Unified Framework of Distributional Regret in Multi-Armed Bandits and Reinforcement Learning

Harin Lee LEEHARIN@CS.WASHINGTON.EDU
University of Washington
Min-hwan Oh MINOH@SNU.AC.KR
Seoul National University

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the distribution of regret in stochastic multi-armed bandits and episodic reinforcement learning through a unified framework. We formalize a *distributional regret bound* as a probabilistic guarantee that holds *uniformly* over all confidence levels $\delta \in (0, 1]$, thereby characterizing the regret distribution across the full range of δ. We present a simple UCBVI-style algorithm with exploration bonus $\min\{c_{1,k}/N,\ c_{2,k}/\sqrt{N}\}$, where N denotes the visit count and $(c_{1,k}, c_{2,k})$ are user-specified parameters. For arbitrary parameter sequences, we derive general gap-independent and gap-dependent distributional regret bounds, yielding a principled characterization of how the parameters control the trade-off between expected performance, tail risk, and instance-dependent behavior. In particular, our bounds achieve optimal trade-offs between expected and distributional regret in both minimax and instance-dependent regimes. As a special case, for multi-armed bandits with A arms and horizon T, we obtain a distributional regret bound of order $\mathcal{O}\big(\sqrt{AT}\log(1/\delta)\big)$, confirming the conjecture of Lattimore and Szepesvári (2020, Section 17.1) for the first time.

Keywords: distributional regret, multi-armed bandit, reinforcement learning

1. Introduction

In online decision-making problems—including stochastic multi-armed bandits (MAB) (Lattimore and Szepesvári, 2020) and reinforcement learning (RL) (Sutton and Barto, 2018)—an agent repeatedly interacts with an unknown environment by selecting actions and observing stochastic rewards. A standard measure of performance is *regret*, defined as the gap between the cumulative reward achieved by an optimal strategy and that obtained by the agent. Since the environment is stochastic and the agent may employ a randomized action-selection strategy, regret is a random variable; consequently, performance is typically assessed via expected regret or via high-probability guarantees at a prescribed confidence level.

While expectation and fixed-confidence bounds are useful quantities, they do not fully describe the distribution of regret. Recent work has highlighted that tail behavior can be subtle even in classical settings: Fan and Glynn (2025) show that asymptotically optimal bandit algorithms can exhibit heavy-tailed regret distributions of $\mathbb{P}(\mathrm{Regret} > x) \approx \frac{1}{x}$. Simchi-Levi et al. (2023a, 2025) propose bonus designs that accelerates the decay of the tail to $\mathcal{O}(\exp(-x^\gamma))$ for a tunable parameter $\gamma \in (0, 1]$ and further investigate the trade-off between the tail distribution and the instance-dependent regret. For RL, the distributional picture is considerably less complete: while a recent work (Khodadadian and Moharrami, 2025) studies a notion of distributional regret in RL, near-optimal characterizations and the corresponding optimal trade-offs remain unknown. Even in MAB, the optimality of existing distributional guarantees is not fully resolved.

In this work, we provide a unified framework of distributional regret analysis for MAB and RL, and we establish regret bounds that meet existing lower bounds in multiple regimes.

Main Contribution. We propose a simple and flexible algorithm for MAB and RL, `EQO+`, and analyze the distribution of its cumulative regret. The algorithm is a `UCBVI`-type method with a bonus term of $\frac{c_{1,k}}{N^k(s,a)} \wedge \frac{c_{2,k}}{\sqrt{N^k(s,a)}}$, where $c_{1,k}$ and $c_{2,k}$ are input parameters and $N^k(s,a)$ is the visit count of the state-action pair (s,a). Our theoretical guarantees offer the following novelties.

- We study a distributional regret bound, defined as a function of δ that upper-bounds the cumulative regret with probability at least $1-\delta$ *simultaneously for all* $\delta \in (0,1]$ (see Section 3). The resulting δ-dependence directly captures the distributional properties of the regret. A distributional regret bound implies (i) high-probability bounds at any prescribed confidence level, (ii) an expected regret bound via integration over δ, and (iii) the light-tailed risk notion of Simchi-Levi et al. (2025). In particular, converting our uniform-in-δ bounds to expectation avoids the extra $\log K$ factors that commonly arise when one derives expected regret, where K is the number of episodes (see Corollary 11).
- We introduce a regularity assumption that bounds the sub-exponential norm of the reward and the optimal value of the next state. This assumption strictly generalizes (and is not limited to) the standard sub-Gaussian noise assumption in stochastic MAB and the bounded-reward assumption in RL, enabling a single unified framework for regret analysis that covers both settings (see Section 4).
- We establish both gap-independent and gap-dependent distributional regret bounds for arbitrary input parameters $\{c_{1,k}\}_{k=1}^{\infty}, \{c_{2,k}\}_{k=1}^{\infty}$ (Theorems 8 and 9). Our results rigorously characterize how these two bonus parameters balance the trade-offs between the expected bound and the tail distribution of worst-case and instance-dependent regret, and we show that our analyses achieve optimal trade-offs in both.
- In the MAB setting with A arms and horizon T, we obtain a distributional regret bound of $\mathcal{O}(\sqrt{AT}\log(1/\delta))$ together with an expected regret bound of $\mathcal{O}(\sqrt{AT})$ (Theorem 4), matching minimax lower bounds up to constant factors. To the best of our knowledge, this is the tightest known regret guarantee for MAB, and it confirms the conjecture of Lattimore and Szepesvári (2020, Section 17.1).

1.1. Related Works

MAB and RL with Expected and High-probability Regret Guarantees. Near-optimal expected and high-probability regret bounds have been established in numerous works for MAB (Auer et al., 2002; Audibert and Bubeck, 2009; Agrawal and Goyal, 2012; Bubeck et al., 2012; Degenne and Perchet, 2016; Ménard and Garivier, 2017; Lattimore, 2018; Lattimore and Szepesvári, 2020; Jin et al., 2023) and episodic RL (Azar et al., 2017; Zanette and Brunskill, 2019; Simchowitz and Jamieson, 2019; Dann et al., 2021; Zhang et al., 2021; Tiapkin et al., 2022; Zhou et al., 2023; Zhang et al., 2024; Lee and Oh, 2025). However, the distributional behavior of regret has been far less studied.

Comparison with Lee and Oh (2025). The work most closely related to ours is Lee and Oh (2025), and our results generalize the analysis of Lee and Oh (2025). The primary focus of Lee and Oh (2025) is to achieve a minimax optimal high-probability regret bound under fixed input parameters. We generalize and improve their analysis framework, providing distributional regret bounds and instance-dependent bounds under arbitrary input parameters.

Distributional Regret Bound. Neu (2015) shows that in adversarial bandits, `EXP3-IX` (Kocák et al., 2014) achieves a high-probability regret bound of $\mathcal{O}\big(\sqrt{AT}\big(\sqrt{\log A}+\frac{\log(1/\delta)}{\sqrt{\log A}}\big)\big)$ uniformly for all $\delta \in (0,1]$. Simchi-Levi et al. (2022, 2023b) and their extensions (Simchi-Levi et al., 2025, 2023a) study stochastic MAB. These works design specific bonus terms for UCB-type algorithms and upper-bound $\mathbb{P}(\text{Regret} > x)$, yielding nearly exponential decay. Their analyses extend to instance-dependent bounds and provide lower bounds for the trade-off between the regret distribution and expected regret. Zhu and Simchi-Levi (2025) apply similar ideas to Thompson sampling, obtaining exponential decay for regret and error rates in best-arm identification. However, these approaches typically incur additional logarithmic factors in expected regret and offer limited flexibility in balancing distributional and expected guarantees; moreover, their analyses are based on an independent framework disabling unified analysis. Khodadadian and Moharrami (2025) extend related ideas to RL, but their bound appears loose, incurring both $\text{gap}_{\min}^{-1}$ and $\sqrt{K}$ terms simultaneously (where K is the number of episodes), as well as a particularly large $\mathcal{O}(H^6)$ dependence on the horizon length H.

Comparison with Simchi-Levi et al. (2023a, 2025). Our framework and results generalize those of Simchi-Levi et al. (2023a, 2025), most notably by extending the problem setting from MAB to episodic RL. While our algorithm shares structural similarities with the bonus design in Simchi-Levi et al. (2023a) (and also Lee and Oh (2025)), our regret analysis is substantially different. Moreover, our theory accommodates arbitrary input parameter sequences, whereas the guarantees in Simchi-Levi et al. (2023a, 2025) are derived for specific parameter choices. Although a recent preprint version of Simchi-Levi et al. (2023a) relaxes the constraint on $c_{2,k}$, the other parameter $c_{1,k}$ still remains less flexible. Finally, in the MAB setting, our bounds sharpen the guarantees in these prior works.

2. Preliminaries

2.1. Markov Decision Process

We consider an episodic Markov decision (MDP) process $M=(\mathcal{S},\mathcal{A},P,r,H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ is the transition probability, $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is the reward function, and H is the horizon length in each episode. We consider the tabular case, where $\mathcal{S}$ and $\mathcal{A}$ have finite cardinalities of S and A, respectively. The agent interacts with an MDP for iterations of episodes, where each episode consists of H time steps. At the beginning of the k-th episode, the environment chooses the initial state s_1^k, possibly adaptively. For time steps $h=1,\dots,H$, the agent observes the current state s_h^k, takes an action a_h^k, and receives a stochastic reward of $R_h^k\in\mathbb{R}$ with mean $r(s_h^k,a_h^k)$. Then, the next state s_{h+1}^k is sampled from $P(s_h^k,a_h^k)$.
A policy $\pi=\{\pi_h\}$ is a sequence of mappings $\pi_h:\mathcal{S}\to\mathcal{A}$ from the current state to an action. We denote the set of all deterministic policies by Π. For given policy π and $h\in[H]$, we define the value function as $V_h^\pi(s):=\mathbb{E}_{\pi(\cdot|s_h=s)}[\sum_{j=h}^H r(s_j,a_j)]$, the expectation is taken over the trajectory starting from $s_h=s$ with the j-th action being $a_j=\pi_j(s_j)$. Similarly, we define the action value function $Q_h^\pi(s,a):=\mathbb{E}_{\pi(\cdot|s_h=s,a_h=a)}[\sum_{j=h}^H r(s_j,a_j)]$. The optimal value function is defined as $V_h^*(s):=\max_{\pi\in\Pi}V_h^\pi(s)$ and $Q_h^*(s,a):=\max_{\pi\in\Pi}Q_h^\pi(s,a)$. The optimal policy π^* is defined as the policy that satisfies $V_h^{\pi^*}(s)=V_h^*(s)$ for all $h\in[H]$ and $s\in\mathcal{S}$. Given an algorithm Alg that chooses $\pi^1,\pi^2,\dots,\pi^k,\dots$ based on the prior observations, the cumulative regret of Alg in M over K episodes is defined as $\text{Reg}_M^{\mathsf{Alg}}(K):=\sum_{k=1}^K\big(V_1^*(s_1^k)-V_1^{\pi^k}(s_1^k)\big)$.

2.2. Multi-Armed Bandits

We consider MAB instances as a special case of episodic MDPs with $H = S = 1$ and no transitions. When focusing on the MAB setting, we denote the time steps and the horizon by t and T, respectively. The interaction is simplified to choosing an action $a_t \in \mathcal{A}$ (with $|\mathcal{A}| = A$) and observing a reward R_t with mean $r(a_t)$ for time steps $t = 1, 2, \ldots$. We define the optimal action $a^* = \operatorname{argmax}_{a \in \mathcal{A}} r(a)$ as the action with the highest reward. The cumulative regret of Alg in M over T time steps can be written as $\mathrm{Reg}_M^{\mathsf{Alg}}(T) := \sum_{t=1}^{T}(r(a^*) - r(a_t))$.

2.3. Definitions and Notations

For functions $\widetilde{P} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\mathcal{S}}$ and $V : \mathcal{S} \to \mathbb{R}$, we denote the expectation of V under $\widetilde{P}(s,a)$ by $\widetilde{P}V(s,a) := \sum_{s' \in \mathcal{S}} \widetilde{P}(s' \mid s,a)V(s')$. We denote the variance of V under the true transition probability $P(s,a)$ by $\mathrm{Var}(V)(s,a) := \sum_{s' \in \mathcal{S}} P(s' \mid s,a)(V(s') - PV(s,a))^2$. We define a filtration $\{\mathcal{F}_h^k\}_{k,h}$ as $\mathcal{F}_h^k := \sigma(s_1^1, a_1^1, R_1^1, \ldots, s_h^k, a_h^k)$. Note that $\mathcal{F}_{H+1}^k = \mathcal{F}_0^{k+1} := \sigma(s_1^1, a_1^1, R_1^1, \ldots, s_H^k, a_H^k, r_H^k, s_{H+1}^k)$. In the MAB setting, the gap of an action is defined as $\mathrm{gap}(a) := r(a^*) - r(a)$. In the RL setting, the gap of a state-action pair at time step h is defined as $\mathrm{gap}_h(s,a) := V_h^*(s) - Q_h^*(s,a)$, where we denote $\mathrm{gap}(s,a) := \min_{h \in [H]} \mathrm{gap}_h(s,a)$. In addition, we denote the minimum non-zero gap by $\mathrm{gap}_{\min} := \min_{(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}, \mathrm{gap}_h(s,a) > 0} \mathrm{gap}_h(s,a)$. For a natural number $N \in \mathbb{N}$, let $[N] := \{1, 2, \ldots, N\}$. For two real numbers a and b, we define $a \vee b$ as $\max\{a, b\}$ and $a \wedge b$ as $\min\{a, b\}$. We also write $(a)_+$ for $a \vee 0$.

3. Distributional Regret

In this section, we define *distributional regret*, which corresponds to the upper-quantile function (equivalently, the inverse complementary cumulative distribution function) of the regret.

Definition 1 (Distributional regret) *A **distributional regret** $\mathrm{Reg}_M^{\mathsf{Alg}}(K, \delta)$ is a deterministic real-valued function of an algorithm Alg, an MDP M, the number of episodes $K \in \mathbb{N}$, and a failure probability $\delta \in (0, 1]$, defined as $\mathrm{Reg}_M^{\mathsf{Alg}}(K, \delta) := \inf\{x \in \mathbb{R} \mid \mathbb{P}(\mathrm{Reg}_M^{\mathsf{Alg}}(K) > x) \le \delta\}$.*

Distributional regret bound. Equivalently, for any $\delta \in (0, 1]$, the regret satisfies $\mathrm{Reg}_M^{\mathsf{Alg}}(K) \le \mathrm{Reg}_M^{\mathsf{Alg}}(K, \delta)$ with probability at least $1 - \delta$, and $\mathrm{Reg}_M^{\mathsf{Alg}}(K, \delta)$ is the smallest value with this property. Our goal is to provide an explicit upper bound on $\mathrm{Reg}_M^{\mathsf{Alg}}(K, \delta)$ as a function of the same inputs; we refer to such an upper bound as a *distributional regret bound*.

Generality of distributional regret. The δ-dependence of a distributional regret bound captures the tail distribution of regret at any given level δ. For instance, light-tailed risk defined in Simchi-Levi et al. (2025) translates to $\mathrm{Reg}_M^{\mathsf{Alg}}(K, \delta) = \mathrm{poly}\log(\frac{1}{\delta})$. A distributional regret bound implies high-probability bounds for any failure probability. It also implies an expected regret bound via the identity $\mathbb{E}[\mathrm{Reg}_M^{\mathsf{Alg}}(K)] = \int_0^1 \mathrm{Reg}_M^{\mathsf{Alg}}(K, \delta)\, d\delta$. This integration technique replaces $\log \frac{1}{\delta}$ factors by constant factors in the expected regret bound, which shaves off the $\log K$ factor that typically arises in high-probability-based approaches where δ is often set as $\delta = \frac{1}{K}$. To the best of our knowledge, Corollary 11 achieves the sharpest logarithmic factor in the worst-case expected regret bound for RL using this technique.

4. Assumptions

In this section, we present the assumptions for our analysis. They encompass the standard sub-Gaussian noise assumption from the MAB literature and the bounded reward assumption standard in RL. The first assumption regularizes the scale of the value function.

Assumption 1 (Boundedness) *There exists a known value $V_{\max} \geq 0$ such that for all $h \in [H]$, $s \in \mathcal{S}$, and policy $\pi \in \Pi$, we have $0 \leq V_h^\pi(s) \leq V_{\max}$.*

The second assumption is for the concentration of measure. The standard assumptions in recent MAB and RL literature are slightly different. The bounded reward assumption in the RL literature assumes $\sum_{h=1}^H R_h \in [0, V_{\max}]$ (Zanette and Brunskill, 2019) or $R_h \in [0, V_{\max}]$ (Lee and Oh, 2025), which allows leveraging the reward variance via Bernstein's inequality, and simultaneously guarantees that the sum of the variances over a trajectory is at most $V_{\max}^2$ without H dependence, incentivizing the use of variances. However, this property is difficult to capture with a uniform sub-Gaussian assumption, making the two frameworks seemingly incompatible. Addressing this issue, we propose a unified setting that assumes bounded sub-exponential norms on the random variables $R_h^k + V_{h+1}^*(s_{h+1}^k)$, whose means are $Q_h^*(s_h^k, a_h^k)$. We first provide the definition of sub-exponential random variables, and then formally present the assumption.

Definition 2 (Sub-exponential random variable (Wainwright, 2019)) *Let X be a random variable, $\mathcal{F}$ be a σ-algebra, and $\sigma, \alpha \geq 0$ be $\mathcal{F}$-measurable random variables. X is **$\mathcal{F}$-conditionally (σ, α)-sub-exponential** if $\mathbb{E}[\exp(\lambda X) \mid \mathcal{F}] \leq \exp(\frac{\sigma^2\lambda^2}{2})$ holds almost surely for all $\lambda \in [-\frac{1}{\alpha}, \frac{1}{\alpha}]$. If $\alpha = 0$, we assume $[-\frac{1}{0}, \frac{1}{0}] := \mathbb{R}$, and we say X is **$\mathcal{F}$-conditionally σ^2-sub-Gaussian**.*

Assumption 2 (Conditional sub-exponentiality) *There exist a variance proxy function $\sigma_{\exp} : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$, and a constant V_α unknown to the agent such that $R_h^k + V_{h+1}^*(s_{h+1}^k) - Q^*(s_h^k, a_h^k)$ is $\mathcal{F}_h^k$-conditionally $(\sigma_{\exp}(h, s_h^k, a_h^k), V_\alpha)$-sub-exponential.*

The sub-exponential norm of a bounded random variable is proportional to its variance when α is set to the range of the random variable (see Lemma 50). Hence, under the bounded reward assumption, $\sigma_{\exp}^2(h, s, a)$ can be set as the variance of $R_h^k + V_{h+1}^*$ with V_α proportional to $V_{\max}$. When the reward noise is independently σ^2-sub-Gaussian, $\sigma_{\exp}^2(h, s, a)$ can be set as $\sigma^2 + 2\operatorname{Var}(V_{h+1}^*)(s, a)$. Furthermore, our assumption extends to sub-exponential reward noise, such as the exponential, chi-squared, and Poisson distributions. Analogous to cumulative variance in RL, we define the sum of variance proxies.

Definition 3 (Total variance proxy function) *For $h \in [H]$, $s \in \mathcal{S}$, and $\pi \in \Pi$, the **total variance proxy function** is defined as $W_h^\pi(s) := \mathbb{E}_{\pi(\cdot|s_h=s)}[\sum_{j=h}^H \frac{1}{2}\sigma_{\exp}^2(j, s_j, a_j)]$.*

We further define additional notations related to this function. We denote the maximum of the function by $\mathbb{W}^* := \max_{h\in[H], s\in\mathcal{S}, \pi\in\Pi} W_h^\pi(s)$. Since we also consider the range of W_h^π, we define the minimum of the function $W_{h,\min}^\pi := \min_{s\in\mathcal{S}} W_h^\pi(s)$ for fixed $h \in [H]$ and $\pi \in \Pi$, the difference from the minimum $W_{h,\text{diff}}^\pi(s) := W_h^\pi(s) - W_{h,\min}^\pi$, and the maximum range of the function $\mathbb{W}_{\text{diff}}^* := \max_{h\in[H], s\in\mathcal{S}, \pi\in\Pi} W_{h,\text{diff}}^\pi(s)$. Table 1 provides example bounds for $\sigma_{\exp}^2(h, s, a)$, V_α, $\mathbb{W}^*$, and $\mathbb{W}_{\text{diff}}^*$ under the conventional settings. Their derivation is presented in Appendix B. We additionally define a constant $\sigma_{\max} := \max_{h\in[H],(s,a)\in\mathcal{S}\times\mathcal{A}} 2\sigma_{\exp}(h, s, a) \vee \sqrt{2V_\alpha V_{\max}}$ that serves as a threshold for $c_{2,k}$, which arises from Lemma 52. Note that we do not assume the agent knows these values.

Table 1: Examples of upper bounds on the values σ^2_{exp}, V_α, $\mathbb{W}^*$, and $\mathbb{W}^*_{\text{diff}}$.

Setting	$\sigma^2_{\text{exp}}(h,s,a)$	V_α	$\mathbb{W}^*$	$\mathbb{W}^*_{\text{diff}}$
$0 \le R_h \le V_{\max}$	$2\,\text{Var}(R_h + V^*_{h+1})(s,a)$	$2V_{\max}$	$2V^2_{\max}$	$2V^2_{\max}$
R_h σ^2-sub-Gaussian	$\sigma^2 + 2\,\text{Var}(V^*_{h+1})(s,a)$	$V_{\max}$	$\sigma^2 H + V^2_{\max}$	$V^2_{\max}$
R_h σ^2-sub-Gaussian, MAB	σ^2	0	σ^2	0

5. Algorithm: EQO+

We introduce EQO+ algorithm with a flexible bonus term, taking a sequence $\{(c_{1,k}, c_{2,k})\}_{k=1}^\infty$ as input and using the bonus term as $\frac{c_{1,k}}{N^k(s,a)} \wedge \frac{c_{2,k}}{\sqrt{N^k(s,a)}}$, where $N^k(s,a)$ is the visit count of the state-action pair. Algorithm 1 describes the specific procedure for RL. Its simpler version for MAB is analogous, where Lines 4, 6, 7, and 11 are unnecessary. We provide a refined description for MAB in Appendix C. EQO+ is an extension of EQO in Lee and Oh (2025), where their algorithm uses the $\frac{c_{1,k}}{N^k(s,a)}$ term only. We recover EQO by setting $c_{2,k} = \infty$, which disables the $\frac{1}{\sqrt{N}}$-bonus term.

Algorithm 1 EQO+ (Exploration via Quasi-Optimism Plus)

Input: $\{(c_{1,k}, c_{2,k})\}_{k=1}^\infty$

1 **for** $k = 1, 2, \ldots,$ **do**

2 $\quad N^k(s,a) \leftarrow \sum_{i=1}^{k-1} \sum_{h=1}^H \mathbb{1}\{(s^i_h, a^i_h) = (s,a)\}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

3 $\quad \hat{r}^k(s,a) \leftarrow \frac{1}{N^k(s,a)} \sum_{i=1}^{k-1} \sum_{h=1}^H R^i_h \mathbb{1}\{(s^i_h, a^i_h) = (s,a)\}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

4 $\quad \hat{P}^k(s'|s,a) \leftarrow \frac{1}{N^k(s,a)} \sum_{i=1}^{k-1} \sum_{h=1}^H \mathbb{1}\{(s^i_h, a^i_h, s^i_{h+1}) = (s,a,s')\}, \forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$

5 $\quad b^k(s,a) \leftarrow \frac{c_{1,k}}{N^k(s,a)} \wedge \frac{c_{2,k}}{\sqrt{N^k(s,a)}}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

6 $\quad V^k_{H+1}(s) \leftarrow 0$ for all $s \in \mathcal{S}$

7 $\quad$ **for** $h = H, H-1, \ldots, 1$ **do**

8 $\quad\quad$ **foreach** $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

9 $\quad\quad\quad Q^k_h(s,a) \leftarrow \begin{cases} \left(\left(\hat{r}^k(s,a) + \hat{P}^k V^k_{h+1}(s,a)\right)_+ + b^k(s,a)\right) \wedge V_{\max} & \text{if } N^k(s,a) > 0 \\ V_{\max} & \text{if } N^k(s,a) = 0 \end{cases}$

10 $\quad\quad$ **end**

11 $\quad\quad V^k_h(s) \leftarrow \max_{a \in \mathcal{A}} Q^k_h(s,a)$ for all $s \in \mathcal{S}$

12 $\quad\quad \pi^k_h(s) \leftarrow \text{argmax}_{a \in \mathcal{A}} Q^k_h(s,a)$ for all $s \in \mathcal{S}$

13 $\quad$ **end**

14 $\quad$ Execute π^k and obtain $(s^k_1, a^k_1, R^k_1, \ldots, s^k_H, a^k_H, R^k_H, s^k_{H+1})$

15 **end**

Generality of EQO+ algorithm. EQO+ recovers several existing algorithms through specific choices of the input parameters. When $c_{1,k} = \widetilde{\mathcal{O}}\big(V_{\max}(\frac{k}{SA} \log \frac{1}{\delta_0})^{1/2}\big)$ and $c_{2,k} = \infty$ for some specific $\delta_0 \in (0,1]$, we recover the bonus term considered in Lee and Oh (2025). When $c_{1,k} = \infty$ and $c_{2,k} = \widetilde{\mathcal{O}}\big(V_{\max}(\log \frac{1}{\delta_0})^{1/2}\big)$, we obtain UCBVI-CH in Azar et al. (2017), whose confidence bound is also known as the Hoeffding-style bound. When $c_{1,k} = \infty$ and $c_{2,k} = \mathcal{O}(S + K^\alpha)$ for some $\alpha \in [0,1]$, we have the algorithm by Khodadadian and Moharrami (2025). In the MAB setting,

assigning $c_{1,t} = \mathcal{O}\big((\frac{t}{A})^{\alpha}\sqrt{\log A}\big)$ and $c_{2,t} = \sqrt{t^{\beta}}$ for some constants $0 < \beta \leq \alpha < 1$ recovers the bonus term in Simchi-Levi et al. (2023a).

6. Main Results

In this section, we present distributional regret bounds for EQO+. In Section 6.1, we first present key results for the MAB setting, which also illustrate the role of the input parameters. Then, in Section 6.2, we turn to general distributional regret bounds for the RL setting, which also apply to the MAB setting, and allow arbitrary input parameters. In Section 6.3, we consider several representative parameter choices and discuss their implications.

Throughout this section, the algorithm under consideration is $\mathsf{Alg} := \texttt{EQO+}(\{(c_{1,k}, c_{2,k})\}_{k=1}^{\infty})$, where the values of $\{(c_{1,k}, c_{2,k})\}_{k=1}^{\infty}$ are specified in each theorem.

6.1. Distributional Regret Bounds for Multi-Armed Bandits

In this section, we provide distributional regret bounds for the MAB setting. Let $\mathcal{B}(\sigma)$ denote the set of MAB instances with σ^2-sub-Gaussian reward noise and $V_{\max} = 1$. We present two special cases of input parameters that provide insights into their roles, which carry over to the RL setting. The first result shows how the parameter $c_{1,t}$ controls the minimax regret. Extended theorems and proofs are provided in Appendix C.

Theorem 4 *Suppose $M \in \mathcal{B}(\sigma)$. Set $c_{1,t} = c_1$ for a constant $c_1 > 0$ and $c_{2,t} = \infty$, meaning that we use bonus term $\frac{c_1}{N}$ only. Then, the distributional regret of EQO+ is bounded as*

$$\mathrm{Reg}_M^{\mathsf{Alg}}(T, \delta) \leq \frac{\sigma^2 T \log \frac{4}{\delta}}{c_1} + \frac{3}{2} c_1 A + \sum_{a \in \mathcal{A}} \mathrm{gap}(a).$$

Taking $c_{1,t} = \sigma\sqrt{T/A}$ yields $\mathrm{Reg}_M^{\mathsf{Alg}}(T, \delta) = \mathcal{O}(\sigma\sqrt{AT}\log\frac{1}{\delta})$ and $\mathbb{E}[\mathrm{Reg}_M^{\mathsf{Alg}}(T)] = \mathcal{O}(\sigma\sqrt{AT})$.

Discussion of Theorem 4. The distributional regret bound in Theorem 4 shows that c_1 balances the $(T \log \frac{1}{\delta})/c_1$ term and the $c_1 A$ term. To reduce the coefficient of the $\log \frac{1}{\delta}$ term (i.e., to obtain stronger distributional guarantees), c_1 must increase. However, a larger c_1 increases the total regret through the $c_1 A$ term, and hence increases the expected regret bound linearly in c_1. This inverse trade-off between the coefficient of the $\log \frac{1}{\delta}$ factor and the expected regret is optimal for all choices of $c_1 = \Omega(\sqrt{T/A})$ by Theorem 17.1 of Lattimore and Szepesvári (2020).
In particular, setting $c_1 = \sigma\sqrt{T/A}$ yields a distributional bound of $\mathcal{O}\big(\sqrt{AT}\log\frac{1}{\delta}\big)$ and an expected bound of $\mathcal{O}(\sqrt{AT})$, without any logarithmic factors in A or T. The expected regret bound is minimax-optimal up to only constant factors, and the distributional bound is optimal given this expected bound. To the best of our knowledge, these rates are the tightest possible for MAB (up to constant factors). The existence of an algorithm achieving them was conjectured based on the lower bound in Lattimore and Szepesvári (2020, Section 17.1) but had not been established. To the best of our knowledge, this is the first result to attain these bounds, thereby confirming that conjecture.

The next theorem shows how $\frac{c_{2,t}}{\sqrt{N}}$-type bonus term achieves an optimal gap-dependent regret bound when used together with a proper confidence bound. To specify $c_{2,t}$, we first define the following concept.

Definition 5 (Time-uniform UCB) *A function* $u(t,\delta) : \mathbb{R}_{\geq 0} \times (0,1] \to \mathbb{R}_{\geq 0}$ *induces a* ***time-uniform upper confidence bound (UCB)*** *if* $\mathbb{P}\big(\exists t \in \mathbb{N}, \exists a \in \mathcal{A} : |\hat{r}^t(a) - r(a)| > \frac{u(t,\delta)}{\sqrt{N^t(a)}}\big) \leq \delta$ *always holds.*

While many different functions can induce time-uniform UCBs, we refer to those with asymptotically optimal constants and logarithmic factors as *tight time-uniform UCBs*; we provide a formal definition and an example in Appendix C.2. The following theorem considers $c_{2,t} = u(t,\delta_t)$ for a sequence $\{\delta_t\}_{t=1}^{\infty}$.

Theorem 6 *Suppose* $M \in \mathcal{B}(\sigma)$ *and* $u(t,\delta)$ *induces a time-uniform UCB. For a decreasing sequence* $\{\delta_t\}_{t=1}^{\infty}$*, set* $c_{2,t} = u(t,\delta_t)$ *and* $c_{1,t} = \infty$*, meaning that the bonus term is* $\frac{u(t,\delta_t)}{\sqrt{N}}$*. Then,*

$$\mathrm{Reg}_M^{\mathsf{Alg}}(T,\delta) \leq \tau_2(\delta) \wedge T + \sum_{a\in\mathcal{A}} \mathrm{gap}(a) + \sum_{a\in\mathcal{A},\mathrm{gap}(a)\neq 0} \frac{(c_{2,T} + u(t,\delta))^2}{\mathrm{gap}(a)},$$

where $\tau_2(\delta) := \max\{t \in \mathbb{N} : \delta_t > \delta\}$*. If* $u(t,\delta)$ *is tight and* δ_t *is appropriately decreasing, e.g.,* $\delta_t = (t\log t)^{-1}$*, we have* $\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_M^{\mathsf{Alg}}(T)]}{\log T} \leq \sum_{a\in\mathcal{A},\mathrm{gap}(a)\neq 0} \frac{2\sigma^2}{\mathrm{gap}(a)}$.

Discussion of Theorem 6. Theorem 6 shows how $c_{2,t}$ governs the gap-dependent distributional regret, exhibiting behavior different from that of $c_{1,t}$. When $c_{2,t} < u(t,\delta)$ (i.e., for the first $\tau_2(\delta)$ time steps), no non-trivial guarantee is provided. After $c_{2,t}$ passes the threshold, the bound increases with $\sum_a c_{2,t}^2/\mathrm{gap}(a)$, showing a more discrete behavior than the case of $c_{1,t}$.
The sequence $\{\delta_t\}_t$ can be understood as an intermediate parameter that balances $c_{2,t}$ and $\tau_2(\delta)$. Slowly diminishing $\{\delta_t\}_t$ decreases $c_{2,t}$ but increases $\tau_2(\delta)$, and vice versa. For instance, setting $\delta_t = (t\log t)^{-1}$ yields the tightest constant factor in the asymptotic regret, but results in a super-linear dependence on $\frac{1}{\delta}$ as $\tau_2(\delta) \approx \frac{1}{\delta}\log\frac{1}{\delta}$. If we accelerate the decay as $\delta_t = t^{-p}$ for some $p > 1$, then we obtain smaller $\tau_2(\delta) = (1/\delta)^{\frac{1}{p}}$, but the total regret scales linearly in p. Considering $\delta_t = \exp(-t^{\beta})$ for some $\beta > 0$ achieves a poly-logarithmic dependence on δ as $\tau_2(\delta) = \mathcal{O}((\log(1/\delta))^{\frac{1}{\beta}})$ but the total regret scales with T^{β}. This is consistent with the result of Simchi-Levi et al. (2023a), where they show that this order of trade-off is optimal.

Remark 7 *Theorem 6 states that a UCB algorithm with any tight time-uniform UCB and properly decreasing failure probabilities achieves the asymptotically optimal regret bound of Lai and Robbins (1985), which may be of independent interest. Refer to Appendix C.2 for the exact conditions.*

6.2. Distributional Regret Bounds for Reinforcement Learning

In this section, we provide distributional regret bounds for EQO+ with arbitrary input in the RL setting, which naturally applies also to MAB due to our unified framework. The proofs of the theorems are provided in Appendix D. Let $\ell(\delta) := \log\frac{cHSA\log k}{\delta}$ denote certain logarithmic factor with an absolute constant c, where the dependence on δ is emphasized. We define a quantity $\kappa(\delta)$ that is analogous to $\tau_2(\delta)$ in Theorem 6. For $\delta \in (0,1]$, let

$$\kappa(\delta) := \max\left\{k \in \mathbb{N} : c_{1,k} < (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell(\delta) \text{ or } c_{2,k} < \left(\sigma_{\max} + \frac{6\mathbb{W}^*_{\mathrm{diff}}\ell(\delta)}{c_{1,k}}\right)\sqrt{\ell(\delta)}\right\}.$$

$\kappa(\delta)$ is the number of episodes where the parameters $c_{1,k}$ and $c_{2,k}$ are too small to provide guarantees. Such a term is unavoidable since for any fixed input parameters and K, we can take δ small enough to invalidate the analysis; note that the thresholds in the definition go to infinity as $\delta \to 0$.

We provide gap-independent and gap-dependent bounds for EQO+ under arbitrary positive, non-decreasing input sequences $\{c_{1,k}\}_{k=1}^{\infty}$ and $\{c_{2,k}\}_{k=1}^{\infty}$.

Theorem 8 *For any $K \in \mathbb{N}$ and $\delta \in (0,1]$, the distributional regret of Alg satisfies*

$$\mathrm{Reg}_{\mathcal{M}}^{\mathsf{Alg}}(K,\delta) \le V_{\max}(\kappa(\delta) \wedge K) + (16 c_{1,K} SA \log KH)) \wedge \left(16\sqrt{2} c_{2,K} \sqrt{HSAK}\right) + \sum_{k=\kappa(\delta)+1}^{K} \frac{18 \mathbb{W}^* \ell(\delta)}{c_{1,k}} + 72 V_{\max} S^2 A \ell(\delta) \log 2KH \,.$$

Theorem 9 *Define $\kappa_{\mathrm{gap}}(\mathrm{gap}_{\min}, \delta) := \max\{k \in \mathbb{N} : c_{1,k} < \frac{36 \mathbb{W}^* \ell(\delta)}{\mathrm{gap}_{\min}}\}$. For any $K \in \mathbb{N}$ and $\delta \in (0,1]$, the distributional regret of Alg satisfies*

$$\mathrm{Reg}_{\mathcal{M}}^{\mathsf{Alg}}(K,\delta) \le V_{\max}(\kappa(\delta) \wedge K) + \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(\frac{4096 c_{2,K}^2}{\mathrm{gap}(s,a) \vee \frac{\mathrm{gap}_{\min}}{H}} \right) \wedge \left(64 c_{1,K} \log \frac{64 c_{1,K}}{\mathrm{gap}(s,a) \vee \frac{\mathrm{gap}_{\min}}{H}} \right) + \sum_{k=\kappa(\delta)+1}^{\kappa_{\mathrm{gap}}(\mathrm{gap}_{\min},\delta) \wedge K} \frac{144 \mathbb{W}^* \ell(\delta)}{c_{1,k}} + 288 V_{\max} S^2 A \ell(\delta) \log 2KH$$

Discussion of Theorems 8 and 9. Both bounds involve the minimum of $c_{1,K}$-related and $c_{2,K}$-related terms, which arises from the bonus term $(c_{1,k}/N) \wedge (c_{2,k}/\sqrt{N})$. For most parameter choices we consider, the $c_{1,K} SA \log KH$ term is smaller in Theorem 8, whereas the $\sum_{s,a} c_{2,K}^2 / \mathrm{gap}(s,a)$ term is smaller in Theorem 9, dominating their respective bounds. In this sense, we see that $c_{1,k}$ mainly controls the worst-case behavior, while $c_{2,k}$ mainly controls the instance-dependent behavior, corroborating the presentation of Section 6.1.

In Theorem 8, the parameter $c_{1,k}$ balances the sum $\sum_k (\mathbb{W}^* \ell(\delta)/c_{1,k})$ and the term $c_{1,K} SA (\log KH)$, creating a trade-off parallel to Theorem 4. We expect this trade-off to be optimal up to logarithmic factors, considering that certain MDP instances behave analogous to MAB instances with SA actions (Domingues et al., 2021). In addition, setting $c_{1,k} = \widetilde{\Theta}(V_{\max} \sqrt{k/(SA)})$ yields the minimax optimal regret bound of $\widetilde{\mathcal{O}}(V_{\max} \sqrt{SAK})$ (see Corollary 11).

The parameter $c_{2,k}$ affects both bounds in a manner analogous to Theorem 6, where its growth rate balances the δ-dependence in $\kappa(\delta)$ and the overall distributional regret bound.

In Theorem 9, the sum $\sum_k \mathbb{W}^* \ell(\delta)/c_{1,k}$ appears, which might seem problematic as it scales with $\sqrt{K}$ when $c_{1,k} = \Theta_k(\sqrt{k})$. We show that the summand stops affecting the order of the distributional regret bound once $\mathbb{W}^* \ell(\delta)/c_{1,k} \lesssim \mathrm{gap}_{\min}^{-1}$ holds, implying that the sum becomes independent of K. When $c_{1,k} = \Theta_k(\sqrt{k})$, this sum scales moderately with $\frac{1}{\mathrm{gap}_{\min}}$ (see Corollary 12).

Remark 10 *The condition that $\{c_{1,k}\}_{k=1}^{\infty}$ and $\{c_{2,k}\}_{k=1}^{\infty}$ be increasing can be easily relaxed by redefining $\kappa(\delta)$ accordingly and by replacing $c_{1,K}$ and $c_{2,K}$ in the bounds with $\max_{k \in [K]} c_{1,k}$ and $\max_{k \in [K]} c_{2,k}$, respectively. Then, the bounds apply to arbitrary positive input parameters.*

6.3. Corollaries for Specific Parameter Choices

In this section, we propose exemplary parameter choices and present the resulting bounds. Full versions of the corollaries and their proofs are presented in Appendix E.
We let $\mathcal{M}$ be a set of MDPs with the time horizon H, S states, A actions, and $V_{\max}$ as the range of the value function, and that further satisfy $\mathbb{W}^* \le 2V_{\max}^2$ and $V_\alpha \le 2V_{\max}$. $\mathcal{M}$ includes MDPs with $V_{\max}$-bounded reward or σ^2-sub-Guassian reward noise with $\sigma^2 H \le 2V_{\max}^2$.

First, we present a standard choice of $c_{1,k} \approx V_{\max}\sqrt{k/(SA)}$ with logarithmic factors.

Corollary 11 *Set $c_{1,k} = c_1 V_{\max}\sqrt{\frac{k\ell(1)}{SA\log KH}}$ for some constant $c_1 > 0$ and $c_{2,k} = \infty$. Then,*

$$\sup_{M\in\mathcal{M}} \mathrm{Reg}_M^{\mathit{Alg}}(K,\delta) \le \left(\frac{36\log\frac{1}{\delta}}{c_1 \log HSA} + \frac{36}{c_1} + 8c_{1,k}\right) V_{\max}\sqrt{KSA\ell(1)\log KH} + 72V_{\max}S^2A\ell(\delta)\log 2KH\,. \tag{1}$$

The expected regret is bounded as

$$\sup_{M\in\mathcal{M}} \mathbb{E}[\mathrm{Reg}_M^{\mathit{Alg}}(K)] = \mathcal{O}\Big(V_{\max}\sqrt{SAK(\log KH)(\log HSA(\log K))} + V_{\max}S^2A(\log KH)(\log HSA(\log K))\Big)\,.$$

Furthermore, if K is known and $c_{1,k} = c_1 V_{\max}\sqrt{\frac{K\log HSA}{SA\log KH}}$, then the $\log HSA(\log K)$ factor in the square root reduces to $\log HSA$.

This distributional regret bound achieves the minimax optimal $\widetilde{\mathcal{O}}(V_{\max}\sqrt{SAK})$ bound, while simultaneously achieving a linear dependence on $\log\frac{1}{\delta}$. The logarithmic factor in the leading term of the expected bound is $\sqrt{(\log KH)(\log HSA(\log K))}$, where the $\sqrt{\log\log K}$ factor can be removed when K is known. To the best of our knowledge, this $\sqrt{\log K}$-dependence is the sharpest among the minimax optimal RL algorithms.

The next corollary shows that we can combine Hoeffding's bound in addition to Corollary 11.

Corollary 12 *Assume that $\sigma_{\max}^2 \le 2V_{\max}^2$, and set $c_{1,k} = c_1 V_{\max}\sqrt{\frac{k\ell(1)}{SA\log KH}}$ for some constant $c_1 > 0$ and $c_{2,k} = c_2 V_{\max}\sqrt{\log 32HSAk}$ for some constant $c_2 \ge 2$. Then, we have*

$$\mathrm{Reg}_M^{\mathit{Alg}}(K,\delta) = V_{\max}\exp\left(\mathcal{O}\left(\frac{1}{c_2^2}\log\frac{1}{\delta}\right)\right) + \mathcal{O}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{c_2^2 V_{\max}^2 \log HSAK}{\mathrm{gap}(s,a)\vee\frac{\mathrm{gap}_{\min}}{H}}\right) + \mathcal{O}\left(\frac{V_{\max}^2 SA(\log KH)(\ell(\delta))^2}{c_1^2\mathrm{gap}_{\min}(\log HSA)} + V_{\max}S^2A(\log KH)\ell(\delta)\right)\,.$$

The worst-case bound is the sum of the bound in Eq. (1) and $\exp(\mathcal{O}(c_2^{-2}\log(1/\delta)))$ term.

Corollary 12 shows an interesting result: by taking the minimum of the bonus terms of `EQO` (Lee and Oh, 2025) and `UCBVI-CH` (Azar et al., 2017), one can achieve both minimax optimal and

instance-dependent $\log K$ regret bounds. As a trade-off, the distributional bound grows polynomially in $1/\delta$, which is consistent with Theorem 6. We note that the orders of $V_{\max}$, $\log K$, and $\mathrm{gap}(s,a)$ match the lower bound of Proposition 2.2 in Simchowitz and Jamieson (2019).
The $\mathcal{O}\left(\frac{V_{\max}^2 SA(\log KH)(\ell(\delta))^2}{c_1^2 \mathrm{gap}_{\min}(\log HSA)}\right)$ term comes from the sum of $\frac{\mathbb{W}^*\ell(\delta)}{c_{1,k}}$ terms, where the $\log K$ factor can be improved to K-independent logarithmic factors. Compared to the bound of Simchi-Levi et al. (2023a), we improve the $\mathrm{gap}_{\min}^{-2}$ dependence to $\mathrm{gap}_{\min}^{-1}$. Despite this, this term may still appear large and could potentially be the leading term. We show that the $\mathrm{gap}_{\min}$ dependence can be improved for MDPs with specific structures in Appendix D.4.

We also consider the parameter choice by Simchi-Levi et al. (2023a).

Corollary 13 *Let $c_1 > 0, c_2 > 0$, $\alpha \in [\frac{1}{2}, 1]$ and $0 < \beta \le \alpha$ be constants. Set $c_{1,k} = c_1 V_{\max}(\frac{k}{SA})^\alpha$ and $c_{2,k} = c_2\sqrt{k^\beta}$. Then, we have*

$$\sup_{M\in\mathcal{M}} \mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) = \widetilde{\mathcal{O}}_{K,\delta}\left(\left(\log\frac{1}{\delta}\right)^{\frac{1}{\beta}} + K^{1-\alpha}\log\frac{1}{\delta} + K^\alpha \wedge K^{\beta+\frac{1}{2}}\right) \qquad \textit{and}$$

$$\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) = \widetilde{\mathcal{O}}_{K,\delta,\mathrm{gap}}\left(\left(\log\frac{1}{\delta}\right)^{\frac{1}{\beta}} + \left(\frac{1}{\mathrm{gap}_{\min}}\right)^{\frac{1}{\alpha}-1}\left(\log\frac{1}{\delta}\right)^{\frac{1}{\alpha}} + \sum_{s,a}\frac{K^\beta}{\mathrm{gap}(s,a)\vee\frac{\mathrm{gap}_{\min}}{H}}\right).$$

The hyper-parameters α and β provide control over the trade-offs between different types of regret bounds we have discussed. α balances the coefficient of $\log\frac{1}{\delta}$ and the expected regret bound in the gap-independent case, whereas β controls the orders of the $\log\frac{1}{\delta}$ factor and the instance-dependent regret. When restricted to the MAB setting, our results achieve the same order of the gap-independent bound as Simchi-Levi et al. (2023a) while improving the gap-dependent bound. Specifically, we improve the previous $(\frac{1}{\mathrm{gap}_{\min}})^{\frac{1}{\alpha}}$ factor to $(\frac{1}{\mathrm{gap}_{\min}})^{\frac{1}{\alpha}-1}$, and eliminate a $\mathcal{O}((\log\frac{1}{\delta})\sum_{s,a}\frac{1}{\mathrm{gap}(s,a)})$ term. A more detailed comparison with Simchi-Levi et al. (2023a) and Khodadadian and Moharrami (2025) is provided in Appendix F.

7. Proof Sketch

In this section, we provide a sketch of the proof of Theorems 8 and 9. At the end of the section, we additionally sketch the technique used in Theorem 4 for the tightest logarithmic factors.

The main challenge in deriving distributional regret bounds is to remove the δ-dependence in the algorithm while maintaining the high-probability guarantees. Standard UCB-based analyses are invalidated since confidence bounds require specified failure probabilities. In Theorem 6, we proposed a novel method of using decreasing failure probabilities and considering the corresponding warm-up time. While this technique provides optimal trade-offs between the δ-dependence and the instance-dependent regret bound in some regimes, we require additional techniques to obtain the optimal trade-off for the minimax regret bound. To this end, we combine and extend the *quasi-optimism* analysis by Lee and Oh (2025), where quasi-optimism means that the value estimates are almost optimistic, allowing for potential underestimation.
We fix $\delta \in (0,1]$ and set a good event $\mathcal{E}(\delta)$ whose probability is at least $1-\delta$. We do not make any guarantees for the first $\kappa(\delta)$ episodes of warm-up time, incurring regret of at most $V_{\max}\kappa(\delta)$. For $k > \kappa(\delta)$, the input parameters $c_{1,k}$ and $c_{2,k}$ become large enough to serve as (quasi-)optimistic bonus terms. Specifically, we have the following lemma.

Lemma 14 (Quasi-optimism (informal)) *Under $\mathcal{E}(\delta)$, the value estimate $V_h^k(s)$ computed in Algorithm 1 satisfies $V_h^*(s) - V_h^k(s) \lesssim \frac{\mathbb{W}^*\ell(\delta)}{c_{1,k}}$ for all $h \in [H]$, $s \in \mathcal{S}$, and $k > \kappa(\delta)$.*

Lemma 14 is the main tool that allows us to use δ-independent bonus terms while maintaining high-probability guarantees. Instead of guaranteeing overestimation of the algorithm as in the usual optimism-based analyses, it provides a distributional bound on the amount of potential underestimation of the given bonus terms. We note that the quasi-optimism analysis in Lee and Oh (2025, Lemma 2) only applies to a fixed value of δ since their bonus term depends on it, whereas our result holds simultaneously for all $\delta \in (0, 1]$, which is an important distinction for deriving the distributional regret bounds. In addition, our analysis improves the scaling of the underestimation from $V_{\max}^2/c_{1,k}$ to $\mathbb{W}^*/c_{1,k}$, showing that a smaller variance of an MDP automatically leads to less underestimation. Due to our unified framework, it is immediately derived that it may also scale with $\sigma^2 H$ when the reward noise is σ^2-sub-Gaussian.

For Theorem 8, the remaining steps of bounding $V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)$ follow standard techniques. We bound the regret of one episode by $V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \lesssim \frac{\mathbb{W}^*\ell(\delta)}{c_{1,k}} + U_1^k(s_1^k)$, where $U_h^k(s) \approx \mathbb{E}_{\pi(\cdot|s_h=s)}[\sum_{j=h}^H b^k(s_j, a_j) + \frac{V_{\max} S\ell(\delta)}{N^k(s_j,a_j)}]$. We also derive a useful tool for bounding the expected visit counts by the visit counts of the sampled trajectory, which is a generalization of Lemma 15 in Lee and Oh (2025).

Lemma 15 *Let $\{X_h^k\}_{k,h}$ be a sequence of non-negative random variables adapted to a filtration $\{\mathcal{F}_h^k\}_{k,h}$. Let $c > 0$ be a constant. Recursively define $\{J_h^k\}_{h,k}$ as $J_{H+1}^k := 0$ and $J_h^k := \left(X_h^k + \mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k]\right) \wedge c$ for all $k \in \mathbb{N}$ and $h \in [H]$. Then, for any $\delta \in (0, 1]$, it holds that $\sum_{k=1}^K J_1^k \le 2\sum_{k=1}^K \sum_{h=1}^H X_h^k + 6c \log \frac{2}{\delta}$ for all $K \in \mathbb{N}$ with probability at least $1 - \delta$.*

We bound $\sum_k U_1^k(s_1^k)$ by taking $X_h^k \approx b^k(s_h^k, a_h^k) + \frac{V_{\max} S\ell(\delta)}{N^k(s_h^k,a_h^k)}$, which yields Theorem 8.

The remaining steps for the proof of Theorem 9 are more subtle. We incorporate the clipping framework of Simchowitz and Jamieson (2019); Dann et al. (2021) that shows that the bonus term $b^k(s_h^k, a_h^k)$ may not be added to the regret bound once $b^k(s_h^k, a_h^k) \lesssim \frac{1}{2}\mathrm{gap}_h(s_h^k, a_h^k)$ holds. However, this framework is designed specifically for optimistic algorithms, so the quasi-optimism term requires additional attention. The main idea is to show that the quasi-optimism term $\frac{\mathbb{W}^*\ell(\delta)}{c_{1,k}}$ is also negligible beyond a certain threshold $k \ge \kappa'$, where κ' should be as small as possible.

One method is to define the minimum non-zero regret $\mathrm{Reg}_{\min} := \min_{s\in\mathcal{S},\pi\in\Pi,V_1^*(s)-V_1^\pi(s)>0} V_1^*(s) - V_1^\pi(s)$, then let κ' be the time step where $\frac{\mathbb{W}^*\ell(\delta)}{c_{1,k}} \lesssim \frac{1}{2}\mathrm{Reg}_{\min}$, implying $V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \lesssim 2U_1^k(s_1^k)$ for $k > \kappa'$. While $\mathrm{Reg}_{\min}$ coincides with $\mathrm{gap}_{\min}$ in the bandit setting, it may be arbitrarily smaller than $\mathrm{gap}_{\min}$ in an MDP when the policy only makes a sub-optimal action at a state that is reachable with an arbitrarily small probability, hence κ' could be arbitrarily large. To address this issue, we derive a bound that probabilistically adds the quasi-optimism term only when the agent makes a sub-optimal action. Since the sub-optimality of that action would be at least $\mathrm{gap}_{\min}$, we can subsume the quasi-optimism term once it goes below $\frac{1}{2}\mathrm{gap}_{\min}$. Specifically, defining B as the time step of the first sub-optimal action, we show that $V_1^*(s_1) - V_1^{\pi^k}(s_1) = \mathbb{E}_{\pi^k}[V_B^*(s_B) - V_B^{\pi^k}(s_B)] \lesssim \mathbb{E}_{\pi^k}[2U_B^{\pi^k}(s_B)]$ when $\frac{\mathbb{W}^*\ell(\delta)}{c_{1,k}} \lesssim \frac{1}{2}\mathrm{gap}_{\min}$. This gives rise to the definition of $\kappa_{\mathrm{gap}}(\delta, \mathrm{gap}_{\min})$ in Theorem 9. By carefully setting X_h^k in Lemma 15 to properly capture this property in a way that depends on the history within the same episode, we derive Theorem 9.

Proof Sketch of Theorem 4. We additionally provide a sketch of the technique that shaves off logarithmic factors in Theorem 4. For each $a \in \mathcal{A}$, denote the last time step that the action a is taken by t_a. Then, at the t_a-th time step, we have

$$\begin{aligned} r(a^*) &\leq \frac{\sigma^2 \log \frac{4}{\delta}}{2c_1} + \hat{r}^{t_a}(a^*) + \frac{c_1}{N^{t_a}(a^*)} \\ &\leq \frac{\sigma^2 \log \frac{4}{\delta}}{2c_1} + \hat{r}^{t_a}(a) + \frac{c_1}{N^{t_a}(a)}\,, \end{aligned}$$

where the first inequality is due to quasi-optimism and the second inequality is from the action-selection rule. There is no logarithmic factor of A or T since quasi-optimism only requires the concentration of the optimal action's noise and the inequality is time-uniform. We then have $\mathrm{gap}(a) \leq \frac{\sigma^2 \log \frac{4}{\delta}}{2c_1} + \hat{r}^{t_a}(a) - r(a) + \frac{c_1}{N^{t_a}(a)}$. If we individually bound $\hat{r}^t(a) - r(a)$ for each $a \in \mathcal{A}$, we must take the union bound over $a \in \mathcal{A}$, which incurs a $\log A$ factor. We observe that the union bound is overly pessimistic here, as it assumes that violations of the confidence bounds are mutually exclusive for all arms. We avoid this by bounding the sum of noises over $a \in \mathcal{A} \setminus \{a^*\}$, requiring only two concentration inequalities instead of A. Taking the sum of $\mathrm{gap}(a)N^{t_a}(a)$ over $a \in \mathcal{A} \setminus \{a^*\}$, we have

$$\sum_{a \in \mathcal{A} \setminus \{a^*\}} \mathrm{gap}(a) N^{t_a}(a) \leq \frac{\sigma^2 T \log \frac{4}{\delta}}{2c_1} + c_1 A + \sum_{a \in \mathcal{A} \setminus \{a^*\}} N^{t_a}(a)(\hat{r}^{t_a}(a) - r(a))\,.$$

The last sum represents the sum of reward noises from all time steps with sub-optimal action selections, excluding the last noise of each action, which is sampled at time step t_a. We show that this sum is bounded by $\mathcal{O}(\sqrt{AT \log \frac{1}{\delta}})$ using concentration results for the total noise from sub-optimal actions and the noise specifically at the final time steps. Although the latter requires taking the union bound over all possible such time steps, we show that it does not affect the leading term. Combining the terms yields the regret bound of Theorem 4.

8. Conclusion

In this paper, we study the distributional properties of regret under the algorithm `EQO+`. We provide a distributional regret bound that holds for arbitrary failure probability. We provide very generic theorems for arbitrary input, which allow us to study the trade-off controlled by the input parameters. We also propose a framework that unifies the bandit and RL settings. While we achieve optimal results in the bandit setting, the lack of lower bound results specific to the RL settings leaves an open question of whether our guarantee is optimal. Extending the work to function approximation would be interesting future work.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant both funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023-00222663, RS-2025-25463302, RS-2026-25507282).

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26. PMLR, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Annual Conference on Learning Theory*, pages 217–226, 2009.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR, 2017.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34:10849–10861, 2021.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595. PMLR, 2016.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

Lin Fan and Peter W Glynn. The fragility of optimized bandit algorithms. *Operations Research*, 73 (6):3173–3198, 2025.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Huiwen Jia, Cong Shi, and Siqian Shen. Multi-armed bandit with sub-exponential rewards. *Operations Research Letters*, 49(5):728–733, 2021.

Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration is fast and optimal. In *International Conference on Machine Learning*, pages 15239–15261. PMLR, 2023.

Sajad Khodadadian and Mehrdad Moharrami. Tail distribution of regret in optimistic reinforcement learning. *arXiv preprint arXiv:2511.18247*, 2025.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems*, 27, 2014.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(20):1–32, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Harin Lee and Min-hwan Oh. Minimax optimal reinforcement learning with quasi-optimism. In *The Thirteenth International Conference on Learning Representations*, 2025.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, 2017.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. A simple and optimal policy design for online learning with safety against heavy-tailed risk. *Advances in Neural Information Processing Systems*, 35:33795–33805, 2022.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. Regret distribution in stochastic bandits: Optimal trade-off between expectation and tail risk. *arXiv preprint arXiv:2304.04341*, 2023a.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. Stochastic multi-armed bandits: Optimal trade-off among optimality, consistency, and tail risk. *Advances in Neural Information Processing Systems*, 36:35619–35630, 2023b.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. A simple and optimal policy design with safety against heavy-tailed risk for stochastic bandits. *Management Science*, 71(7):6298–6318, 2025.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Daniil Tiapkin, Denis Belomestny, Eric Moulines, Alexey Naumov, Sergey Samsonov, Yunhao Tang, Michal Valko, and Pierre Ménard. From dirichlet to rubin: Optimistic exploration in rl without bonuses. In *International Conference on Machine Learning*, pages 21380–21431. PMLR, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

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

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

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

Runlong Zhou, Zhang Zihan, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*, pages 42878–42914. PMLR, 2023.

Feng Zhu and David Simchi-Levi. Adaptive variance inflation in thompson sampling: Efficiency, safety, robustness, and beyond. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Appendix A. Additional Definitions and Notations

In this section, we define additional notations for the analysis.

We define logarithmic factors $\ell_1(i,\delta) := \frac{32HSAi^2}{\delta}$ and $\ell_2(k,\delta) = \frac{32HSA(\log e^2kH)^2}{\delta}$. Using these logarithmic factors, we provide a more precise definitions of $\kappa(\delta)$ and $\kappa_{\mathrm{gap}}(\mathrm{gap},\delta)$.

$$\kappa(\delta) := \max\left\{k \in \mathbb{N} : c_{1,k} < (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta) \text{ or } c_{2,k} < \left(\sigma_{\max} + \tfrac{6\mathbb{W}^*_{\mathrm{diff}}\ell_1(\iota_k,\delta)}{c_{1,k}}\right)\sqrt{\ell_2(k,\delta)}\right\},$$

$$\kappa_{\mathrm{gap}}(\mathrm{gap},\delta) := \max\left\{k \in \mathbb{N} : c_{1,k} < \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{\mathrm{gap}}\right\}.$$

We let $N^k_h(s,a) := N^k(s,a) + \sum_{j=1}^h \mathbb{1}\{(s^k_j,a^k_j) = (s,a)\}$, which is the visit count of (s,a) up to the h-th time step of the k-th episode. Let $\eta^k \in [H+1]$ be a stopping time with respect to $\{\mathcal{F}^k_h\}_{h=1}^{H+1}$ defined as the first time step h within the k-th episode such that $N^k_h(s^k_h,a^k_h) \geq 2N^k(s^k_h,a^k_h)$, where $\eta^k = H+1$ if such a time step does not exist. This definition is from Lee and Oh (2025) and is useful for handling the possibility that one state-action pair may be visited multiple times within an episode.

A.1. Auxiliary Sequence for Analysis

In this section, we define auxiliary sequences $\{\lambda_i\}_{i=1}^\infty$ and $\{\iota_k\}_{k=1}^\infty$ for the proof that depends on $\{c_{1,k}\}_{k=1}^\infty$ and δ. For most parameter choices we consider, we have $\iota_k \approx \log k$ and $\lambda_{\iota_k} \approx \frac{\ell_1(\iota_k,\delta)}{c_{1,k}}$. If $c_{1,k}$ is a fixed constant, then we have $\iota_k = 1$.

Lemma 16 *Suppose $\delta \in (0,1]$ and a positive sequence $\{c_{1,k}\}_{k=1}^\infty$ is given. Then, there exist sequences $\{\lambda_i\}_{i=1}^\infty$ and $\{\iota_k\}_{k=1}^\infty$ such that (i) $0 < \lambda_i \leq \frac{1}{V_\alpha}$ for all $i \in \mathbb{N}$, and (ii) for all $k \in \mathbb{N}$ with $c_{1,k} \geq (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta)$, we have*

$$\frac{\ell_1(\iota_k,\delta)}{\lambda_{\iota_k}} + 13\lambda_{\iota_k}\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta) \leq c_{1,k} \leq \frac{4\ell_1(\iota_k,\delta)}{\lambda_{\iota_k}}. \tag{2}$$

Furthermore, if $c_{1,k}$ is non-decreasing, then we have $\iota_k \leq k \wedge (3 + \log_2(c_{1,k}/c_{1,1}))$. For $k \in \mathbb{N}$ with $c_{1,k} < (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta)$, we define $\iota_k := 1$.

Proof Our goal is to construct $\{\lambda_i\}_{i=1}^\infty$ such that the union of ranges $\cup_i[\frac{\ell_1(i,\delta)}{\lambda_i} + 13\lambda_i\mathbb{W}^*_{\mathrm{diff}}\ell(1,\delta), \frac{4\ell_1(i,\delta)}{\lambda_i}]$ covers all $c_{1,k}$ with $c_{1,k} \geq (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta)$. We iteratively choose λ_i to be the value λ that satisfies $\frac{\ell_1(i,\delta)}{\lambda} + 13\lambda\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta) = c_{1,k}$ for the least $c_{1,k}$ that is not covered by the previous $\lambda_{i'}$, which is equivalent to the least $c_{1,k}$ with $c_{1,k} \geq \frac{4\ell_1(i-1,\delta)}{\lambda_{i-1}}$ for $i \geq 2$. The existence of such λ is guaranteed by Lemma 17. We also have $\lambda \leq \sqrt{\frac{\ell_1(i,\delta)}{13\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta)}}$ by Lemma 17, which implies $\frac{\ell_1(i,\delta)}{\lambda} + 13\lambda\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta) \leq \frac{2\ell_1(i,\delta)}{\lambda}$. Then, the right ends of the intervals $\frac{4\ell_1(i,\delta)}{\lambda_i}$ increase at least

exponentially by the following reasoning:

$$\begin{aligned}\frac{4\ell_1(i,\delta)}{\lambda_i} &\geq 2\cdot\frac{2\ell_1(i,\delta)}{\lambda_i}\\ &\geq 2\left(\frac{\ell_1(i,\delta)}{\lambda_i}+13\lambda_i\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)\right)\\ &\geq 2\cdot\frac{4\ell_1(i-1,\delta)}{\lambda_{i-1}}\,,\end{aligned}$$

where the last inequality holds by the choice of λ_i. Hence, for any $c_{1,k}$, there exists $i\in\mathbb{N}$ such that $c_{1,k}\leq\frac{4\ell_1(i,\delta)}{\lambda_i}$, and by the construction of λ_i, the least such i also satisfies $\frac{\ell_1(i,\delta)}{\lambda_i}+13\lambda_i\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)\leq c_{1,k}$. By setting $\iota_k=i$, Eq. (2) is satisfied.
Now, suppose $c_{1,k}$ is non-decreasing. Since each range covered by λ_i covers at least one $c_{1,k}$, we obtain $\iota_k\leq k$. Also, the fact that $\frac{4\ell_1(i,\delta)}{\lambda_i}$ increases exponentially implies $\iota_k\leq 3+\log_2(c_k/c_1)$. Formally, let $\kappa(\delta)+1$ be the least $k\in\mathbb{N}$ that satisfies $c_{1,k}\geq(2\sqrt{13\mathbb{W}^*_{\text{diff}}}\vee 2V_\alpha)\ell_1(1,\delta)$. Then, for $k>\kappa(\delta)$, we have

$$\begin{aligned}c_{1,k}\geq\frac{\ell_1(\iota_k,\delta)}{\lambda_{\iota_k}} &\geq 2^{\iota_k-1}\cdot\frac{\ell_1(1,\delta)}{\lambda_1}\\ &= 2^{\iota_k-3}\cdot\frac{4\ell_1(1,\delta)}{\lambda_1}\\ &\geq 2^{\iota_k-3}c_{1,\kappa(\delta)+1}\,.\end{aligned}$$

The inequality above implies that $\iota_k\leq 3+\log_2(c_{1,k}/c_{1,\kappa(\delta)+1})\leq 3+\log_2(c_{1,k}/c_{1,1})$. For $k\leq\kappa(\delta)$, we have $\iota_k=1\leq 3+\log_2(c_{1,k}/c_{1,1})$. ■

Lemma 17 *For given $i\in\mathbb{N}$, $\delta\in(0,1]$, and $c\geq(2\sqrt{13\mathbb{W}^*_{\text{diff}}}\vee 2V_\alpha)\ell_1(1,\delta)$, there exists λ such that $0<\lambda\leq\frac{1}{V_\alpha}\wedge\sqrt{\frac{\ell_1(i,\delta)}{13\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)}}$ and $\frac{\ell_1(i,\delta)}{\lambda}+13\lambda\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)=c$.*

Proof The proof is simply using that $f(\lambda):=\frac{\ell_1(i,\delta)}{\lambda}+13\lambda\mathbb{W}^*_{\text{diff}}\ell_1(i,\delta)$ is continuous.
Take $\gamma:=\frac{1}{V_\alpha}\wedge\sqrt{\frac{\ell_1(i,\delta)}{13\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)}}$. From $\gamma\leq\sqrt{\frac{\ell_1(i,\delta)}{13\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)}}$, we have $13\gamma\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)\leq\frac{\ell_1(i,\delta)}{\gamma}$, and hence $f(\gamma)\leq\frac{2\ell_1(i,\delta)}{\gamma}$. Then, we have.

$$f(\gamma)\leq\frac{2\ell_1(i,\delta)}{\gamma}\leq\left(2V_\alpha\vee 2\sqrt{\frac{13\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)}{\ell_1(i,\delta)}}\right)\ell_1(i,\delta)\leq(2V_\alpha\vee 2\sqrt{13\mathbb{W}^*_{\text{diff}}})\ell_1(i,\delta)\,.$$

Since $\lim_{\lambda\to 0}f(\gamma)=\infty$ and f is continuous, the intermediate value theorem implies that for any $c\geq(2V_\alpha\vee 2\sqrt{13\mathbb{W}^*_{\text{diff}}})\ell_1(i,\delta)$, there exists $\lambda\in(0,\gamma]$ such that $f(\lambda)=c$. ■

A.2. Properties of Maximums and Minimums

In this paper, we frequently do operations with maximum and minimum over values. We state the following facts in case the readers find some steps in the proof non-trivial. When ambiguous, we assume that the priorities of the $\wedge$ and $\vee$ operators are in between addition and multiplication. For example, we have $a \wedge b + c = \min\{a, b\} + c$ and $ab \vee c = \max\{ab, c\}$.

Fact 1 *For real numbers a, b, c, d, and a non-negative real number x, the followings are true:*

(i) $(a \wedge b) + (c \wedge d) \le (a \wedge c) \wedge (b \wedge d)$.

(ii) $(a + c) \vee (b + d) \le (a \vee b) + (c \vee d)$.

(iii) $(a + c) \wedge (b + d) \le (a \vee b) + (c \wedge d) \le (a + c) \vee (b + d)$

(iv) $(a + x) \wedge b \le a \wedge b + x$.

(v) $(a + x) \vee b \le a \vee b + x$.

(vi) $a \wedge b - x \le (a - x) \wedge b$.

(vii) $a \vee b - x \le (a - x) \vee b$.

(viii) $(a \wedge b) \vee c = (a \vee c) \wedge (b \vee c)$.

Appendix B. Derivation of Table 1

In this section, we show how the values in Table 1 can be derived.

Bounded reward $0 \le R_h \le V_{\max}$: We define $R_h(s_h, a_h)$ as the distribution of R_h given s_h, a_h, which may depend on the history. For given $(s, a) \in \mathcal{S} \times \mathcal{A}$, the random variable $R_h + V^*_{h+1}(s')$ with $R_h \sim R_h(s, a)$ and $s' \sim P(s, a)$ lies in $[0, 2V_{\max}]$. By Lemma 50, $R_h + V^*_{h+1}(s')$ is $(2\operatorname{Var}(R_h + V^*_{h+1})(s, a), 2V_{\max})$-sub-exponential. $\mathbb{W}^*_{\text{diff}} \le \mathbb{W}^*$ is trivial, so it remains to prove $\mathbb{W}^* \le 2V^2_{\max}$. For any $h \in [H]$, $s \in \mathcal{S}$, and $\pi \in \Pi$, we have

$$
\begin{aligned}
W^\pi_h(s) &= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H} \operatorname{Var}(R_j + V^*_{j+1})(s_j, a_j)\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}\left(\mathop{\mathbb{E}}_{\substack{R_j\sim R(s_j,a_j)\\ s'\sim P(s_j,a_j)}}[(R_j + V^*_{j+1}(s'))^2] - \mathop{\mathbb{E}}_{\substack{R_j\sim R(s_j,a_j)\\ s'\sim P(s_j,a_j)}}[R_j + V^*_{j+1}(s')]^2\right)\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}(R_j + V^*_{j+1}(s_{j+1}))^2 - \sum_{j=h}^{H}(Q^*_j(s_j, a_j))^2\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}\left((R_j + V^*_{j+1}(s_{j+1}))^2 - (Q^*_{j+1}(s_{j+1}, a_{j+1}))^2\right)\right] - \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[(Q_h(s_h, a_h))^2\right],
\end{aligned}
$$

where the third equality uses the law of total expectation and that $Q_j^*(s_j, a_j) = \mathbb{E}_{\substack{R_j \sim R(s_j,a_j) \\ s' \sim P(s_j,a_j)}}[R_j + V_{j+1}^*(s')]$. The summand can be modified as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[(R_j + V_{j+1}^*(s_{j+1}))^2 - (Q_{j+1}^*(s_{j+1}, a_{j+1}))^2\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[(R_j + V_{j+1}^*(s_{j+1}) + Q_{j+1}^*(s_{j+1}, a_{j+1}))(R_j + V_{j+1}^*(s_{j+1}) - Q_{j+1}^*(s_{j+1}, a_{j+1}))\right] \\
&\le \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[3V_{\max}(R_j + V_{j+1}^*(s_{j+1}) - Q_{j+1}^*(s_{j+1}, a_{j+1}))\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[3V_{\max}(Q_j^*(s_j, a_j) - Q_{j+1}^*(s_{j+1}, a_{j+1}))\right] ,
\end{aligned}$$

where the inequality holds since $0 \le R_j + V_{j+1}^*(s_{j+1}) + Q_{j+1}^*(s_{j+1}, a_{j+1}) \le 3V_{\max}$ and $R_j + V_{j+1}^*(s_{j+1}) - Q_{j+1}^*(s_{j+1}, a_{j+1}) = R_j + Q_{j+1}^*(s_{j+1}, a_{j+1}^*) - Q_{j+1}^*(s_{j+1}, a_{j+1}) \ge 0$. Taking the sum of $\mathbb{E}_{\pi(\cdot|s_h=s)}\left[3V_{\max}(Q_j^*(s_j, a_j) - Q_{j+1}^*(s_{j+1}, a_{j+1}))\right]$ and telescoping yields

$$\begin{aligned}
W_h^\pi(s) &\le \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[\sum_{j=h}^{H} 3V_{\max}(Q_j^*(s_j, a_j) - Q_{j+1}^*(s_{j+1}, a_{j+1}))\right] - \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[(Q_h(s_h, a_h))^2\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[3V_{\max}Q_h^*(s_h, a_h) - (Q_h(s_h, a_h))^2\right] \\
&\le 2V_{\max}^2 ,
\end{aligned}$$

where the last inequality holds due to that $3V_{\max}x - x^2 \le 2V_{\max}^2$ for all $x \in [0, V_{\max}]$.

Sub-Gaussian reward noise: If R_h is σ^2-sub-Gaussian and is independent of s_{h+1}, then for any $\lambda \in [-\frac{1}{V_{\max}}, \frac{1}{V_{\max}}]$, we have

$$\begin{aligned}
\mathbb{E}_{s_h,a_h}[\exp(\lambda(R_h + V_{h+1}^*(s_{h+1})))] &= \mathbb{E}_{s_h,a_h}[\exp(\lambda R_h)]\mathbb{E}_{s_h,a_h}[\exp(\lambda V_{h+1}^*(s_{h+1}))] \\
&\le \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \exp\left(\lambda^2 \operatorname{Var}(V_{h+1}^*)(s_h, a_h)\right) \\
&= \exp\left(\frac{\lambda^2}{2}\left(\sigma^2 + 2\operatorname{Var}(V_{h+1}^*)(s_h, a_h)\right)\right) ,
\end{aligned}$$

where the first equality uses independence, and the following inequality uses the sub-Gaussianity of R_h and Lemma 50. Therefore, we have that $R_h + V_{h+1}^*(s_{h+1})$ is $\left(\sqrt{\sigma^2 + 2\operatorname{Var}(V_{h+1}^*)(s_h, a_h)}, V_{\max}\right)$-sub-exponential. Then, the total variance proxy function is bounded as

$$\begin{aligned}
W_h^\pi(s) &= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[\sum_{j=h}^{H} \left(\frac{\sigma^2}{2} + \operatorname{Var}(V_{j+1}^*)(s_j, a_j)\right)\right] \\
&= \frac{\sigma^2(H-h+1)}{2} + \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)} \left[\sum_{j=h}^{H} \operatorname{Var}(V_{j+1}^*)(s_j, a_j)\right] .
\end{aligned}$$

The expected sum is bounded similarly to the first case.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}\mathrm{Var}(V_{j+1}^*)(s_j,a_j)\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}(V_{j+1}^*(s_{j+1}))^2-\sum_{j=h}^{H}(PV_{j+1}^*(s_j,a_j))^2\right] \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}\left((V_j^*(s_j))^2-(PV_{j+1}^*(s_j,a_j))^2\right)\right]-(V_h^*(s_h))^2 \\
&= \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}(V_j^*(s_j)+PV_{j+1}^*(s_j,a_j))(V_j^*(s_j)-PV_{j+1}^*(s_j,a_j))\right]-(V_h^*(s_h))^2 \\
&\le \mathop{\mathbb{E}}_{\pi(\cdot|s_h=s)}\left[\sum_{j=h}^{H}2V_{\max}(V_j^*(s_j)-PV_{j+1}^*(s_j,a_j))\right]-(V_h^*(s_h))^2 \\
&= 2V_{\max}V_h^*(s_h)-(V_h^*(s_h))^2 \\
&\le V_{\max}^2\,,
\end{aligned}
$$

where we use that $V_j^*(s_j)-PV_{j+1}^*(s_j,a_j)\ge Q_j^*(s_j,a_j^*)-PV_{j+1}^*(s_j,a_j)=r(s_j,a_j^*)\ge 0$, where $r(s,a)\ge 0$ is guaranteed by Assumption 1 since $r(s,a)=V_H^{\pi'}(s)\ge 0$ for a policy π' with $\pi'_H(s)=a$. We have derived that $\frac{\sigma^2(H-h+1)}{2}\le W_h^\pi(s)\le\frac{\sigma^2(H-h+1)}{2}+V_{\max}^2$, hence we have $\mathbb{W}^*\le\frac{\sigma^2 H}{2}+V_{\max}^2$ and $\mathbb{W}^*_{\mathrm{diff}}\le V_{\max}^2$.

Sub-Gaussian noise MAB: The proof follows directly from the definitions.

Appendix C. Proof of Bandit Theorems in Section 6.1

In this section, we provide a bandit version of EQO+ in Algorithm 2 and prove formal versions of Theorems 4 and 6.

C.1. Proof of Theorem 4

We present a formal version of Theorem 4.

Theorem 18 (Restatement of Theorem 4) *Suppose $M\in\mathcal{B}(\sigma)$. Take $c_{1,t}=c_1$ and $c_{2,t}=\infty$ for all $t\in\mathbb{N}$ for some constant $c_1>0$. Then, we have*

$$
\mathrm{Reg}_M^{\mathsf{Alg}}(T,\delta)\le\frac{\sigma^2 T\log\frac{4}{\delta}}{2c_1}+c_1(A-1)+\sigma\sqrt{AT\log\frac{4}{\delta}}+\sum_{a\in\mathcal{A}}\mathrm{gap}(a)\,.
$$

In particular, when $c_1=\sigma\sqrt{\frac{T}{A}}$, the distributional regret is bounded as

$$
\mathrm{Reg}_M^{\mathsf{Alg}}(T,\delta)\le\frac{1}{2}\sigma\sqrt{AT}\log\frac{4e^2}{\delta}+\sigma\sqrt{AT\log\frac{4}{\delta}}+\sum_{a\in\mathcal{A}}\mathrm{gap}(a)
$$

Algorithm 2 EQO+ for bandits

Input: $\{(c_{1,k}, c_{2,k})\}_{k=1}^{\infty}$

for $t = 1, 2, \ldots, A$ **do**
  Take t-th action in $\mathcal{A}$
end
for $t = A+1, A+2, \ldots,$ **do**
  $N^t(a) \leftarrow \sum_{i=1}^{t-1} \mathbb{1}\{a_i = a\}$ for all $a \in \mathcal{A}$
  $\hat{r}^t(a) \leftarrow \frac{1}{N^t(a)} \sum_{i=1}^{t-1} R_i \mathbb{1}\{a_i = a\}$ for all $a \in \mathcal{A}$
  $b^t(a) \leftarrow \left(\frac{c_{1,t}}{N^t(a)}\right) \wedge \left(\frac{c_{2,t}}{\sqrt{N^t(a)}}\right)$ for all $a \in \mathcal{A}$
  $a_t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \hat{r}^t(a) + b^t(a)$
  Take action a_t and observe reward R_t
end

and the expected regret bound is bounded as

$$\mathbb{E}[\mathrm{Reg}_M^{\mathsf{Alg}}(T)] \leq 4\sigma\sqrt{AT} + \sum_{a \in \mathcal{A}} \mathrm{gap}(a)\,.$$

Remark 19 *In Theorem 4, we further bounded* $\sigma\sqrt{AT\log\frac{4}{\delta}} \leq \frac{\sigma^2 T \log\frac{4}{\delta}}{2c_1} + \frac{c_1 A}{2}$ *using the AM-GM inequality for a simpler form. We note that* $\sigma\sqrt{AT\log\frac{4}{\delta}}$ *can be tightened to* $\sigma\sqrt{2T\log\frac{4}{\delta}} + \sigma\sqrt{A^2\log\frac{eT}{A} + A\log\frac{2}{\delta}}$*, although the order of the bound does not change.*

The high-probability event we impose is a union of three concentration inequalities, where we take a novel approach to avoid incurring a $\log A$ or $\log T$ factor.

Lemma 20 *Define* $\hat{\mathcal{E}}_1(\frac{\delta}{4})$ *as the event that*

$$r(a^*) - \hat{r}^t(a^*) \leq \frac{\sigma^2 \log\frac{4}{\delta}}{2c_1} + \frac{c_1}{N^t(a^*)}\,.$$

holds for all $t \in \mathbb{N}$*. Then,* $\mathbb{P}(\hat{\mathcal{E}}_1(\frac{\delta}{4})) \geq 1 - \frac{\delta}{4}$.

Proof Let $\{R_i^*\}_i$ be the sequence of sampled rewards from the optimal action. We apply Lemma 48 to $\{R_i^*\}_i$ with $\lambda = \frac{\log\frac{4}{\delta}}{c_1}$ and obtain that with probability at least $1 - \frac{\delta}{4}$, the following inequality holds for all $n \in \mathbb{N}$:

$$\sum_{i=1}^{n} (r(a^*) - R_i^*) \leq \frac{\sigma^2 n \log\frac{4}{\delta}}{c_1} + c_1\,.$$

Dividing both sides by n and plugging in $n = N^t(a^*)$ completes the proof. ■

Lemma 21 *For fixed $T \in \mathbb{N}$, define $\hat{\mathcal{E}}_2(\frac{\delta}{4})$ as the event that*

$$\sum_{t=1}^{T} \mathbb{1}\{a_t \neq a^*\}(R_t - r(a_t)) \leq \sigma\sqrt{2T \log \frac{4}{\delta}}$$

Then, $\mathbb{P}(\hat{\mathcal{E}}_2(\frac{\delta}{4})) \geq 1 - \frac{\delta}{4}$.

Proof We apply Lemma 49 to $X_t = \mathbb{1}\{a_t \neq a^*\}(R_t - r(a_t))$. ■

Lemma 22 *For fixed $T \in \mathbb{N}$, define $\hat{\mathcal{E}}_3(\frac{\delta}{2})$ as the event that*

$$\sum_{t \in \mathcal{T}} (r(a_t) - R_t) \leq \sigma\sqrt{|\mathcal{T}| T \log \frac{4}{\delta}}\,.$$

holds for all subsets $\mathcal{T} \subset [T]$. Then, $\mathbb{P}(\hat{\mathcal{E}}_3(\frac{\delta}{2})) \geq 1 - \frac{\delta}{2}$.

Proof For a fixed set $\mathcal{T}$, we apply Lemma 49 and obtain that

$$\sum_{t \in \mathcal{T}} (r(a_t) - R_t) \leq \sigma\sqrt{|\mathcal{T}| \log \frac{2}{\delta}}$$

holds with probability at least $1 - \frac{\delta}{2}$. We take the union bound over all $\mathcal{T}$, where there are at most 2^T subsets. Then, the logarithm term is bounded as $\log \frac{2^T \cdot 2}{\delta} \leq T \log \frac{4}{\delta}$, which completes the proof. ■

Proof [Proof of Theorem 18] For each $a \in \mathcal{A}$, let $t(a) \in [T]$ be the last time step the action a is taken, so that $a_{t(a)} = a$ and $N^{T+1}(a) = N^{t(a)} + 1$. Assuming $t(a) > A$, the following inequality must have held for the agent to take action a:

$$\hat{r}^{t(a)}(a^*) + \frac{c_1}{N^{t(a)}(a^*)} \leq \hat{r}^{t(a)}(a) + \frac{c_1}{N^{t(a)}(a)}\,.$$

Under $\hat{\mathcal{E}}_1(\frac{\delta}{4})$, we have $r(a^*) - \hat{r}^{t(a)}(a^*) \leq \frac{\sigma^2 \log \frac{4}{\delta}}{2c_1} + \frac{c_1}{N^{t(a)}(a^*)}$, which implies that

$$r(a^*) \leq \frac{\sigma^2 \log \frac{4}{\delta}}{2c_1} + \hat{r}^{t(a)}(a^*) + \frac{c_1}{N^{t(a)}(a^*)} \leq \frac{\sigma^2 \log \frac{4}{\delta}}{2c_1} + \hat{r}^{t(a)}(a) + \frac{c_1}{N^{t(a)}(a)}\,.$$

Subtracting both sides by $r(a)$, we obtain that

$$\mathrm{gap}(a) \leq \frac{\sigma^2 \log \frac{4}{\delta}}{2c_1} + \frac{c_1}{N^{t(a)}(a)} + (\hat{r}^{t(a)} - r)(a)\,.$$

We multiply both sides by $N^{t(a)}(a)$.

$$\begin{aligned}
\mathrm{gap}(a) N^{t(a)}(a) &\leq \frac{\sigma^2 N^{t(a)}(a) \log \frac{4}{\delta}}{2c_1} + c_1 + (\hat{r}^{t(a)} - r)(a) N^{t(a)}(a) \\
&= \frac{\sigma^2 N^{t(a)}(a) \log \frac{3}{\delta}}{2c_1} + c_1 + \sum_{t=1}^{t(a)-1} \mathbb{1}\{a_t = a\}(R_t - r(a))\,.
\end{aligned}$$

Note that this inequality also holds when $t(a) \leq A$, since it implies that the action a is taken only once and hence $N^{t(a)}(a) = 0$. Taking the sum over $a \in \mathcal{A} \setminus \{a^*\}$, we obtain that

$$\sum_{a \in \mathcal{A} \setminus \{a^*\}} \mathrm{gap}(a) N^{t(a)}(a) \leq \frac{\sigma^2 T \log \frac{4}{\delta}}{2c_1} + c_1(A-1) + \sum_{t=1}^{T} \mathbb{1}\{a_t \neq a^*, t \neq t(a_t)\}(R_t - r(a_t))\,,$$

where we use that $\sum_{a \in \mathcal{A} \setminus \{a^*\}} N^{t(a)}(a) \leq T$. We bound the last sum as follows:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{1}\{a_t \neq a^*, t \neq t(a_t)\}(R_t - r(a_t)) \\
&= \sum_{t=1}^{T} \mathbb{1}\{a_t \neq a^*\}(R_t - r(a_t)) - \sum_{t \in \{t(a) | a \in \mathcal{A} \setminus \{a^*\}\}} (R_t - r(a_t)) \\
&\leq \sigma \sqrt{2T \log \frac{4}{\delta}} + \sigma \sqrt{(A-1) T \log \frac{4}{\delta}} \\
&\leq 2\sigma \sqrt{AT \log \frac{4}{\delta}}\,,
\end{aligned}$$

where the first inequality holds under $\hat{\mathcal{E}}_2(\frac{\delta}{4})$ and $\hat{\mathcal{E}}_3(\frac{\delta}{2})$. Therefore, under an event whose probability is at least $1 - \delta$, we have

$$\begin{aligned}
\mathrm{Reg}_M^{\mathsf{Alg}}(T) &= \sum_{a \in \mathcal{A} \setminus \{a^*\}} \mathrm{gap}(a) N^{T+1}(a) \\
&= \sum_{a \in \mathcal{A} \setminus \{a^*\}} \mathrm{gap}(a)(N^{t(a)}(a) + 1) \\
&\leq \frac{\sigma^2 T \log \frac{4}{\delta}}{2c_1} + c_1(A-1) + 2\sigma \sqrt{AT \log \frac{4}{\delta}} + \sum_{a \in \mathcal{A}} \mathrm{gap}(a)\,.
\end{aligned}$$

■

C.2. Proof of Theorem 6

Recall that we defined a time-uniform UCB as a function $u(t, \delta) : \mathbb{R}_{\geq 0} \times (0, 1] \to \mathbb{R}_{\geq 0}$ that satisfies

$$\mathbb{P}\left(\exists t > A : |\hat{r}^t(a) - r(a)| > \frac{u(t, \delta)}{\sqrt{N^t(a)}}\right) \leq \delta\,.$$

A sensible concentration inequality would yield a form of $u(t, \delta) = \sigma \sqrt{f(t) \log \frac{1}{\delta} + g(t)}$ for some real-valued functions $f(t)$ and $g(t)$, where we omit dependence on A. We will say $u(t, \delta)$ is (asymptotically) *tight* if we have $\lim_{t \to \infty} f(t) = 2$ and $\lim_{t \to \infty} \frac{g(t)}{\log t} = 0$, which is tight in the sense of meeting the lower bound of Lai and Robbins (1985) by the following theorem. An example of such a concentration inequality is provided in Lemma 23.

Lemma 23 *The following function $u(t,\delta)$ is a tight time-uniform UCB:*

$$u(t,\delta) = \sigma\sqrt{2\left(1+\frac{1}{\log(1+t)}\right)\left(\log\frac{1}{\delta}+\log A+6\log\log(1+t)+8\right)}\,.$$

Proof Apply Lemma 54 with $\eta = \frac{1}{\log(1+t)}$. Then, for fixed $a \in \mathcal{A}$, with probability at least $1-\delta$, it holds that for all $t \in \mathbb{N}$,

$$\begin{aligned}&\left|\hat{r}^t(a) - r(a)\right| \\ &\le \sigma\sqrt{2\left(1+\frac{1}{\log(1+t)}\right)\left(\log\frac{1}{\delta}+\log 4\left(2+\log\log(1+t)\right)^2\left(1+2e(\log(1+t))(\log N^t(a))\right)^2\right)}\,.\end{aligned}$$

We bound $\log N^t(a) \le \log(1+t)$. We further modify the logarithmic term as follows:

$$\begin{aligned}&\log 4\left(2+\log\log(1+t)\right)^2\left(1+2e(\log(1+t))(\log N^t(a))\right)^2 \\ &\le \log 4 + 2\log(2+\log\log(1+t)) + 2\log(1+2e(\log(1+t))^2)\,.\end{aligned}$$

For the second term, we use $\log(1+x) \le x$ and bound

$$2\log(2+\log\log(1+t)) \le 2(1+\log\log(1+t))\,.$$

For the third term, we have

$$\begin{aligned}2\log(1+2e(\log(1+t))^2) &\le 2\log\left(\left(\frac{1}{(\log 2)^2}+2e\right)(\log(1+t))^2\right) \\ &\le 4\log\log(1+t) + 2\log 8\,.\end{aligned}$$

Taking the sum of three terms, the logarithmic term is upper bounded by $6\log\log t + 8$. By taking the union bound over $a \in \mathcal{A}$, we incur an additional $\log A$ factor and fully recover $u(t,\delta)$.
We have shown that $u(t,\delta)$ is a time-uniform UCB, and the fact that it is tight follows from direct computation. ■

Theorem 24 (Formal extension of Theorem 6) *Suppose $M \in \mathcal{B}(\sigma)$. Suppose $u(t,\delta)$ is a time-uniform UCB. For any decreasing sequence δ_t, let $c_{1,t} = \infty$ and $c_{2,t} = u(t,\delta_t)$ for all $t \in \mathbb{N}$ Let $\tau_2(\delta) := \mathrm{argmax}_{t\in\mathbb{N}}\,\delta_t > \delta$. Then, we have*

$$\mathrm{Reg}_M^{\mathsf{Alg}}(T,\delta) \le \tau_2(\delta) \wedge T + \sum_{\substack{a\in\mathcal{A}\\ \mathrm{gap}(a)\neq 0}} \mathrm{gap}(a) + \sum_{\substack{a\in\mathcal{A}\\ \mathrm{gap}(a)\neq 0}} \frac{\left(c_{2,T}+u(T,\delta)\right)^2}{\mathrm{gap}(a)}\,.$$

If $u(t,\delta)$ admits a form of $\sigma\sqrt{f(t)\log\frac{1}{\delta}+g(t)}$ for some functions f and g, then the expected regret is bounded as

$$\mathbb{E}[\mathrm{Reg}_M^{\mathsf{Alg}}(T)] \le \sum_{t=1}^{T}\delta_t + \sum_{a\in\mathcal{A}}\mathrm{gap}(a) + \sum_{\substack{a\in\mathcal{A}\\ \mathrm{gap}(a)\neq 0}} \frac{\left(c_{2,T}+u(T,e^{-1})\right)^2}{\mathrm{gap}(a)}\,.$$

Furthermore, if $\delta_t = o(t^{-1})$ and $\delta_t = \omega(t^{-\alpha})$ for all $\alpha > 1$, e.g., $\delta_t = (t \log t)^{-1}$, and if $u(t, \delta)$ is tight, then we have

$$\limsup_{T \to \infty} \frac{\mathbb{E}[\mathrm{Reg}_M^{\mathsf{Alg}}(T)]}{\log T} \leq \sum_{\substack{a \in \mathcal{A} \\ \mathrm{gap}(a) \neq 0}} \frac{2\sigma^2}{\mathrm{gap}(a)} .$$

Proof Fix $\delta \in (0, 1]$ and assume the event that $|\hat{r}^t(a) - r(a)| \leq \frac{u(t,\delta)}{\sqrt{N^t(a)}}$ holds for all $a \in \mathcal{A}$ and $t \in \mathbb{N}$, whose probability is at least $1 - \delta$ by the definition of time-uniform UCB. Suppose $t > \tau_2(\delta)$, which implies $\delta \geq \delta_t$ and consequently $c_{2,t} \geq u(t, \delta)$. Then, the UCB estimate of the optimal arm is at least its true mean as

$$r(a^*) \leq \hat{r}^t(a^*) + \frac{u(t,\delta)}{\sqrt{N^t(a^*)}} \leq \hat{r}^t(a^*) + \frac{c_{2,t}}{\sqrt{N^t(a^*)}} .$$

Hence, for a sub-optimal arm a to be taken, it must satisfy

$$r(a^*) \leq \hat{r}^t(a^*) + \frac{c_{2,t}}{\sqrt{N^t(a^*)}} \leq \hat{r}^t(a) + \frac{c_{2,t}}{\sqrt{N^t(a)}} .$$

Subtracting $r(a)$ from both sides, we have

$$\begin{aligned} \mathrm{gap}(a) &= r(a^*) - r(a) \\ &\leq \hat{r}^t(a) - r(a) + \frac{c_{2,t}}{\sqrt{N^t(a)}} \\ &\leq \frac{u(t,\delta)}{\sqrt{N^t(a)}} + \frac{c_{2,t}}{\sqrt{N^t(a)}} . \end{aligned}$$

This inequality implies that if an action $a \in \mathcal{A}$ is taken at a time step $t > \tau_2(\delta)$, then it satisfies $N^t(a) \leq \frac{(u(t,\delta)+c_{2,t})^2}{\mathrm{gap}(a)^2}$. Fix T and denote the last time step an action a is taken by $t(a)$. Then, we have either $t(a) \leq \tau_2(\delta)$ or $N^{t(a)} \leq \frac{(u(T,\delta)+c_{2,t})^2}{\mathrm{gap}(a)^2}$. Noting that $N^{t(a)}(a) + 1 = N^{T+1}(a)$ by the definition of $t(a)$, we have that

$$\begin{aligned} \mathrm{Reg}_M^{\mathsf{Alg}}(T, \delta) &= \sum_{\substack{a \in \mathcal{A} \\ \mathrm{gap}(a) \neq 0}} \mathrm{gap}(a) N^{T+1}(a) \\ &= \sum_{\substack{a \in \mathcal{A} \\ \mathrm{gap}(a) \neq 0 \\ t(a) \leq \tau_2(\delta)}} \mathrm{gap}(a) N^{T+1}(a) + \sum_{\substack{a \in \mathcal{A} \\ \mathrm{gap}(a) \neq 0 \\ t(a) > \tau_2(\delta)}} \mathrm{gap}(a) N^{T+1}(a) \\ &\leq \tau_2(\delta) \wedge T + \sum_{\substack{a \in \mathcal{A} \\ \mathrm{gap}(a) \neq 0 \\ t(a) > \tau_2(\delta)}} \left(\mathrm{gap}(a) + \frac{(c_{2,T} + u(T,\delta))^2}{\mathrm{gap}(a)} \right) \\ &\leq \tau_2(\delta) \wedge T + \sum_{a \in \mathcal{A}} \mathrm{gap}(a) + \sum_{\substack{a \in \mathcal{A} \\ \mathrm{gap}(a) \neq 0}} \frac{(c_{2,T} + u(T,\delta))^2}{\mathrm{gap}(a)} . \end{aligned}$$

This proves the first part of the lemma.
The second part of the lemma is simply taking the integral of the first bound with respect to δ. First, note that $(\tau_2(\delta) \wedge T) - (\tau_2(\delta) \wedge (T-1)) = \mathbb{1}\{\tau_2(\delta) \geq T\} = \mathbb{1}\{\delta_T > \delta\}$. Integrating this indicator function, we have $\int_0^1 \mathbb{1}\{\delta_T > \delta\}\, d\delta = \delta_T$, and hence we obtain that

$$\int_0^1 \tau_2(\delta) \wedge T = \sum_{t=1}^{T} \delta_t \,.$$

To integrate $(c_{2,T} + u(T,\delta))^2$, we expand it as follows:

$$\int_0^1 (c_{2,T} + u(T,\delta))^2 \, d\delta = \int_0^1 c_{2,T}^2 + 2c_{2,T}u(T,\delta) + (u(T,\delta))^2 \, d\delta \,.$$

Using that $u(T,\delta) = \sigma\sqrt{f(T)\log\frac{1}{\delta} + g(T)}$, we have

$$\begin{aligned}
\int_0^1 (u(T,\delta))^2 \, d\delta &= \int_0^1 \sigma^2 \left(f(T)\log\frac{1}{\delta} + g(T) \right) d\delta \\
&= \sigma^2 \left(f(T) + g(T) \right) \\
&= (u(T, e^{-1}))^2 \,,
\end{aligned}$$

where we use that $\int_0^1 \log\frac{1}{\delta}\, d\delta = 1$. We also have

$$\begin{aligned}
\int_0^1 u(T,\delta)\, d\delta &\leq \sqrt{\int_0^1 (u(T,\delta))^2 \, d\delta} \\
&\leq u(T, e^{-1}) \,,
\end{aligned}$$

where we use the Cauchy-Schwarz inequality for the first inequality. Therefore, we obtain that

$$\begin{aligned}
\int_0^1 (c_{2,T} + u(T,\delta))^2 \, d\delta &\leq c_{2,T}^2 + 2c_{2,T}u(T, e^{-1}) + (u(T, e^{-1}))^2 \\
&= (c_{2,T} + u(T, e^{-1}))^2 \,.
\end{aligned}$$

This proves the second part of the lemma.
For the last part, $\delta_t = o(t^{-1})$ implies that $\lim_{T\to\infty} \frac{1}{\log T} \sum_{t=1}^{T} \delta_t = 0$. By the condition of tightness, we have $\lim_{T\to\infty} \frac{u(T,e^{-1})}{\log T} = 0$. The condition $\delta_t = \omega(t^{-\alpha})$ for all $\alpha > 1$ implies that $\lim_{T\to\infty} \frac{\log\frac{1}{\delta_T}}{\log T} = 1$, which implies that

$$\begin{aligned}
\limsup_{T\to\infty} \frac{c_{2,T}^2}{\log T} &= \limsup_{T\to\infty} \frac{\sigma^2(f(T)\log\frac{1}{\delta_T} + g(T))}{\log T} \\
&= 2\sigma^2 \,.
\end{aligned}$$

We note that the limits of $c_{2,T}$ and $u(T,\delta)$ imply that $\lim_{T\to\infty} \frac{c_{2,T}u(T,\delta)}{\log T} = 2\sigma^2 \cdot 0 = 0$. Therefore, we conclude that

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_M^{\mathsf{Alg}}(T)]}{\log T} \\
&\le \limsup_{T\to\infty} \frac{1}{\log T}\left(\sum_{t=1}^{T}\delta_t + \sum_{a\in\mathcal{A}}\mathrm{gap}(a) + \sum_{\substack{a\in\mathcal{A}\\ \mathrm{gap}(a)\neq 0}} \frac{1}{\mathrm{gap}(a)}\left(c_{2,T}+u(T,e^{-1})\right)^2\right) \\
&= \limsup_{T\to\infty} \frac{1}{\log T} \sum_{\substack{a\in\mathcal{A}\\ \mathrm{gap}(a)\neq 0}} \frac{c_{2,T}^2}{\mathrm{gap}(a)} \\
&= \sum_{\substack{a\in\mathcal{A}\\ \mathrm{gap}(a)\neq 0}} \frac{2\sigma^2}{\mathrm{gap}(a)}\,.
\end{aligned}$$

■

Appendix D. Proof of RL Theorems in Section 6.2

In this section, we prove Theorems 8 and 9. In Appendix D.1, we provide and prove high-probability events that constitute $\mathcal{E}(\delta)$, which is the event that the distributional regret bounds hold. In Appendix D.2, we provide general properties of EQO+, including the quasi-optimism lemma (Lemma 31, the formal version of Lemma 14) and its proof. In Appendix D.3, we state and prove Theorem 34, which is the formal version of Theorem 8. In Appendix D.4, we state and prove Theorem 37, which is the formal version of Theorem 9.

D.1. High-probability Events

In this section, we provide high-probability events under which the analysis is conducted. In Appendix D.1.1, we define the events and provide lemmas that state these events occur with high probabilities. The proofs of those lemmas are provided in Appendices D.1.2–D.1.4.

D.1.1. Definition of Events

Lemma 25 *Define $\mathcal{E}_1(\frac{\delta}{4})$ as the event that*

$$\begin{aligned}
&\left|\hat{r}^k(s,a) + \hat{P}^k V^*_{h+1}(s,a) - Q^*_{h+1}(s,a)\right| \wedge V_{\max} \\
&\qquad \le \frac{\lambda_{\iota_k}\sigma^2_{\exp}(h,s,a)}{2} + \frac{\ell_1(\iota_k,\delta)}{\lambda_{\iota_k}N^k(s,a)} \wedge \sigma_{\max}\sqrt{\frac{\ell_2(k,\delta)}{N^k(s,a)}}
\end{aligned}$$

holds for all $h\in[H]$, $(s,a)\in\mathcal{S}\times\mathcal{A}$, and $k\in\mathbb{N}$. Then, we have $\mathbb{P}(\mathcal{E}_1(\frac{\delta}{4}))\ge 1-\frac{\delta}{4}$.

The proof is presented in Appendix D.1.2.

Lemma 26 *Define $\mathcal{E}_2(\frac{\delta}{4})$ as the event that*

$$(\hat{P}^k - P)\left(2W^{\pi^*}_{h+1} - \frac{1}{2\mathbb{W}^*_{\text{diff}}}\left(W^{\pi^*}_{h+1,\text{diff}}\right)^2\right)(s,a)$$
$$\leq \frac{1}{2\mathbb{W}^*_{\text{diff}}}\text{Var}(W^{\pi^*}_{h+1,\text{diff}})(s,a) + \frac{13\mathbb{W}^*_{\text{diff}}\ell_1(1,\delta)}{N^k(s,a)} \wedge \frac{3\mathbb{W}^*_{\text{diff}}}{2}\sqrt{\frac{\ell_2(k,\delta)}{N^k(s,a)}}$$

holds for all $k \in \mathbb{N}$, $h \in [H]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then, we have $\mathbb{P}(\mathcal{E}_2(\frac{\delta}{4})) \geq 1 - \frac{\delta}{4}$.

The proof is presented in Appendix D.1.3.

Lemma 27 (Lemma 29 in Lee and Oh (2025)) *There exists an event $\mathcal{E}_3(\frac{\delta}{4})$ whose probability is at least $1 - \frac{\delta}{4}$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $k \in \mathbb{N}$, and constants $c, \rho > 0$, the following inequality holds for any functions $V : \mathcal{S} \to [-c, c]$ under $\mathcal{E}_3(\frac{\delta}{4})$:*

$$\left|(\hat{P}^k - P)V(s,a)\right| \leq \frac{1}{c\rho}\text{Var}(V)(s,a) + \frac{c(\rho + \frac{2}{3})S\ell_2(k,\delta)}{N^k(s,a)}.$$

Lemma 28 *Suppose $\{X^k_h\}_{k,h}$ is a sequence of non-negative random variables adapted to filtration $\{\mathcal{F}^k_h\}_{k,h}$ for $k \in \mathbb{N}$ and $h \in [H]$. Let $c > 0$ be a constant and $\{J^k_h\}_{k,h}$ a sequence of random variables recursively defined as $J^k_{H+1} := 0$ and*

$$J^k_h := \left(X^k_h + \mathbb{E}[J^k_{h+1} \mid \mathcal{F}^k_h]\right) \wedge c$$

for all $k \in \mathbb{N}$ and $h \in [H]$. Recall that η^k is a stopping time defined in Appendix A. Let $\mathcal{E}_4(\{X^k_h\}_{k,h}, c, \frac{\delta}{4})$ be the event that

$$\sum_{k=1}^{K} J^k_1 \leq 2\sum_{K=1}^{K}\sum_{h=1}^{\eta^k - 1} X^k_h + 6cSA\log\frac{16H}{\delta}$$

holds for all $K \in \mathbb{N}$. Then, $\mathbb{P}(\mathcal{E}_4(\{X^k_h\}_{k,h}, c, \frac{\delta}{4})) \geq 1 - \frac{\delta}{4}$.

The proof is presented in Appendix D.1.4.

D.1.2. PROOF OF LEMMA 25

Proof [Proof of Lemma 25] Fix $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ and $\delta' \in (0,1]$. Let $\{(R^j, s^j)\}_j$ be the sequence of the observed reward and next state pairs when (s,a) is selected. By Assumption 2, $\{(R^j + V^*_{h+1}(s^j) - Q^*_h(s,a)\}_j$ is $(\sigma_{\text{exp}}(h,s,a), V_\alpha)$-sub-exponential conditioned on the previous observations. For fixed $i \in \mathbb{N}$, the following inequality holds for all $n \in \mathbb{N}$ with probability at least $1 - \frac{\delta'}{2i^2}$ by Lemma 48:

$$\left|\sum_{j=1}^{n}\left(R^j + V^*_{h+1}(s^j) - Q^*_{h+1}(s,a)\right)\right| \leq \frac{\lambda_i\sigma^2_{\text{exp}}(h,s,a)n}{2} + \frac{1}{\lambda_i}\log\frac{4i^2}{\delta'},$$

Taking the union bound and using that $\sum_{i=1}^{\infty} \frac{1}{2i^2} \leq 1$, the inequality above holds for all $i \in \mathbb{N}$ with probability at least $1 - \delta'$. By Lemma 52, the following inequality also holds for all $n \in \mathbb{N}$ with probability at least $1 - \delta'$:

$$\left| \sum_{j=1}^{n} \left(R^j + V_{h+1}^*(s^j) - Q_{h+1}^*(s,a) \right) \right| \wedge n V_{\max} \leq \sigma_{\max} \sqrt{n \log \frac{4(\log e^2 n)^2}{\delta'}}$$

Then, with probability at least $1 - 2\delta'$, the following inequality holds for all $i \in \mathbb{N}$ and $n \in \mathbb{N}$:

$$\begin{aligned}
&\left| \sum_{j=1}^{n} \left(R^j + V_{h+1}^*(s^j) - Q_{h+1}^*(s,a) \right) \right| \wedge n V_{\max} \\
&\leq \left(\frac{\lambda_i \sigma_{\exp}^2(h,s,a) n}{2} + \frac{1}{\lambda_i} \log \frac{4i^2}{\delta'} \right) \wedge \sigma_{\max} \sqrt{n \log \frac{4(\log e^2 n)^2}{\delta'}} \\
&\leq \frac{\lambda_i \sigma_{\exp}^2(h,s,a) n}{2} + \left(\frac{1}{\lambda_i} \log \frac{4i^2}{\delta'} \right) \wedge \sigma_{\max} \sqrt{n \log \frac{4(\log e^2 n)^2}{\delta'}} \,,
\end{aligned}$$

where the last inequality uses that $\frac{\lambda_i \sigma_{\exp}^2(h,s,a) n}{2} \geq 0$. Dividing both sides by n, plugging in $n = N^k(s,a)$ and $i = \iota_k$, and noting that $\frac{1}{N^k(s,a)} \sum_{j=1}^{N^k(s,a)} R^j = \hat{r}^k(s,a)$ and $\frac{1}{N^k(s,a)} \sum_{j=1}^{N^k(s,a)} V_{h+1}^*(s^j) = \hat{P}^k V_{h+1}^*(s,a)$, we have

$$\begin{aligned}
&\left| \hat{r}^k(s,a) + \hat{P}^k V_{h+1}^*(s,a) - Q_{h+1}^*(s,a) \right| \wedge V_{\max} \\
&\leq \frac{\lambda_{\iota_k} \sigma_{\exp}^2(h,s,a)}{2} + \left(\frac{1}{\lambda_{\iota_k} N^k(s,a)} \log \frac{4{\iota_k}^2}{\delta'} \right) \wedge \sigma_{\max} \sqrt{\frac{\log \frac{4(\log e^2 N^k(s,a))^2}{\delta'}}{N^k(s,a)}} \,.
\end{aligned}$$

The proof is completed by bounding $\log N^k(s,a) \leq \log kH$, taking the union bound over $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, and plugging in $\delta' = \frac{\delta}{8HSA}$. ■

D.1.3. PROOF OF LEMMA 26

Proof [Proof of Lemma 26] Fix $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ and $\delta' \in (0,1]$. For simplicity, we define $W(s') := 2W_{h+1,\text{diff}}^{\pi^*}(s') - \frac{1}{2\mathbb{W}_{\text{diff}}^*} \left(W_{h+1,\text{diff}}^{\pi^*} \right)^2 (s')$ for $s' \in \mathcal{S}$. First, note that $W_{h+1}^{\pi^*}(s') = W_{h+1,\text{diff}}^{\pi^*}(s') + W_{h+1,\min}^{\pi^*}$ and $W_{h+1,\min}^{\pi^*}$ does not depend on s', so we have

$$\begin{aligned}
&(\hat{P}^k - P) \left(2W_{h+1}^{\pi^*} - \frac{1}{2\mathbb{W}_{\text{diff}}^*} \left(W_{h+1,\text{diff}}^{\pi^*} \right)^2 \right) (s,a) \\
&= (\hat{P}^k - P) \left(2W_{h+1,\text{diff}}^{\pi^*} - \frac{1}{2\mathbb{W}_{\text{diff}}^*} \left(W_{h+1,\text{diff}}^{\pi^*} \right)^2 \right) (s,a) + (\hat{P}^k - P) W_{h+1,\min}^{\pi^*} \\
&= (\hat{P}^k - P) W(s,a) + 0 \,.
\end{aligned}$$

Now, we bound $(\hat{P}^k - P)W(s,a)$ in two ways and take the minimum. Note that by $0 \leq W_{h+1,\text{diff}}^{\pi^*}(s') \leq \mathbb{W}_{\text{diff}}^*$, we have $0 \leq W(s') \leq \frac{3\mathbb{W}_{\text{diff}}^*}{2}$ for all $s' \in \mathcal{S}$. By Hoeffding's lemma (Lemma 62), $W(s')$ is

$(\frac{3\mathbb{W}^*_{\text{diff}}}{4}, 0)$-sub-exponential, and by Lemma 52, we have that with probability at least $1-\delta'$, for all $k \in \mathbb{N}$,

$$(\hat{P}^k - P)W(s,a) \le \frac{3\mathbb{W}^*_{\text{diff}}}{2}\sqrt{\frac{1}{N^k(s,a)}\log\frac{2(\log e^2 N^k(s,a))^2}{\delta'}} \le \frac{3\mathbb{W}^*_{\text{diff}}}{2}\sqrt{\frac{1}{N^k(s,a)}\log\frac{2(\log e^2 kH)^2}{\delta'}}, \tag{3}$$

where we bound $\log e^2 N^k(s,a) \le \log e^2 kH$. Next, we apply Freedman's inequality. We obtain a bound on the variance $\text{Var}_{s'\sim P(\cdot|s,a)}(W(s'))$ by Lemma 58 as follows:

$$\begin{aligned}
\underset{s'\sim P(\cdot|s,a)}{\text{Var}}\left(W(s')\right) &= \underset{s'\sim P(\cdot|s,a)}{\text{Var}}\left(2W^{\pi^*}_{h+1,\text{diff}}(s') - \frac{1}{2\mathbb{W}^*_{\text{diff}}}\left(W^{\pi^*}_{h+1,\text{diff}}\right)^2(s')\right)\\
&= \underset{s'\sim P(\cdot|s,a)}{\text{Var}}\left(W^{\pi^*}_{h+1,\text{diff}}(s')\left(2 - \frac{1}{2\mathbb{W}^*_{\text{diff}}}W^{\pi^*}_{h+1,\text{diff}}(s')\right)\right)\\
&\le 2\underset{s'\sim P(\cdot|s,a)}{\text{Var}}\left(W^{\pi^*}_{h+1,\text{diff}}(s')\right)\cdot 2^2 + 2{\mathbb{W}^*_{\text{diff}}}^2\underset{s'\sim P(\cdot|s,a)}{\text{Var}}\left(2 - \frac{1}{2\mathbb{W}^*_{\text{diff}}}W^{\pi^*}_{h+1,\text{diff}}(s')\right)\\
&= \frac{17}{2}\underset{s'\sim P(\cdot|s,a)}{\text{Var}}\left(W^{\pi^*}_{h+1,\text{diff}}(s')\right)
\end{aligned}$$

From Lemma 50, we obtain that $W(s') - PW(s,a)$ is $(13\,\text{Var}(W^{\pi^*}_{h+1,\text{diff}})(s,a), \frac{3}{2}\mathbb{W}^*_{\text{diff}})$-sub-exponential, where the constant 13 comes from $(e-2)\cdot 2\cdot\frac{17}{2} \le 13$. Applying Lemma 48 with $\lambda = \frac{1}{13\mathbb{W}^*_{\text{diff}}}$, we have that for all $k \in \mathbb{N}$, the following inequality holds with probability at least $1-\delta'$:

$$(\hat{P}^k - P)W(s,a) \le \frac{1}{2\mathbb{W}^*_{\text{diff}}}\text{Var}(W^{\pi^*}_{h+1,\text{diff}})(s,a) + \frac{13\mathbb{W}^*_{\text{diff}}}{N^k(s,a)}\log\frac{1}{\delta'}. \tag{4}$$

Taking the minimum over the bounds of Eq. (3) and Eq. (4), and by the union bound, we obtain that for all $k \in \mathbb{N}$, with probability at least $1-2\delta'$, we have

$$\begin{aligned}
&(\hat{P}^k - P)W(s,a)\\
&\le \left(\frac{1}{2\mathbb{W}^*_{\text{diff}}}\text{Var}(W^{\pi^*}_{h+1,\text{diff}})(s,a) + \frac{13\mathbb{W}^*_{\text{diff}}}{N^k(s,a)}\log\frac{1}{\delta'}\right) \wedge \frac{3\mathbb{W}^*_{\text{diff}}}{2}\sqrt{\frac{1}{N^k(s,a)}\log\frac{2(\log e^2 kH)^2}{\delta'}}\\
&\le \frac{1}{2\mathbb{W}^*_{\text{diff}}}\text{Var}(W^{\pi^*}_{h+1,\text{diff}})(s,a) + \left(\frac{13\mathbb{W}^*_{\text{diff}}}{N^k(s,a)}\log\frac{1}{\delta'}\right) \wedge \frac{3\mathbb{W}^*_{\text{diff}}}{2}\sqrt{\frac{1}{N^k(s,a)}\log\frac{2(\log e^2 kH)^2}{\delta'}},
\end{aligned}$$

where the last inequality uses that $\frac{1}{2\mathbb{W}^*_{\text{diff}}}\text{Var}(W^{\pi^*}_{h+1,\text{diff}})(s,a) \ge 0$. The proof is completed by plugging in $\delta' = \frac{\delta}{8HSA}$ and taking the union bound over $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$. ■

D.1.4. PROOF OF LEMMA 28

Before proving Lemma 28, we present the following lemma. This lemma is a minor generalization of Lemma 15 in Lee and Oh (2025), based on an observation that its proof holds for any random variables adapted to $\{\mathcal{F}_h^k\}_{k,h}$. While the proof directly follows that of Lee and Oh (2025), we provide it here for completeness.

Lemma 29 (Restatement of Lemma 15) *Let $\{X_h^k\}_{k,h}$ be a sequence of non-negative random variables adapted to filtration $\{\mathcal{F}_h^k\}_{k,h}$. Let $c > 0$ be a constant. Recursively define $\{J_h^k\}_{h,k}$ as $J_{H+1}^k := 0$ and*

$$J_h^k := \left(X_h^k + \mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k]\right) \wedge c$$

for all $k \in \mathbb{N}$ and $h \in [H]$. Then, for any $\delta \in (0, 1]$, the following inequality holds for all $K \in \mathbb{N}$ with probability at least $1 - \delta$:

$$\sum_{k=1}^{K} J_1^k \leq 2 \sum_{k=1}^{K} \sum_{h=1}^{H} X_h^k + 6c \log \frac{2}{\delta}.$$

Proof [Proof of Lemma 29] We expand J_1^k as follows:

$$\begin{aligned} J_1^k &= \sum_{h=1}^{H} J_h^k - \sum_{h=2}^{H} J_h^k \\ &\leq \sum_{h=1}^{H} \left(X_h^k + \mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k]\right) - \sum_{h=2}^{H} J_h^k \\ &= \sum_{h=1}^{H} X_h^k + \sum_{h=1}^{H} \left(\mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k] - J_{h+1}^k\right). \end{aligned}$$

Taking the sum over $k \in [K]$, we obtain

$$\sum_{k=1}^{K} J_1^k \leq \sum_{k=1}^{K} \sum_{h=1}^{H} X_h^k + \sum_{k=1}^{K} \sum_{h=1}^{H} \left(\mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k] - J_{h+1}^k\right). \tag{5}$$

Note that $\{\mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k] - J_h^k\}_{h,k}$ is a martingale sequence with a finite range $[-c, c]$. By Lemma 51 with $\lambda = \frac{1}{4(e-2)c}$, the following inequality holds for all $K \in \mathbb{N}$ with probability at least $1 - \frac{\delta}{2}$:

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left(\mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k] - J_{h+1}^k\right) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{4c} \operatorname{Var}(J_{h+1}^k \mid \mathcal{F}_h^k) + 4(e-2)c \log \frac{2}{\delta}. \tag{6}$$

By Lemma 55, we have $\mathrm{Var}(J_{h+1}^k \mid \mathcal{F}_h^k) \le \mathbb{E}[(J_{h+1}^k)^2 \mid \mathcal{F}_h^k] - (J_h^k)^2 + 2cX_h^k$, hence the sum of the variances is bounded as

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \mathrm{Var}(J_{h+1}^k \mid \mathcal{F}_h^k) \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \left(\mathbb{E}[(J_{h+1}^k)^2 \mid \mathcal{F}_h^k] - (J_h^k)^2 + 2cX_h^k\right) \\
&\le \sum_{k=1}^{K} \left((J_{H+1}^k)^2 - (J_1^k)^2\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \left(\mathbb{E}[(J_{h+1}^k)^2 \mid \mathcal{F}_h^k] - (J_{h+1}^k)^2\right) + 2c\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \left(\mathbb{E}[(J_{h+1}^k)^2 \mid \mathcal{F}_h^k] - (J_{h+1}^k)^2\right) + 2c\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k\,. \qquad (7)
\end{aligned}$$

Note that $\{\mathbb{E}[(J_{h+1}^k)^2 \mid \mathcal{F}_h^k] - (J_{h+1}^k)^2\}_{k,h}$ is a martingale difference sequence with a finite range $[-c^2, c^2]$. Again, by Lemma 51 with $\lambda = \frac{1}{8(e-2)c^2}$, the following inequality holds for all $k \in K$ with probability at least $1 - \frac{\delta}{2}$:

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \left(\mathbb{E}[(J_{h+1}^k)^2 \mid \mathcal{F}_h^k] - (J_{h+1}^k)^2\right) \le \frac{1}{8c^2}\sum_{k=1}^{K}\sum_{h=1}^{H} \mathrm{Var}((J_{h+1}^k)^2 \mid \mathcal{F}_h^k) + 8(e-2)c^2 \log\frac{2}{\delta}\,. \qquad (8)$$

By Lemma 58, we have $\mathrm{Var}((J_{h+1}^k)^2 \mid \mathcal{F}_h^k) \le 4c^2\,\mathrm{Var}(J_{h+1}^k \mid \mathcal{F}_h^k)$. Combining this bound together with inequalities (7) and (8), we obtain that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \mathrm{Var}(J_{h+1}^k \mid \mathcal{F}_h^k) \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \left(\mathbb{E}[(J_{h+1}^k)^2 \mid \mathcal{F}_h^k] - (J_{h+1}^k)^2\right) + 2c\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k \\
&\le \frac{1}{8c^2}\sum_{k=1}^{K}\sum_{h=1}^{H} \mathrm{Var}((J_{h+1}^k)^2 \mid \mathcal{F}_h^k) + 8(e-2)c^2\log\frac{2}{\delta} + 2c\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k \\
&\le \frac{1}{2}\sum_{k=1}^{K}\sum_{h=1}^{H} \mathrm{Var}(J_{h+1}^k \mid \mathcal{F}_h^k) + 2c\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k + 8(e-2)c^2\log\frac{2}{\delta}\,,
\end{aligned}$$

which implies that

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathrm{Var}(J_h^k \mid \mathcal{F}_{h-1}^k) \le 4c\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k + 16(e-2)c^2\log\frac{2}{\delta}\,.$$

Plugging this bound into inequality (6) and then into inequality (5), we conclude that

$$\begin{aligned}\sum_{k=1}^{K} J_1^k &\le \sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k + \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k] - J_{h+1}^k\right) \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k + \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{4c}\operatorname{Var}(J_{h+1}^k \mid \mathcal{F}_h^k) + 4(e-2)c\log\frac{2}{\delta} \\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k + 8(e-2)c\log\frac{2}{\delta} \\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H} X_h^k + 6c\log\frac{2}{\delta}\,.\end{aligned}$$

Taking the union bound over the events of inequality (6) and inequality (8), we obtain that the inequality above holds for all $K \in \mathbb{N}$ with probability at least $1-\delta$. ■

Proof [Proof of Lemma 28] Define another sequence of random variable $\widetilde{X}_h^k$ as

$$\widetilde{X}_h^k := \begin{cases} X_h^k & (h < \eta^k) \\ c & (h = \eta^k) \\ 0 & (h > \eta^k) \end{cases}.$$

It is clear that $\widetilde{X}_h^k$ is also adapted to $\mathcal{F}_h^k$. Define $\widetilde{J}_{H+1}^k := 0$ and $\widetilde{J}_h^k := (\widetilde{X}_h^k + \mathbb{E}[\widetilde{J}_{h+1}^k \mid \mathcal{F}_h^k]) \wedge c$ for $h \in [H]$.
First, we prove $J_h^k \mathbb{1}\{h \le \eta^k\} \le \widetilde{J}_h^k \mathbb{1}\{h \le \eta^k\}$ with backward induction on h. This inequality implies $J_1^k \le \widetilde{J}_1^k$ in particular since $\eta^k \ge 1$ always holds. The inequality is trivial for $h = H+1$ or for $h > \eta^k$ as $0 \le 0$. Suppose $h = \eta^k < H+1$. Then, we have

$$J_h^k \mathbb{1}\{h \le \eta^k\} \le c = \widetilde{J}_h^k \mathbb{1}\{h \le \eta^k\}\,,$$

where the first inequality is by definition, and the second equality is due to that $X_h^k = c$ when $h = \eta^k$ and hence $\widetilde{J}_h^k = (X_h^k + \mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k]) \wedge c = (c + \mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k]) \wedge c = c$. When $h < \eta^k$, we have

$$J_h^k = (X_h^k + \mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k]) \wedge c = (\widetilde{X}_h^k + \mathbb{E}[J_{h+1}^k \mid \mathcal{F}_h^k]) \wedge c \le (\widetilde{X}_h^k + \mathbb{E}[\widetilde{J}_{h+1}^k \mid \mathcal{F}_h^k]) \wedge c = \widetilde{J}_h^k\,,$$

where the equalities are by definitions, and the inequality is due to the induction hypothesis. Therefore, we conclude that $J_h^k \mathbb{1}\{h \le \eta^k\} \le \widetilde{J}_h^k \mathbb{1}\{h \le \eta^k\}$, and in particular $J_1^k \le \widetilde{J}_1^k$ holds for all $k \in \mathbb{N}$.
Then, we bound $\sum_{k=1}^{K} \widetilde{J}_1^k$. By Lemma 29, with probability at least $1 - \frac{\delta}{4}$, the following holds for

all $K \in \mathbb{N}$.

$$\begin{aligned}\sum_{K=1}^{K} \widetilde{J}_1^k &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H} \widetilde{X}_h^k + 6c\log\frac{8}{\delta} \\ &= 2\sum_{k=1}^{K}\sum_{h=1}^{H}\left(X_h^k \mathbb{1}\{h < \eta^k\} + c\mathbb{1}\{h = \eta^k\}\right) + 6c\log\frac{8}{\delta} \\ &= 2\sum_{k=1}^{K}\sum_{h=1}^{\eta^k - 1} X_h^k + \sum_{k=1}^{K} c\mathbb{1}\{\eta^k < H+1\} + 6c\log\frac{8}{\delta}\,.\end{aligned}$$

By Lemma 59, we have $\sum_{k=1}^{K}\mathbb{1}\{\eta^k < H+1\} \le SA\log_2 2H \le 2SA\log 2H$. The proof is completed by that $2cSA\log 2H + 6c\log\frac{1}{\delta} \le 3cSA\log 2H + 3cSA\log\frac{8}{\delta} \le 6cSA\log\frac{16H}{\delta}$. ■

D.2. General Properties of EQO+

In this section, we prove the following lemma that bounds the instantaneous regret of EQO+ of an episode. The proof is simply combining Lemma 31 and Lemma 32, where Lemma 31 is a formal version of Lemma 14. We also prove these lemmas in this section.

First, we formally define a function $U_h^k(s)$. Let $\beta^k(s,a) := 2b^k(s,a) + \frac{9V_{\max}S\ell_2(k,\delta)}{N^k(s,a)}$. Then, $U_h^k(s)$ is iteratively defined starting from $U_{H+1}^k(s) := 0$ and for $a = \pi_h^k(s)$,

$$U_h^k(s) := \left(2\beta^k(s,a) + PU_{h+1}^k(s,a)\right) \wedge 2V_{\max}\,.$$

Lemma 30 states that the instantaneous regret is bounded by a quasi-optimism term and $U_1^k(s_1^k)$.

Lemma 30 *Fix any $\delta \in (0,1]$ and $k \in \mathbb{N}$. Assume the event of $\mathcal{E}_1(\frac{\delta}{4}) \cap \mathcal{E}_2(\frac{\delta}{4}) \cap \mathcal{E}_3(\frac{\delta}{4})$. Suppose $k > \kappa(\delta)$. Then, the instantaneous regret of episode k is bounded as*

$$V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \le \frac{18\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} + U_1^k(s_1^k)\,.$$

Proof We present the following two key lemmas.

Lemma 31 (Quasi-optimism, formal version of Lemma 14) *Fix $\delta \in (0,1]$ and $k > \kappa(\delta)$. Then, under the event $\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)$, the estimated value $V_h^k(s)$ in Algorithm 1 satisfies*

$$V_h^*(s) - V_h^k(s) \le (2\lambda_{\iota_k}\mathbb{W}^*) \wedge V_{\max}$$

for all $h \in [H]$ and $s \in \mathcal{S}$.

The proof of Lemma 31 is provided in Appendix D.2.1.

Lemma 32 *Take any $\delta \in (0,1]$, $h \in [H]$, $s \in \mathcal{S}$, and $k > \kappa(\delta)$. Then, under the event of $\mathcal{E}_1(\frac{\delta}{4}) \cap \mathcal{E}_3(\frac{\delta}{4})$, we have*

$$V_h^k(s) - V_h^{\pi^k}(s) \le \frac{5}{2}\lambda_{\iota_k}\mathbb{W}^* + U_h^k(s)\,.$$

The proof of Lemma 32 is provided in Appendix D.2.2.

By combining Lemmas 31 and 32, we obtain that

$$\begin{aligned} V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) &\le 2\lambda_{\iota_k}\mathbb{W}^* + V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \\ &\le 2\lambda_{\iota_k}\mathbb{W}^* + \frac{5}{2}\lambda_{\iota_k}\mathbb{W}^* + U_1^k(s_1^k) \\ &= \frac{9}{2}\lambda_{\iota_k}\mathbb{W}^* + U_1^k(s_1^k)\,. \end{aligned}$$

By Lemma 16, we have $\lambda_{\iota_k} \le \frac{4\ell_1(\iota_k,\delta)}{c_{1,k}}$, which yields that

$$V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \le \frac{18\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} + U_1^k(s_1^k)\,.$$

■

D.2.1. PROOF OF LEMMA 31

Proof [Proof of Lemma 31] We prove the following stronger inequality by backward induction on h.

$$V_h^*(s) - V_h^k(s) \le \lambda_{\iota_k}\left(2W_h^{\pi^*}(s) - \frac{1}{2\mathbb{W}^*_{\text{diff}}}(W_{h,\text{diff}}^{\pi^*}(s))^2\right) \wedge V_{\max}$$

The inequality holds trivially for $h = H+1$ as $0 \le 0$. Now, suppose the inequality holds for $h+1$. If $V_h^k(s) = V_{\max}$, then the inequality holds since we have $V_h^*(s) - V_h^k(s) \le 0$ and

$$2W_h^{\pi^*}(s) - \frac{1}{2\mathbb{W}^*_{\text{diff}}}(W_{h,\text{diff}}^{\pi^*}(s))^2 \ge 2W_{h,\text{diff}}^{\pi^*}(s) - \frac{1}{2\mathbb{W}^*_{\text{diff}}}(W_{h,\text{diff}}^{\pi^*}(s))^2 \ge 0\,,$$

where the last inequality uses that $2x - \frac{x^2}{2c} \ge 0$ holds for all $x \in [0,c]$, where we plug in $x = W_{h,\text{diff}}^{\pi^*}$ and $c = \mathbb{W}^*_{\text{diff}}$.

Suppose $V_h^k(s) < V_{\max}$. It implies that for $a := \pi_h^k(s)$ and $a^* := \pi_h^*(s)$, we have $V_h^k(s) = Q_h^k(s,a) \ge Q_h^k(s,a^*) = (\hat{r}^k(s,a^*) + \hat{P}^k V_{h+1}^k(s,a^*))_+ + b^k(s,a^*)$, where the condition $V_h^k(s) < V_{\max}$ is used for the last equality to ensure that the clipping by $V_{\max}$ did not happen. Then, we have

$$\begin{aligned} V_h^*(s) - V_h^k(s) &\le V_h^*(s) - \left(\hat{r}^k(s,a^*) + \hat{P}^k V_{h+1}^k(s,a^*)\right)_+ - b^k(s,a^*) \\ &= V_h^*(s) \wedge V_{\max} + \left(-\hat{r}^k(s,a^*) - \hat{P}^k V_{h+1}^k(s,a^*)\right) \wedge 0 - b^k(s,a^*) \\ &\le \underbrace{\left(V_h^*(s) - \hat{r}^k(s,a^*) - \hat{P}^k V_{h+1}^k(s,a^*)\right)}_{I_1} \wedge V_{\max} - b^k(s,a^*)\,. \end{aligned} \tag{9}$$

We focus on bounding I_1. By adding and subtracting $\hat{P}^k V_{h+1}^*(s,a^*)$, we obtain that

$$I_1 = V_h^*(s) - \hat{r}^k(s,a^*) - \hat{P}^k V_{h+1}^*(s,a^*) + \underbrace{\hat{P}^k(V_{h+1}^* - V_{h+1}^k)(s,a^*)}_{I_2}\,. \tag{10}$$

We bound I_2 as follows. Using the induction hypothesis, we have $V^*_{h+1}(s') - V^k_{h+1}(s') \le \lambda_{\iota_k}(2W^{\pi^*}_{h+1}(s') - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}(W^{\pi^*}_{h+1,\mathrm{diff}}(s'))^2)$ for all $s' \in \mathcal{S}$, and hence

$$I_2 \le \lambda_{\iota_k} \hat{P}^k \left(2W^{\pi^*}_{h+1} - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\left(W^{\pi^*}_{h+1,\mathrm{diff}}\right)^2\right)(s,a^*)\,. \tag{11}$$

Under $\mathcal{E}_2(\frac{\delta}{4})$ (Lemma 26), we have

$$\begin{aligned}
&(\hat{P}^k - P)\left(2W^{\pi^*}_{h+1} - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\left(W^{\pi^*}_{h+1,\mathrm{diff}}\right)^2\right)(s,a^*) \\
&\le \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\mathrm{Var}(W^{\pi^*}_{h+1,\mathrm{diff}})(s,a^*) + \frac{13\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta)}{N^k(s,a^*)} \wedge \frac{3\mathbb{W}^*_{\mathrm{diff}}}{2}\sqrt{\frac{\ell_2(k,\delta)}{N^k(s,a^*)}} \\
&=: \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\mathrm{Var}(W^{\pi^*}_{h+1,\mathrm{diff}})(s,a^*) + b^k_2(s,a^*)\,,
\end{aligned} \tag{12}$$

where we denote $\frac{13\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta)}{N^k(s,a^*)} \wedge \frac{3\mathbb{W}^*_{\mathrm{diff}}}{2}\sqrt{\frac{\ell_2(k,\delta)}{N^k(s,a^*)}}$ by $b^k_2(s,a^*)$. We apply Lemma 55 to $X = W^{\pi^*}_{h,\mathrm{diff}}$, $Y = \frac{1}{2}\sigma^2_{\exp}(h,s,a^*) - W^{\pi^*}_{h,\min} + W^{\pi^*}_{h+1,\min}$, and $Z = W^{\pi^*}_{h+1,\mathrm{diff}}(s')$ with $s' \sim P(s,a^*)$, and obtain that

$$\mathrm{Var}(W^{\pi^*}_{h+1,\mathrm{diff}})(s,a^*) \le P(W^{\pi^*}_{h+1,\mathrm{diff}})^2(s,a^*) - (W^{\pi^*}_{h,\mathrm{diff}}(s))^2 + \mathbb{W}^*_{\mathrm{diff}}\sigma^2_{\exp}(h,s,a^*)\,,$$

where we bound $Y \le \frac{1}{2}\sigma^2_{\max}(h,s,a^*)$ using that $W^{\pi^*}_{h,\min} = \min_{s'\in\mathcal{S}} W^{\pi^*}_h(s') \ge \min_{s'\in\mathcal{S}} PW^{\pi^*}_{h+1}(s',\pi^*_h(s')) \ge W^{\pi^*}_{h+1,\min}$. Then, when taking the sum of the variance term and the expectation of the amount of underestimation, $\frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}P(W^{\pi^*}_{h+1,\mathrm{diff}})^2(s,a^*)$ terms cancel out as follows:

$$\begin{aligned}
&\frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\mathrm{Var}(W^{\pi^*}_{h+1,\mathrm{diff}})(s,a^*) + P\left(2W^{\pi^*}_{h+1} - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\left(W^{\pi^*}_{h+1,\mathrm{diff}}\right)^2\right)(s,a^*) \\
&\le \frac{1}{2}\sigma^2_{\exp}(h,s,a^*) + 2PW^{\pi^*}_{h+1}(s,a^*) - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}(W^{\pi^*}_{h,\mathrm{diff}}(s))^2 \\
&= W^{\pi^*}_h(s) + PW^{\pi^*}_{h+1}(s,a^*) - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}(W^{\pi^*}_{h,\mathrm{diff}}(s))^2\,.
\end{aligned}$$

Plugging this bound into inequality (12) and rearranging, we obtain that

$$\begin{aligned}
&\hat{P}^k\left(2W^{\pi^*}_{h+1} - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\left(W^{\pi^*}_{h+1,\mathrm{diff}}\right)^2\right)(s,a^*) \\
&\le P\left(2W^{\pi^*}_{h+1} - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\left(W^{\pi^*}_{h+1,\mathrm{diff}}\right)^2\right)(s,a^*) + \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}\mathrm{Var}(W^{\pi^*}_{h+1,\mathrm{diff}})(s,a^*) + b^k_2(s,a^*) \\
&\le W^{\pi^*}_h(s) + PW^{\pi^*}_{h+1}(s,a^*) - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}(W^{\pi^*}_{h,\mathrm{diff}}(s))^2 + b^k_2(s,a^*) \\
&=: I'_2\,,
\end{aligned}$$

where we denote the final bound by I'_2. Then, we have $I_2 \le \lambda_{\iota_k} I'_2$ by inequality (11). We note that I'_2 is always greater than 0 since we have $W^{\pi^*}_h(s) - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}(W^{\pi^*}_{h,\mathrm{diff}}(s))^2 \ge W^{\pi^*}_{h,\mathrm{diff}}(s) -$

$\frac{1}{2\mathbb{W}^*_{\text{diff}}}(W^{\pi^*}_{h,\text{diff}}(s))^2 \geq 0$, which comes from that $x - \frac{x^2}{2c} \geq 0$ for $x \in [0, c]$, and the other terms are also greater than 0. Applying the bound $I_2 \leq \lambda_{\iota_k} I_2'$ to inequality (10), we obtain that

$$\begin{aligned} I_1 \wedge V_{\max} &\leq (V^*_h(s) - \hat{r}^k(s, a^*) - \hat{P}^k V^*_{h+1}(s, a^*) + \lambda_{\iota_k} I_2') \wedge V_{\max} \\ &\leq (V^*_h(s) - \hat{r}^k(s, a^*) - \hat{P}^k V^*_{h+1}(s, a^*)) \wedge V_{\max} + \lambda_{\iota_k} I_2' \,. \end{aligned}$$

The remaining terms are bounded under $\mathcal{E}_1(\frac{\delta}{4})$ (Lemma 25) as follows.

$$\begin{aligned} &(V^*_h(s) - \hat{r}^k(s, a^*) - \hat{P}^k V^*_{h+1}(s, a^*)) \wedge V_{\max} \\ &\leq \left| Q^*_h(s, a^*) - \hat{r}^k(s, a^*) - \hat{P}^k V^*_{h+1}(s, a^*) \right| \wedge V_{\max} \\ &\leq \frac{\lambda_{\iota_k} \sigma^2_{\text{exp}}(h, s, a^*)}{2} + \frac{\ell_1(\iota_k, \delta)}{\lambda_{\iota_k} N^k(s, a^*)} \wedge \sigma_{\max} \sqrt{\frac{\ell_2(k, \delta)}{N^k(s, a^*)}} \\ &=: \frac{\lambda_{\iota_k} \sigma^2_{\text{exp}}(h, s, a^*)}{2} + b^k_1(s, a^*) \,, \end{aligned} \tag{13}$$

where we define $b^k_1(s, a^*) := \frac{\ell_1(\iota_k, \delta)}{\lambda_{\iota_k} N^k(s, a^*)} \wedge \sigma_{\max} \sqrt{\frac{\ell_2(k, \delta)}{N^k(s, a^*)}}$. Therefore, we obtain that

$$\begin{aligned} &I_1 \wedge V_{\max} \\ &\leq \frac{\lambda_{\iota_k} \sigma^2_{\text{exp}}(h, s, a^*)}{2} + b^k_1(s, a^*) + \lambda_{\iota_k} I_2' \\ &= \frac{\lambda_{\iota_k} \sigma^2_{\text{exp}}(h, s, a^*)}{2} + b^k_1(s, a^*) \\ &\quad + \lambda_{\iota_k} \left(W^{\pi^*}_h(s) + P W^{\pi^*}_{h+1}(s, a^*) - \frac{1}{2\mathbb{W}^*_{\text{diff}}} (W^{\pi^*}_{h,\text{diff}}(s))^2 + b^k_2(s, a^*) \right) \\ &= \lambda_{\iota_k} \left(2W^{\pi^*}_h(s) - \frac{1}{2\mathbb{W}^*_{\text{diff}}} (W^{\pi^*}_{h,\text{diff}}(s))^2 \right) + b^k_1(s, a^*) + \lambda_{\iota_k} b^k_2(s, a^*) \,. \end{aligned}$$

Plugging this bound into inequality (9), we obtain that

$$\begin{aligned} V^*_h(s) - V^k_h(s) \leq \lambda_{\iota_k} &\left(2W^{\pi^*}_h(s) - \frac{1}{2\mathbb{W}^*_{\text{diff}}} (W^{\pi^*}_{h,\text{diff}}(s))^2 \right) && (14) \\ &+ b^k_1(s, a^*) + \lambda_{\iota_k} b^k_2(s, a^*) - b^k(s, a^*) \,. && (15) \end{aligned}$$

Now, we show that $b^k(s, a^*) \geq b^k_1(s, a^*) + \lambda_{\iota_k} b^k_2(s, a^*)$. We first have

$$\begin{aligned} b^k_1(s, a^*) + \lambda_{\iota_k} b^k_2(s, a^*) &\leq \frac{\ell_1(\iota_k, \delta)}{\lambda_{\iota_k} N^k(s, a^*)} \wedge \sigma_{\max} \sqrt{\frac{\ell_2(k, \delta)}{N^k(s, a^*)}} \\ &\quad + \frac{13 \lambda_{\iota_k} \mathbb{W}^*_{\text{diff}} \ell_1(1, \delta)}{N^k(s, a^*)} \wedge \frac{3 \lambda_{\iota_k} \mathbb{W}^*_{\text{diff}}}{2} \sqrt{\frac{\ell_2(k, \delta)}{N^k(s, a^*)}} \\ &\leq \frac{1}{N^k(s, a^*)} \left(\frac{\ell_1(\iota_k, \delta)}{\lambda_{\iota_k}} + 13 \lambda_{\iota_k} \mathbb{W}^*_{\text{diff}} \ell_1(1, \delta) \right) \\ &\quad \wedge \sqrt{\frac{\ell_2(k, \delta)}{N^k(s, a^*)}} \left(\sigma_{\max} + \frac{3 \lambda_{\iota_k} \mathbb{W}^*_{\text{diff}}}{2} \right) \end{aligned}$$

From $k > \kappa(\delta)$, we have $c_{1,k} \geq (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta)$. Then, we can use Lemma 16 to guarantee that $\frac{\ell_1(\iota_k,\delta)}{\lambda_{\iota_k}} + 13\lambda_{\iota_k}\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta) \leq c_{1,k}$ and $\lambda_{\iota_k} \leq \frac{4\ell_1(\iota_k,\delta)}{c_{1,k}}$. We also have $c_{2,k} \geq (\sigma_{\max} + \frac{6\mathbb{W}^*_{\mathrm{diff}}\ell_1(\iota_k,\delta)}{c_{1,k}})\sqrt{\ell_2(k,\delta)}$ when $k > \kappa(\delta)$, so we obtain that

$$\begin{aligned}
&\frac{1}{N^k(s,a^*)}\left(\frac{\ell_1(\iota_k,\delta)}{\lambda_{\iota_k}} + 13\lambda_{\iota_k}\mathbb{W}^*_{\mathrm{diff}}\ell_1(1,\delta)\right) \wedge \sqrt{\frac{\ell_2(k,\delta)}{N^k(s,a^*)}}\left(\sigma_{\max} + \frac{3\lambda_{\iota_k}\mathbb{W}^*_{\mathrm{diff}}}{2}\right) \\
&\leq \frac{c_{1,k}}{N^k(s,a^*)} \wedge \sqrt{\frac{\ell_2(k,\delta)}{N^k(s,a^*)}}\left(\sigma_{\max} + \frac{6\mathbb{W}^*_{\mathrm{diff}}\ell_1(\iota_k,\delta)}{c_{1,k}}\right) \\
&\leq \frac{c_{1,k}}{N^k(s,a^*)} \wedge \frac{c_{2,k}}{N^k(s,a)} \\
&= b^k(s,a^*)\,.
\end{aligned}$$

We have proved that $b^k(s,a^*) \geq b^k_1(s,a^*) + \lambda_{\iota_k} b^k_2(s,a^*)$, so from inequality (15), we obtain that

$$V^*_h(s) - V^k_h(s) \leq \lambda_{\iota_k}\left(2W^{\pi^*}_h(s,a^*) - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}(W^{\pi^*}_{h,\mathrm{diff}}(s))^2\right)\,.$$

Lastly, we apply the trivial bound of $V_{\max}$ that comes from $V^*_h(s) \leq V_{\max}$ and $V^k_h(s) \geq 0$ and obtain

$$V^*_h(s) - V^k_h(s) \leq \lambda_{\iota_k}\left(2W^{\pi^*}_h(s,a^*) - \frac{1}{2\mathbb{W}^*_{\mathrm{diff}}}(W^{\pi^*}_{h,\mathrm{diff}}(s))^2\right) \wedge V_{\max}\,,$$

which completes the induction step. ■

D.2.2. PROOF OF LEMMA 32

To prove the lemma, we require the following lemma. It is an analogue of Lemma 13 in Lee and Oh (2025), but we obtain improved dependence on the variance term $\sigma^2_{\exp}(h,s,a)$.

Lemma 33 *Take any $h \in [H]$, $s \in \mathcal{S}$, and $k > \kappa(\delta)$. Let $a := \pi^k_h(s)$ and $\varepsilon := 2\lambda_{\iota_k}\mathbb{W}^* \wedge V_{\max}$. Recall that $\beta^k(s,a) := 2b^k(s,a) + \frac{9V_{\max}S\ell_2(k,\delta)}{N^k(s,a)}$. Then, under $\mathcal{E}_1(\frac{\delta}{4}) \cap \mathcal{E}_3(\frac{\delta}{4})$, we have*

$$\begin{aligned}
&V^k_h(s) - r(s,a) - PV^k_{h+1}(s,a) \\
&\leq \lambda_{\iota_k}\sigma^2_{\exp}(h,s,a) + 2\beta^k(s,a) \\
&\quad + \frac{1}{4V_{\max}}\left(-(V^k_h(s) - V^*_h(s) + \varepsilon)^2 + P(V^k_{h+1} - V^*_{h+1} + \varepsilon)^2(s,a)\right)
\end{aligned}$$

Proof Starting from the definition, we have

$$\begin{aligned}
&V^k_h(s) - r(s,a) - PV^k_{h+1}(s,a) \\
&= \left((\hat{r}^k(s,a) + \hat{P}^kV^k_{h+1}(s,a))_+ + b^k(s,a)\right) \wedge V_{\max} - r(s,a) - PV^k_{h+1}(s,a) \\
&\leq (\hat{r}^k(s,a) + \hat{P}^kV^k_{h+1}(s,a))_+ \wedge V_{\max} - r(s,a) - PV^k_{h+1}(s,a) + b^k(s,a) \\
&\leq \underbrace{(\hat{r}^k(s,a) + \hat{P}^kV^k_{h+1}(s,a) - r(s,a) - PV^k_{h+1}(s,a))_+}_{I_1} \wedge V_{\max} + b^k(s,a)\,, \qquad (16)
\end{aligned}$$

where we use that $r(s,a) + PV_{h+1}^k(s,a) \geq 0$ for the last inequality. We focus on bounding I_1. By adding and subtracting $(\hat{P}^k - P)V_{h+1}^*(s,a)$, we obtain that

$$I_1 = \underbrace{\hat{r}^k(s,a) + \hat{P}^k V_{h+1}^*(s,a) - r(s,a) - PV_{h+1}^*(s,a)}_{I_2} + \underbrace{(\hat{P}^k - P)(V_{h+1}^k(s,a) - V_{h+1}^*)(s,a)}_{I_3} .$$

Under $\mathcal{E}_3(\frac{\delta}{4})$, using Lemma 27 with $\rho = 8$ and $c = V_{\max}$, we bound I_3 as follows:

$$\begin{aligned} I_3 &\leq \frac{1}{8V_{\max}} \operatorname{Var}(V_{h+1}^k - V_{h+1}^*)(s,a) + \frac{9V_{\max} S\ell_2(k,\delta)}{N^k(s,a)} \\ &=: I_3' . \end{aligned}$$

Noting that $I_3' \geq 0$, we have

$$\begin{aligned} (I_1 \vee 0) \wedge V_{\max} &= ((I_2 + I_3) \vee 0) \wedge V_{\max} \\ &\leq \left(\left(I_2 + I_3'\right) \vee 0\right) \wedge V_{\max} \\ &\leq (I_2 \vee 0) \wedge V_{\max} + I_3' \\ &= (I_2 \wedge V_{\max}) \vee 0 + I_3' . \end{aligned}$$

$I_2 \wedge V_{\max}$ is bounded under $\mathcal{E}_1(\frac{\delta}{4})$ (Lemma 25) as follows:

$$\begin{aligned} I_2 \wedge V_{\max} &\leq \frac{\lambda_{\iota_k} \sigma_{\exp}^2(s,a,V_{h+1}^*)}{2} + \frac{\ell_1(\iota_k,\delta)}{\lambda_{\iota_k} N^k(s,a)} \wedge \sigma_{\max} \sqrt{\frac{\ell_2(k,\delta)}{N^k(s,a)}} \\ &\leq \frac{\lambda_{\iota_k} \sigma_{\exp}^2(s,a,V_{h+1}^*)}{2} + b^k(s,a) , \end{aligned}$$

where we use $k > \kappa(\delta)$ for the last inequality. Combining the bounds, we obtain that

$$(I_1 \vee 0) \wedge V_{\max} \leq \frac{\lambda_{\iota_k} \sigma_{\exp}^2(s,a,V_{h+1}^*)}{2} + b^k(s,a) + \frac{1}{8V_{\max}} \operatorname{Var}(V_{h+1}^k - V_{h+1}^*)(s,a) + \frac{9V_{\max} S\ell_2(k,\delta)}{N^k(s,a)} .$$

Plugging this bound into inequality (16), we derive that

$$\begin{aligned} &V_h^k(s) - r(s,a) - PV_{h+1}^k(s,a) \\ &\leq \frac{\lambda_{\iota_k} \sigma_{\exp}^2(s,a,V_{h+1}^*)}{2} + \frac{1}{8V_{\max}} \operatorname{Var}(V_{h+1}^k - V_{h+1}^*)(s,a) + \beta^k(s,a) , \end{aligned} \tag{17}$$

where we use that $\beta^k(s,a) = 2b^k(s,a) + \frac{9V_{\max} S\ell_2(k,\delta)}{N^k(s,a)}$. We define $I_4 := (V_h^k(s) - r(s,a) - PV_{h+1}^k(s,a))_+$. Noting that the right-hand side of inequality (17) is at least 0, we have

$$I_4 \leq \frac{\lambda_{\iota_k} \sigma_{\exp}^2(s,a,V_{h+1}^*)}{2} + \frac{1}{8V_{\max}} \operatorname{Var}(V_{h+1}^k - V_{h+1}^*)(s,a) + \beta^k(s,a) \tag{18}$$

Recall that by Lemma 31, we have $V_{h+1}^k(s') - V_{h+1}^*(s') + \varepsilon \geq 0$ for all $s' \in \mathcal{S}$. In addition, we have $V_{h+1}^k(s') - V_{h+1}^*(s') + \varepsilon \leq 2V_{\max}$. By Lemma 55, we have

$$\begin{aligned}
\mathrm{Var}(V_{h+1}^k - V_{h+1}^*)(s,a) &= \mathrm{Var}(V_{h+1}^k - V_{h+1}^* + \varepsilon)(s,a) \\
&\leq -(V_h^k(s) - V_h^*(s) + \varepsilon)^2 + P(V_{h+1}^k - V_{h+1}^k + \varepsilon)^2(s,a) \\
&\quad + 4V_{\max}\left(V_h^k(s) - V_h^*(s) + \varepsilon - P(V_{h+1}^k - V_{h+1}^k + \varepsilon)(s,a)\right)_+ \\
&= -(V_h^k(s) - V_h^*(s) + \varepsilon)^2 + P(V_{h+1}^k - V_{h+1}^* + \varepsilon)^2(s,a) \\
&\quad + 4V_{\max}\left(V_h^k(s) - V_h^*(s) - P(V_{h+1}^k - V_{h+1}^*)(s,a)\right)_+ .
\end{aligned}$$

Then, denoting $a^* := \pi_h^*(s)$, we have

$$-V_h^*(s) + PV_{h+1}^*(s,a) = -Q_h^*(s,a^*) + Q_h^*(s,a) - r(s,a) \leq -r(s,a)\,,$$

and hence

$$\begin{aligned}
\left(V_h^k(s) - V_h^*(s) - P(V_{h+1}^k - V_{h+1}^*)(s,a)\right)_+ &\leq \left(V_h^k(s) - r(s,a) - PV_{h+1}^k(s,a)\right)_+ \\
&= I_4\,.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathrm{Var}(V_{h+1}^k - V_{h+1}^*)(s,a) \\
&\leq -(V_h^k(s) - V_h^*(s) + \varepsilon)^2 + P(V_{h+1}^k - V_{h+1}^* + \varepsilon)^2(s,a) + 4V_{\max}I_4\,.
\end{aligned}$$

Plugging this bound into inequality (18), we obtain

$$\begin{aligned}
I_4 \leq \frac{\lambda_{\iota_k}\sigma_{\exp}^2(s,a,V_{h+1}^*)}{2} + \frac{1}{2}I_4 + \beta^k(s,a) \\
+ \frac{1}{8V_{\max}}\left(-(V_h^k(s) - V_h^*(s) + \varepsilon)^2 + P(V_{h+1}^k - V_{h+1}^* + \varepsilon)^2(s,a)\right)\,.
\end{aligned}$$

Solving the inequality with respect to I_4, we conclude that

$$\begin{aligned}
I_4 \leq \lambda_{\iota_k}\sigma_{\exp}^2(s,a,V_{h+1}^*) + 2\beta^k(s,a) \\
+ \frac{1}{4V_{\max}}\left(-(V_h^k(s) - V_h^*(s) + \varepsilon)^2 + P(V_{h+1}^k - V_{h+1}^* + \varepsilon)^2(s,a)\right)\,.
\end{aligned}$$

■

Proof [Proof of Lemma 32] We use backward induction on h to prove the following stronger inequality:

$$V_h^k(s) - V_h^{\pi^k}(s) \leq 2\lambda_{\iota_k}W_h^{\pi^k}(s) + \frac{1}{4V_{\max}}(\varepsilon^2 - (V_h^k(s) - V_h^*(s) + \varepsilon)^2) + U_h^k(s)\,.$$

This is a stronger inequality since we have $W_h^{\pi^k}(s) \leq \mathbb{W}^*$ and $\frac{\varepsilon^2}{4V_{\max}} \leq \frac{\varepsilon}{4} \leq \frac{\lambda_{\iota_k}\mathbb{W}^*}{2}$, where we use $\varepsilon \leq V_{\max}$ for the first inequality, and $\varepsilon \leq 2\lambda_{\iota_k}\mathbb{W}^*$ for the second inequality. The inequality we

want to prove holds when $h = H + 1$ as $0 \le 0$. Suppose the inequality holds for $h + 1$. Then, we have

$$\begin{aligned}
&V_h^k(s) - V_h^{\pi^k}(s) \\
&= V_h^k(s) - r(s,a) - PV_{h+1}^{\pi^k}(s,a) \\
&\le \underbrace{V_h^k(s) - r(s,a) - PV_{h+1}^k(s,a)}_{I_1} + \underbrace{P(V_{h+1}^k - V_{h+1}^{\pi^k})(s,a)}_{I_2} .
\end{aligned}$$

I_1 is bounded by Lemma 33. I_2 is bounded by the induction hypothesis. Combining the two, we derive that

$$\begin{aligned}
&V_h^k(s) - V_h^{\pi^k}(s) \\
&\le \lambda_{\iota_k}\sigma_{\exp}^2(h,s,a) + 2\beta^k(s,a) + \frac{1}{4V_{\max}}\left(-(V_h^k(s) - V_h^*(s) + \varepsilon)^2 + P(V_{h+1}^k - V_{h+1}^* + \varepsilon)^2(s,a)\right) \\
&\quad + P\left(2\lambda_{\iota_k}W_{h+1}^{\pi^k} + \frac{1}{4V_{\max}}(\varepsilon^2 - (V_{h+1}^k - V_{h+1}^* + \varepsilon)^2) + U_{h+1}^k\right)(s,a) \\
&= 2\lambda_{\iota_k}W_h^k(s) + \frac{1}{4V_{\max}}\left(\varepsilon^2 - (V_h^k(s) - V_h^*(s) + \varepsilon)^2\right) + 2\beta^k(s,a) + PU_{h+1}^k(s,a) .
\end{aligned}$$

We need a proper clipping for the $2\beta^k(s,a) + PU_{h+1}^k(s,a)$ term to bound it with $U_h^k(s)$. We note that $2\lambda_{\iota_k}W_h^k(s) \ge 0$ and $\frac{1}{4V_{\max}}(\varepsilon^2 - (V_h^k(s) - V_h^*(s) + \varepsilon)^2) \ge -\frac{1}{4V_{\max}}(V_h^k(s) - V_h^*(s) + \varepsilon)^2 \ge -\frac{1}{4V_{\max}}(2V_{\max})^2 = -V_{\max}$. Therefore, we have that

$$\begin{aligned}
&V_h^k(s) - V_h^{\pi^k}(s) \\
&= (V_h^k(s) - V_h^{\pi^k}(s)) \wedge V_{\max} \\
&\le \left(2\lambda_{\iota_k}W_h^k(s) + \frac{1}{4V_{\max}}\left(\varepsilon^2 - (V_h^k(s) - V_h^*(s) + \varepsilon)^2\right) + 2\beta^k(s,a) + PU_{h+1}^k(s,a)\right) \wedge V_{\max} \\
&\le 2\lambda_{\iota_k}W_h^k(s) + \frac{1}{4V_{\max}}\left(\varepsilon^2 - (V_h^k(s) - V_h^*(s) + \varepsilon)^2\right) \vee (-V_{\max}) + (2\beta^k(s,a) + PU_{h+1}^k(s,a)) \wedge 2V_{\max} \\
&= 2\lambda_{\iota_k}W_h^k(s) + \frac{1}{4V_{\max}}\left(\varepsilon^2 - (V_h^k(s) - V_h^*(s) + \varepsilon)^2\right) + U_h^k(s) , .
\end{aligned}$$

where we use the third property in Fact 1. The induction step is proved, so the proof is complete. ■

D.3. Proof of Theorem 8

In this section, we restate Theorem 8 with specific logarithmic factors and provide its proof.

Theorem 34 (Formal statement of Theorem 8) *Let $\{c_{1,k}\}_{k=1}^\infty$ and $\{c_{2,k}\}_{k=1}^\infty$ be positive non-decreasing sequences. Fix any $\delta \in (0,1]$ and $K \in \mathbb{N}$. The regret bound of Alg satisfies*

$$\begin{aligned}
\mathrm{Reg}_{\mathcal{M}}^{\mathsf{Alg}}(K,\delta) \le 18\mathbb{W}^* \sum_{k=1}^{K} \frac{\ell_1(\iota_k,\delta)}{c_{1,k}} + (16c_{1,k}SA\log KH)) \wedge \left(16\sqrt{2}c_{2,k}\sqrt{HSAK}\right) \\
+ V_{\max}(\kappa(\delta) \wedge K) + 72V_{\max}S^2A\ell_2(K,\delta)\log 2KH .
\end{aligned}$$

Proof [Proof of Theorem 34] By Lemma 30, we have that for $k > \kappa(\delta)$,

$$V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \le \frac{18\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} + U_1^k(s_1^k)\,.$$

We bound the instantaneous regret of the first $\kappa(\delta)$ episodes by $V_{\max}$, and use Lemma 30 for the following episodes. Then, we obtain that under $\mathcal{E}_1(\frac{\delta}{4}) \cap \mathcal{E}_2(\frac{\delta}{4}) \cap \mathcal{E}_3(\frac{\delta}{4})$, it holds that

$$\mathrm{Reg}_M^{\mathsf{Alg}}(K) \le V_{\max}(\kappa(\delta) \wedge K) + 18\mathbb{W}^* \sum_{k=\kappa(\delta)+1}^{K} \frac{\ell_1(\iota_k,\delta)}{c_{1,k}} + \sum_{k=\kappa(\delta)+1}^{K} U_1^k(s_1^k)\,.$$

By applying Lemma 35, which bounds the sum of $U_1^k(s_1^k)$, we obtain that under an additional event of $\mathcal{E}_4(\{2\beta^k(s_h^k,a_h^k)\}_{k,h}, 2V_{\max}, \frac{\delta}{4})$, we have

$$\begin{aligned}\mathrm{Reg}_M^{\mathsf{Alg}}(K) \le 18\mathbb{W}^* \sum_{k=\kappa(\delta)+1}^{K} \frac{\ell_1(\iota_k,\delta)}{c_{1,k}} + (16c_{1,K}SA\log KH)) \wedge \left(16\sqrt{2}c_{2,K}\sqrt{HSAK}\right) \\ + V_{\max}(\kappa(\delta)\wedge K) + 72V_{\max}S^2A\ell_2(K,\delta)\log 2KH\,.\end{aligned}$$

Taking the union bound over the events, we conclude that the bound holds with probability at least $1-\delta$. Also, note that the right-hand side is no longer a random variable. Therefore, this bound is an upper bound for $\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta)$. ■

The following lemma is an analogue of Lemma 4 in Lee and Oh (2025).

Lemma 35 *Under $\mathcal{E}_4(\{2\beta^k(s_h^k,a_h^k)\}_{k,h}, 2V_{\max}, \frac{\delta}{4})$, the following inequality holds for all $K \in \mathbb{N}$.*

$$\begin{aligned}\sum_{k=1}^{K} U_1^k(s_1^k) \le (16c_{1,K}SA\log KH)) \wedge \left(16\sqrt{2}c_{2,K}\sqrt{HSAK}\right) \\ + 72V_{\max}S^2A\ell_2(K,\delta)\log 2KH\,.\end{aligned}$$

Proof By Lemma 28, under $\mathcal{E}_4(\{2\beta^k(s_h^k,a_h^k)\}_{k,h}, 2V_{\max}, \frac{\delta}{4})$, it holds that

$$\sum_{k=1}^{K} U_1^k(s_1^k) \le 4\sum_{k=1}^{K}\sum_{h=1}^{\eta^k-1} \beta^k(s_h^k,a_h^k) + 12V_{\max}SA\log\frac{16H}{\delta}\,.$$

Recall that

$$\begin{aligned}\beta^k(s,a) &= 2b^k(s,a) + \frac{9V_{\max}S\ell_2(k,\delta)}{N^k(s,a)} \\ &= \frac{2c_{1,k}}{N^k(s,a)} \wedge \frac{2c_{2,k}}{\sqrt{N^k(s,a)}} + \frac{9V_{\max}S\ell_2(k,\delta)}{N^k(s,a)}\,.\end{aligned}$$

Furthermore, using that $N^k(s_h^k,a_h^k) \ge \frac{1}{2}N_h^k(s_h^k,a_h^k)$ when $h < \eta^k$, where $N_h^k(s,a)$ and η^k are defined in Appendix A, we have

$$\beta^k(s_h^k,a_h^k)\mathbb{1}\{h<\eta^k\} \le \frac{4c_{1,k}}{N_h^k(s,a)} \wedge \frac{2\sqrt{2}c_{2,k}}{\sqrt{N_h^k(s,a)}} + \frac{18V_{\max}S\ell_2(k,\delta)}{N_h^k(s,a)}\,. \tag{19}$$

When $h < \eta^k$, we have $N^k(s_h^k, a_h^k)+1 \le N_h^k(s_h^k, a_h^k) \le 2N^k(s_h^k, a_h^k)$, which implies $N_h^k(s_h^k, a_h^k) \ge 2$ for such h. Now, we take the sum over $(h, k) \in [H] \times [K]$. Note that one has

$$\sum_{n=2}^{N} \frac{1}{n} \le \log(1 \vee N)\,, \qquad\qquad \sum_{n=2}^{N} \frac{1}{\sqrt{n}} \le 2\sqrt{N}\,.$$

Hence, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{\mathbb{1}\{h < \eta^k\}}{N_h^k(s_h^k, a_h^k)} \le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{n=2}^{N^{K+1}(s,a)} \frac{1}{n} \le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \log(1 \vee N^{K+1}(s,a)) \le SA \log KH$$

and

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{\mathbb{1}\{h < \eta^k\}}{\sqrt{N_h^k(s_h^k, a_h^k)}} &\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{n=2}^{N^{K+1}(s,a)} \frac{1}{\sqrt{n}} \\
&\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} 2\sqrt{N^{K+1}(s,a)} \\
&\le 2\sqrt{SA \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N^{K+1}(s,a)} \\
&= 2\sqrt{HSAK}\,,
\end{aligned}$$

where the third inequality uses the Cauchy-Schwarz inequality. Therefore, the sum of $\beta^k(s,a)\mathbb{1}\{h < \eta^k\}$ is bounded as

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \beta^k(s,a)\mathbb{1}\{h < \eta^k\} \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \left(\frac{4c_{1,k}}{N_h^k(s_h^k, a_h^k)} \wedge \frac{2\sqrt{2}c_{2,k}}{\sqrt{N_h^k(s_h^k, a_h^k)}} + \frac{18V_{\max}S\ell_2(k,\delta)}{N_h^k(s_h^k, a_h^k)} \right) \mathbb{1}\{h < \eta^k\} \\
&\le \left(4c_{1,K} \sum_{k=1}^{K}\sum_{h=1}^{H} \frac{\mathbb{1}\{h < \eta^k\}}{N_h^k(s_h^k, a_h^k)} \right) \wedge \left(2\sqrt{2}c_{2,K} \sum_{k=1}^{K}\sum_{h=1}^{H} \frac{\mathbb{1}\{h < \eta^k\}}{\sqrt{N_h^k(s_h^k, a_h^k)}} \right) \\
&\qquad + 18V_{\max}S\ell_2(K,\delta) \sum_{k=1}^{K}\sum_{h=1}^{H} \frac{\mathbb{1}\{h < \eta^k\}}{N_h^k(s_h^k, a_h^k)} \\
&\le (4c_{1,K}SA \log KH) \wedge \left(4\sqrt{2}c_{2,K}\sqrt{HSAK} \right) + 18V_{\max}S^2A\ell_2(K,\delta) \log KH\,,
\end{aligned}$$

where the second inequality uses that $\{c_{1,k}\}_{k=1}^{\infty}$ and $\{c_{2,k}\}_{k=1}^{\infty}$ are non-decreasing and $\sum_{i=1}^{n} a_i \wedge b_i \le (\sum_{i=1}^{n} a_i) \wedge (\sum_{i=1}^{n} b_i)$ for any sequences $\{a_i\}_{i=1}^{n}$ and $\{b_i\}_{i=1}^{n}$. Finally, we obtain the following

bound on the sum $\sum_{k=1}^{K} U_1^k(s_1^k)$:

$$\begin{aligned}\sum_{k=1}^{K} U_1^k(s_1^k) &\le 4\sum_{k=1}^{K}\sum_{h=1}^{\eta^k-1}\beta^k(s_h^k,a_h^k) + 12V_{\max}SA\log\frac{16H}{\delta} \\ &\le (16c_{1,K}SA\log KH) \wedge \left(16\sqrt{2}c_{2,K}\sqrt{HSAK}\right) \\ &\quad + 72V_{\max}S^2A\ell_2(K,\delta)\log KH + 12V_{\max}SA\log\frac{16H}{\delta}\,.\end{aligned}$$

To obtain a simpler form, we bound the last term by $12V_{\max}SA\log\frac{16H}{\delta} \le 12V_{\max}S^2A\ell_2(K,\delta) \le 72(\log 2)V_{\max}S^2A\ell_2(K,\delta)$, which bounds the last two terms as

$$72V_{\max}S^2A\ell_2(K,\delta)\log KH + 12V_{\max}SA\log\frac{16H}{\delta} \le 72V_{\max}S^2A\ell_2(K,\delta)\log 2KH\,.$$

■

D.4. Proof for Theorem 9

In this section, we restate Theorem 9 with specific logarithmic factors and provide its proof.

Dann et al. (2021) show that the true gap $\mathrm{gap}_h(s,a) := V_h^*(s) - Q_h^*(s,a)$ can be replaced with another function, which may improve the overall regret bound depending on reachability conditions of the MDP. We also use the following generalized definition, which comes from the condition of Proposition 3.3 in Dann et al. (2021). First, we define a stopping time $B = \min(\{h \in \{1,2,\dots,H\} : \mathrm{gap}_h(s_h,a_h) > 0\} \cup \{H+1\})$, which is the first time step that the policy takes a sub-optimal action. We denote the B value at the k-th episode by B^k. In addition, we define $\mathcal{S}_{\text{initial}}$ as the set of possible initial states, that is, the environment choose s_1^k from $\mathcal{S}_{\text{initial}}$.

Definition 36 (Effective gap) *A sequence of functions $\overline{\mathrm{gap}}_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}_{>0}$ for $h \in [H]$ is called **effective gap** if it satisfies*

$$\mathop{\mathbb{E}}_{\pi(\cdot|s_1=s)}\left[\sum_{h=B}^{H}\overline{\mathrm{gap}}_h(s_h,a_h)\right] \le V_1^*(s) - V_1^\pi(s)$$

for all $s \in \mathcal{S}_{\text{initial}}$, and $\pi \in \Pi$.

There can be many different functions that are effective gaps, and we note that it is better to set the effective gaps as large as possible. A common choice of effective gap is $\overline{\mathrm{gap}}_h(s,a) := \frac{1}{2}\mathrm{gap}_h(s,a) \vee \frac{1}{2H}\mathrm{gap}_{\min}$, where $\mathrm{gap}_{\min} := \min\{\mathrm{gap}_h(s,a) \mid \mathrm{gap}_h(s,a) > 0, (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}\}$. Dann et al. (2021) provide another choice of effective gap named *return gap*, defined as

$$\overline{\mathrm{gap}}_h(s,a) := \frac{1}{2}\mathrm{gap}_h(s,a) \vee \min_{\substack{\pi\in\Pi \\ \mathbb{P}_\pi((s_h,a_h)=(s,a))>0 \\ \mathbb{E}_\pi[\sum_{h'=1}^{h}\mathrm{gap}_{h'}(s_{h'},a_{h'})|(s_h,a_h)=(s,a)]>0}} \frac{1}{2H}\mathop{\mathbb{E}}_{\pi}\left[\sum_{h'=1}^{h}\mathrm{gap}_{h'}(s_{h'},a_{h'})\middle|(s_h,a_h)=(s,a)\right].$$

It is possible to define effective gap to even depend on the policy of the agent, but we do not go further in this direction. Refer to Dann et al. (2021) for more discussions. We define $\overline{\mathrm{gap}}(s,a) := \min_{h\in[H]} \overline{\mathrm{gap}}_h(s,a)$.

In addition we let $v_{\mathrm{gap}} := \min_{\pi\in\Pi} \mathrm{gap}_B(s_B, a_B)$. This is the minimum regret the agent can incur by making one suboptimal selection. While this may be equal to $\mathrm{gap}_{\min}$ is most cases, it may be larger when state-action pairs with gap $\mathrm{gap}_h(s,a) = \mathrm{gap}_{\min}$ can be reached only if the agent takes another sub-optimal action before time step h. We note that $v_{\mathrm{gap}} \geq \mathrm{gap}_{\min}$. We replace $\kappa_{\mathrm{gap}}(\mathrm{gap}_{\min}, \delta)$ in Theorem 9 with $\kappa_{\mathrm{gap}}(v_{\mathrm{gap}}, \delta)$, leading to a potential improvement.

We restate Theorem 9 using the effective gap and v_{gap}, which leads to a more concise and potentially improved result.

Theorem 37 (Formal statement of Theorem 9) *Let $\{c_{1,k}\}_{k=1}^{\infty}$ and $\{c_{2,k}\}_{k=1}^{\infty}$ be positive non-decreasing sequences. Fix any $\delta \in (0,1]$ and $K \in \mathbb{N}$. Then, we have*

$$\mathrm{Reg}_{\mathcal{M}}^{\mathsf{Alg}}(K,\delta) \leq V_{\max}(\kappa(\delta) \wedge K) + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(\frac{2048 c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)} \wedge 64 c_{1,K} \log \frac{32 c_{1,K}}{\overline{\mathrm{gap}}(s,a)} \right) + \sum_{k=\kappa(\delta)+1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K} \frac{144 \mathbb{W}^* \ell_1(\iota_k,\delta)}{c_{1,k}} + 288 V_{\max} S^2 A \ell_2(K,\delta) \log 2KH$$

Proof The proof has two main steps. First, we define a variant of the U_h^k function. For $k \in \mathbb{N}$, define an event $E^k := \{\omega : V_{B^k}^*(s_{B^k}^k) - V_{B^k}^{\pi^k}(s_{B^k}^k) > \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\}$. Define $\ddot{\beta}_h^k(s,a)$ in the following way:

$$\ddot{\beta}_h^k(s,a) = \begin{cases} \frac{72\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} & (h = B^k, (E^k)^{\mathsf{c}}) \\ (8\beta^k(s,a) - \overline{\mathrm{gap}}_h(s,a))_+ & (h \geq B^k, E^k) \\ 0 & (\text{otherwise})\,. \end{cases}$$

In words, $\ddot{\beta}_h^k$ is 0 until the agent makes an sub-optimal action. If a sub-optimal selection happens for the first time at time step h, where $h = B^k$ is implied by definition, denote the difference between the optimal value and the policy's value by $\mathsf{Gap} := V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k)$. If $\frac{1}{2}\mathsf{Gap} < \frac{18\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}$, we set $\ddot{\beta}_h^k(s,a)$ to $\frac{72\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}$, and all subsequent $\ddot{\beta}_{h'}^k$ in the same episode are set to 0. Otherwise, the following sequence of $\ddot{\beta}_{h'}^k$ are set to $(8\beta^k(s,a) - \overline{\mathrm{gap}}_{h'}(s,a))_+$ for $h' \geq h$. We define the clipped sum $\ddot{U}_h^k(s)$ of $\ddot{\beta}_h^k$ iteratively starting from $\ddot{U}_{H+1}^k(s) := 0$ and

$$\ddot{U}_h^k(s) := (\ddot{\beta}_h^k(s,a) + P\ddot{U}_{h+1}^k(s,a)) \wedge 8V_{\max}\,.$$

Then, we have the following lemma:

Lemma 38 *Assume the event of $\mathcal{E}_1(\frac{\delta}{4}) \cap \mathcal{E}_2(\frac{\delta}{4}) \cap \mathcal{E}_3(\frac{\delta}{4})$ and assume that $k > \kappa(\delta)$. Then, for all $s \in \mathcal{S}$, it holds that*

$$V_1^*(s) - V_1^{\pi^k}(s) \leq \ddot{U}_1^k(s)\,.$$

The proof of Lemma 38 is deferred to Appendix D.4.1.
By Lemma 38, we have

$$\mathrm{Reg}_{\mathcal{M}}^{\mathsf{Alg}}(K,\delta) \le V_{\max}\kappa(\delta) + \sum_{k=\kappa(\delta)+1}^{K} \ddot{U}_1^k(s_1^k)\,.$$

Then, we bound the sum of $\ddot{U}_1^k(s_1^k)$ using the following lemma.

Lemma 39 *Under the event of* $\mathcal{E}_4(\{\ddot{\beta}_h^k\}_{k,h}, 8V_{\max}, \frac{\delta}{4})$*, we have that*

$$\sum_{k=\kappa(\delta)+1}^{K} \ddot{U}_1^k(s_1^k) \le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(\frac{2048c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)} \wedge 64c_{1,K}\log\frac{32c_{1,K}}{\overline{\mathrm{gap}}(s,a)}\right) + \sum_{k=\kappa(\delta)+1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K} \frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} + 288V_{\max}S^2A\ell_2(K,\delta)\log 2KH\,.$$

The proof of Lemma 39 is deferred to Appendix D.4.2. The theorem is proved by combining the two lemmas. ■

D.4.1. PROOF OF LEMMA 38

We first prove the following two simple lemmas.

Lemma 40 *For* $h\in[H]$ *and* $s\in\mathcal{S}$*, let* $E_h^k(s) := \{\omega\in\Omega : V_h^*(s) - V_h^{\pi^k}(s) \ge \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\}$*. Then, under the event of Lemma 30, we have*

$$V_h^*(s) - V_h^{\pi^k}(s) \le \mathbb{1}\{E_h^k(s)\}\cdot 2U_h^k(s) + \mathbb{1}\{E_h^k(s)^{\mathsf{C}}\}\left(\frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} \wedge 4V_{\max}\right).$$

Proof By Lemma 30, we have

$$V_h^*(s) - V_h^{\pi^k}(s) \le U_h^k(s) + \frac{18\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\,.$$

The lemma we have to prove is stating that

$$V_h^*(s) - V_h^{\pi^k}(s) \le \begin{cases} 2U_h^k(s) & (V_h^*(s) - V_h^{\pi^k}(s) \ge \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}) \\ \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} \wedge 4V_{\max} & (V_h^*(s) - V_h^{\pi^k}(s) < \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}) \end{cases}.$$

The first case is true since otherwise we would have $V_h^*(s) - V_h^{\pi^k} > \frac{1}{2}(2U_h^k(s) + \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}) = U_h^k(s) + \frac{18\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}$, which is a contradiction. The latter case is trivial. ■

Lemma 41 *Let* $\dot{\beta}_h^k(s,a) = (8\beta^k(s,a) - \overline{\mathrm{gap}}_h(s,a))_+$ *and* $\dot{U}_h^k(s)$ *be defined iteratively starting from* $\dot{U}_{H+1}^k(s) := 0$ *and*

$$\dot{U}_h^k(s) := (\dot{\beta}_h^k(s,a) + P\dot{U}_{h+1}(s,a)) \wedge 8V_{\max}\,.$$

Then, for all $h \in [H]$ *and* $s \in \mathcal{S}$*, we have*

$$4U_h^k(s) - \mathop{\mathbb{E}}_{\pi^k(\cdot|s_h=s)}\left[\sum_{h'=h}^{H} \overline{\mathrm{gap}}_{h'}(s_{h'},a_{h'})\right] \le \dot{U}_h^k(s)\,.$$

Proof We use backward induction on h. The inequality is trivial for $h = H+1$ as $0 \le 0$. Suppose the inequality holds for $h+1$, Then, denoting $a := \pi_h^k(s)$, we have

$$\begin{aligned}
&4U_h^k(s) - \mathop{\mathbb{E}}_{\pi^k(\cdot|s_h=s)}\left[\sum_{h'=h}^{H} \overline{\mathrm{gap}}_{h'}(s_{h'},a_{h'})\right] \\
&\le 4(2\beta^k(s,a) + PU_{h+1}^k(s,a)) - \mathop{\mathbb{E}}_{\pi^k(\cdot|s_h=s)}\left[\sum_{h'=h}^{H} \overline{\mathrm{gap}}_{h'}(s_{h'},a_{h'})\right] \\
&= (8\beta^k(s,a) - \overline{\mathrm{gap}}_h(s,a)) + \mathop{\mathbb{E}}_{\pi^k(\cdot|s_h=s)}\left[4U_{h+1}^k(s_{h+1}) - \sum_{h'=h+1}^{H} \overline{\mathrm{gap}}_{h'}(s_{h'},a_{h'})\right] \\
&\le (8\beta^k(s,a) - \overline{\mathrm{gap}}_h(s,a)) + \mathop{\mathbb{E}}_{\pi^k(\cdot|s_h=s)}\left[\dot{U}_{h+1}^k(s_{h+1})\right] \\
&\le \dot{\beta}_h^k(s,a) + P\dot{U}_{h+1}^k(s,a)\,.
\end{aligned}$$

In addition, since $U_h^k(s) \le 2V_{\max}$, we have $4U_h^k(s) - \mathbb{E}_{\pi^k(\cdot|s_h=s)}\left[\sum_{h'=h}^{H} \overline{\mathrm{gap}}_{h'}(s_{h'},a_{h'})\right] \le 8V_{\max}$. Combining these, we conclude that

$$4U_h^k(s) - \mathop{\mathbb{E}}_{\pi^k(\cdot|s_h=s)}\left[\sum_{h'=h}^{H} \overline{\mathrm{gap}}_{h'}(s_{h'},a_{h'})\right] \le (\dot{\beta}_h^k(s,a) + P\dot{U}_{h+1}^k(s,a)) \wedge 8V_{\max} = \dot{U}_h^k(s)\,.$$

■

Proof [Proof of Lemma 38] For any policy π, We have

$$\begin{aligned}
V_h^*(s) - V_h^\pi(s) &= Q_h^*(s,\pi_h^*(s)) - Q_h^\pi(s,\pi_h(s)) \\
&= Q_h^*(s,\pi_h^*(s)) - Q_h^*(s,\pi_h(s)) + Q_h^*(s,\pi_h(s)) - Q_h^\pi(s,\pi_h(s)) \\
&= \mathrm{gap}_h(s,\pi_h(s)) + P(V_{h+1}^* - V_{h+1}^\pi)(s,\pi_h(s))\,.
\end{aligned}$$

Since $\mathrm{gap}_h(s_h,a_h) = 0$ for $h < B$ by the definition of B, iteratively expanding this equation up to $h = B$ and using the optional stopping theorem yields

$$V_1^*(s) - V_1^\pi(s) = \mathop{\mathbb{E}}_{\pi(\cdot|s_1=s)}\left[V_B^*(s_h) - V_B^\pi(s_h)\right]\,.$$

Lemma 40 states that for all $h \in [H]$ and $s \in \mathcal{S}$, we have

$$V_h^*(s) - V_h^{\pi^k}(s) \le \mathbb{1}\{E_h^k(s)\} \cdot 2U_h^k(s) + \mathbb{1}\{E_h^k(s)^\mathsf{c}\} \left(\frac{36\mathbb{W}^* \ell_1(\iota_k, \delta)}{c_{1,k}} \wedge 4V_{\max} \right).$$

Observe that E^k can be regarded as $E_{B^k}^k(s_{B^k}^k)$, or more precisely, $E^k = \cup_{h,s}(E^k(h,s) \cap \{h = B^k, s = s_{B^k}^k\})$. Then,

$$\begin{aligned}
&V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \\
&= \mathop{\mathbb{E}}_{\pi^k}\left[V_{B^k}^*(s_h^k) - V_{B^k}^{\pi}(s_h^k) \right] \\
&\le \mathop{\mathbb{E}}_{\pi^k}\left[\left(\frac{36\mathbb{W}^* \ell_1(\iota_k, \delta)}{c_{1,k}} \wedge 4V_{\max} \right) \mathbb{1}\{(E^k)^\mathsf{c}, B^k \le H\} + 2U_{B^k}^k(s_{B^k}^k)\mathbb{1}\{E^k\} \right].
\end{aligned}$$

From the definition of effective gap, we have $\mathbb{E}_{\pi^k}\left[\sum_{h=B^k}^H \overline{\mathrm{gap}}_h(s_h^k, a_h^k)\right] \le V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)$. Then, it holds that

$$\begin{aligned}
&V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \\
&= 2(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)) - (V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)) \\
&\le 2 \mathop{\mathbb{E}}_{\pi^k}\left[\left(\frac{36\mathbb{W}^* \ell_1(\iota_k, \delta)}{c_{1,k}} \wedge 4V_{\max} \right) \mathbb{1}\{(E^k)^\mathsf{c}, B^k \le H\} + 2U_{B^k}^k(s_{B^k}^k)\mathbb{1}\{E^k\} \right] - \mathbb{E}_{\pi^k}\left[\sum_{h=B^k}^H \overline{\mathrm{gap}}_h(s_h^k, a_h^k) \right] \\
&= \mathop{\mathbb{E}}_{\pi^k}\left[\left(\frac{72\mathbb{W}^* \ell_1(\iota_k, \delta)}{c_{1,k}} \wedge 8V_{\max} \right) \mathbb{1}\{(E^k)^\mathsf{c}, B^k \le H\} + 4U_{B^k}^k(s_{B^k}^k)\mathbb{1}\{E^k\} - \sum_{h=B^k}^H \overline{\mathrm{gap}}_h(s_h^k, a_h^k) \right].
\end{aligned}$$

We bound the last two terms as

$$\begin{aligned}
\mathop{\mathbb{E}}_{\pi^k}\left[4U_{B^k}^k(s_{B^k}^k)\mathbb{1}\{E^k\} - \sum_{h=B^k}^H \overline{\mathrm{gap}}_h(s_h^k, a_h^k) \right] &\le \mathop{\mathbb{E}}_{\pi^k}\left[\mathbb{1}\{E^k\} \left(4U_{B^k}^k(s_{B^k}^k) - \sum_{h=B^k}^H \overline{\mathrm{gap}}_h(s_h^k, a_h^k) \right) \right] \\
&\le \mathop{\mathbb{E}}_{\pi^k}\left[\mathbb{1}\{E^k\} \dot{U}_{B^k}^k(s_{B^k}^k) \right],
\end{aligned}$$

where we use Lemma 41 for the last inequality. Therefore, we obtain that

$$V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \le \mathop{\mathbb{E}}_{\pi^k}\left[\left(\frac{72\mathbb{W}^* \ell_1(\iota_k, \delta)}{c_{1,k}} \wedge 8V_{\max} \right) \mathbb{1}\{(E^k)^\mathsf{c}, B^k \le H\} + \mathbb{1}\{E^k\} \dot{U}_{B^k}^k(s_{B^k}^k) \right].$$

The proof is completed by noting that the right-hand side is equivalent to $\ddot{U}_1^k(s_1^k)$. ■

D.4.2. PROOF OF LEMMA 39

Proof [Proof of Lemma 39] Under $\mathcal{E}_4(\{\ddot{\beta}_h^k\}_{k,h}, 8V_{\max}, \delta)$ (Lemma 28), we have

$$\sum_{k=\kappa(\delta)+1}^K \ddot{U}_1^k(s_1^k) \le 2 \sum_{k=\kappa(\delta)+1}^K \sum_{h=1}^{\eta^k - 1} \ddot{\beta}_h^k(s_h^k, a_h^k) + 48 V_{\max} SA \log \frac{16H}{\delta}. \tag{20}$$

Note that $\sum_{h=1}^{\eta^k-1} \ddot{\beta}_h^k(s_h^k, a_h^k)$ is either $\frac{72\mathbb{W}^*\ell_1(\iota_k\delta)}{c_{1,k}}$ or $\sum_{h=B^k}^{\eta^k-1}(8\beta^k(s_h^k,a_h^k) - \overline{\mathrm{gap}}_h(s_h^k,a_h^k))_+$. For large enough $c_{1,k}$, specifically if $v_{\mathrm{gap}} > \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}$, then E^k always occur. Therefore, we bound the sum of $\ddot{\beta}_h^k(s_h^k, a_h^k)$ as

$$\sum_{h=1}^{\eta^k-1} \ddot{\beta}_h^k(s_h^k, a_h^k) \le \mathbb{1}\left\{v_{\mathrm{gap}} \le \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\right\}\frac{72\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} + \sum_{h=B^k}^{\eta^k-1}(8\beta^k(s_h^k,a_h^k) - \overline{\mathrm{gap}}_h(s_h^k,a_h^k))_+ \tag{21}$$

By the definition of $\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)$, we have

$$\sum_{k=\kappa(\delta)+1}^{K} \mathbb{1}\left\{v_{\mathrm{gap}} \le \frac{36\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\right\}\frac{72\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} \le \sum_{k=\kappa(\delta)+1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K}\frac{72\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}. \tag{22}$$

Now, we rearrange the sum of $(8\beta^k(s_h^k,a_h^k) - \overline{\mathrm{gap}}_h(s_h^k,a_h^k))_+$ by grouping the same state-action pairs as follows:

$$\begin{aligned}
&\sum_{k=\kappa(\delta)+1}^{K}\sum_{h=B^k}^{\eta^k-1}(8\beta^k(s_h^k,a_h^k) - \overline{\mathrm{gap}}_h(s_h^k,a_h^k))_+ \\
&\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=2}^{N^{K+1}(s,a)}\left(8\left(\frac{4c_{1,K}}{n}\wedge\frac{2\sqrt{2}c_{2,K}}{n} + \frac{18V_{\max}S\ell_2(K,\delta)}{n}\right) - \overline{\mathrm{gap}}(s,a)\right)_+,
\end{aligned}$$

where we use Eq. (19) to bound $\beta^k(s,a)\mathbb{1}\{h<\eta^k\}$. Focusing on one state-action pair, we have

$$\begin{aligned}
&\sum_{n=2}^{N^{K+1}(s,a)}\left(8\left(\frac{4c_{1,K}}{n}\wedge\frac{2\sqrt{2}c_{2,K}}{\sqrt{n}} + \frac{18V_{\max}S\ell_2(K,\delta)}{n}\right) - \overline{\mathrm{gap}}(s,a)\right)_+ \\
&\le \sum_{n=2}^{N^{K+1}(s,a)}\left(\frac{32c_{1,K}}{n}\wedge\frac{16\sqrt{2}c_{2,K}}{\sqrt{n}} - \overline{\mathrm{gap}}(s,a)\right)_+ + \sum_{n=2}^{N^{K+1}(s,a)}\frac{144V_{\max}S\ell_2(K,\delta)}{n}.
\end{aligned} \tag{23}$$

Using properties of maximums and minimums, we have

$$\begin{aligned}
&\sum_{n=2}^{N^{K+1}(s,a)}\left(\frac{32c_{1,K}}{n}\wedge\frac{16\sqrt{2}c_{2,K}}{\sqrt{n}} - \overline{\mathrm{gap}}(s,a)\right)_+ \\
&= \sum_{n=2}^{N^{K+1}(s,a)}\left(\left(\frac{32c_{1,K}}{n} - \overline{\mathrm{gap}}\right)\wedge\left(\frac{16\sqrt{2}c_{2,K}}{\sqrt{n}} - \overline{\mathrm{gap}}(s,a)\right)\right)_+ \\
&= \sum_{n=2}^{N^{K+1}(s,a)}\left(\frac{32c_{1,K}}{n} - \overline{\mathrm{gap}}\right)_+\wedge\left(\frac{16\sqrt{2}c_{2,K}}{\sqrt{n}} - \overline{\mathrm{gap}}(s,a)\right)_+ \\
&\le \left(\sum_{n=2}^{N^{K+1}(s,a)}\left(\frac{32c_{1,K}}{n} - \overline{\mathrm{gap}}(s,a)\right)_+\right)\wedge\left(\sum_{n=2}^{N^{K+1}(s,a)}\left(\frac{16\sqrt{2}c_{2,K}}{\sqrt{n}} - \overline{\mathrm{gap}}(s,a)\right)_+\right).
\end{aligned}$$

Using Lemma 57, the first sum is bounded as

$$\sum_{n=2}^{N^{K+1}(s,a)} \left(\frac{32c_{1,K}}{n} - \overline{\mathrm{gap}}(s,a) \right)_+ \leq 32c_{1,K} \log \frac{32c_{1,K}}{\overline{\mathrm{gap}}(s,a)} ,$$

and the second sum is bounded as

$$\sum_{n=2}^{N^{K+1}(s,a)} \left(\frac{16\sqrt{2}c_{2,K}}{\sqrt{n}} - \overline{\mathrm{gap}}(s,a) \right)_+ \leq \frac{1024c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)} .$$

Now, we bound the remaining sum in Eq. (23) as follows:

$$\sum_{n=2}^{N^{K+1}(s,a)} \frac{144V_{\max}S\ell_2(K,\delta)}{n} \leq 144V_{\max}S\ell_2(K,\delta)\log KH .$$

Plugging these bounds back to Eq. (23), we obtain that the sum of $(8\beta^k(s_h^k,a_h^k) - \overline{\mathrm{gap}}_h(s,a))_+$ for one state-action pair is bounded as

$$\left(32c_{1,K}\log\frac{32c_{1,K}}{\overline{\mathrm{gap}}(s,a)}\right) \wedge \left(\frac{1024c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)}\right) + 144V_{\max}S\ell_2(K,\delta)\log KH ,$$

and taking the sum over all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we obtain that

$$\begin{aligned}
&\sum_{k=\kappa(\delta)+1}^{K} \sum_{h=B^k}^{\eta^k-1} (8\beta^k(s_h^k,a_h^k) - \overline{\mathrm{gap}}(s_h^k,a_h^k))_+ \\
&\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(32c_{1,K}\log\frac{32c_{1,K}}{\overline{\mathrm{gap}}(s,a)}\right) \wedge \left(\frac{1024c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)}\right) + 144V_{\max}S^2A\ell_2(K,\delta)\log KH . \quad (24)
\end{aligned}$$

Combining Eq. (20), Eq. (21), Eq. (22), and Eq. (24), we obtain that

$$\begin{aligned}
&\sum_{k=\kappa(\delta)+1}^{K} \ddot{U}_1^k(s_1^k) \\
&\leq \sum_{k=\kappa(\delta)+1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K} \frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} \\
&\quad + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(64c_{1,K}\log\frac{32c_{1,K}}{\overline{\mathrm{gap}}(s,a)}\right) \wedge \left(\frac{2048c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)}\right) \\
&\quad + 288V_{\max}S^2A\ell_2(K,\delta)\log KH + 48V_{\max}SA\log\frac{16H}{\delta} .
\end{aligned}$$

We bound the last term as $48V_{\max}SA\log\frac{16H}{\delta} \leq 48V_{\max}S^2A\ell_2(K,\delta) \leq 288(\log 2)V_{\max}S^2A\ell_2(K,\delta)$, which leads to

$$288V_{\max}S^2A\ell_2(K,\delta)\log KH + 48V_{\max}SA\log\frac{16H}{\delta} \leq 288V_{\max}S^2A\ell_2(K,\delta)\log 2KH .$$

■

D.5. Possibility of Obtaining Better Second-order Terms

We note that the $\mathcal{O}(V_{\max} S^2 A \ell_2(K,\delta) \log KH)$ terms in Theorems 34 and 37, and consequently those in Corollaries 11, 43, and 44, can be replaced with $\mathcal{O}(V_{\max} SA(S \log KH + \log \frac{HSA}{\delta}) \log KH)$ by substituting Lemma 27 with Lemmas 3 and 10 in Ménard et al. (2021). Specifically, a $\mathcal{O}(S \log \frac{HSA \log K}{\delta})$ factor can be replaced by a $\mathcal{O}(S \log KH + \log \frac{HSA}{\delta})$ factor. This modification improves the coefficient of the $\log(1/\delta)$ term by a factor of S at the cost of an additional $\mathcal{O}(S \log KH)$ term. While further improvement to $\mathcal{O}(S \log \log KH + \log \frac{HSA}{\delta})$ may be possible, which would yield a strict improvement over our current results, we do not pursue optimizing the second-order terms in this work.

Appendix E. Proof of Corollaries in Section 6.3

In this section, we prove Corollaries 11–13.

While plugging in the values of $c_{1,K}$ and $c_{2,K}$ is straightforward, bounding $\kappa(\delta)$ is non-trivial. First, we present a simple technique that circumvents the direct computation of $\kappa(\delta)$ in certain cases. For simplicity, we denote the condition in the definition of $\kappa(\delta)$ by $I(k,\delta) := \mathbb{1}\{c_{1,k} < (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta)$ or $c_{2,k} < \left(\sigma_{\max} + \frac{6\mathbb{W}^*_{\mathrm{diff}} \ell_1(\iota_k,\delta)}{c_{1,k}}\right)\sqrt{\ell_2(k,\delta)}\}$. Then, we have

$$
\begin{aligned}
&V_{\max}(\kappa(\delta) \wedge K) \\
&= \sum_{k=1}^{K} V_{\max} I(k,\delta) \\
&= \sum_{k=1}^{\kappa(\delta)} \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{c_{1,k}} + \sum_{k=1}^{K} \left(V_{\max} - \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{c_{1,k}}\right) I(k,\delta) \\
&\le \sum_{k=1}^{\kappa(\delta)} \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{c_{1,k}} + \sum_{k=1}^{K} \left(V_{\max} - \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{c_{1,k}}\right)_+ I(k,\delta)\,.
\end{aligned}
$$

Now, we consider a new condition $\mathbb{1}\{V_{\max} \ge \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{c_{1,k}}\} I(k,\delta)$. The resulting condition on $c_{1,k}$ can be written as $c_{1,k} < (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta)$ and $c_{1,k} \ge \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{V_{\max}}$. When considering the worst-case regret, we must upper-bound $\mathbb{W}^*$ by $2V^2_{\max}$, and this condition becomes impossible to satisfy since $2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha \le 11 V_{\max} \le \frac{18\mathbb{W}^*}{V_{\max}}$. The condition on $c_{2,k}$ can be written as $c_{2,k} < (\sigma_{\max} + \frac{6\mathbb{W}^*_{\mathrm{diff}} \ell_1(\iota_k,\delta)}{c_{1,k}})\sqrt{\ell_2(k,\delta)}$ and $c_{1,k} \ge \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{V_{\max}}$. Using that $\mathbb{W}^*_{\mathrm{diff}} \le \mathbb{W}^*$, these two conditions imply that $c_{2,k} < (\sigma_{\max} + \frac{V_{\max}}{3})\sqrt{\ell_2(k,\delta)}$, which leads to a simpler calculation. We summarize this discussion by the following lemma:

Lemma 42 *Let*

$$
\begin{aligned}
\kappa_1(\delta) &:= \max\left\{k \in \mathbb{N} : \frac{18\mathbb{W}^* \ell_1(\iota_k,\delta)}{V_{\max}} < c_{1,k} < (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}} \vee 2V_\alpha)\ell_1(1,\delta))\right\} \\
\kappa_2(\delta) &:= \max\left\{k \in \mathbb{N} : c_{2,k} < (\sigma_{\max} + \frac{V_{\max}}{3})\sqrt{\ell_2(k,\delta)}\right\}\,.
\end{aligned}
$$

Then, we have

$$V_{\max}\kappa(\delta) \le \sum_{k=1}^{\kappa(\delta)} \frac{18\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} + V_{\max}(\kappa_1(\delta) \vee \kappa_2(\delta))\,.$$

E.1. Proof of Corollary 11

Proof [Proof of Corollary 11] By Lemma 42, we upper bound $V_{\max}\kappa(\delta)$ by $\sum_{k=1}^{\kappa(\delta)} \frac{18V_{\max}^2}{c_{1,k}}$, where plugging in $\mathbb{W}^* = 2V_{\max}^2$ and $c_{2,k} = \infty$ yields $\kappa_1(\delta) = \kappa_2(\delta) = 0$. Then, Theorem 34 implies that

$$\mathrm{Reg}_{\mathcal{M}}^{\mathsf{Alg}}(K,\delta) \le \sum_{k=1}^{K} \frac{18V_{\max}^2\ell_1(\iota_k,\delta)}{c_{1,k}} + 16c_{1,K}SA\log KH + 72V_{\max}S^2A\ell_2(K,\delta)\log 2KH$$

We plug in $c_{1,k} = c_1 V_{\max}\sqrt{\frac{K\ell_1(\iota_k,1)}{SA\log kH}}$ and bound each term. The first sum is bounded as

$$\begin{aligned}
\sum_{k=1}^{K} \frac{18V_{\max}^2\ell_1(\iota_k,\delta)}{c_{1,k}} &\le 18V_{\max}\sqrt{SA\log KH}\sum_{k=1}^{K}\frac{\ell_1(\iota_k,\delta)}{\sqrt{k\ell(\iota_k,1)}} \\
&\le 18V_{\max}\sqrt{SA\log KH}\sum_{k=1}^{K}\left(\frac{\log\frac{1}{\delta}}{c_1\sqrt{k\ell(1,1)}} + \frac{\sqrt{\ell_1(\iota_k,1)}}{c_1\sqrt{k}}\right) \\
&\le 18V_{\max}\sqrt{SA\log KH}\cdot 2\sqrt{K}\left(\frac{\log\frac{1}{\delta}}{c_1\sqrt{\ell_1(1,1)}} + \frac{\sqrt{\ell_1(\iota_K,1)}}{c_1}\right) \\
&= 36\left(\frac{\log\frac{1}{\delta}}{c_1\sqrt{\ell_1(1,1)}} + \frac{\sqrt{\ell_1(\iota_K,1)}}{c_1}\right)V_{\max}\sqrt{SAK\log KH}\,.
\end{aligned}$$

The second term becomes

$$8c_{1,K}SA\log KH = 8c_1V_{\max}\sqrt{SAK(\log KH)\ell_1(\iota_K,1)}\,.$$

Therefore, we derive the total bound as

$$\begin{aligned}
\mathrm{Reg}_{M}^{\mathsf{Alg}}(K,\delta) \le &\left(\frac{36\log\frac{1}{\delta}}{c_1\ell(1,1)} + \frac{36}{c_1} + 8c_1\right)V_{\max}\sqrt{KSA\ell(\iota_K,1)\log KH} \\
&+ 72V_{\max}S^2A\ell_2(K,\delta)\log 2KH\,.
\end{aligned}$$

■

E.2. Proof of Corollary 12

Corollary 43 (Restatement of Corollary 12) *Assume that* $\sigma_{\max}^2 \le 2V_{\max}^2$. *Set* $c_{1,k} = c_1 V_{\max}\sqrt{\frac{k\ell_1(k,1)}{SA\log KH}}$ *and* $c_{2,k} = c_2 V_{\max}\sqrt{\log 32HSAk}$ *for some constants* $c_1 > 0$ *and* $c_2 \ge 2$. *Then, we have*

$$\begin{aligned}\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) \le V_{\max}\exp\left(\mathcal{O}\left(\frac{1}{c_2^2}\log\frac{1}{\delta}\right)\right) + \mathcal{O}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{c_2^2 V_{\max}^2\log HSAK}{\overline{\mathrm{gap}}(s,a)}\right)\\ + \mathcal{O}\left(\left(\frac{\mathbb{W}^{*2}}{v_{\mathrm{gap}}V_{\max}^2} + \frac{\mathbb{W}_{\mathrm{diff}}^* \vee V_\alpha^2}{V_{\max}}\right)\left(\frac{(\ell_1(\iota_K,\delta))^2}{c_1^2\ell_1(1,1)}\right)SA(\log KH)\right)\\ + \mathcal{O}\left(V_{\max}S^2A\ell_2(K,\delta)\log KH\right).\end{aligned}$$

Proof By Theorem 34, we have

$$\begin{aligned}\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) &= V_{\max}(\kappa(\delta)\wedge K) + \sum_{k=\kappa(\delta)+1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K}\frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\\ &\quad + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(\frac{2048c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)}\wedge 64c_{1,K}\log\frac{32c_{1,K}}{\overline{\mathrm{gap}}(s,a)}\right) + 288V_{\max}S^2A\ell_2(K,\delta)\log 2KH\\ &= \mathcal{O}\Bigg(V_{\max}(\kappa(\delta)\wedge K) + \sum_{k=\kappa(\delta)}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K}\frac{\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\\ &\quad + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{V_{\max}^2\log HSAK}{\overline{\mathrm{gap}}(s,a)} + V_{\max}S^2A\ell_2(K,\delta)\log KH\Bigg).\end{aligned}$$

We first bound $\kappa(\delta)\wedge K$ using Lemma 42. First, $c_{1,k} = c_1V_{\max}\sqrt{\frac{k\ell_1(\iota_k,1)}{SA\log kH}} < (2\sqrt{13\mathbb{W}_{\mathrm{diff}}^*}\vee 2V_\alpha)\ell_1(1,\delta)$ and $k\le K$ implies

$$\begin{aligned}k &< \frac{SA\log kH}{\ell_1(\iota_k,1)}\left(\frac{(2\sqrt{13\mathbb{W}_{\mathrm{diff}}^*}\vee 2V_\alpha)\ell_1(1,\delta)}{c_1V_{\max}}\right)^2\\ &= \mathcal{O}\left(\frac{SA(\mathbb{W}_{\mathrm{diff}}^*\vee V_\alpha^2)\log KH}{\ell_1(1,1)}\left(\frac{\ell_1(1,\delta)}{c_1V_{\max}}\right)^2\right),\end{aligned}$$

which implies that $\kappa_1(\delta)\wedge K = \mathcal{O}\left(\frac{SA(\mathbb{W}_{\mathrm{diff}}^*\vee V_\alpha^2)\log KH}{\ell_1(1,1)}\left(\frac{\ell_1(1,\delta)}{c_1V_{\max}}\right)^2\right)$. For $\kappa_2(\delta)$, we may relax the condition on $c_{2,k}$ using the assumption $\sigma_{\max}^2 \le 2V_{\max}^2$ to $c_{2,k} = c_2V_{\max}\sqrt{\log 32HSAk} < 2V_{\max}\sqrt{\log\frac{32HSA(2+\log kH)^2}{\delta}}$. Using that $c_2\ge 2$, this condition implies that $\log k - 2\log(2+\log kH) < \frac{4}{c_2^2}\log\frac{1}{\delta}$, which in turn yields $\kappa_2(\delta) = \exp(\mathcal{O}(\frac{1}{c_2^2}\log\frac{1}{\delta}))$. Hence, we have

$$V_{\max}(\kappa(\delta)\wedge K) = V_{\max}\exp\left(\mathcal{O}\left(\frac{1}{c_2^2}\log\frac{1}{\delta}\right)\right) + \mathcal{O}\left(\frac{SA(\mathbb{W}_{\mathrm{diff}}^*\vee V_\alpha^2)(\log KH)(\ell_1(1,\delta))^2}{c_1^2V_{\max}\ell_1(1,1)}\right).$$

Now, we bound $\mathbb{W}^* \sum_{k=1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K} \frac{\ell_1(\iota_k,\delta)}{c_{1,k}}$. First, for $k \le K$, we have

$$\begin{aligned}\frac{\ell_1(\iota_k,\delta)}{c_{1,k}} &= \frac{(\ell_1(\iota_k,1)+\log\frac{1}{\delta})\sqrt{SA\log kH}}{c_1 V_{\max}\sqrt{k\ell_1(\iota_k,1)}} \\ &\le \frac{\sqrt{SA\log KH}}{c_1 V_{\max}}\left(\sqrt{\ell_1(\iota_K,1)}+\frac{\log\frac{1}{\delta}}{\sqrt{\ell_1(1,1)}}\right)\frac{1}{\sqrt{k}}\,.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\sum_{k=1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K} \frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} \\ &\le \sum_{k=1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K} \frac{144\mathbb{W}^*\sqrt{SA\log KH}}{c_1 V_{\max}}\left(\sqrt{\ell_1(\iota_K,1)}+\frac{\log\frac{1}{\delta}}{\sqrt{\ell_1(1,1)}}\right)\frac{1}{\sqrt{k}}\,.\end{aligned}$$

Applying Lemma 57 with $\varepsilon = 4v_{\mathrm{gap}}$ and $\alpha = \frac{1}{2}$, we have

$$\begin{aligned}&\sum_{k=1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K} \frac{144\mathbb{W}^*\sqrt{SA\log KH}}{c_1 V_{\max}}\left(\sqrt{\ell_1(\iota_K,1)}+\frac{\log\frac{1}{\delta}}{\sqrt{\ell_1(1,1)}}\right)\frac{1}{\sqrt{k}} \\ &= \mathcal{O}\left(\frac{\mathbb{W}^{*2}SA(\log KH)(\ell_1(\iota_K,\delta))^2}{c_1^2 v_{\mathrm{gap}}V_{\max}^2\ell_1(1,1)}\right).\end{aligned}$$

Combining all the bounds, we obtain that

$$\begin{aligned}\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) \le V_{\max}\exp\left(\mathcal{O}\left(\frac{1}{c_2^2}\log\frac{1}{\delta}\right)\right) &+ \mathcal{O}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{V_{\max}^2\log HSAK}{\overline{\mathrm{gap}}(s,a)}\right) \\ &+ \mathcal{O}\left(\left(\frac{\mathbb{W}^{*2}}{v_{\mathrm{gap}}V_{\max}^2}+\frac{\mathbb{W}^*_{\mathrm{diff}}\vee V_\alpha^2}{V_{\max}}\right)\left(\frac{(\ell_1(\iota_K,\delta))^2}{c_1^2\ell_1(1,1)}\right)SA(\log KH)\right) \\ &+ \mathcal{O}\left(V_{\max}S^2A\ell_2(K,\delta)\log KH\right).\end{aligned}$$

■

E.3. Proof of Corollary 13

Corollary 44 (Formal statement of Corollary 13) *Let $c_1 > 0, c_2 > 0$, $\alpha \in [\frac{1}{2},1]$, and $0 < \beta \le \alpha$ be constants. Set $c_{1,k} = c_1 V_{\max}(\frac{k}{SA})^\alpha$ and $c_{2,k} = c_2\sqrt{k^\beta}$. Then, we have the worst-case*

distributional regret bound of

$$\begin{aligned}\sup_{M\in\mathcal{M}} \mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) &\le V_{\max}\left(\frac{4V_{\max}}{c_2}\right)^{\frac{2}{\beta}} \ell_2(K,\delta)^{\frac{1}{\beta}} \\ &\quad + \frac{36V_{\max}}{c_1}\left(\frac{1}{1-\alpha}\wedge\log K\right) K^{1-\alpha}(SA)^{\alpha}\ell_1(\iota_K,\delta) \\ &\quad + \left(16c_1V_{\max}K^{\alpha}(SA)^{1-\alpha}\log KH\right) \wedge \left(16\sqrt{2}c_2\sqrt{HSA}K^{\beta+\frac{1}{2}}\right) \\ &\quad + 72V_{\max}S^2A\ell_2(K,\delta)\log 2KH \\ &= \widetilde{\mathcal{O}}_{K,\delta}\left(\left(\log\frac{1}{\delta}\right)^{\frac{1}{\beta}} + K^{1-\alpha}\log\frac{1}{\delta} + K^{\alpha}\wedge K^{\beta+\frac{1}{2}}\right)\end{aligned}$$

and the instance-dependent distributional regret bound of

$$\begin{aligned}\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) &\le V_{\max}\left(\frac{4V_{\max}}{c_2}\right)^{\frac{2}{\beta}} \ell_2(K,\delta)^{\frac{1}{\beta}} \\ &\quad + 4SA\left(\frac{1}{1-\alpha}\wedge\log K\right)\left(\frac{36\mathbb{W}^*\ell_1(\iota_K,\delta)}{c_1V_{\max}}\right)^{\frac{1}{\alpha}}\left(\frac{1}{v_{\mathrm{gap}}}\right)^{\frac{1}{\alpha}-1} \\ &\quad + \sum_{(s,a)\times\mathcal{S}\times\mathcal{A}} \frac{2048c_2^2K^{\beta}}{\overline{\mathrm{gap}}(s,a)} + 288V_{\max}S^2A\ell_2(K,\delta)\log 2KH \\ &= \widetilde{\mathcal{O}}_{K,\delta,\overline{\mathrm{gap}}}\left(\left(\log\frac{1}{\delta}\right)^{\frac{1}{\beta}} + \left(\frac{1}{v_{\mathrm{gap}}}\right)^{\frac{1}{\alpha}-1}\left(\log\frac{1}{\delta}\right)^{\frac{1}{\alpha}} + \sum_{s,a}\frac{K^{\beta}}{\overline{\mathrm{gap}}(s,a)}\right).\end{aligned}$$

Proof Worst-case bound. By Theorem 34, we have

$$\begin{aligned}\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) &\le V_{\max}(\kappa(\delta)\wedge K) + \sum_{k=\kappa(\delta)+1}^{K}\frac{36V_{\max}^2\ell_1(\iota_k,\delta)}{c_{1,k}} \\ &\quad + (16c_{1,k}SA\log KH)\wedge\left(16\sqrt{2}c_{2,K}\sqrt{HSAK}\right) \\ &\quad + 72V_{\max}S^2\ell_2(K,\delta)\log 2KH\,.\end{aligned}$$

Using Lemma 42, we bound the first two terms as

$$V_{\max}(\kappa(\delta)\wedge K) + \sum_{k=\kappa(\delta)+1}^{K}\frac{36V_{\max}^2\ell_1(\iota_k,\delta)}{c_{1,k}} \le V_{\max}(\kappa_2(\delta)\wedge K) + \sum_{k=1}^{K}\frac{36V_{\max}^2\ell_1(\iota_k,\delta)}{c_{1,k}}\,.$$

We first bound $\kappa_2(\delta)$. $c_{2,k} = c_2\sqrt{k^{\beta}} < (\sigma_{\max}+\frac{V_{\max}}{3})\sqrt{\ell_2(k,\delta)} \le 2V_{\max}\sqrt{\ell_2(k,\delta)}$ and $k\le K$ implies that $k \le \left(\left(\frac{2V_{\max}}{c_2}\right)^2\ell_2(K,\delta)\right)^{\frac{1}{\beta}}$, so we have $\kappa_2(\delta)\wedge K \le \left(\frac{2V_{\max}}{c_2}\right)^{\frac{2}{\beta}}(\ell_2(K,\delta))^{\frac{1}{\beta}}$.

Next, we bound $\sum_{k=1}^{K}\frac{36V_{\max}^2\ell_1(\iota_k,\delta)}{c_{1,k}}$. The summand is upper-bounded by $\frac{36V_{\max}(SA)^{\alpha}\ell_1(\iota_K,\delta)}{c_1k^{\alpha}}$. By Lemma 56, we have

$$\sum_{k=1}^{K}\frac{36V_{\max}(SA)^{\alpha}\ell_1(\iota_K,\delta)}{c_1k^{\alpha}} \le \left(\frac{1}{1-\alpha}\wedge\log K\right)\frac{36V_{\max}(SA)^{\alpha}K^{1-\alpha}\ell_1(\iota_K,\delta)}{c_1}\,.$$

For the third term, we have $16c_{1,k}SA\log KH = 16c_1V_{\max}(SA)^{1-\alpha}K^{\alpha}\log KH$ and $16\sqrt{2}c_{2,K}\sqrt{HSAK} = 16\sqrt{2}c_2\sqrt{HSA}K^{\beta+\frac{1}{2}}$, the the fourth term does not need further modification.

Instance-dependent bound. By Theorem 37, we have

$$\mathrm{Reg}_{\mathcal{M}}^{\mathsf{Alg}}(K,\delta) \le V_{\max}(\kappa(\delta)\wedge K) + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{2048c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)} + \sum_{k=\kappa(\delta)+1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K}\frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} + 288V_{\max}S^2A\ell_2(K,\delta)\log 2KH\,.$$

We use Lemma 42 to bound $\kappa(\delta)\wedge K$. We have $\kappa_2(\delta)\wedge K \le \left(\frac{2V_{\max}}{c_2}\right)^{\frac{2}{\beta}}(\ell_2(K,\delta))^{\frac{1}{\beta}}$ from the previous case. For $\kappa_1(\delta)$, we have that $c_{1,k} = c_1V_{\max}(\frac{k}{SA})^{\alpha} < (2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}}\vee 2V_{\alpha})\ell_1(\iota_k,\delta)$ and $k\le K$ implies $k < SA\left(\frac{(2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}}\vee 2V_{\alpha})\ell_1(\iota_K,\delta)}{c_1V_{\max}}\right)^{\frac{1}{\alpha}}$, which yields $\kappa_1(\delta)\wedge K < SA\left(\frac{(2\sqrt{13\mathbb{W}^*_{\mathrm{diff}}}\vee 2V_{\alpha})\ell_1(\iota_K,\delta)}{c_1V_{\max}}\right)^{\frac{1}{\alpha}}$.

Next, we bound $\sum_{k=1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K}\frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}$. We have

$$\begin{aligned}
\sum_{k=1}^{\kappa_{\mathrm{gap}}(v_{\mathrm{gap}},\delta)\wedge K}\frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} &= \sum_{k=1}^{K}\frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}}\mathbb{1}\left\{\frac{144\mathbb{W}^*\ell_1(\iota_k,\delta)}{c_{1,k}} \ge 4v_{\mathrm{gap}}\right\}\\
&\le \sum_{k=1}^{K}\frac{144\mathbb{W}^*(SA)^{\alpha}\ell_1(\iota_K,\delta)}{c_1V_{\max}k^{\alpha}}\mathbb{1}\left\{\frac{144\mathbb{W}^*(SA)^{\alpha}\ell_1(\iota_K,\delta)}{c_1V_{\max}k^{\alpha}} \ge 4v_{\mathrm{gap}}\right\}\\
&\le SA\left(\frac{1}{1-\alpha}\wedge\log K\right)\left(\frac{144\mathbb{W}^*\ell_1(\iota_K,\delta)}{c_1V_{\max}}\right)^{\frac{1}{\alpha}}\left(\frac{1}{4v_{\mathrm{gap}}}\right)^{\frac{1}{\alpha}-1}\\
&= 4SA\left(\frac{1}{1-\alpha}\wedge\log K\right)\left(\frac{36\mathbb{W}^*\ell_1(\iota_K,\delta)}{c_1V_{\max}}\right)^{\frac{1}{\alpha}}\left(\frac{1}{v_{\mathrm{gap}}}\right)^{\frac{1}{\alpha}-1},
\end{aligned}$$

where the last inequality is due to Lemma 57.

For the remaining terms, we have $\sum_{s,a}\frac{2048c_{2,K}^2}{\overline{\mathrm{gap}}(s,a)} = \sum_{s,a}\frac{2048c_2^2K^{\beta}}{\overline{\mathrm{gap}}(s,a)}$ and $288V_{\max}S^2A\ell_2(K,\delta)\log 2KH$ does not need further modification. ■

Appendix F. Comparison with Existing Distributional Regret Bounds

In this section, we discuss the distributional regret bounds in Simchi-Levi et al. (2023a) and Khodadadian and Moharrami (2025) in more details and compare them with our results. The bounds in Simchi-Levi et al. (2022); Khodadadian and Moharrami (2025) are expressed in the form of $\mathbb{P}(\mathrm{Reg}_M^{\mathsf{Alg}}(K) > x)$ as a function of x, so we restate their results to be consistent with our presentation, inverting their bounds from a function of x to a function of δ.

We present Theorem 3 in Simchi-Levi et al. (2023a), which considers the MAB setting. We note that this result is more general than the results in Simchi-Levi et al. (2025).

Theorem 45 (Theorem 3 in Simchi-Levi et al. (2023a), restated with respect to δ) *Suppose $M\in\mathcal{B}(\sigma)$. Set $c_{1,t} = \eta_1\left(\frac{t}{A}\right)^{\alpha}\sqrt{\log A}$ and $c_{2,t} = \eta_2\sqrt{t^{\beta}}$ for some constants $\eta_1,\eta_2>0$, $\alpha\in[\frac{1}{2},1)$, and*

$0 < \beta \leq \alpha$. *Then, the worst-case distributional regret bound is given as*

$$\begin{aligned}\sup_{M \in \mathcal{B}(\sigma)} \mathrm{Reg}_M^{\mathsf{Alg}}(T, \delta) &= \mathcal{O}\Bigg(A + \sigma\sqrt{AT \log \frac{A}{\delta}} + \left(\sigma^2 \log \frac{A}{\delta}\right)^{\frac{1}{\beta}} \\ &\qquad + \frac{1}{1-\alpha} A^{1-\alpha} T^{\alpha} \sqrt{\log A} + \frac{1}{1-\alpha} A^{\alpha} T^{1-\alpha} \frac{\log \frac{A}{\delta}}{\sqrt{\log A}} \Bigg) \\ &= \widetilde{\mathcal{O}}_{T,\delta}\left(\left(\log \frac{1}{\delta}\right)^{\frac{1}{\beta}} + T^{1-\alpha} \log \frac{1}{\delta} + T^{\alpha} \right)\end{aligned}$$

and the instance-dependent distributional regret bound is given as

$$\begin{aligned}\mathrm{Reg}_M^{\mathsf{Alg}}(T, \delta) &= \mathcal{O}\Bigg(A + \sum_{a \in \mathcal{A}} \frac{\sigma^2 \log \frac{A}{\delta} + T^{\beta}}{\mathrm{gap}(a)} \\ &\qquad + A \left(\frac{\sigma^2 \log \frac{A}{\delta}}{\mathrm{gap}_{\min} \sqrt{\log A}} \right)^{\frac{1}{\alpha}} + \left(\sigma^2 \log \frac{A}{\delta}\right)^{\frac{1}{\beta}} \Bigg) \\ &= \widetilde{\mathcal{O}}_{T,\delta,\mathrm{gap}} \left(\left(\log \frac{1}{\delta}\right)^{\frac{1}{\beta}} + \left(\frac{1}{\mathrm{gap}_{\min}}\right)^{\frac{1}{\alpha}} \left(\log \frac{1}{\delta}\right)^{\frac{1}{\alpha}} + \sum_{a \in \mathcal{A}} \frac{\log \frac{1}{\delta} + T^{\beta}}{\mathrm{gap}(a)} \right).\end{aligned}$$

Comparison of Corollary 44 and Theorem 45. We compare Corollary 44 and Theorem 45, which take the same parameters, assuming that the instance considered in Corollary 44 is an MAB instance. We achieve the same order of regret in the worst-case bound up to logarithmic factors. When considering the logarithmic factors, while Corollary 44 includes an additional $\log T$ factor, our analysis for bandits (Theorem 18) shows that the logarithmic factor can be improved in the MAB setting, even achieving a better logarithmic dependence than Theorem 45.
We make several improvements to the instance dependent bound. First, we reduce the $(\frac{1}{\mathrm{gap}_{\min}})^{\frac{1}{\alpha}}$ dependence to $(\frac{1}{\mathrm{gap}_{\min}})^{\frac{1}{\alpha}-1}$, which is an improvement from $(\frac{1}{\mathrm{gap}_{\min}})^2$ to $\frac{1}{\mathrm{gap}_{\min}}$ under the standard choice of $\alpha = \frac{1}{2}$. Additionally, we do not incur a $\sum_{a \in \mathcal{A}} \frac{\log \frac{1}{\delta}}{\mathrm{gap}(a)}$ term, thereby improving the coefficient of the $\log \frac{1}{\delta}$ term.

Remark 46 *The recent preprint of Simchi-Levi et al. (2023a) considers a more general parameter choice of $c_{2,t} = \sqrt{f(t)}$ for any increasing function $f(t) = \omega(\log t)$. Their bounds for $\mathbb{P}(\mathrm{Reg}_M^{\mathsf{Alg}}(T) > x)$ involve an integral $\int_0^T \exp(-\frac{f(x \vee y)}{2\sigma^2})\, dy$. While this term is difficult to invert directly, a coarse approximation of $T \exp(-\frac{f(x)}{2\sigma^2})$ yields that it translates to $f^{-1}(2\sigma^2 \log \frac{T}{\delta})$, which corresponds with our $\tau_2(\delta)$ or $\kappa(\delta)$ in the sense that it coincides with the number of time steps required for $c_{2,t} \geq \sigma\sqrt{2 \log \frac{T}{\delta}}$ to hold. We take $f(t) = t^{\beta}$ in Theorem 45, an example considered in Simchi-Levi et al. (2023a), for a more concrete comparison, and our improvement is valid even if we consider arbitrary parameters of $c_{2,k}$.*

We present Theorem 1 in Khodadadian and Moharrami (2025), which considers the RL setting.

Theorem 47 (Theorem 1 in Khodadadian and Moharrami (2025), restated with respect to δ) *Let $V_{\max} = H$. Take $c_{1,k} = \infty$ and $c_{2,k} = H\sqrt{(0.5\log 2)S + \mu(1+k)^{\beta}}$ for some constants $\mu > 0$ and $\beta \in [0,1]$. Then, it holds that*

$$\mathrm{Reg}_M^{\mathsf{Alg}}(K,\delta) = \mathcal{O}\left(\frac{H^6 SA(S+\mu K^{\beta})}{\mathrm{gap}_{\min}} + H^2\sqrt{HK\log\frac{1}{\delta}} + H\left(\frac{1}{\mu}\log\frac{1}{\delta}\right)^{\frac{1}{\beta}}\right).$$

Comparison of Corollary 44 and Theorem 47. While we do not show the bounds for the exact same parameter choice with Khodadadian and Moharrami (2025), we compare their results with our Corollary 44. Taking $c_1 = \infty$ and $c_2 = H$ recovers their parameter except for an S factor, which is not necessary in our work. We observe multiple improvements over their results. First, the bound of Theorem 47 has a term linear in $\frac{SA}{\mathrm{gap}_{\min}}$, whereas we provide a fine-grained gap dependence of $\sum_{s,a}\frac{1}{\mathrm{gap}(s,a)}$. Furthermore, while we have at most H^2 dependence when $V_{\max} = H$, the order of H is H^6 in Theorem 47. In addition, their bound further incurs a $H^2\sqrt{HK\log\frac{1}{\delta}}$ factor, which does not appear in our bound.

Appendix G. Concentration Inequalities

In this section, we provide general concentration inequalities for sub-exponential random variables.

Lemma 48 *Let $\{X_t\}_{t=1}^{\infty}$ be a martingale difference sequence adapted to filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Suppose there exists a sequence of predictable random variables $\{\sigma_t\}_{t=1}^{\infty}$ with respect to $\{\mathcal{F}_t\}_{t=0}^{\infty}$ and a constant $\alpha \geq 0$ such that X_t is $\mathcal{F}_{t-1}$-conditionally (σ_t,α)-sub-exponential for all $t \in \mathbb{N}$. Then, for any $\lambda \in (0, 1/\alpha]$ and $\delta \in (0,1]$, it holds that*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \sum_{t=1}^{n} X_t \geq \frac{\lambda}{2}\sum_{t=1}^{n}\sigma_t^2 + \frac{1}{\lambda}\log\frac{1}{\delta}\right) \leq \delta\,.$$

Proof The proof is a standard supermartingale method using Ville's inequality.
Let $M_n = \exp\left(\sum_{t=1}^{n}\left(\lambda X_t - \frac{\lambda^2\sigma_t^2}{2}\right)\right)$. Since X_n is $\mathcal{F}_{n-1}$-conditionally (σ_n,α)-sub-exponential, we have $\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}\mathbb{E}[\exp(\lambda X_n - \frac{\lambda^2\sigma_n^2}{2}) \mid \mathcal{F}_{n-1}] \leq M_{n-1}$, showing that $\{M_n\}_{n=0}^{\infty}$ is a non-negative supermartingale. By Ville's inequality, we have $\mathbb{P}(\exists n \in \mathbb{N} : M_n \geq \frac{1}{\delta}) \leq \delta$. Rearranging $M_n \geq \frac{1}{\delta}$ yields $\sum_{t=1}^{n} X_t \geq \frac{\lambda}{2}\sum_{t=1}^{n}\sigma_t^2 + \frac{1}{\lambda}\log\frac{1}{\delta}$, which completes the proof. ■

Hoeffding's inequality can be viewed as a special case of Lemma 48.

Lemma 49 (Hoeffding's inequality) *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence of $\mathcal{F}_{t-1}$-conditionally σ^2-sub-Gaussian random variables adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$, where $\sigma^2 \geq 0$ is a constant. For fixed $n \in \mathbb{N}$, we have*

$$\mathbb{P}\left(\sum_{t=1}^{n} X_t \geq \sigma\sqrt{2n\log\frac{1}{\delta}}\right) \leq \delta\,.$$

The following lemma shows that bounded random variables are sub-exponential.

Lemma 50 *Suppose a random variable X lies in an interval $[-c, c]$ almost surely for some constant $c \geq 0$. Suppose $\mathbb{E}[X] = 0$ and denote $V := \mathrm{Var}(X)$. Then, for any $\alpha > 0$, the random variable X is $(\frac{e^{c/\alpha} - 1 - (c/\alpha)}{(c/\alpha)^2} \cdot 2V, \alpha)$-sub-exponential. In particular, if $\alpha = c$, then X is $(2(e-2)V, c)$-sub-exponential.*

Proof By Lemma 61, we have $\mathbb{E}[\exp(\lambda'(X/c))] \leq \exp((e^{\lambda'} - 1 - \lambda')\,\mathrm{Var}(X/c))$ for $\lambda' \geq 0$. It is also possible to obtain $\mathbb{E}[\exp(-\lambda'(X/c))] \leq \exp((e^{\lambda'} - 1 - \lambda')\,\mathrm{Var}(X/c))$ by applying the lemma to $-X/c$. By defining $\lambda = \lambda'/c$ in the first case and $\lambda = -\lambda'/c$ in the latter case, we obtain that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{e^{c|\lambda|} - 1 - c|\lambda|}{c^2}\,\mathrm{Var}(X)\right)$$

for all $\lambda \in \mathbb{R}$. By Lemma 60, one has $(e^{c|\lambda|} - 1 - c|\lambda|)/(c\lambda)^2 \leq (e^{c/\alpha} - 1 - c\alpha)/(c/\alpha)^2$ when $|\lambda| \leq \frac{1}{\alpha}$. Therefore, for all $\lambda \in [-1/\alpha, 1/\alpha]$, it holds that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda^2 \cdot \frac{e^{\frac{c}{\alpha}} - 1 - \frac{c}{\alpha}}{\left(\frac{c}{\alpha}\right)^2} V\right),$$

which proves that X is $(\frac{e^{c/\alpha} - 1 - (c/\alpha)}{(c/\alpha)^2} \cdot 2V, \alpha)$-sub-exponential. ■

By combining Lemma 48 and Lemma 50, we obtain the well-known variant of Freedman's inequality.

Lemma 51 *For a constant $c \geq 0$, let $\{X_t\}_{t=1}^{\infty}$ be a martingale difference sequence adapted to filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$ that satisfies $X_t \in [-c, c]$ almost surely for all $t \in \mathbb{N}$. Then, for any $\lambda \in (0, 1/c]$ and $\delta \in (0, 1]$, it holds that*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \sum_{t=1}^{T} X_t \geq (e-2)\lambda \sum_{t=1}^{n} \mathrm{Var}(X_t \mid \mathcal{F}_{t-1}) + \frac{1}{\lambda}\log\frac{1}{\delta}\right).$$

The following inequality allows us to use a Hoeffding-like concentration bound by clipping by a constant c.

Lemma 52 *Let $\{X_t\}_{t=1}^{\infty}$ be a martingale difference sequence adapted to filtration $\{\mathcal{F}_t\}_{t=1}^{\infty}$. Suppose there exist constants $\sigma, \alpha > 0$ such that X_t is $\mathcal{F}_{t-1}$-conditionally (σ, α)-sub-exponential for all $t \in \mathbb{N}$. Then, for any $c > 0$ and $\delta \in (0, 1]$, it holds that*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) \wedge c \geq 2\left(\sigma \vee \sqrt{\frac{\alpha c}{2}}\right)\sqrt{\frac{1}{n}\log\frac{2(\log e^2 n)^2}{\delta}}\right) \leq \delta.$$

Remark 53 *It is common to use $\frac{\sigma}{\sqrt{n}} + \frac{\alpha}{n}$-type bounds for sub-exponential random variables (Jia et al., 2021). While this lemma provides a simpler bound, we note that its rate is not necessarily optimal, especially when the value of α is large and known. We use this lemma for a simpler design of the algorithm.*

Proof Let $\sigma' = \sigma \vee \sqrt{\frac{\alpha c}{2}}$, $r = \sqrt{\frac{2\alpha}{c} \vee \frac{1}{\log \frac{2}{\delta}}}$, and $\alpha' = \sigma' r$. Note that $\sigma' \geq \sqrt{\frac{\alpha c}{2}}$ and $r \geq \sqrt{\frac{2\alpha}{c}}$ imply $\alpha' \geq \sqrt{\frac{\alpha c}{2}} \cdot \sqrt{\frac{2\alpha}{c}} = \alpha$. We apply Lemma 48 to a sequence of $\{(\lambda_j, \delta_j)\}_{j=0}^{\infty}$ with $\delta_j := \delta/(2(j+1)^2)$ and $\lambda_j := \frac{e^{-j/2}}{\alpha'}$. By the union bound, we have

$$\mathbb{P}\left(\exists n \in \mathbb{N}, \exists j \in \mathbb{N} \cup \{0\} : \sum_{t=1}^{n} X_t \geq \frac{\lambda_j \sigma^2 n}{2} + \frac{1}{\lambda_j} \log \frac{1}{\delta_j}\right) \leq \delta\,.$$

The right-hand side is equal to

$$\frac{\lambda_j \sigma^2 n}{2} + \frac{1}{\lambda_j} \log \frac{1}{\delta_j} = \frac{e^{-j/2}\sigma^2 n}{2\alpha'} + e^{j/2}\alpha' \log \frac{2(j+1)^2}{\delta}\,. \tag{25}$$

Suppose $n \geq r^2 \log \frac{2}{\delta} = (\frac{2\alpha}{c} \log \frac{2}{\delta}) \vee 1$. Then, there exists an integer $j_n \geq 0$ such that

$$r^2 e^{j_n} \log \frac{2(j_n+1)^2}{\delta} \leq n < r^2 e^{j_n+1} \log \frac{2(j_n+2)^2}{\delta}\,. \tag{26}$$

By the first part of inequality (26), we bound the last term in Eq. (25) as

$$\begin{aligned} e^{j_n/2}\alpha' \log \frac{2(j_n+1)^2}{\delta} &= \left(e^{j_n}\sigma'^2 r^2 \log \frac{2(j_n+1)^2}{\delta}\right)^{\frac{1}{2}} \sqrt{\log \frac{2(j_n+1)^2}{\delta}} \\ &\leq \sigma' \sqrt{n \log \frac{2(j_n+1)^2}{\delta}}\,. \end{aligned}$$

By the second part of inequality (26), we bound the first term in the right-hand side of Eq. (25) as

$$\begin{aligned} \frac{e^{-j_n/2}\sigma^2 n}{2\alpha'} &\leq \frac{e^{-j_n/2}\sigma'^2 n}{2\alpha'} \\ &= \frac{\sigma' n}{2}\left(e^{j_n} r^2\right)^{-\frac{1}{2}} \\ &\leq \frac{\sigma' n}{2}\sqrt{\frac{e}{n}\log \frac{2(j_n+2)^2}{\delta}} \\ &= \frac{\sqrt{e}\sigma'}{2}\sqrt{n \log \frac{2(j_n+2)^2}{\delta}}\,. \end{aligned}$$

In addition, the first part of inequality (26) also implies $j_n \leq \log n$ as

$$e^{j_n} \leq e^{j_n} \cdot \frac{\log \frac{2(j_n+1)^2}{\delta}}{\log \frac{2}{\delta}} \leq e^{j_n} r^2 \log \frac{2(j_n+1)^2}{\delta} \leq n\,.$$

Therefore, for $n \geq r^2 \log \frac{2}{\delta}$, we have

$$\begin{aligned} \frac{\lambda_{j_n}\sigma^2 n}{2} + \frac{1}{\lambda_{j_n}} \log \frac{1}{\delta_{j_n}} &\leq \frac{\sqrt{e}\sigma'}{2}\sqrt{n \log \frac{2(j_n+2)^2}{\delta}} + \sigma' \sqrt{n \log \frac{2(j_n+1)^2}{\delta}} \\ &\leq 2\sigma' \sqrt{n \log \frac{2(j_n+2)^2}{\delta}} \\ &\leq 2\sigma' \sqrt{n \log \frac{2(\log n+2)^2}{\delta}}\,. \end{aligned}$$

It implies that

$$\begin{aligned}
&\mathbb{P}\left(\exists n \in \mathbb{N}, n \geq r^2 \log \frac{2}{\delta} : \sum_{t=1}^{n} X_t \geq 2\sigma' \sqrt{n \log \frac{2(\log n + 2)^2}{\delta}}\right) \\
&\leq \mathbb{P}\left(\exists n \in \mathbb{N}, n \geq r^2 \log \frac{2}{\delta} : \sum_{t=1}^{n} X_t \geq \frac{\lambda_{j_n} \sigma^2 n}{2} + \frac{1}{\lambda_{j_n}} \log \frac{1}{\delta_{j_n}}\right) \\
&\leq \mathbb{P}\left(\exists n \in \mathbb{N}, \exists j \in \mathbb{N} \cup \{0\} : \sum_{t=1}^{n} X_t \geq \frac{\lambda_j \sigma^2 n}{2} + \frac{1}{\lambda_j} \log \frac{1}{\delta_j}\right) \\
&\leq \delta
\end{aligned}$$

Finally, suppose $n < r^2 \log \frac{2}{\delta} = (\frac{2\alpha}{c} \log \frac{2}{\delta}) \vee 1$, which is equivalent to $n < \frac{2\alpha}{c} \log \frac{2}{\delta}$ since we must have $n \geq 1$. Then, we have

$$\begin{aligned}
2\sigma' \sqrt{\frac{1}{n} \log \frac{2(\log n + 2)^2}{\delta}} &> 2\sigma' \sqrt{\frac{c \log \frac{2(\log n + 2)^2}{\delta}}{2\alpha \log \frac{2}{\delta}}} \\
&\geq 2\sqrt{\frac{\alpha c}{2}} \sqrt{\frac{c}{2\alpha}} \\
&= c\,.
\end{aligned}$$

Hence, $\left(\frac{1}{n} \sum_{t=1}^{n} X_t\right) \wedge c \geq 2\sigma' \sqrt{\frac{1}{n} \log \frac{2(\log n + 2)^2}{\delta}}$ is possible only when $n \geq r^2 \log \frac{2}{\delta}$. Finally, we have

$$\begin{aligned}
&\mathbb{P}\left(\exists n \in \mathbb{N} : \left(\frac{1}{n} \sum_{t=1}^{n} X_t\right) \wedge c \geq 2\sigma' \sqrt{\frac{1}{n} \log \frac{2(\log n + 2)^2}{\delta}}\right) \\
&= \mathbb{P}\left(\exists n \in \mathbb{N}, n \geq r^2 \log \frac{2}{\delta} : \left(\frac{1}{n} \sum_{t=1}^{n} X_t\right) \wedge c \geq 2\sigma' \sqrt{\frac{1}{n} \log \frac{2(\log n + 2)^2}{\delta}}\right) \\
&\leq \mathbb{P}\left(\exists n \in \mathbb{N}, n \geq r^2 \log \frac{2}{\delta} : \sum_{t=1}^{n} X_t \geq 2\sigma' \sqrt{n \log \frac{2(\log n + 2)^2}{\delta}}\right) \\
&\leq \delta\,.
\end{aligned}$$

■

The following is a time-uniform concentration inequality for sub-Gaussian random variables whose constant factor is asymptotically tight.

Lemma 54 *Let $\{X_t\}_{t=1}^{\infty}$ be a martingale difference sequence adapted to filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Suppose X_t is $\mathcal{F}_{t-1}$-conditionally 1-sub-Gaussian for all $t \in \mathbb{N}$, meaning that for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda X_t) \mid \mathcal{F}_{t-1}] \leq \frac{\lambda^2}{2}$ almost surely. Then, for a constant $\alpha > 1$ and $\eta \in (0, e)$, we have*

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \sum_{t=1}^{n} X_t \geq \sqrt{2(1+\eta) n \log \frac{4(2 + \log \frac{1}{\eta})^2 (1 + \frac{2e}{\eta} \log n)^2}{\delta}}\right) \leq \delta\,.$$

Proof For fixed $\lambda > 0$, Lemma 48 implies that the following inequality holds for all $n \geq \mathbb{N}$ with probability at least $1 - \delta$:

$$\sum_{t=1}^{n} X_t \leq \frac{\lambda n}{2} + \frac{1}{\lambda} \log \frac{1}{\delta} .$$

We first fix $\eta \in (0, 1]$ and apply the lemma for a sequence of $\{(\lambda_j, \delta_j)\}_{j=1}^{\infty}$ with $\lambda_j = \sqrt{2(1+\eta)^{-j+1} \log \frac{2j^2}{\delta}}$ and $\delta_j = \frac{\delta}{2j^2}$. Taking the union bound, we obtain that $\sum_{t=1}^{n} X_t \leq \frac{\lambda_j n}{2} + \frac{1}{\lambda_j} \log \frac{2j^2}{\delta}$ holds for all $n \in \mathbb{N}$ and $j \in \mathbb{N}$ with probability at least $1 - \delta$, which is equivalent to

$$\sum_{t=1}^{n} X_t \leq \left(\frac{n}{(1+\eta)^{\frac{j-1}{2}}} + (1+\eta)^{\frac{j-1}{2}} \right) \sqrt{\frac{1}{2} \log \frac{2j^2}{\delta}} . \tag{27}$$

Taking $j_n = \lceil \log_{1+\eta} n \rceil$, we have $(1+\eta)^{j_n} \geq n$ and $(1+\eta)^{j_n - 1} \leq n$. Then, we have

$$\begin{aligned} \frac{n}{(1+\eta)^{\frac{j_n-1}{2}}} + (1+\eta)^{\frac{j_n-1}{2}} &\leq \sqrt{(1+\eta)n} + \sqrt{n} \\ &\leq 2\sqrt{(1+\eta)n} . \end{aligned}$$

In addition, we have $j_n \leq 1 + \log_{1+\eta} n = 1 + \frac{\log n}{\log(1+\eta)} \leq 1 + \frac{2}{\eta} \log n$, where we use that $\frac{x}{2} \leq \log(1+x)$ for $x \in (0, 1]$. Hence, under the event of Eq. (27), we have

$$\sum_{t=1}^{n} X_t \leq \sqrt{2(1+\eta) n \log \frac{2(1 + \frac{2}{\eta} \log n)^2}{\delta}} .$$

Now, we take a sequence $\eta_i = e^{-i+1}$ for $i \in \mathbb{N}$ and take the union bound over i, where we assign probability $\frac{\delta}{2i^2}$ for each i. Then, with probability at least $1 - \delta$, the following inequality holds for all $n \in \mathbb{N}$ and $i \in \mathbb{N}$:

$$\sum_{t=1}^{n} X_t \leq \sqrt{2(1+\eta_i) n \log \frac{4i^2(1 + \frac{2}{\eta_i} \log n)^2}{\delta}} .$$

For any given $\eta \in (0, e)$, we take $i^* = 1 + \lceil \log \frac{1}{\eta} \rceil$, so that $\frac{\eta}{e} \leq e^{-i^*+1} \leq \eta$. Then, we have

$$\begin{aligned} \sum_{t=1}^{n} X_t &\leq \sqrt{2(1+\eta_{i^*}) n \log \frac{4(i^*)^2(1 + \frac{2}{\eta_{i^*}} \log n)^2}{\delta}} \\ &\leq \sqrt{2(1+\eta) n \log \frac{4(2 + \log \frac{1}{\eta})^2(1 + \frac{2e}{\eta} \log n)^2}{\delta}} . \end{aligned}$$

■

Appendix H. Technical Lemmas

The following lemma encapsulates a procedure that appears multiple times in this paper when bounding the variance of a random variable, and is a minor generalization of Lemma 27 in Lee and Oh (2025).

Lemma 55 *Let $c \geq 0$ be a constant and $\mathcal{F}$ be a σ-algebra. Let Z be a random variable such that $0 \leq \mathbb{E}[Z \mid \mathcal{F}] \leq c$ holds almost surely. Let X and Y be random variables that satisfy*

$$X = (Y + \mathbb{E}[Z \mid \mathcal{F}]) \wedge c$$

and $X \geq 0$. Then, the variance of Z is upper bounded by

$$\mathrm{Var}(Z \mid \mathcal{F}) \leq \mathbb{E}[Z^2 \mid \mathcal{F}] - X^2 + 2c(Y \vee 0)\,.$$

Proof We have

$$\begin{aligned}
\mathrm{Var}(Z \mid \mathcal{F}) &= \mathbb{E}[Z^2 \mid \mathcal{F}] - (\mathbb{E}[Z \mid \mathcal{F}])^2 \\
&= \mathbb{E}[Z^2 \mid \mathcal{F}] - X^2 + X^2 - (\mathbb{E}[Z \mid \mathcal{F}])^2 \\
&= \mathbb{E}[Z^2 \mid \mathcal{F}] - X^2 + (X + \mathbb{E}[Z \mid \mathcal{F}])(X - \mathbb{E}[Z \mid \mathcal{F}])\,.
\end{aligned}$$

Note that we have $0 \leq X + \mathbb{E}[Z \mid \mathcal{F}] \leq 2c$. If $X - \mathbb{E}[Z \mid \mathcal{F}] \leq 0$, the last term is at most 0. If $X - \mathbb{E}[Z \mid \mathcal{F}] \geq 0$, then we have $0 \leq X - \mathbb{E}[Z \mid \mathcal{F}] \leq Y$. Combining these cases, we obtain $(X + \mathbb{E}[Z \mid \mathcal{F}])(X - \mathbb{E}[Z \mid \mathcal{F}]) \leq 2c(Y \vee 0)$, completing the proof. ■

Lemma 56 *For $\alpha \in [1/2, 1]$ and $n \geq 2$, one has*

$$\sum_{t=2}^{n} t^{-\alpha} \leq \left(\frac{1}{1-\alpha} \wedge \log n \right) n^{1-\alpha}\,.$$

Proof The result holds for $\alpha = 1$ since $\sum_{t=2}^{n} \frac{1}{t} \leq \log n$. Suppose $\alpha \in [1/2, 1)$. By the integration technique, we have

$$\begin{aligned}
\sum_{t=2}^{n} t^{-\alpha} &\leq \int_{1}^{n} x^{-\alpha}\, dx \\
&= \frac{n^{1-\alpha} - 1}{1-\alpha}\,.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\sum_{t=2}^{n} t^{-\alpha} &\leq \sum_{t=2}^{n} (1 + (1-\alpha)\log t) t^{-\alpha} \\
&\leq \int_{1}^{n} \frac{1 + (1-\alpha)\log x}{x^{\alpha}}\, dx \\
&= n^{1-\alpha} \log n\,,
\end{aligned}$$

where the second inequality holds since $\frac{1+(1-\alpha)\log x}{x^\alpha}$ is decreasing on $x \geq 1$ when $\alpha \in (1/2, 1]$, and the last inequality is due to $\frac{d}{dx}(x^{1-\alpha}\log x) = (1+(1-\alpha)\log x)x^{-\alpha}$. Therefore, we have

$$\sum_{t=2}^{n} t^{-\alpha} \leq \frac{n^{1-\alpha}-1}{1-\alpha} \wedge n^{1-\alpha}\log n \leq \left(\frac{1}{1-\alpha} \wedge \log n\right) n^{1-\alpha}\,.$$

■

Lemma 57 *For $\alpha \in [1/2, 1]$, $\varepsilon > 0$, $c > 0$, and $N \in \mathbb{N}$, the following inequality holds:*

$$\sum_{n=1}^{N} \frac{c}{n^\alpha} \mathbb{1}\left\{\frac{c}{n^\alpha} \geq \varepsilon\right\} \leq \left(\frac{1}{1-\alpha} \wedge \left(\frac{1}{\alpha}\log\left(\frac{c}{\varepsilon}\right)\right) \wedge \log N\right) c^{\frac{1}{\alpha}} \left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}-1}\,.$$

Proof Let $n_0 = \lfloor (c/\varepsilon)^{1/\alpha} \rfloor \wedge N$. Then, we have

$$\begin{aligned}
\sum_{n=1}^{N} \frac{c}{n^\alpha} \mathbb{1}\left\{\frac{c}{n^\alpha} \geq \varepsilon\right\} &= \sum_{n=0}^{n_0} \frac{c}{n^\alpha} \\
&\leq \left(\frac{1}{1-\alpha} \wedge \log n_0\right) c n_0^{1-\alpha} \\
&\leq \left(\frac{1}{1-\alpha} \wedge \log\left(\left(\frac{c}{\varepsilon}\right)^{\frac{1}{\alpha}} \wedge N\right)\right) c \left(\frac{c}{\varepsilon}\right)^{\frac{1-\alpha}{\alpha}} \\
&= \left(\frac{1}{1-\alpha} \wedge \left(\frac{1}{\alpha}\log\frac{c}{\varepsilon}\right) \wedge \log N\right) c^{\frac{1}{\alpha}} \left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}-1}\,,
\end{aligned}$$

where the first inequality is due to Lemma 56. ■

Appendix I. Auxiliary Lemmas

Lemma 58 (Lemma 30 in Chen et al. (2021)) *For any two random variables X, Y, we have:*

$$\mathrm{Var}(XY) \leq 2\,\mathrm{Var}(X)\|Y\|_\infty^2 + 2\mathbb{E}[X]^2\,\mathrm{Var}(Y)\,.$$

Lemma 59 (Lemma 30 in Lee and Oh (2025)) *For any sequence of K trajectories, we have*

$$\sum_{k=1}^{K} \mathbb{1}\{\eta^k < H+1\} \leq SA\log_2 2H\,.$$

Lemma 60 (Lemma (3.1) in Freedman (1975)) *Let $g(0) = \frac{1}{2}$ and $g(x) = (e^x - 1 - x)/x^2$ for $x \neq 0$. Then, g is increasing.*

Lemma 61 (Equation (3.5) in Freedman (1975)) *For $\lambda \geq 0$ and a random variable X satisfying $X \leq 1$ and $\mathbb{E}[X] \leq 0$, we have $\mathbb{E}[\exp(\lambda X)] \leq \exp((e^\lambda - 1 - \lambda)\,\mathrm{Var}(X))$.*

Lemma 62 (Hoeffding's lemma, Eq. (3.16) in Hoeffding (1963)) *Let X be a real-valued random variable that satisfies $a \leq X \leq b$ for some real numbers a and b, and assume $\mathbb{E}[X] = 0$. Then, X is $(\frac{b-a}{2})^2$-sub-Gaussian.*