. Note that $1 - w_k \sim \mathrm{Beta}(\frac{n_0+k}{\kappa}, \frac{H+1}{\kappa})$, directly from the definition 1. Thus, By (11), we have

$$\mathbb{E}\prod_{k=i+1}^{m-1}(1-w_k)^p = \prod_{k=i+2}^{m}\frac{(\frac{n_0+k-1}{\kappa})_p}{(\frac{H+n_0+k}{\kappa})_p} \leq \prod_{r=0}^{p-1}\frac{n_0+i+1+r\kappa}{H+n_0+m+r\kappa} \leq p!\left(\frac{n_0+i+1+\kappa}{H+n_0+m+\kappa}\right)^p,$$

where the last inequality is from Lemma 2. Therefore, we can obtain

$$\begin{aligned}
\left(\mathbb{E}[\langle M\rangle_{-m}^{p/2}]\right)^{2/p} &\leq 2\sum_{i=0}^{m-1}\frac{\kappa(H+1)(n_0+i)}{(H+1+n_0+i)^2(H+1+n_0+i+\kappa)}(p!)^{2/p}\left(\frac{n_0+i+1+\kappa}{H+n_0+m+\kappa}\right)^2\\
&\leq 2\kappa p^2\frac{H+1+\kappa}{(H+n_0+m+\kappa)^2}\sum_{i=0}^{m-1}\frac{n_0+i+\kappa}{H+1+n_0+i+\kappa}\\
&\leq 2\kappa p^2\frac{H+1+\kappa}{H+n_0+m+\kappa} \leq 2(\kappa+1)^2p^2\frac{H+1}{H+n_0+m}
\end{aligned}$$

where the first and last inequality are due to Lemma 2 (i.e., (13)), the second inequality uses the facts that $p! \leq p^p$ and $\frac{n_0+i+1+\kappa}{H+1+n_0+i+\kappa} \leq 1$ for every $i = 0, \ldots, m-1$, and the third inequality is from the fact that the summation term is less than m.

Therefore,

$$\left(\mathbb{E}[\langle M\rangle_{-m}^{p/2}]\right)^{1/p} \leq 2p(\kappa+1)\sqrt{\frac{(H+1)}{H+n_0+m}}. \tag{34}$$

We turn to bound $\mathbb{E}[|M_{-i} - M_{-(i-1)}|^p]$ for $i \in [m]$. It is clear from (31) that

$$\begin{aligned}\mathbb{E}[|M_{-i} - M_{-(i-1)}|^p] &\le \mathbb{E}\left[\left|2|w_{m-i} - \mathbb{E}[w_{m-i}]| \prod_{k=m-(i-1)}^{m-1}(1-w_k)\right|^p\right] \\ &\le 8^p \mathbb{E}[w_{m-i}^p] \prod_{k=m-(i-1)}^{m-1} \mathbb{E}[(1-w_k)^p] \\ &= 8^p \mathbb{E}[(W_m^{m-i+1})^p]\end{aligned}$$

where the second inequality uses $|w_{m-i} - \mathbb{E}[w_{m-i}]| \le 2^{p+1} w_{m-i}^p$, and the equality is from (25) where $W_m^{m-i+1} = w_{m-i}\prod_{k=m-(i-1)}^{m-1}(1-w_k)$ for $i \in [m]$.

From Proposition 1, we have that for any $i \in [m]$,

$$\begin{aligned}\mathbb{E}\left[(W_m^i)^p\right] &= \left(\prod_{j=i+1}^{m} \frac{(\frac{n_0+j-1}{\kappa})_p}{(\frac{H+n_0+j}{\kappa})_p}\right) \cdot \frac{(\frac{H+1}{\kappa})_p}{(\frac{H+n_0+i}{\kappa})_p} \\ &\le \prod_{r=0}^{p-1}\left(\frac{H+1+r\kappa}{H+n_0+i+r\kappa} \cdot \prod_{j=i+1}^{m}\frac{n_0+j-1+r\kappa}{H+n_0+j+r\kappa}\right) \\ &\le \prod_{r=0}^{p-1}\frac{r\kappa(H+1)}{H+n_0+m} \le p!\left(\frac{(H+1)\kappa}{H+n_0+m}\right)^p\end{aligned}$$

Thus,

$$\left(\sum_{i=1}^{m}\mathbb{E}[|M_{-i} - M_{-(i-1)}|^p]\right)^{1/p} \le \left(8^p\sum_{i=1}^{m}\mathbb{E}[(W_m^{m-i+1})^p]\right)^{1/p} \le 8m^{1/p}p\frac{(H+1)\kappa}{H+n_0+m}. \quad (35)$$

Substituting (34) and (35) into (28) leads to

$$\mathbb{E}^{1/p}[|M_{-m} - M_0|^p] \le 2C\sqrt{\frac{p^3(\kappa+1)^2(H+1)}{H+n_0+m}} + 8C\cdot m^{1/p}p^2\frac{(H+1)\kappa}{H+n_0+m}.$$

Note that by definition, $m = n_h^t(s,a) \le T$ always holds for any $(h,t,s,a) \in [H]\times[T]\times\mathcal{S}\times\mathcal{A}$. Let $p = \lceil\log(T/\delta)\rceil \ge 2$ and thus $m^{1/p} \le \mathrm{e}$ since $m^{1/p} \le \mathrm{e}^{(\log T)/p}$. Finally, by Markov inequality with $t = 2\mathrm{e}C\sqrt{\frac{(H+1)(\kappa+1)^2\log^3(T/\delta)}{H+n_0+m}} + 8\mathrm{e}^2C\frac{(H+1)\kappa\log^2(T/\delta)}{H+n_0+m}$, we have

$$\mathbb{P}\left[\left|\sum_{i=0}^{m}\lambda_i\left(W_m^i - \mathbb{E}[W_m^i]\right)\right| \ge t\right] \le \left(\frac{\mathbb{E}^{1/p}\left[\left|\sum_{i=0}^{m}\lambda_i\left(W_m^i - \mathbb{E}[W_m^i]\right)\right|^p\right]}{t}\right)^p \le \left(\frac{1}{\mathrm{e}}\right)^{\lceil\log(T/\delta)\rceil} \le \frac{\delta}{T},$$

which finishes the proof. □

In addition, we also have the following lemma regarding the concentration of $W_q^\flat$ for any stage $q \ge 0$. We omit its proof since it is similar to Lemma 5 .

Lemma 6. *Let $\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_{e_q}$ be nonnegative real numbers such that $\lambda_i \le 1$, $i = 1, 2, \cdots, e_q$. Then, with probability at least $1 - \delta/T$, we have*

$$\left|\sum_{i=0}^{e_q}\lambda_i\left(W_q^{\flat,i} - \mathbb{E}[W_q^{\flat,i}]\right)\right| \le c_1^\flat\sqrt{\frac{(\kappa^\flat)^2\log^3(T/\delta)}{1+n_0^\flat+e_q}} + c_2^\flat\frac{\kappa^\flat\log^2(T/\delta)}{1+n_0^\flat+e_q},$$

where $c_1^\flat$ and $c_2^\flat$ are some positive universal constants.

C Analysis: Gap-independent Regret Bound

In this section, we present the detailed proof of Theorem 1. Before proceeding, we first rewrite the update of policy Q-function by unrolling the updates of temporary Q-functions (i.e., equations (17) and (19)) as

$$Q_h^t(s_h^t,a_h^t) = \eta_{\ell^{n_h^t},h} \max_{j\in[J]} \left\{\widetilde{Q}_h^{j,t}(s_h^t,a_h^t)\right\} + (1-\eta_{\ell^{n_h^t},h})\cdot \widetilde{Q}_h^{\flat,t}(s_h^t,a_h^t) \tag{36}$$

$$= r_h(s_h^t,a_h^t) + \eta_{\ell^{n_h^t},h} \max_{j\in[J]} \left\{\sum_{i=0}^{n_h^t} W_{j,n_h^t}^i \widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i})\right\}$$

$$+ (1-\eta_{\ell^{n_h^t},h})\cdot \max_{j\in[J]} \left\{\sum_{i=0}^{e_{q-1}} W_{j,q-1}^{\flat,i} \widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}})\right\}, \tag{37}$$

for each $(t,h)\in[T]\times[H]$, where in the notations we omit the dependency of q regarding the step h and the episode t for simplicity and let n_h^t be the number of visits of the state-action pair (s_h^t,a_h^t) before episode t at step h such that $\ell^{n_h^t}$ be the index of the last visit of the state-action pair (s_h^t,a_h^t) at step h. The properties and the concentration inequality of the aggregated weights (i.e., W_{j,n_h^t} and $W_{j,q-1}^{\flat}$), proved in Appendix B.1 and B.2, will play a crucial role in the following analysis.

C.1 Proof of Theorem 1

To control the total regret, we first present the following lemma regarding the optimism of the policy Q-function Q_h^t for any $(t,h,s,a)\in[T]\times[H]\times\mathcal{S}\times\mathcal{A}$, where the detailed proof is postponed to Appendix C.2.

Lemma 7 (Optimism). *Consider $\delta\in(0,1)$. Assume that $J=\lceil c\cdot\log(SAHT/\delta)\rceil$, $\kappa^{\flat}=c\cdot(\log(SAH/\delta)+\log(T))$, and $n_0^{\flat}=\lceil c\cdot\log(T)\cdot\kappa^{\flat}\rceil$, where c is some universal constant. Let the initialized value function $V_h^0=2(H-h+1)$ and the mixing rate $\eta_{t,h}\in(0,1)$ for every $(t,h)\in\{0,1,\ldots,T\}\times[H]$. Then, for any $(t,h,s,a)\in[T]\times[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, the following event holds*

$$0<(1-\eta_{\ell^{n_h^t},h})\cdot Q_h^\star(s,a)\le Q_h^t(s,a).$$

Define

$$V_h^t(s_h^t) := Q_h^t(s_h^t,\pi_h^t(s_h^t)) = \max_{a\in\mathcal{A}} Q_h^t(s_h^t,a). \tag{38}$$

Note that Lemma 7 implies that for every $t\in[T]$,

$$V_1^\star(s_1^t) = \max_a Q_1^\star(s_1^t,a) \le \frac{1}{1-\eta_{\ell^{n_h^t},h}} \max_a Q_1^t(s_1^t,a)$$

$$= \frac{1}{1-\eta_{\ell^{n_h^t},h}} Q_1^t(s_1^t,a_1^t) = \frac{1}{1-\eta_{\ell^{n_h^t},h}} V_1^t(s_1^t). \tag{39}$$

Thus, we can decompose the total regret as

$$\mathrm{Regret}_T = \sum_{t=1}^T (V_1^\star - V_1^{\pi^t})(s_1^t) \le \sum_{t=1}^T (\frac{1}{1-\eta_{\ell^{n_1^t},1}} V_1^t - V_1^{\pi^t})(s_1^t) = \sum_{t=1}^T \delta_1^t + \sum_{t=0}^{T-1} \frac{\eta_{t,1}}{1-\eta_{t,1}} H,$$

where we denote the performance gap as $\delta_h^t = (V_h^t - V_h^{\pi^t})(s_h^t)$ for every $(t,h)\in[T]\times[H]$.

Note that $Q_h^{\pi^t}(s_h^t, a_h^t) = V_h^{\pi^t}(s_h^t)$. Then, for any fixed episode $t \in [T]$ and step $h \in [H]$, we decompose the performance gap δ_h^t:

$$\begin{aligned}
\delta_h^t &\le \left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \left(Q_h^\star - Q_h^{\pi_t}\right)(s_h^t, a_h^t) \\
&\le \left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + P_{h,s_h^t,a_h^t}\left(V_{h+1}^\star - V_{h+1}^{\pi^t}\right) \\
&\le \left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \underbrace{\left(V_{h+1}^t - V_{h+1}^{\pi_t}\right)(s_{h+1}^t)}_{=:\delta_{h+1}^t} - \underbrace{\left(V_{h+1}^t - V_{h+1}^\star\right)(s_{h+1}^t)}_{=:\xi_{h+1}^t} \\
&\quad + \underbrace{P_{h,s_h^t,a_h^t}\left(V_{h+1}^\star - V_{h+1}^{\pi^t}\right) - \left(V_{h+1}^\star - V_{h+1}^{\pi^t}\right)(s_{h+1}^t)}_{=:\tau_{h+1}^t},
\end{aligned}$$

where the second inequality is from the Bellman equations (2) and (3). Together with the update (36),

$$\begin{aligned}
\delta_h^t \le \eta_{\ell^{n_h^t},h} \underbrace{\max_{j\in[J]}\left\{\widetilde{Q}_h^{j,t}(s_h^t, a_h^t) - Q_h^\star(s_h^t, a_h^t) + H\right\}}_{=:\zeta_h^t} + (1 - \eta_{\ell^{n_h^t},h}) \underbrace{\left\{\widetilde{Q}_h^{\flat,t}(s_h^t, a_h^t) - Q_h^\star(s_h^t, a_h^t)\right\}}_{=:\zeta_h^{\flat,t}} \\
+ \delta_{h+1}^t - \xi_{h+1}^t + \tau_{h+1}^t - \eta_{\ell^{n_h^t},h} H. \qquad (40)
\end{aligned}$$

The main idea of the proof is to show that the performance gap δ_h^t can be upper bounded by some quantities from the next step $h+1$, and correspondingly, the total regret can be controlled by rolling out the performance gap over all episodes and steps.

Next, the following lemmas present a recursive bound for ζ_h^t and $\zeta_h^{\flat,t}$, respectively. The proofs are deferred to Appendix C.3 and C.4.

Lemma 8 (Recursive bound for ζ_h^t)**.** *Consider $\delta \in (0,1)$. For any $i \in [m]$, let $\alpha_m^0 := \prod_{k=1}^m \frac{n_0+k-1}{H+n_0+k}$ and $\alpha_m^i := \frac{H+1}{H+n_0+i}\prod_{k=i+1}^m \frac{n_0+k-1}{H+n_0+k}$, where $m = n_h^t(s_h^t, a_h^t)$. Then, for any $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$, with probability $1-\delta$, we have*

$$0 \le \zeta_h^t \le 2\alpha_m^0 V_{h+1}^0 + \sum_{i=1}^m \alpha_m^i \left(\left(\widetilde{V}_{h+1}^{\ell^i} - V_{h+1}^\star\right)(s_{h+1}^{\ell^i}) + H\right) + \tilde{b}_h^t, \qquad (41)$$

where $\tilde{b}_h^t = \widetilde{O}\left(\sqrt{\frac{(H+1)^3}{H+n_0+m}} + \frac{(H+1)^2}{H+n_0+m}\right)$.

Lemma 9 (Recursive bound for $\zeta_h^{\flat,t}$)**.** *Consider $\delta \in (0,1)$. For any $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$, we let $q = q_h^t(s,a)$ represent the index of the current stage and e_{q-1} be the length of the prior stage for $q \ge 1$. Then for any $t \in [T]$, with probability at least $1-\delta$, we have*

$$0 \le \zeta_h^{\flat,t} \le \left(\sum_{i=0}^{e_{q-1}} \frac{1}{1+e_{q-1}}\left(\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}} - V_{h+1}^\star\right)(s_{h+1}^{\ell_{q-1}^{\flat,i}}) + \tilde{b}_h^{\flat,t}\right)\cdot \mathbb{1}\{q \ge 1\} + V_{h+1}^0 \cdot \mathbb{1}\{q = 0\}, \qquad (42)$$

where $\tilde{b}_h^{\flat,t} = \widetilde{O}\left(\sqrt{\frac{(H+1)^2}{e_{q-1}}}\right)$.

We denote $\tilde{\xi}_h^t = \left(\widetilde{V}_h^t - V_h^\star\right)(s_h^t) + H$ and $\tilde{\xi}_h^{\flat,t} = \left(\widetilde{V}_h^{\flat,t} - V_h^\star\right)(s_h^t)$. Combining Lemma 8 and Lemma 9 with (40), we have that for any $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$ within the stage $q_h^t \ge 1$,

$$\begin{aligned}
\delta_h^t \le& 2\alpha_{n_h^t}^0 V_{h+1}^0 + \eta_{\ell^{n_h^t},h}\sum_{i=1}^{n_h^t}\alpha_{n_h^t}^i \tilde{\xi}_{h+1}^{\ell^i} + \frac{1-\eta_{\ell^{n_h^t},h}}{1+e_{q-1}}\sum_{i=0}^{e_{q-1}} \tilde{\xi}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}} + \eta_{\ell^{n_h^t},h}\tilde{b}_h^t + (1-\eta_{\ell^{n_h^t},h})\tilde{b}_h^{\flat,t} \\
&+ (\delta_{h+1}^t - \tilde{\xi}_{h+1}^t + \tau_{h+1}^t) - \eta_{\ell^{n_h^t},h} H,
\end{aligned}$$

and during the initial stage $q_h^t = 0$

$$\delta_h^t \le 2\alpha_{n_h^t}^0 V_{h+1}^0 + \eta_{\ell^{n_h^t},h}\sum_{i=1}^{n_h^t}\alpha_{n_h^t}^i \tilde{\xi}_{h+1}^{\ell^i} + \tilde{b}_h^t + (1-\eta_{\ell^{n_h^t},h})V_{h+1}^0 + (\delta_{h+1}^t - \tilde{\xi}_{h+1}^t + \tau_{h+1}^t) - \eta_{\ell^{n_h^t},h} H.$$

Define $b_h^t = \eta_{\ell^{n_h^t},h}\tilde{b}_h^t + (1-\eta_{\ell^{n_h^t},h})\cdot \tilde{b}_h^{\flat,t}\mathbb{1}\{q_h^t \geq 1\}$. Thus, summing over t from 1 to T leads to

$$\begin{aligned}\sum_{t=1}^{T}\delta_h^t &\leq \sum_{t:q_h^t=0}^{T}\delta_h^t + \sum_{t:q_h^t\geq 1}^{T}\delta_h^t \\ &\leq 2\sum_{t=1}^{T}\alpha_{n_h^t}^0 V_{h+1}^0 + \sum_{t=1}^{T}\eta_{\ell^{n_h^t},h}\sum_{i=1}^{m}\alpha_{n_h^t}^i\tilde{\xi}_{h+1}^{\ell^i} + \sum_{t=1}^{T}\frac{(1-\eta_{\ell^{n_h^t},h})\mathbb{1}\{q_h^t\geq 1\}}{1+e_{q-1}}\sum_{i=0}^{e_{q-1}}\tilde{\xi}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}} \\ &\quad + \sum_{t=1}^{T}(b_h^t + \delta_{h+1}^t - \tilde{\xi}_{h+1}^t + \tau_{h+1}^t - \eta_{\ell^{n_h^t},h}H) + \sum_{t:q_h^t=0}^{T}(1-\eta_{\ell^{n_h^t},h})V_{h+1}^0. \end{aligned} \tag{43}$$

The first term on the right-hand-side of (43) can be bounded by

$$2\sum_{t=1}^{T}\alpha_{n_h^t}^0 V_{h+1}^0 = 2\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{m=1}^{n_h^T(s,a)}\alpha_m^0 V_{h+1}^0 \leq 2\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{m=1}^{\infty}\alpha_{n_h^t}^0 V_{h+1}^0 \leq 2n_0 SAV_{h+1}^0, \tag{44}$$

where the last inequality is from

$$\sum_{m=1}^{\infty}\alpha_{n_h^t}^0 = \frac{n_0}{H+n_0}\left(1+\frac{n_0+1}{H+n_0+1}+\cdots\right) = \frac{n_0}{H+n_0}\frac{H+n_0}{H-1} \leq \frac{n_0}{H-1}.$$

For any $(t,h)\in[T]\times[H]$, let the mixing rate be chosen as

$$\eta_{t,h} := \frac{1}{\sqrt{e_{q_h^t}}+1} = \frac{1}{\sqrt{(1+1/H)^{q_h^t}H}+1}, \tag{45}$$

which is non-increasing in t along the visits to any fixed (h,s,a).

For the second term of (43), we have

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{n_h^t}\eta_{\ell^{n_h^t},h}\,\alpha_{n_h^t}^i\,\tilde{\xi}_{h+1}^{\ell^i} &= \sum_{t'=1}^{T}\tilde{\xi}_{h+1}^{t'}\sum_{t=t'}^{T}\eta_{\ell^{n_h^t},h}\sum_{i=1}^{n_h^t}\alpha_{n_h^t}^i\,\mathbb{1}\{\ell^i=t'\} \\ &= \sum_{t'=1}^{T}\tilde{\xi}_{h+1}^{t'}\sum_{t=t'}^{T}\eta_{\ell^{n_h^t},h}\,\alpha_{n_h^t}^{n_h^{t'}} \\ &\leq \sum_{t'=1}^{T}\tilde{\xi}_{h+1}^{t'}\Big(\max_{t\geq t'}\eta_{\ell^{n_h^t},h}\Big)\sum_{m=n_h^{t'}}^{\infty}\alpha_m^{n_h^{t'}} \\ &\leq \left(1+\frac{1}{H}\right)\sum_{t'=1}^{T}\eta_{\ell^{n_h^{t'}},h}\,\tilde{\xi}_{h+1}^{t'}, \end{aligned} \tag{46}$$

where the penultimate inequality uses the non-increasing property of $\{\eta_{t,h}\}_{t=0}^T$ along the visits to any fixed (h,s,a) and the last inequality follows from Proposition 1. Note that $\tilde{\xi}_{h+1}^t \leq 3H$ for any $t\in[T]$. Thus,

$$\sum_{t=1}^{T}\sum_{i=1}^{n_h^t}\eta_{\ell^{n_h^t},h}\,\alpha_{n_h^t}^i\,\tilde{\xi}_{h+1}^{\ell^i} \lesssim H\sum_{t=1}^{T}\eta_{\ell^{n_h^t},h}$$

To control the third term of (43), note that by (45) ensures, the sequence $\{\eta_{t,h}\}$ is constant within each stage for any fixed (h,s,a). Following Zhang et al. [7], we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=0}^{e_{q_h^t}-1}\frac{(1-\eta_{\ell^{n_h^t},h})\mathbb{1}\{q_h^t\geq 1\}}{1+e_{q_h^t-1}}\tilde{\xi}_{h+1}^{\flat,\ell^{\flat,i}_{q_h^t-1}}\\
&=\sum_{t=1}^{T}\frac{(1-\eta_{\ell^{n_h^t},h})\mathbb{1}\{q_h^t\geq 1\}}{1+e_{q_h^t-1}}\sum_{t'=1}^{T}\tilde{\xi}_{h+1}^{\flat,t'}\sum_{i=0}^{e_{q_h^t-1}}\mathbb{1}\{\ell^{\flat,i}_{q_h^t-1}=t'\}\\
&=\sum_{t'=1}^{T}\tilde{\xi}_{h+1}^{\flat,t'}\sum_{t=1}^{T}\frac{(1-\eta_{\ell^{n_h^t},h})\mathbb{1}\{q_h^t\geq 1\}}{1+e_{q_h^t-1}}\sum_{i=0}^{e_{q_h^t-1}}\mathbb{1}\{\ell^{\flat,i}_{q_h^t-1}=t'\}\\
&\leq(1+\frac{1}{H})^2\sum_{t'=1}^{T}(1-\eta_{t',h})\tilde{\xi}_{h+1}^{\flat,t'},
\end{aligned}\tag{47}$$

where the proof of the last inequality is deferred to Appendix C.6.

For the last term in the right-hand side of (43), $(t,h)\in[T]\times[H+1]$ with $q_h^t=0$, we have

$$\sum_{t:q_h^t=0}^{T}(1-\eta_{\ell^{n_h^t},h})V_{h+1}^0\leq SAHV_{h+1}^0\leq 2SAH^2.\tag{48}$$

since there are at most H episodes during the initial stage for any fixed (h,s,a).

Moreover, it is easy to verify that $(1-\eta_{t',h})\tilde{\xi}_{h+1}^{\flat,t}\leq\xi_{h+1}^t+\eta_{\ell^{n_{h+1}^t},h+1}2H$, for any $(t,h)\in[T]\times[H]$ with $q_h^t\geq 1$ since

$$\begin{aligned}
(1-\eta_{t,h})\cdot\widetilde{V}_{h+1}^{\flat,t}(s_{h+1}^t)&\leq\eta_{\ell^{n_{h+1}^t},h+1}\max_{j\in[J]}\{\widetilde{Q}_{h+1}^{j,t}(s_{h+1}^t,\pi_{h+1}^{\flat,t}(s_{h+1}^t))\}\\
&\quad+(1-\eta_{\ell^{n_{h+1}^t},h+1})\widetilde{Q}_{h+1}^{\flat,t}(s_{h+1}^t,\pi_{h+1}^{\flat,t}(s_{h+1}^t))+\eta_{\ell^{n_{h+1}^t},h+1}\widetilde{V}_{h+1}^{\flat,t}(s_{h+1}^t)\\
&\leq\eta_{\ell^{n_{h+1}^t},h+1}\max_{j\in[J]}\{\widetilde{Q}_{h+1}^{j,t}(s_{h+1}^t,\pi_{h+1}^t(s_{h+1}^t))\}\\
&\quad+(1-\eta_{\ell^{n_{h+1}^t},h+1})\cdot\widetilde{Q}_{h+1}^{\flat,t}(s_{h+1}^t,\pi_{h+1}^t(s_{h+1}^t))+\eta_{\ell^{n_{h+1}^t},h+1}2H\\
&=Q_{h+1}^t(s_{h+1}^t,\pi_{h+1}^t(s_{h+1}^t))+\eta_{\ell^{n_{h+1}^t},h+1}2H\\
&=V_{h+1}^t(s_{h+1}^t)+\eta_{\ell^{n_{h+1}^t},h+1}2H,
\end{aligned}\tag{49}$$

where the first inequality follows from Line 21 in Algorithm 1, and the second inequality is due to the definition of the greedy policy π^t and the first equation is from (36).

By substituting (44)-(48) to (43) and using (49), we have

$$\begin{aligned}
\sum_{t=1}^{T}\delta_h^t&\lesssim SAH^2+H\sum_{t=1}^{T}\eta_{\ell^{n_h^t},h}+(1+\frac{1}{H})^2\sum_{t=1}^{T}(1-\eta_{\ell^{n_h^t},h})\tilde{\xi}_{h+1}^{\flat,t}\\
&\quad+\sum_{t=1}^{T}(b_h^t+\delta_{h+1}^t-\xi_{h+1}^t+\tau_{h+1}^t)\\
&\lesssim SAH^2+H\sum_{t=1}^{T}\left(\eta_{\ell^{n_h^t},h}+\eta_{\ell^{n_{h+1}^t},h+1}\right)+(1+\frac{1}{H})^2\sum_{t=1}^{T}\xi_{h+1}^t\\
&\quad+\sum_{t=1}^{T}(b_h^t+\delta_{h+1}^t-\xi_{h+1}^t+\tau_{h+1}^t)\\
&\lesssim SAH^2+H\sum_{t=1}^{T}\left(\eta_{\ell^{n_h^t},h}+\eta_{\ell^{n_{h+1}^t},h+1}\right)+(1+\frac{1}{H})^2\sum_{t=1}^{T}\delta_{h+1}^t+\sum_{t=1}^{T}(b_h^t+\tau_{h+1}^t),
\end{aligned}$$

where the last line is from the fact that $\xi_{h+1}^t \le \delta_{h+1}^t$, since $V^\star \ge V^\pi$ for any policy π.

Note that $(1+\frac{1}{H})^{2H} \le e^2$. Thus, by unrolling the above inequality until $h=1$, we obtain

$$\sum_{t=1}^{T} \delta_1^t \le \widetilde{O}\left(SAH^3 + \sum_{t=1}^{T}\sum_{h=1}^{H} (b_h^t + \tau_{h+1}^t + \eta_{\ell^{n_h^t},h} H) \right). \tag{50}$$

Recall that $b_h^t = \eta_{\ell^{n_h^t},h} \tilde{b}_h^t + (1-\eta_{\ell^{n_h^t},h}) \cdot \tilde{b}_h^{\flat,t} \mathbb{1}\{q_h^t \ge 1\}$. Before proceeding, we note that $\frac{e_q}{\sqrt{e_{q-1}}} \le 2\sqrt{e_{q-1}}$ for any $q \ge 1$ and we denote $Q_{h,s,a} = q_h^{T+1}(s,a)$ for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ such that $\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \sum_{q=1}^{Q_{h,s,a}} e_{q-1} \le TH$. Also, let $Q = \max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} Q_{h,s,a} \le \frac{\log(T/H)}{\log(1+\frac{1}{H})} \le 4H\log(T/H)$, where the last inequality is due to the fact that $\log(1+\frac{1}{H}) \ge \frac{1}{4H}$ for $H \ge 1$. Thus, one has

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{1}\{q_h^t \ge 1\}}{\sqrt{e_{q-1}}} &\le \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \sum_{q=1}^{Q_{h,s,a}} \frac{e_q}{\sqrt{e_{q-1}}} \le 2 \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \sum_{q=1}^{Q_{h,s,a}} \sqrt{e_{q-1}} \\
&\le 4\sqrt{SAH^2\log(T/H)} \sqrt{\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \sum_{q=1}^{Q_{h,s,a}} e_{q-1}} \\
&\le O(1)\sqrt{SAH^2\log(T)\cdot TH} \le \widetilde{O}(\sqrt{SAH^3T}),
\end{aligned}$$

where the penultimate line is from the Cauchy-Schwarz inequality.

In addition, we have

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H} \tilde{b}_h^t &\le \widetilde{O}(1) \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \sum_{m=1}^{n_h^T(s,a)} \sqrt{\frac{H^3}{m}} \\
&\le \widetilde{O}(1)\sqrt{SAH\cdot T/SA} \sqrt{\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \sum_{m=1}^{\frac{T}{SA}} \frac{H^3}{m}} \\
&\le \widetilde{O}(\sqrt{SAH^5T}),
\end{aligned}$$

where the penultimate inequality holds since the left-hand side is maximized when $n_h^T(s,a) = \frac{T}{SA}$ for every $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$. Thus, one has

$$\sum_{h=1}^{H}\sum_{t=1}^{T} b_h^t = \sum_{h=1}^{H}\sum_{t=1}^{T} \left(\eta_{\ell^{n_h^t},h} \tilde{b}_h^t + (1-\eta_{\ell^{n_h^t},h}) \sum_{h=1}^{H}\sum_{t=1}^{T} \mathbb{1}\{q_h^t \ge 1\} \cdot \tilde{b}_h^{\flat,t} \right) \le \widetilde{O}(\sqrt{SAH^5T}). \tag{51}$$

Moreover, we have

$$\tau_{h+1}^t = P_{h,s_h^t,a_h^t}\left(V_{h+1}^\star - V_{h+1}^{\pi^t}\right) - \left(V_{h+1}^\star - V_{h+1}^{\pi^t}\right)(s_{h+1}^t)$$

is a martingale-difference sequence with respect to the filtration $\mathcal{F}_{t,h}$ that contains all the random variables before the step $h+1$ at the t-th episode. By the Hoeffding's inequality, we have

$$\left| \sum_{h=1}^{H}\sum_{t=1}^{T} \tau_{h+1}^t \right| \le \widetilde{O}(\sqrt{H^3T}) \tag{52}$$

with probability $1-\delta$. Note that

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H} \eta_{\ell^{n_h^t},h} = \sum_{s,a,h} \sum_{q}^{Q_{h,s,a}} \frac{e_q}{\sqrt{e_q}+1} \le \sum_{s,a,h} \sum_{q}^{Q_{h,s,a}} \sqrt{e_q} &\lesssim \sqrt{SAHQ} \sqrt{\sum_{s,a,h} \sum_{q=0}^{Q_{h,s,a}} e_q} \\
&\le \widetilde{O}(\sqrt{SAH^3T})
\end{aligned}$$

where we use the fact $Q \le 4H\log(T/H)$ again.

The last part is to show the upper bound of $\sum_{t=0}^{T-1}\frac{\eta_{t,1}}{1-\eta_{t,1}}H$. By the choice of $\eta_{t,h}$ defined in (45), we have

$$\begin{aligned}\sum_{t=0}^{T-1}\frac{\eta_{t,1}}{1-\eta_{t,1}} \le \sum_{a\in\mathcal{A}}\sum_{q=0}^{Q}\frac{e_q}{\sqrt{e_q}} &\le \sum_{a\in\mathcal{A}}\sum_{q=1}^{Q}\sqrt{e_q}\\ &\le \sqrt{AQ}\sqrt{\sum_{a\in\mathcal{A}}\sum_{q=1}^{Q}e_q}\\ &\le \widetilde{O}(\sqrt{AHT}).\end{aligned}$$

Finally, substituting (51) and (52) into (50) and rescaling δ to $\delta/4$ complete the proof.

C.2 Proof of Lemma 7

For any $(s,a,h,t)\in\mathcal{S}\times\mathcal{A}\times[H]\times[T]$, we let $m=n_h^t(s,a)$ and $q=q_h^t(s,a)$ for simplicity. From (37), one has

$$Q_h^t(s,a)=\eta_{\ell^{n_h^t},h}\max_{j\in[J]}\Big\{\sum_{i=0}^{m}W_{j,m}^i\widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i})+r_h(s,a)\Big\}+(1-\eta_{\ell^{n_h^t},h})\widetilde{Q}_h^{\flat,t}(s,a), \tag{53}$$

where $\{\ell^i\}_{i=0}^m$ represent the episode index of the i-th visit of (h,s,a) before t, and

$$\widetilde{Q}_h^{\flat,t}(s,a)=r_h(s,a)+\max_{j\in[J]}\left\{\sum_{i=0}^{e_{q-1}}W_{j,q-1}^{\flat,i}\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}})\right\}.$$

Before proceeding, we first claim that for any $(s,a,t,h)\in\mathcal{S}\times\mathcal{A}\times[T]\times[H]$, we have $\widetilde{Q}_h^{\flat,t}(s,a)\ge Q_h^\star(s,a)$ if the following relationship holds

$$\max_{j\in[J]}\left\{\sum_{i=0}^{e_{q-1}}W_{j,q-1}^{\flat,i}V_{h+1}^\star(s_{h+1}^{\ell_{q-1}^{\flat,i}})\right\}\ge P_{h,s,a}V_{h+1}^\star, \tag{54}$$

where we leave the detailed proof of this claim to the end of this subsection.

Next, the following lemma shows that (54) holds with high probability, which implies

$$Q_h^t(s,a)\ge(1-\eta_{\ell^{n_h^t},h})\widetilde{Q}_h^{\flat,t}(s,a)\ge(1-\eta_{\ell^{n_h^t},h})Q_h^\star(s,a)$$

and completes the proof of Lemma 7. The detailed proof of the following lemma is postponed to Appendix C.5.

Lemma 10. *Consider $\delta\in(0,1)$. Assume that $J=\lceil c\cdot\log(SAHT/\delta)\rceil$, $\kappa^\flat=c\cdot(\log(SAH/\delta)+\log(T))$, and $n_0^\flat=\lceil c\cdot\log(T)\cdot\kappa^\flat\rceil$, where $c>0$ is some universal constant. Let $V_h^0=2(H-h+1)$. Then, for any $(s,a,h,t)\in\mathcal{S}\times\mathcal{A}\times[H]\times[T]$ with stage index $q\ge1$, the equation (54) holds with probability at least $1-\delta$.*

Proof of the claim: Assuming that (54) holds, we will first show by induction that

$$\widetilde{Q}_h^{\flat,t}(s,a)\ge Q_h^\star(s,a)$$

holds correspondingly, for any $(s,a,h,t)\in\mathcal{S}\times\mathcal{A}\times[H]\times[T]$. To begin with, when $h'=H+1$, $\widetilde{Q}_{H+1}^{\flat,t}(s,a)=Q_{H+1}^\star=0$ holds naturally. When $h'=h+1\le H$, suppose that $\widetilde{Q}_{h+1}^{\flat,t}(s,a)\ge Q_{h+1}^\star(s,a)$, for any $(t,s,a)\in[T]\times\mathcal{S}\times\mathcal{A}$. By this hypothesis, we also have $\widetilde{V}_{h+1}^{\flat,t}(s_{h+1}^t)=\max_a\widetilde{Q}_{h+1}^{\flat,t}(s_{h+1}^t,a)\ge\widetilde{Q}_{h+1}^{\flat,t}(s_{h+1}^t,\pi_h^\star(s_{h+1}^t))\ge Q_{h+1}^\star(s_{h+1}^t,\pi_h^\star(s_{h+1}^t))=V_{h+1}^\star(s_{h+1}^t)$. By induction, when $h'=h$, we have

$$\begin{aligned}\widetilde{Q}_h^{\flat,t}(s,a)&=r_h(s,a)+\max_{j\in[J]}\left\{\sum_{i=0}^{e_{q-1}}W_{j,q-1}^{\flat,i}\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}})\right\}\\&\ge r_h(s,a)+\max_{j\in[J]}\left\{\sum_{i=0}^{e_{q-1}}W_{j,q-1}^{\flat,i}V_{h+1}^\star(s_{h+1}^{\ell_{q-1}^{\flat,i}})\right\}\end{aligned}$$

Thus, if (54) holds, we have $\widetilde{Q}_h^{\flat,t}(s,a) \geq r_h(s,a) + P_{h,s,a}V_{h+1}^\star = Q_h^\star(s,a)$ for any $(s,a,t,h) \in \mathcal{S}\times\mathcal{A}\times[T]\times[H]$, by the Bellman optimality equation (3).

C.3 Proof of Lemma 8

For any $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$, denote $m = n_h^t(s,a)$ as the number of visits on (h,s,a) before the t-th episode. Also, let $\alpha_m^0 := \prod_{k=1}^m \frac{n_0+k-1}{H+n_0+k}$ and $\alpha_m^i := \frac{H+1}{H+n_0+i}\prod_{k=i+1}^m \frac{n_0+k-1}{H+n_0+k}$.

From (17) , for any $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$, we have

$$\widetilde{Q}_h^{j,t}(s,a) = r_h(s,a) + \sum_{i=0}^m W_{j,m}^i \widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i}) \leq r_h(s,a) + 2H\cdot\sum_{i=0}^m W_{j,m}^i \frac{\widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i})}{2H}.$$

Note that $\frac{\widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i})}{2H} \leq 1$ for any $i = 0,\ldots,m$. Thus, we can apply Lemma 5 and Proposition 1

$$\begin{aligned}
\max_{j\in[J]} \widetilde{Q}_h^{j,t}(s,a) &\leq r_h(s,a) + 2H\left(\sum_{i=0}^m \mathbb{E}[W_{j,m}^i]\frac{\widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i})}{2H} + \frac{c_1}{2}\sqrt{\frac{(H+1)\kappa^2\log^3(2SAHTJ/\delta)}{H+n_0+m}}\right.\\
&\quad\left. + \frac{c_2}{2}\frac{(H+1)\kappa\log^2(2SAHTJ/\delta)}{H+n_0+m}\right)\\
&\leq r_h(s,a) + \alpha_m^0\widetilde{V}_{h+1}^0 + \sum_{i=1}^m \alpha_m^i \widetilde{V}_{h+1}^{\ell^i}(s_{h+1}^{\ell^i})\\
&\quad + c_1\sqrt{\frac{(H+1)^3\kappa^2\log^3(2SAHTJ/\delta)}{H+n_0+m}} + c_2\frac{(H+1)^2\kappa\log^2(2SAHTJ/\delta)}{H+n_0+m},
\end{aligned}$$

with probability at least $1-\delta/2$, where c_1, c_2 are universal constants. By the Bellman optimality equation (3), we have

$$\begin{aligned}
\zeta_h^t &\leq \alpha_m^0\widetilde{V}_{h+1}^0 + \sum_{i=1}^m \alpha_m^i\left(\left(\widetilde{V}_{h+1}^{\ell^i} - V_{h+1}^\star\right)(s_{h+1}^{\ell^i}) + H\right) + \sum_{i=1}^m \alpha_m^i\left(V_{h+1}^\star(s_{h+1}^{\ell^i}) - P_{h,s,a}V_{h+1}^\star\right)\\
&\quad + c_1\sqrt{\frac{(H+1)^3\kappa^2\log^3(2SAHTJ/\delta)}{H+n_0+m}} + c_2\frac{(H+1)^2\kappa\log^2(2SAHTJ/\delta)}{H+n_0+m} + H - \sum_{i=1}^m \alpha_m^i H\\
&\leq 2\alpha_m^0\widetilde{V}_{h+1}^0 + \sum_{i=1}^m \alpha_m^i\left(\left(\widetilde{V}_{h+1}^{\ell^i} - V_{h+1}^\star\right)(s_{h+1}^{\ell^i}) + H\right) + \sum_{i=1}^m \alpha_m^i\left(V_{h+1}^\star(s_{h+1}^{\ell^i}) - P_{h,s,a}V_{h+1}^\star\right)\\
&\quad + c_1\sqrt{\frac{(H+1)^3\kappa^2\log^3(2SAHTJ/\delta)}{H+n_0+m}} + c_2\frac{(H+1)^2\kappa\log^2(2SAHTJ/\delta)}{H+n_0+m},
\end{aligned} \tag{55}$$

where the last line uses the fact $H \leq \widetilde{V}_{h+1}^0$ and the equation (20) such that

$$H - \sum_{i=1}^m \alpha_m^i H = \alpha_m^0 H.$$

In addition, we denote $\mathcal{F}_i$ as the filtration containing all the random variables before the episode $\ell_h^i(s,a)$, such that $\alpha_m^i\left(V_{h+1}^\star(s_{h+1}^{\ell^i}) - P_{h,s,a}V_{h+1}^\star\right)$ is a martingale difference sequence w.r.t. $\{\mathcal{F}_i\}_{i\geq 0}$ for any $i \leq m$. Following [6] and by Hoeffding's inequality and Proposition 1 (i.e., (22)), with probability at least $1-\delta/2$, we have

$$\begin{aligned}
\left|\sum_{i=1}^m \alpha_m^i\left(V_{h+1}^\star(s_{h+1}^{\ell^i}) - P_{h,s,a}V_{h+1}^\star\right)\right| &\leq c_3 H\sqrt{\sum_{i=1}^m(\alpha_m^i)^2\log(2SAHT/\delta)}\\
&\leq c_3\sqrt{\frac{(H+1)^3\kappa}{H+n_0+m}\log(2SAHT/\delta)}
\end{aligned}$$

for any $(t,h,s,a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$ and some universal constant $c_3 > 0$. Substituting the above inequality into (55) gives

$$\zeta_h^t \le 2\alpha_m^0 V_{h+1}^0 + \sum_{i=1}^m \alpha_m^i \left(\left(\widetilde{V}_{h+1}^{\ell^i} - V_{h+1}^\star \right)(s_{h+1}^{\ell^i}) + H \right) + (c_1 + c_3)\sqrt{\frac{(H+1)^3\kappa^2 \log^3(2SAHTJ/\delta)}{H+n_0+m}} + c_2 \frac{(H+1)^2 \kappa \log^2(2SAHTJ/\delta)}{H+n_0+m}.$$

By letting $\tilde{b}_h^t = \widetilde{O}\left(\sqrt{\frac{(H+1)^3}{H+n_0+m}} + \frac{(H+1)^2}{H+n_0+m} \right)$, we complete the proof.

C.4 Proof of Lemma 9

Note that during the initial stage, for any $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$ within the stage $q_h^t(s,a) = 0$, we have

$$\widetilde{Q}_h^{\flat,t}(s,a) - Q_h^\star(s,a) \le V_{h+1}^0.$$

From (19) and Lemma 7, for any $(t,h,s,a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$ with $q_h^t(s,a) \ge 1$, we have

$$0 \le \widetilde{Q}_h^{\flat,t}(s,a) - Q_h^\star(s,a) \le \max_{j\in[J]} \Big\{ \sum_{i=0}^{e_{q-1}} W_{j,q-1}^{\flat,i} \widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}}) \Big\} - P_{h,s,a} V_{h+1}^\star. \tag{56}$$

Then, applying Lemma 6 leads to

$$\begin{aligned}
&\sum_{i=0}^{e_{q-1}} W_{j,q-1}^{\flat,i} \widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}}) \\
&= 2H \left(\sum_{i=0}^{e_{q-1}} W_{j,q-1}^{\flat,i} \frac{\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}})}{2H} \right) \\
&\le 2H \left(\frac{1}{1+e_{q-1}} \sum_{i=0}^{e_{q-1}} \frac{\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}})}{2H} + c_1^\flat \sqrt{\frac{(\kappa^\flat)^2 \log^3(SAHTJ/\delta)}{e_{q-1}}} + c_2^\flat \frac{\kappa^\flat \log^2(SAHTJ/\delta)}{e_{q-1}} \right) \\
&\le \frac{1}{1+e_{q-1}} \sum_{i=0}^{e_{q-1}} \widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}}(s_{h+1}^{\ell_{q-1}^{\flat,i}}) + 2H \left(c_1^\flat \sqrt{\frac{(\kappa^\flat)^2 \log^3(SAHTJ/\delta)}{e_{q-1}}} + c_2^\flat \frac{\kappa^\flat \log^2(SAHTJ/\delta)}{e_{q-1}} \right),
\end{aligned}$$

for every $j \in [J]$ with probability at least $1 - \delta/2$. Following the similar procedure in Appendix C.3 and by Hoeffding's inequality, with probability at least $1-\delta/2$, we have

$$\left| \sum_{i=0}^{e_{q-1}} \frac{1}{1+e_{q-1}} \left(V_{h+1}^\star(s_{h+1}^{\ell_{q-1}^{\flat,i}}) - P_{h,s,a} V_{h+1}^\star \right) \right| \le c_4 \sqrt{\frac{H^2}{e_{q-1}} \log(2SAHT/\delta)}.$$

Thus, we obtain that the following holds with probability $1-\delta$,

$$\begin{aligned}
&\widetilde{Q}_h^{\flat,t}(s,a) - Q_h^\star(s,a) \\
&\le \frac{1}{1+e_{q-1}} \sum_{i=0}^{e_{q-1}} \left(\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}} - V_{h+1}^\star \right)(s_{h+1}^{\ell_{q-1}^{\flat,i}}) + \frac{1}{1+e_{q-1}} \sum_{i=0}^{e_{q-1}} \left(V_{h+1}^\star(s_{h+1}^{\ell_{q-1}^{\flat,i}}) - P_{h,s,a} V_{h+1}^\star \right) \\
&\quad + 2H \left(c_1^\flat \sqrt{\frac{(\kappa^\flat)^2 \log^3(SAHTJ/\delta)}{e_{q-1}}} + c_2^\flat \frac{\kappa^\flat \log^2(SAHTJ/\delta)}{e_{q-1}} \right) \\
&\le \frac{1}{1+e_{q-1}} \sum_{i=1}^{e_{q-1}} \left(\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}} - V_{h+1}^\star \right)(s_{h+1}^{\ell_{q-1}^{\flat,i}}) + \tilde{b}_h^{\flat,t},
\end{aligned}$$

where $\widetilde{b}_h^{\flat,t} = \widetilde{O}\left(\sqrt{\frac{H^2}{e_{q-1}}}\right)$.

Combining two cases of $q_h^t(s,a) \geq 1$ and $q_h^t(s,a) = 0$ leads to

$$\zeta_h^{\flat,t} \leq \mathbb{1}\{q_h^t(s,a) \geq 1\} \left(\sum_{i=0}^{e_{q-1}} \frac{1}{1+e_{q-1}} \left(\widetilde{V}_{h+1}^{\flat,\ell_{q-1}^{\flat,i}} - V_{h+1}^\star \right) (s_{h+1}^{\ell_{q-1}^{\flat,i}}) + \widetilde{b}_h^{\flat,t} \right) + \mathbb{1}\{q_h^t(s,a) = 0\} \cdot V_{h+1}^0,$$

which completes the proof.

C.5 Proof of Lemma 10

Following [10], let $\mathcal{E}^\star(\delta)$ be the event containing all $(t,h,s,a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$, such that

$$\mathcal{K}_{\inf}\left(\frac{1}{e_q} \sum_{i=1}^{e_q} \delta_{V_{h+1}^\star(s_{h+1}^{\ell_q^{\flat,i}})}, P_{h,s,a} V_{h+1}^\star \right) \leq \frac{\beta^\star(\delta, e_q)}{e_q}, \tag{57}$$

where $q = q_h^t(s,a)$ and $\beta^\star(\delta, n) := \log(2SAH/\delta) + 3\log(\mathrm{e}\pi(2n+1))$. The following lemma shows that $\mathcal{E}^\star(\delta)$ holds with probability $1 - \frac{\delta}{2}$

Lemma 11 (Lemma 4 in [10]). *Consider $\delta \in (0,1)$. With probability $1 - \frac{\delta}{2}$, the following event holds*

$$\mathcal{K}_{\inf}\left(\frac{1}{e_q} \sum_{i=1}^{e_q} \delta_{V_{h+1}^\star(s_{h+1}^{\ell_q^i})}, P_{h,s,a} V_{h+1}^\star \right) \leq \frac{\beta^\star(\delta, e_q)}{e_q}, \quad \forall (t,h,s,a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}, \tag{58}$$

where $q = q_h^t(s,a)$ and $\beta^\star(\delta, e_q) := \log(2SAH/\delta) + 3\log(e\pi(2e_q+1))$.

Consider some fixed $(t,h,s,a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$ within the stage $q = q_h^t(s,a)$. To construct an anti-concentration inequality bound of the weighted sum $W_{j,q}^{\flat,i} V_{h+1}^\star(s_{h+1}^{\ell_q^{\flat,i}})$, we first recall the following two lemmas provided in Tiapkin et al. [10].

Lemma 12 (Lemma 3 in Tiapkin et al. [10]). *For any stage $q \geq 0$ and $j \in [J]$, the aggregated weights $W_{j,q}^\flat$ follows a standard Dirichlet distribution* $\mathrm{Dir}(n_0^\flat/\kappa^\flat, 1/\kappa^\flat, \ldots, 1/\kappa^\flat)$.

Lemma 13 (Lemma 10 in Tiapkin et al. [10]). *For any $\alpha = (\alpha_0 + 1, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_{++}^{m+1}$, define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell / \overline{\alpha}$, $\ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Also define a measure $\overline{\nu} = \sum_{i=0}^m \overline{p}(i) \cdot \delta_{f(i)}$. Let $\varepsilon \in (0,1)$. Assume that $\alpha_0 \geq c_0 + \log_{17/16}(2(\overline{\alpha} - \alpha_0))$ for some universal constant c_0. Then for any $f : \{0, \ldots, m\} \to [0, b_0]$ such that $f(0) = b_0$, $f(j) \leq b \leq b_0/2$, $j \in [m]$, and any $\mu \in (0, b)$*

$$\mathbb{P}_{w \sim \mathrm{Dir}(\alpha)}[wf \geq \mu] \geq (1-\varepsilon) \mathbb{P}_{g \sim \mathcal{N}(0,1)} \left[g \geq \sqrt{2\overline{\alpha} \mathcal{K}_{\mathit{inf}}(\overline{\nu}, \mu)} \right].$$

According to Lemma 12 and applying Lemma 13 with $\alpha_0 = n_0^\flat/\kappa^\flat - 1$, $\alpha_i = 1/\kappa^\flat$, $\forall i \in [e_q]$, $r_0 = 2$, $b_0 = 2(H-h+1)$, and $\overline{\nu}_q = \frac{n_0^\flat - \kappa^\flat}{e_q + n_0^\flat - \kappa^\flat} \delta_{V_{h+1}^\star(s_0)} + \sum_{i=1}^{e_q} \frac{1}{e_q + n_0^\flat - \kappa^\flat} \delta_{V_{h+1}^\star(s_{h+1}^{\ell_q^{\flat,i}})}$, we have that conditioned on the event $\mathcal{E}^\star(\delta)$ holds,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=0}^{e_q} W_{j,q}^{\flat,i} V_{h+1}^\star(s_{h+1}^{\ell_q^{\flat,i}}) \geq P_{h,s,a} V_{h+1}^\star \,\middle|\, \mathcal{E}^\star(\delta) \right) \\
&\geq \frac{1}{2}\left(1 - \Phi\left(\sqrt{\frac{2(e_q + n_0^\flat - \kappa^\flat) \mathcal{K}_{\inf}(\overline{\nu}_q, P_{h,s,a} V_{h+1}^\star)}{\kappa^\flat}} \right) \right) \geq \frac{1}{2}\left(1 - \Phi\left(\sqrt{\frac{2\beta^\star(\delta, T)}{\kappa^\flat}} \right) \right).
\end{aligned}$$

where Φ denotes the CDF of the standard normal distribution. Here the last inequality is from Lemma 4 and Lemma 11.

Then, by selecting $\kappa^\flat = 2\beta^\star(\delta, T)$, we ensure a constant probability of optimism:

$$\mathbb{P}\left(\sum_{i=0}^{e_q} W_{j,q}^{\flat,i} V_{h+1}^\star(s_{h+1}^{\ell_q^{\flat,i}}) \geq P_{h,s,a} V_{h+1}^\star \,\middle|\, \mathcal{E}^\star(\delta) \right) \geq \frac{1 - \Phi(1)}{2} \triangleq \gamma.$$

Now, choosing $J = \left\lceil \log\left(\frac{2SAHT}{\delta}\right) / \log(1/(1-\gamma)) \right\rceil = \lceil c_J \cdot \log(2SAHT/\delta) \rceil$ ensures:

$$\mathbb{P}\left(\max_{j\in[J]} \left\{ \sum_{i=0}^{e_q} W_{j,q}^{\flat,i} V_{h+1}^{\star}(s_{q}^{\ell^{\flat,i}}_{h+1}) \right\} \geq P_{h,s,a} V_{h+1}^{\star} \,\middle|\, \mathcal{E}^{\star}(\delta) \right) \geq 1 - (1-\gamma)^J \geq 1 - \frac{\delta}{2SAHT}.$$

By applying a union bound over $(t,h,s,a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$ and taking expectation on $\mathcal{E}^{\star}(\delta)$, we conclude the proof.

C.6 Proof of equation (47)

Fix any episode $t' \in [T]$ that lies in stage $q-1 \geq 0$ for the triple (h,s,a). Then in (47) we have

$$\sum_{i=0}^{e_{q_h^t - 1}} \mathbb{1}\{\ell^{\flat,i}_{q_h^t - 1} = t'\} = 1$$

holds if and only if the episodes t' and t visit the same triple (h,s,a) and the visit (t,h,s,a) lies in the next stage q of the triple (h,s,a). Let $S_{t',h,s,a} = \left\{ t \in [T] : \sum_{i=0}^{e_{q_h^t - 1}} \mathbb{1}\{\ell^{\flat,i}_{q_h^t - 1} = t'\} \right\}$ represent all the visits of (h,s,a) in the same stage q such that its cardinality is at most e_q. For any $t \in S_{t',h,s,a}$, the mixing rate remains the same.

Then, we can then decompose

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{(1-\eta_{\ell^{n_h^t},h})}{1+e_{q_h^t-1}} \sum_{i=0}^{e_{q_h^t-1}} \mathbb{1}\{\ell^{\flat,i}_{q_h^t-1} = t'\} \\
&\leq (1+\frac{1}{H})(1-\eta_{t',h}) \cdot \sum_{t=1}^{T} \frac{1}{e_{q_h^t-1}} \sum_{i=0}^{e_{q_h^t-1}} \mathbb{1}\{\ell^{\flat,i}_{q_h^t-1} = t'\} \\
&\leq (1+\frac{1}{H})^2 (1-\eta_{t',h})
\end{aligned}$$

where the first inequality is from (45) such that,

$$\begin{aligned}
(1-\eta_{\ell^{n_h^t},h}) \leq 1-\eta_{t,h} = 1 - \frac{1}{\sqrt{e_q}+1} = \frac{\sqrt{e_q}}{\sqrt{e_q}+1} &\leq (1+\frac{1}{H}) \frac{\sqrt{e_{q-1}}}{\sqrt{e_{q-1}}+1} \\
&= (1+\frac{1}{H})(1-\eta_{t',h}),
\end{aligned}$$

and the second one follows the tailored choice of stage splitting such that $e_q / e_{q-1} \leq 1 + \frac{1}{H}$.

D Analysis: Gap-dependent Regret Bound

We begin by decomposing the total regret using the suboptimality gaps defined in Assumption 1. Following Yang et al. [21], we obtain:

$$\begin{aligned}
\text{Regret}_T = \sum_{t=1}^{T} (V_1^{\star} - V_1^{\pi^t})(s_1^t) &= \sum_{t=1}^{T} \left(V_1^{\star}(s_1^t) - Q_1^{\star}(s_1^t, a_1^t) + \left(Q_1^{\star} - Q_1^{\pi^t} \right)(s_1^t, a_1^t) \right) \\
&= \sum_{t=1}^{T} \Delta_1(s_1^t, a_1^t) + \sum_{t=1}^{T} \mathbb{E}_{s_2^t \sim P_{1,s_1^t,a_1^t}} \left[\left(V_2^{\star} - V_2^{\pi^t} \right)(s_2^t) \right] \\
&= \ldots = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{h=1}^{H} \Delta_h(s_h^t, a_h^t) \mid a_h^t = \pi_h^t(s_h^t) \right],
\end{aligned}$$

where the expectation is taken with respect to the underlying transition kernel. Before proceeding, we denote $n_h^t = n_h^t(s_h^t, a_h^t)$ and $q_h^t = q_h^t(s_h^t, a_h^t)$ for notational simplicity. Following the notations

used in Appendix C, we define $\alpha_m^0 = \prod_{k=1}^m \frac{n_0+k-1}{H+n_0+k}$ and $\alpha_m^i := \frac{H+1}{H+n_0+i}\prod_{k=i+1}^m \frac{n_0+k-1}{H+n_0+k}$ for any $i \in [m]$ and $m \in \mathbb{N}^\star$, and let $\tilde{\xi}_h^t = \left(\widetilde{V}_h^t - V_h^\star\right)(s_h^t) + H$ and $\tilde{\xi}_h^{\flat,t} = \left(\widetilde{V}_h^{\flat,t} - V_h^\star\right)(s_h^t)$.

We first introduce the following lemma that characterizes the learning error of the Q-functions, which will be used to control the suboptimality gaps. The proof is deferred to Appendix D.1.

Lemma 14. *Let* $\mathcal{E} := \Big\{\forall(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A} : 0 \le \left(Q_h^t - Q_h^\star\right)(s,a) + \eta_{\ell^{n_h^t},h}H \le 2\alpha_{n_h^t}^0 V_{h+1}^0 + \eta_{\ell^{n_h^t},h}\sum_{i=1}^{n_h^t}\alpha_{n_h^t}^i \tilde{\xi}_{h+1}^{\ell^i} + \frac{1-\eta_{\ell^{n_h^t},h}}{1+e_{q_h^t-1}}\sum_{i=0}^{e_{q_h^t-1}}\tilde{\xi}_{h+1}^{\flat,\ell_{q_h^t-1}^{\flat,i}}\mathbb{1}\{q_h^t \ge 1\} + B_h^t + \eta_{\ell^{n_h^t},h}H\Big\}$, *where* $B_h^t \le c_B\left(\sqrt{\frac{H^3(\kappa^\flat)^2\log^3(SAHT^2)}{n_h^t}} + \frac{H^2\kappa^\flat\log^2(SAHT^2)}{n_h^t}\right) + V_{h+1}^0\cdot\mathbb{1}\{q_h^t = 0\}$ *for some universal constant* $c_B > 0$. *The event* $\mathcal{E}$ *holds with probability at least* $1 - 1/T$.

In addition, we define the operator $\mathrm{clip}[x|c] := x\cdot\mathbb{1}\{x \ge c\}$ for some constant $c \ge 0$, which is commonly used in prior work [21, 22, 35]. By Lemma 14, we have $V_h^\star(s_h^t) = \max_a Q_h^\star(s_h^t,a) \le \max_a Q_h^t(s_h^t,a) + \eta_{\ell^{n_h^t},h}H = Q_h^t(s_h^t,a_h^t) + \eta_{\ell^{n_h^t},h}H$ such that

$$\begin{aligned}\Delta_h(s_h^t,a_h^t) &= \mathrm{clip}[V_h^\star(s_h^t) - Q_h^\star(s_h^t,a_h^t) \mid \Delta_{\min}] \\ &\le \mathrm{clip}[\left(Q_h^t - Q_h^\star\right)(s_h^t,a_h^t) + \eta_{\ell^{n_h^t},h}H \mid \Delta_{\min}].\end{aligned}$$

Thus, by definition, the expected total regret can be written as

$$\begin{aligned}\mathbb{E}[\mathrm{Regret}_T] &= \mathbb{P}(\mathcal{E})\cdot\mathbb{E}\left[\sum_{t=1}^T\sum_{h=1}^H \mathrm{clip}[\left(Q_h^t - Q_h^\star\right)(s_h^t,a_h^t) + \eta_{\ell^{n_h^t},h}H \mid \Delta_{\min}] \mid \mathcal{E}\right] \\ &\quad + \mathbb{P}(\mathcal{E}^c)\cdot\mathbb{E}\left[\sum_{t=1}^T\sum_{h=1}^H \mathrm{clip}[\left(Q_h^t - Q_h^\star\right)(s_h^t,a_h^t) + \eta_{\ell^{n_h^t},h}H \mid \Delta_{\min}] \mid \mathcal{E}^c\right] \\ &\le (1-\frac{1}{T})\mathbb{E}\left[\sum_{t=1}^T\sum_{h=1}^H \mathrm{clip}[\left(Q_h^t - Q_h^\star\right)(s_h^t,a_h^t) + \eta_{\ell^{n_h^t},h}H \mid \Delta_{\min}] \mid \mathcal{E}\right] + \frac{2}{T}\cdot TH^2.\end{aligned} \tag{59}$$

Next, we control the first term in (59) by categorizing the suboptimality gaps into different intervals. Specifically, we split the interval $[\Delta_{\min}, H]$ into N disjoint intervals, i.e., $\mathcal{I}_n := [2^{n-1}\Delta_{\min}, 2^n\Delta_{\min}]$ for $n \in [N-1]$ and $\mathcal{I}_N := [2^{N-1}\Delta_{\min}, 3H]$, where $N = \lceil\log_2(3H/\Delta_{\min})\rceil$. Denote the counter of state-action pair for each interval as $C_n := |\left\{(t,h) : \left(\left(Q_h^t - Q_h^\star\right)(s_h^t,a_h^t) + \eta_{\ell^{n_h^t},h}H\right) \in \mathcal{I}_n\right\}|$. We then upper bound (59) as follows:

$$\mathbb{E}[\mathrm{Regret}_T] \le (1-\frac{1}{T})\sum_{n=1}^N 2^n\Delta_{\min}C_n + 2H^2. \tag{60}$$

The following lemma shows that the counter is bounded in each interval, conditioned on event $\mathcal{E}$.

Lemma 15. *Under* $\mathcal{E}$, *we have that for every* $n \in [N]$, $C_n \le O(\frac{H^6SA(\kappa^\flat)^2\log^3(SAHT)}{4^n\Delta_{\min}^2})$.

The proof is postponed to Appendix D.2. Thus, (60) becomes

$$\mathbb{E}[\mathrm{Regret}_T] \le O\left(\frac{H^6SA(\kappa^\flat)^2\log^3(SAHT)}{\Delta_{\min}}\right),$$

Noting that $\kappa^\flat = O\left(\log(SAHT)\right)$, we complete the proof.

D.1 Proof of Lemma 14

We begin by applying Lemma 7 with $\delta = \frac{1}{2T}$, which guarantees that for an ensemble size $J = \lceil c\cdot\log(4SAHT^2)\rceil$, it holds with probability at least $1 - \frac{1}{2T}$ that

$$Q_h^t(s,a) \ge (1-\eta_{\ell^{n_h^t},h})Q_h^\star(s,a) \ge Q_h^\star(s,a) - \eta_{\ell^{n_h^t},h}H,$$

for any $(t,h,s,a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$, where c is a universal positive constant.

Recalling the definition of Q_h^t from (36) and applying Lemma 8 and Lemma 9, we have, again with probability at least $1 - \frac{1}{2T}$,

$$\begin{aligned}
&\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \eta_{\ell^{n_h^t},h} H \\
&= \eta_{\ell^{n_h^t},h} \max_{j\in[J]} \left\{ \widetilde{Q}_h^{j,t}(s_h^t, a_h^t) - Q_h^\star(s_h^t, a_h^t) + H \right\} + (1 - \eta_{\ell^{n_h^t},h}) \left\{ \widetilde{Q}_h^{\flat,t}(s_h^t, a_h^t) - Q_h^\star(s_h^t, a_h^t) \right\} \\
&\le 2\alpha_{n_h^t}^0 V_{h+1}^0 + \eta_{\ell^{n_h^t},h} \sum_{i=1}^{n_h^t} \alpha_{n_h^t}^i \tilde{\xi}_{h+1}^{\ell^i} + \frac{1 - \eta_{\ell^{n_h^t},h}}{1 + e_{q_h^t - 1}} \sum_{i=0}^{e_{q_h^t-1}} \tilde{\xi}_{h+1}^{\flat,\ell^{\flat,i}_{q_h^t-1}} \mathbb{1}\{q_h^t \ge 1\} + B_h^t. \qquad (61)
\end{aligned}$$

Here,

$$\begin{aligned}
B_h^t \le V_{h+1}^0 \cdot \mathbb{1}\{q_h^t = 0\} + c_B \Bigg(&\sqrt{\frac{H^3\kappa^2 \log^3(SAHT^2)}{n_h^t}} + \frac{H^2\kappa \log^2(SAHT^2)}{n_h^t} \\
&+ \sqrt{\frac{H^2(\kappa^\flat)^2 \log^3(SAHT^2)}{e_{q_h^t-1}}} \cdot \mathbb{1}\{q_h^t \ge 1\} + \frac{H\kappa^\flat \log^2(SAHT^2)}{e_{q_h^t-1}} \cdot \mathbb{1}\{q_h^t \ge 1\} \Bigg), \qquad (62)
\end{aligned}$$

where c_B is a positive constant. To simplify the expression in (62), we note

$$\frac{n_h^t}{e_{q_h^t-1}} \le \frac{\sum_{i=0}^{q_h^t} e_i}{e_{q_h^t-1}} = 1 + \frac{\sum_{i=0}^{q_h^t-2} e_i}{e_{q_h^t-1}} + \frac{e_{q_h^t}}{e_{q_h^t-1}} \le 2 + \frac{1}{H} + 4H \le 8H,$$

where the second inequality uses Zheng et al. [22, Lemma D.3] and the choice of the stage length, i.e., $e_{q_h^t} = (1 + \frac{1}{H}) e_{q_h^t - 1}$.

Thus, we could rewrite (62) as

$$B_h^t \le c_B \left(\sqrt{\frac{H^3(\kappa^\flat)^2 \log^3(SAHT^2)}{n_h^t}} + \frac{H^2\kappa^\flat \log^2(SAHT^2)}{n_h^t} \right) + V_{h+1}^0 \cdot \mathbb{1}\{q_h^t = 0\}. \qquad (63)$$

D.2 Proof of Lemma 15

We first partition each interval $\mathcal{I}_n$ according to the step index h. Specifically, for every $n \in [N]$ and $h \in [H]$, we define:

$$\begin{aligned}
w_{n,h}^t &:= \mathbb{1}\left\{ \left(\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + + \eta_{\ell^{n_h^t},h} H \right) \in \mathcal{I}_n \right\}, \quad \forall t \in [T], \qquad (64) \\
C_{n,h} &:= \sum_{t=1}^{T} w_{n,h}^t.
\end{aligned}$$

Note that $w_{n,h}^t \in \{0,1\}$ for all t, since it is an indicator function. By definition, we have $C_n = \sum_{h=1}^{H} C_{n,h}$, and furthermore, for every $n \in [N]$ and $h \in [H]$,

$$2^{n-1} \Delta_{\min} C_{n,h} \le \sum_{t=1}^{T} w_{n,h}^t \left(\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \eta_{\ell^{n_h^t},h} H \right). \qquad (65)$$

To control the right-hand side of (65), we provide the following lemma, which upper bounds the weighted sum. The proof is postponed to Appendix D.3.

Lemma 16. *Let $\eta_{t,h} = 1/e_{q_h^t}$ for every $(t,h) \in [T] \times [H]$. Under event $\mathcal{E}$, for any $h \in [H]$ and $n \in [N]$, the weights $\{w_{n,h}^t\}_{t=1}^T$ defined in (64) satisfy:*

$$\begin{aligned}
\sum_{t=1}^{T} w_{n,h}^t \left(\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \eta_{\ell^{n_h^t},h} H \right) \le O\left(\sqrt{C_{n,h} SAH^5 (\kappa^\flat)^2 \log^3(SAHT)} \right) \\
+ O\left(SAH^3 \kappa^\flat \log(SAHT) \log(1 + C_{n,h}) \right).
\end{aligned}$$

Combining (68) and Lemma 16 with (65), we obtain the following bound:

$$C_{n,h} \leq O\left(\frac{H^5 SA(\kappa^\flat)^2 \log^3(SAHT)}{4^n \Delta_{\min}^2}\right).$$

Summing over $h \in [H]$, we conclude:

$$C_n = \sum_{h=1}^{H} C_{n,h} \leq O\left(\frac{H^6 SA(\kappa^\flat)^2 \log^3(SAHT)}{4^n \Delta_{\min}^2}\right).$$

D.3 Proof of Lemma 16

Under event $\mathcal{E}$, we recall the following upper bound by Lemma 14:

$$\begin{aligned}
&\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \eta_{\ell^{n_h^t},h} H \\
&\leq 2\alpha_{n_h^t}^0 V_{h+1}^0 + \eta_{\ell^{n_h^t},h} \sum_{i=1}^{n_h^t} \alpha_{n_h^t}^i \tilde{\xi}_{h+1}^{\ell^i} + \underbrace{\frac{1-\eta_{\ell^{n_h^t},h}}{1+e_{q_h^t-1}} \sum_{i=0}^{e_{q_h^t-1}} \tilde{\xi}_{h+1}^{\flat,\ell^{\flat,i}_{q_h^t-1}} \mathbb{1}\{q_h^t \geq 1\}}_{T_h^t} + B_h^t + \eta_{\ell^{n_h^t},h} H,
\end{aligned} \tag{66}$$

where

$$B_h^t \leq c_B \underbrace{\sqrt{\frac{H^3\kappa^2 \log^3(8SAHT^2)}{n_h^t}}}_{B_{h,1}^t} + c_B \underbrace{\frac{H^2\kappa^\flat \log(8SAHT^2)}{n_h^t}}_{B_{h,2}^t} + V_{h+1}^0 \cdot \mathbb{1}\{q_h^t = 0\}.$$

Then, our target is to control

$$\begin{aligned}
&\sum_{t=1}^{T} w_{n,h}^t \left(\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \eta_{\ell^{n_h^t},h} H\right) \\
&\leq 2\sum_{t=1}^{T} w_{n,h}^t \alpha_{n_h^t}^0 V_{h+1}^0 + \sum_{t=1}^{T} w_{n,h}^t \eta_{\ell^{n_h^t},h} \sum_{i=1}^{n_h^t} \alpha_{n_h^t}^i \tilde{\xi}_{h+1}^{\ell^i} + \sum_{t=1}^{T} w_{n,h}^t T_h^t + \sum_{t=1}^{T} w_{n,h}^t B_h^t.
\end{aligned} \tag{67}$$

For the first term on the right-hand side of (67), we follow the equation (44) and obtain:

$$2\sum_{t=1}^{T} w_{n,h}^t \alpha_{n_h^t}^0 V_{h+1}^0 \leq \frac{2n_0 SAV_{h+1}^0}{H-1} \leq 4n_0 SA.$$

For the second term on the right-hand side of (67), we follow the equation (46) and obtain:

$$\sum_{t=1}^{T} w_{n,h}^t \eta_{\ell^{n_h^t},h} \sum_{i=1}^{n_h^t} \alpha_{n_h^t}^i \tilde{\xi}_{h+1}^{\ell^i} \lesssim H \sum_{t=1}^{T} \eta_{\ell^{n_h^t},h} \lesssim H \sum_{s,a} \sum_{q=1}^{Q} \frac{e_{q_h^t}}{e_{q_h^t}} \lesssim SAH^2 \log(T), \tag{68}$$

where the last inequality uses the fact $Q \leq 4H\log(T)$.

For the third term on the right-hand side of (67), we apply the similar arguments in (47) and Appendix C.6, Then, we have

$$\begin{aligned}
\sum_{t=1}^{T} w_{n,h}^t T_h^t &\leq \sum_{t=1}^{T} w_{n,h}^t (1-\eta_{\ell^{n_h^t},h}) \sum_{i=0}^{e_{q_h^t-1}} \frac{\mathbb{1}\{q_h^t \geq 1\}}{1+e_{q_h^t-1}} \tilde{\xi}_{h+1}^{\flat,\ell^{\flat,i}_{q_h^t-1}} \\
&\leq (1+\frac{1}{H}) \sum_{t'=1}^{T} (1-\eta_{t',h}) \tilde{\xi}_{h+1}^{\flat,t'} \cdot \sum_{t=1}^{T} \frac{w_{n,h}^t}{1+e_{q_h^t-1}} \sum_{i=0}^{e_{q_h^t-1}} \mathbb{1}\{\ell^{\flat,i}_{q_h^t-1} = t'\} \\
&= (1+\frac{1}{H}) \sum_{t'=1}^{T} (1-\eta_{t',h}) \tilde{\xi}_{h+1}^{\flat,t'} \cdot \bar{w}_1^{t'}
\end{aligned}$$

where $\bar{w}_1^{t'} = \sum_{t=1}^{T} \frac{w_{n,h}^t}{1+e_{q_h^t-1}} \sum_{i=0}^{e_{q_h^t-1}} \mathbb{1}\{\ell^{\flat,i}_{q_h^t-1} = t'\}$ for every $t' \in [T]$. Note that $\{\bar{w}_1^t\}$ is some sequence satisfying $0 \le \bar{w}_1^t \le (1+\frac{1}{H}), \forall t \in [T]$ and $\sum_{t=1}^{T} \bar{w}_1^t = C_{n,h}$, following the similar arguments in Lemma 4.3 in [21]. From (49) and (68), we further have

$$\sum_{t=1}^{T} w_{n,h}^t T_h^t \lesssim (1+\frac{1}{H}) \sum_{t=1}^{T} \left(\bar{w}_1^t \xi_{h+1}^t + \eta_{\ell^{n_{h+1}^t},h+1} H \right) + SAH^2 \log(T)$$

Note that $V_{h+1}^\star(s_{h+1}^t) \ge Q_{h+1}^\star(s_{h+1}^t, a_{h+1}^t)$ by the definition of the optimal policy in (1), and $V_{h+1}^t(s_{h+1}^t) = Q_{h+1}^t(s_{h+1}^t, a_{h+1}^t)$ by (38). Thus,

$$\sum_{t=1}^{T} w_{n,h}^t T_h^t \lesssim (1+\frac{1}{H}) \sum_{t=1}^{T} \bar{w}_1^t \left(\left(Q_{h+1}^t - Q_{h+1}^\star\right)(s_{h+1}^t, a_{h+1}^t) + \eta_{\ell^{n_{h+1}^t},h+1} H \right) + SAH^2 \log(T)$$

We then consider the bound of $\sum_{t=1}^{T} \bar{w}_i^t B_h^t$ for any sequence $\{\bar{w}_i^t\}$ which satisfies $0 \le \bar{w}_i^t \le (1+\frac{1}{H})^i, \forall t \in [T]$ and $\sum_{t=1}^{T} \bar{w}_i^t = C_{n,h}$ for any $i \in \{0, \dots, H-h\}$. Before proceeding, we first define $C_{n,h,s,a} := \sum_{m=1}^{n_h^T(s,a)} \bar{w}_i^{\ell^m}$. Similar to the steps in Appendix C, we bound $\sum_{t=1}^{T} \bar{w}_i^t B_h^t$ by controlling each term separately. To begin with,

$$\begin{aligned}
\sum_{t=1}^{T} \bar{w}_i^t B_{h,1}^t &\le \sqrt{H^3(\kappa^\flat)^2 \log^3(8SAHT^2)} \sum_{t=1}^{T} \frac{\bar{w}_i^t}{\sqrt{n_h^t}} \\
&\lesssim \sqrt{H^3(\kappa^\flat)^2 \log^3(SAHT^2)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m=1}^{n_h^T(s,a)} \frac{\bar{w}_i^{\ell^m}}{\sqrt{m}} \\
&\lesssim \sqrt{H^3(\kappa^\flat)^2 \log^3(SAHT^2)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m=1}^{\lceil C_{n,h,s,a}/(1+1/H)^i \rceil} \frac{(1+1/H)^i}{\sqrt{m}} \\
&\lesssim \sqrt{H^3(\kappa^\flat)^2 \log^3(SAHT^2)} \sqrt{SA \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} C_{n,h,s,a} \cdot (1+1/H)^i} \\
&\lesssim \sqrt{C_{n,h} SAH^3(\kappa^\flat)^2 (1+1/H)^i \log^3(SAHT^2)},
\end{aligned} \tag{69}$$

where the third inequality holds since the left-hand side is maximized when the first $\lceil C_{n,h,s,a}/(1+1/H)^i \rceil$ occupy the smallest indices, the penultimate line uses Cauchy-Schwarz inequality, and the last line is due to the fact that $\sum_{s,a} C_{n,h,s,a} \le C_{n,h}$.

Similarly,

$$\begin{aligned}
\sum_{t=1}^{T} \bar{w}_i^t B_{h,2}^t &\le \sum_{t=1}^{T} \bar{w}_i^t \frac{H^2 \kappa^\flat \log\big(8SAHT^2\big)}{n_h^t} \\
&\lesssim H^2 \kappa^\flat \log\big(SAHT^2\big) \sum_{t=1}^{T} \frac{\bar{w}_i^t}{n_h^t} \\
&\lesssim H^2 \kappa^\flat \log\big(SAHT^2\big) \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m=1}^{\lceil C_{n,h,s,a}/(1+1/H)^i \rceil} \frac{(1+1/H)^i}{m} \\
&\lesssim H^2 \kappa^\flat \log\big(SAHT^2\big)(1+1/H)^i \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \log(1+C_{n,h,s,a}) \\
&\lesssim SAH^2 \kappa^\flat \log\big(SAHT^2\big)(1+1/H)^i \log(1+C_{n,h}).
\end{aligned} \tag{70}$$

Also, $\sum_{t=1}^{T} \bar{w}_i^t V_{h+1}^0 \cdot \mathbb{1}\{q_h^t = 0\} \le SAH^2(1+1/H)^i$, as there are at most SA state-action paris and each can be visited at most $e_0 = H$ times in stage 0.

Combining with (69) and (70), one has

$$\sum_{t=1}^{T} \bar{w}_i^t B_h^t \lesssim \sqrt{C_{n,h} SAH^3 (\kappa^\flat)^2 (1+1/H)^i \log^3(SAHT^2)} + SAH^2 \kappa^\flat \log\big(SAHT^2\big)(1+1/H)^i \log(1+C_{n,h}) + SAH^2(1+1/H)^i \tag{71}$$

Observe that $\{w_{n,h}^t\}$ directly satisfies $0 \le w_{n,h}^t \le 1$ and $\sum_{t=1}^{T} w_{n,h}^t = C_{n,h}$. Thus, we can apply (71) with $i = 0$ and obtain

$$\begin{aligned}
&\sum_{t=1}^{T} w_{n,h}^t \left(\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \eta_{\ell^{n_h^t},h} H \right) \\
&\lesssim \sqrt{C_{n,h} SAH^3 (\kappa^\flat)^2 \log^3(SAHT^2)} + SAH^2 \kappa^\flat \log\big(SAHT^2\big) \log(1+C_{n,h}) \\
&\quad + (1+\frac{1}{H}) \sum_{t=1}^{T} \bar{w}_1^t \left(\left(Q_{h+1}^t - Q_{h+1}^\star\right)(s_{h+1}^t, a_{h+1}^t) + \eta_{\ell^{n_{h+1}^t},h+1} H \right).
\end{aligned}$$

We now recursively unroll the Q-value difference over future steps and utilize the fact that $(1+\frac{1}{H})^H \le \mathrm{e}$, yielding:

$$\begin{aligned}
&\sum_{t=1}^{T} w_{n,h}^t \left(\left(Q_h^t - Q_h^\star\right)(s_h^t, a_h^t) + \eta_{\ell^{n_h^t},h} H \right) \\
&\lesssim \sum_{i=0}^{H-h} \left((1+\frac{1}{H})^{2i} \sqrt{C_{n,h} SAH^3 (\kappa^\flat)^2 \log^3(SAHT^2)} \right. \\
&\qquad \left. + SAH^2 \kappa^\flat \log\big(SAHT^2\big)(1+\frac{1}{H})^{2i} \log(1+C_{n,h}) \right) \\
&\lesssim H \left(\sqrt{C_{n,h} SAH^3 (\kappa^\flat)^2 \log^3(SAHT^2)} + SAH^2 \kappa^\flat \log\big(SAHT^2\big) \log(1+C_{n,h}) \right),
\end{aligned}$$

which completes the proof.