Tackling Heavy-Tailed Rewards in Reinforcement Learning with Function Approximation: Minimax Optimal and Instance-Dependent Regret Bounds

Jiayi Huang[1] **Han Zhong**[1] **Liwei Wang**[1,2] **Lin F. Yang**[3]
[1]Center for Data Science, Peking University
[2]National Key Laboratory of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[3]University of California, Los Angles
`{jyhuang, hanzhong}@stu.pku.edu.cn`,
`wanglw@cis.pku.edu.cn`, `linyang@ee.ucla.edu`

Abstract

While numerous works have focused on devising efficient algorithms for reinforcement learning (RL) with uniformly bounded rewards, it remains an open question whether sample or time-efficient algorithms for RL with large state-action space exist when the rewards are *heavy-tailed*, i.e., with only finite $(1+\epsilon)$-th moments for some $\epsilon \in (0, 1]$. In this work, we address the challenge of such rewards in RL with linear function approximation. We first design an algorithm, HEAVY-OFUL, for heavy-tailed linear bandits, achieving an *instance-dependent* T-round regret of $\widetilde{O}\big(dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\sum_{t=1}^{T}\nu_t^2} + dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\big)$, the *first* of this kind. Here, d is the feature dimension, and $\nu_t^{1+\epsilon}$ is the $(1+\epsilon)$-th central moment of the reward at the t-th round. We further show the above bound is minimax optimal when applied to the worst-case instances in stochastic and deterministic linear bandits. We then extend this algorithm to the RL settings with linear function approximation. Our algorithm, termed as HEAVY-LSVI-UCB, achieves the *first* computationally efficient *instance-dependent* K-episode regret of $\widetilde{O}(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^*K})$. Here, H is length of the episode, and $\mathcal{U}^*, \mathcal{V}^*$ are instance-dependent quantities scaling with the central moment of reward and value functions, respectively. We also provide a matching minimax lower bound $\Omega(dHK^{\frac{1}{1+\epsilon}} + d\sqrt{H^3K})$ to demonstrate the optimality of our algorithm in the worst case. Our result is achieved via a novel robust self-normalized concentration inequality that may be of independent interest in handling heavy-tailed noise in general online regression problems.

1 Introduction

Designing efficient reinforcement learning (RL) algorithms for large state-action space is a significant challenge within the RL community. A crucial aspect of RL is understanding the reward functions, which directly impacts the quality of the agent's policy. In certain real-world situations, reward distributions may exhibit heavy-tailed behavior, characterized by the occurrence of extremely large values at a higher frequency than expected in a normal distribution. Examples include image noise in signal processing [14], stock price fluctuations in financial markets [12, 18], and value functions in online advertising [11, 19]. However, much of the existing RL literature assumes rewards to be either uniformly bounded or light-tailed (e.g., sub-Gaussian). In such light-tailed settings, the primary challenge lies in learning the transition probabilities, leading most studies to assume deterministic rewards for ease of analysis [5, 22, 15]. As we will demonstrate, the complexity of learning reward

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

functions may dominate in heavy-tailed settings. Consequently, the performance of traditional algorithms may decline, emphasizing the need for the development of new, efficient algorithms specifically designed to handle heavy-tailed rewards.

Heavy-tailed distributions have been extensively studied in the field of statistics [9, 27] and in more specific online learning scenarios, such as bandits [7, 28, 31, 34, 40]. However, there is a dearth of theoretical research in RL concerning heavy-tailed rewards, whose distributions only admit finite $(1+\epsilon)$-th moment for some $\epsilon \in (0, 1]$. One notable exception is Zhuang and Sui [44], which made a pioneering effort in establishing worst-case regret guarantees in *tabular* Markov Decision Processes (MDPs) with heavy-tailed rewards. However, their algorithm cannot handle RL settings with large state-action space. Moreover, their reliance on truncation-based methods is sub-optimal as these methods heavily depend on raw moments, which do not vanish in deterministic cases. Therefore, a natural question arises:

Can we derive sample and time-efficient algorithms for RL with large state-action space that achieve instance-dependent regret in the presence of heavy-tailed rewards?

In this work, we focus on linear MDPs [35, 22] with heavy-tailed rewards and answer the above question affirmatively. We say a distribution is *heavy-tailed* if it only admits finite $(1+\epsilon)$-th moment for some $\epsilon \in (0, 1]$. Our contributions are summarized as follows.

- We first propose a computationally efficient algorithm HEAVY-OFUL for heavy-tailed linear bandits. Such a setting can be regarded as a special case of linear MDPs. HEAVY-OFUL achieves an *instance-dependent* T-round regret of $\widetilde{O}\big(dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\sum_{t=1}^{T}\nu_t^2} + dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\big)$, the *first* of this kind. Here d is the feature dimension and $\nu_t^{1+\epsilon}$ is the $(1+\epsilon)$-th central moment of the reward at the t-th round. The instance-dependent regret bound has a main term that only depends on the summation of central moments, and therefore does not have a $\sqrt{T}$ term. Our regret bound is shown to be minimax optimal in both stochastic and deterministic linear bandits (See Remark 4.2 for details).
- We then extend this algorithm to time-inhomogeneous linear MDPs with heavy-tailed rewards, resulting in a new computationally efficient algorithm HEAVY-LSVI-UCB, which achieves a K-episode regret scaling as $\widetilde{O}(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^*K})$ for the *first* time. Here, H is the length of the episode and $\mathcal{U}^*, \mathcal{V}^*$ are quantities measuring the central moment of the reward functions and transition probabilities, respectively (See Theorem 5.2 for details). Our regret bound is *instance-dependent* since the main term only relies on the instance-dependent quantities, which vanishes when the dynamics and rewards are deterministic. When specialized to special cases, our instance-dependent regret recovers the variance-aware regret in Li and Sun [26] (See Remark 5.3 for details) and improves existing first-order regret bounds [33, 26] (See Corollary 5.6 for details).
- We provide a minimax regret lower bound $\Omega(dHK^{\frac{1}{1+\epsilon}} + d\sqrt{H^3K})$ for linear MDPs with heavy-tailed rewards, which matches the worst-case regret bound implied by our instance-dependent regret, thereby demonstrating the minimax optimality of HEAVY-LSVI-UCB in the worst case.

For better comparisons between our algorithms and state-of-the-art results, we summarize the regrets in Table 1 and 2 for linear bandits and linear MDPs, respectively. More related works are deferred to Appendix A. Remarkably, our results demonstrate that $\epsilon = 1$ (i.e. finite variance) is sufficient to obtain variance-aware regret bounds of the same order as the case where rewards are uniformly bounded for both linear bandits and linear MDPs. The main technique contribution behind our results is a novel robust self-normalized concentration inequality inspired by Sun et al. [32]. To be more specific, it is a non-trivial generalization of adaptive Huber regression from independent and identically distributed (i.i.d.) case to heavy-tailed online regression settings and gives a *self-normalized* bound instead of the ℓ_2-norm bound in Sun et al. [32]. Our result is computationally efficient and only scales with the feature dimension, d, $(1+\epsilon)$-th *central* moment of the noise, ν, and does not depend on the absolute magnitude as in other self-normalized concentration inequalities [42, 41].

Road Map The rest of the paper is organized as follows. Section 2 introduces heavy-tailed linear bandits and linear MDPs. Section 3 presents the robust self-normalized concentration inequality for general online regression problems with heavy-tailed noise. Section 4 and 5 give the main results for heavy-tailed linear bandits and linear MDPs, respectively. We then conclude in Section 6. Related work, experiments and all proofs can be found in Appendix.

Table 1: Comparisons with previous works on linear bandits. d, T, $\{\sigma_t\}_{t\in[T]}$, $\{\nu_t\}_{t\in[T]}$ are feature dimension, the number of rounds, the variance or central moment of the reward at the t-th round.

Algorithm	Regret	Instance-dependent?	Minimax Optimal?	Deterministic-Optimal?	Heavy-Tailed Rewards?
OFUL [1]	$\widetilde{O}\left(d\sqrt{T}\right)$	No	**Yes**	No	No
IDS-UCB [24] Weighted OFUL+ [41] AdaOFUL [26]	$\widetilde{O}\left(d\sqrt{\sum_{t=1}^T \sigma_t^2}+d\right)$	**Yes**	**Yes**	**Yes**	No No $\epsilon=1$
MENU and TOFU [31]	$\widetilde{O}\left(dT^{\frac{1}{1+\epsilon}}\right)$	No	**Yes**	No	**Yes**
HEAVY-OFUL (**Ours**)	$\widetilde{O}\left(dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\sum_{t=1}^T \nu_t^2}+dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)$	**Yes**	**Yes**	**Yes**	**Yes**

Table 2: Comparisons with previous works on time-inhomogeneous linear MDPs. d, H, K, V_1^*, $\mathcal{G}^*$ are feature dimension, the length of the episode, the number of episodes, optimal value function, variance-dependent quantity defined in Li and Sun [26]. $\mathcal{U}^*$, $\mathcal{V}^*$ are defined in Theorem 5.2.

Algorithm	Regret	Central Moment-Dependent?	First-Order?	Minimax Optimal?	Computationally Efficient?	Heavy-Tailed Rewards?
LSVI-UCB[22]	$\widetilde{O}\left(\sqrt{d^3H^4K}\right)$	No	No	No	**Yes**	No
FORCE [33]	$\widetilde{O}\left(\sqrt{d^3H^3V_1^*K}\right)$	No	**Yes**	No	No	No
VOQL [2] LSVI-UCB++ [15]	$\widetilde{O}\left(d\sqrt{H^3K}\right)$	No	No	**Yes**	**Yes**	No
VARA [26]	$\widetilde{O}\left(d\sqrt{H\mathcal{G}^*K}\right)$	**Yes**	**Yes**	**Yes**	**Yes**	$\epsilon=1$
HEAVY-LSVI-UCB (**Ours**)	$\widetilde{O}\left(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}}+d\sqrt{H\mathcal{V}^*K}\right)$	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Notations Let $\|\boldsymbol{a}\| := \|\boldsymbol{a}\|_2$. Let $[t] := \{1,2,\dots,t\}$. Let $\mathcal{B}_d(r) := \{\boldsymbol{x}\in\mathbb{R}^d \mid \|\boldsymbol{x}\|\le r\}$. Let $x_{[a,b]} := \max\{a,\min\{x,b\}\}$ denote the projection of x onto the close interval $[a,b]$. Let $\sigma(\{X_s\}_{s\in[t]})$ be the σ-field generated by random vectors $\{X_s\}_{s\in[t]}$.

2 Preliminaries

2.1 Heavy-Tailed Linear Bandits

Definition 2.1 (Heterogeneous linear bandits with heavy-tailed rewards)**.** Let $\{\mathcal{D}_t\}_{t\ge1}$ denote a series of fixed decision sets, where all $\boldsymbol{\phi}_t\in\mathcal{D}_t$ satisfy $\|\boldsymbol{\phi}_t\|\le L$ for some known upper bound L. At each round t, the agent chooses $\boldsymbol{\phi}_t\in\mathcal{D}_t$, then receives a reward R_t from the environment. We define the filtration $\{\mathcal{F}_t\}_{t\ge1}$ as $\mathcal{F}_t=\sigma(\{\boldsymbol{\phi}_s,R_s\}_{s\in[t]}\cup\{\boldsymbol{\phi}_{t+1}\})$ for any $t\ge1$. We assume $R_t=\langle\boldsymbol{\phi}_t,\boldsymbol{\theta}^*\rangle+\varepsilon_t$ with the unknown coefficient $\boldsymbol{\theta}^*\in\mathcal{B}_d(B)$ for some known upper bound B. The random variable $\varepsilon_t\in\mathbb{R}$ is $\mathcal{F}_t$-measurable and satisfies $\mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}]=0, \mathbb{E}[|\varepsilon_t|^{1+\epsilon}|\mathcal{F}_{t-1}]=\nu_t^{1+\epsilon}$ for some $\epsilon\in(0,1]$ with ν_t being $\mathcal{F}_{t-1}$-measurable.

The agent aims to minimize the T-round *pseudo-regret* defined as $\mathrm{Regret}(T)=\sum_{t=1}^T[\langle\boldsymbol{\phi}_t^*,\boldsymbol{\theta}^*\rangle-\langle\boldsymbol{\phi}_t,\boldsymbol{\theta}^*\rangle]$, where $\boldsymbol{\phi}_t^*=\mathrm{argmax}_{\boldsymbol{\phi}\in\mathcal{D}_t}\langle\boldsymbol{\phi},\boldsymbol{\theta}^*\rangle$.

2.2 Linear MDPs with Heavy-Tailed Rewards

We use a tuple $M=M(\mathcal{S},\mathcal{A},H,\{R_h\}_{h\in[H]},\{\mathbb{P}_h\}_{h\in[H]})$ to describe the *time-inhomogeneous finite-horizon MDP*, where $\mathcal{S}$ and $\mathcal{A}$ are state space and action space, respectively, H is the length of the episode, $R_h:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is the random reward function with expectation $r_h:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$, and $\mathbb{P}_h:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ is the transition probability function. More details can be found in Puterman [30]. A time-dependent *policy* $\pi=\{\pi_h\}_{h\in H}$ satisfies $\pi_h:\mathcal{S}\to\Delta(\mathcal{A})$ for any $h\in[H]$. When the policy is deterministic, we use $\pi_h(s_h)$ to denote the action chosen at the h-th step given s_h by policy π. For any state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$, we define the *state-action value function* $Q_h^\pi(s,a)$ and *state value function* $V_h^\pi(s)$ as follows: $Q_h^\pi(s,a)=\mathbb{E}\left[\sum_{h'=h}^H r(s_{h'},a_{h'})|s_h=s,a_h=a\right]$, $V_h^\pi(s)=Q_h^\pi(s,\pi_h(s))$, where the expectation is taken with respect to the transition probability of

M and the agent's policy π. If π is randomized, then the definition of V should have an expectation. Denote the optimal value functions as $V_h^*(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$.

We introduce the following shorthands for simplicity. At the h-th step, for any value function $V : \mathcal{S} \to \mathbb{R}$, let $[\mathbb{P}_h V](s,a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} V(s')$, $[\mathbb{V}_h V](s,a) = [\mathbb{P}_h V^2](s,a) - [\mathbb{P}_h V]^2(s,a)$ denote the expectation and the variance of the next-state value function at the h-th step given (s,a).

We aim to minimize the K-episode *regret* defined as $\mathrm{Regret}(K) = \sum_{k=1}^K [V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)]$.

In the rest of this section, we introduce linear MDPs with heavy-tailed rewards. We first give the definition of linear MDPs studied in Yang and Wang [35], Jin et al. [22], with emphasis that the rewards in their settings are deterministic or uniformly bounded. Then we focus on the heavy-tailed random rewards.

Definition 2.2. An MDP $M = M(\mathcal{S}, \mathcal{A}, H, \{R_h\}_{h\in[H]}, \{\mathbb{P}_h\}_{h\in[H]})$ is a *time-inhomogeneous finite-horizon linear MDP*, if there exist known feature maps $\boldsymbol{\phi}(s,a) : \mathcal{S} \times \mathcal{A} \to \mathcal{B}_d(1)$, unknown d-dimensional signed measures $\{\boldsymbol{\mu}_h^*\}_{h\in[H]}$ over $\mathcal{S}$ with $\|\boldsymbol{\mu}_h^*(\mathcal{S})\| := \int_{s\in\mathcal{S}} |\boldsymbol{\mu}(s)| \mathrm{d}s \le \sqrt{d}$ and unknown coefficients $\{\boldsymbol{\theta}_h^*\}_{h\in[H]} \subseteq \mathcal{B}_d(B)$ for some known upper bound B such that

$$r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h^* \rangle, \quad \mathbb{P}_h(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h^*(\cdot) \rangle$$

for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and timestep $h \in [H]$.

Assumption 2.3 (Realizable rewards)**.** For all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the random reward $R_h(s,a)$ is independent of next state $s' \sim \mathbb{P}_h(\cdot|s,a)$ and admits the linear structure

$$R_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h^* \rangle + \varepsilon_h(s,a),$$

where $\varepsilon_h(s,a)$ is a mean-zero heavy-tailed random variable specified below.

We introduce the notation $\boldsymbol{\nu}_n[X] = \mathbb{E}[|X - \mathbb{E}X|^n]$ for the n-th central moment of any random variable X. And for any random reward function at the h-th step $R_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, let

$$[\mathbb{E}_h R_h](s,a) = \mathbb{E}[R_h(s_h,a_h) | (s_h,a_h) = (s,a)],$$

$$[\boldsymbol{\nu}_{1+\epsilon} R_h](s,a) = \mathbb{E}[|[R_h - \mathbb{E}_h R_h](s_h,a_h)|^{1+\epsilon} | (s_h,a_h) = (s,a)]$$

denote its expectation and the $(1+\epsilon)$-th central moment given $(s_h,a_h) = (s,a)$ for short.

Assumption 2.4 (Heavy-tailedness of rewards)**.** Random variable $\varepsilon_h(s,a)$ satisfies $[\mathbb{E}_h \varepsilon_h](s,a) = 0$. And for some known $\epsilon, \epsilon' \in (0,1]$ and constants $\nu_R, \nu_{R^\epsilon} \ge 0$, the following unknown moments of $\varepsilon_h(s,a)$ satisfy

$$[\mathbb{E}_h |\varepsilon_h|^{1+\epsilon}](s,a) \le \nu_R^{1+\epsilon}, \quad [\boldsymbol{\nu}_{1+\epsilon'} |\varepsilon_h|^{1+\epsilon}](s,a) \le \nu_{R^\epsilon}^{1+\epsilon'}$$

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

Assumption 2.4 generalizes Assumption 2.2 of Li and Sun [26], which is the weakest moment condition on random rewards in the current literature of RL with function approximation. Setting $\epsilon = 1$ and $\epsilon' = 1$ immediately recovers their settings.

Assumption 2.5 (Realizable central moments)**.** There are some unknown coefficients $\{\boldsymbol{\psi}_h^*\}_{h\in[H]} \subseteq \mathcal{B}_d(W)$ for some known upper bound W such that

$$[\mathbb{E}_h |\varepsilon_h|^{1+\epsilon}](s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\psi}_h^* \rangle$$

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

Remark 2.6. When $\epsilon = 1$, that is the rewards have finite variance, Li and Sun [26] use the fact that $[\boldsymbol{\nu}_2 R_h](s,a) = [\mathbb{V}_h R_h](s,a) = [\mathbb{E}_h R_h^2](s,a) - [\mathbb{E}_h R_h]^2(s,a)$, assume the linear realizability of the second moment $[\mathbb{E}_h R_h^2](s,a)$, and estimate it instead. However, when $\epsilon < 1$, there is no such relationship between the $(1+\epsilon)$-th central moment $[\boldsymbol{\nu}_{1+\epsilon} R_h](s,a)$ and the $(1+\epsilon)$-th raw moment $[\mathbb{E}_h R_h^{1+\epsilon}](s,a)$. Thus, we adopt a new approach to estimate $[\boldsymbol{\nu}_{1+\epsilon} R_h](s,a)$ directly, and bound the error by a novel perturbation analysis of adaptive Huber regression in Appendix C.3.

Assumption 2.7 (Bounded cumulative rewards)**.** For any policy π, let $\{s_h, a_h, R_h\}_{h\in[H]}$ be a random trajectory following policy π. And define $r_\pi = \sum_{h=1}^H [\mathbb{E}_h R_h](s_h,a_h) = \sum_{h=1}^H r_h(s_h,a_h)$. We assume (1) $0 \le r_\pi \le \mathcal{H}$. (2) $\sum_{h=1}^H [\boldsymbol{\nu}_{1+\epsilon} R_h]^{\frac{2}{1+\epsilon}}(s_h,a_h) \le \mathcal{U}$. (3) $\mathrm{Var}(r_\pi) \le \mathcal{V}$.

Here, (1) gives an upper bound of cumulative expected rewards r_π. (2) assumes the summation of $(1+\epsilon)$-th central moment of rewards $[\boldsymbol{\nu}_{1+\epsilon} R_h](s_h,a_h)$ is bounded since $[\sum_{h=1}^H [\boldsymbol{\nu}_{1+\epsilon} R_h](s_h,a_h)]^{\frac{2}{1+\epsilon}} \le \sum_{h=1}^H [\boldsymbol{\nu}_{1+\epsilon} R_h]^{\frac{2}{1+\epsilon}}(s_h,a_h) \le \mathcal{U}$ due to Jensen's inequality. And (3) is to bound the variance of r_π along the trajectory following policy π.

Algorithm 1 Adaptive Huber Regression

Require: Number of total rounds T, confidence level δ, regularization parameter λ, $\sigma_{\min}$, parameters for adaptive Huber regression c_0, c_1, τ_0, estimated central moment $\widehat{\nu}_t$ and moment parameter b that satisfy $\nu_t/\widehat{\nu}_t \le b$ for all $t \le T$.

Ensure: The estimated coefficient $\boldsymbol{\theta}_t$.

1: $\kappa = d\log(1 + TL^2/(d\lambda\sigma_{\min}^2))$.

2: Set $\boldsymbol{H}_{t-1} = \lambda\boldsymbol{I} + \sum_{s=1}^{t-1}\sigma_s^{-2}\boldsymbol{\phi}_s\boldsymbol{\phi}_s^\top$.

3: Set $\sigma_t = \max\left\{\widehat{\nu}_t, \sigma_{\min}, \frac{\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}}{c_0}, \frac{\sqrt{LB}}{c_1^{\frac{1}{4}}(2\kappa b^2)^{\frac{1}{4}}}\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}^{\frac{1}{2}}\right\}$.

4: Set $\tau_t = \tau_0\frac{\sqrt{1+w_t^2}}{w_t}t^{\frac{1-\epsilon}{2(1+\epsilon)}}$ with $w_t = \|\boldsymbol{\phi}_t/\sigma_t\|_{\boldsymbol{H}_{t-1}^{-1}}$.

5: Define the loss function $L_t(\boldsymbol{\theta}) := \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \sum_{s=1}^t \ell_{\tau_s}(\frac{y_s - \langle\boldsymbol{\phi}_s, \boldsymbol{\theta}\rangle}{\sigma_s})$.

6: Compute $\boldsymbol{\theta}_t = \mathrm{argmin}_{\boldsymbol{\theta}\in\mathcal{B}_d(B)} L_t(\boldsymbol{\theta})$.

3 Adaptive Huber Regression

At the core of our algorithms for both heavy-tailed linear bandits and linear MDPs is a new approach – adaptive Huber regression – to handle heavy-tailed noise. Sun et al. [32] imposed adaptive Huber regression to handle i.i.d. heavy-tailed noise by utilizing Huber loss [17] as a surrogate of squared loss. Li and Sun [26] modified adaptive Huber regression for heterogeneous online settings, where the variances in each round are different. However, it is not readily applicable to deal with heavy-tailed noise. Our contribution in this section is to construct a new self-normalized concentration inequality for general online regression problems with heavy-tailed noise.

We first give a brief introduction to Huber loss function and its properties.

Definition 3.1 (Huber loss)**.** *Huber loss* is defined as

$$\ell_\tau(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \le \tau, \\ \tau|x| - \frac{\tau^2}{2} & \text{if } |x| > \tau, \end{cases} \tag{3.1}$$

where $\tau > 0$ is referred as a robustness parameter.

Huber loss is first proposed by Huber [17] as a robust version of squared loss while preserving the convex property. Specifically, Huber loss is a quadratic function of x when $|x|$ is less than the threshold τ, while becomes linearly dependent on $|x|$ when $|x|$ grows larger than τ. It has the property of strongly convex near zero point and is not sensitive to outliers. See Appendix C.1 for more properties of Huber loss.

Next, we define general online regression problems with heavy-tailed noise, which include heavy-tailed linear bandits as a special case. Then we utilize Huber loss to estimate $\boldsymbol{\theta}^*$. Below we give the main theorem to bound the deviation of the estimated $\boldsymbol{\theta}_t$ in Algorithm 1 from the ground truth $\boldsymbol{\theta}^*$.

Definition 3.2. Let $\{\mathcal{F}_t\}_{t\ge1}$ be a filtration. For all $t > 0$, let random variables y_t, ε_t be $\mathcal{F}_t$-measurable and random vector $\boldsymbol{\phi}_t \in \mathcal{B}_d(L)$ be $\mathcal{F}_{t-1}$-measurable. Suppose $y_t = \langle\boldsymbol{\phi}_t, \boldsymbol{\theta}^*\rangle + \varepsilon_t$, where $\boldsymbol{\theta}^* \in \mathcal{B}_d(B)$ is an unknown coefficient and

$$\mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}] = 0, \quad \mathbb{E}[|\varepsilon_t|^{1+\epsilon}|\mathcal{F}_{t-1}] = \nu_t^{1+\epsilon}$$

for some $\epsilon \in (0, 1]$. The goal is to estimate $\boldsymbol{\theta}^*$ at any round t given the realizations of $\{\boldsymbol{\phi}_s, y_s\}_{s\in[t]}$.

Theorem 3.3. For the online regression problems in Definition 3.2, we solve for $\boldsymbol{\theta}_t$ by adaptive Huber regression in Algorithm 1 with c_0, c_1, τ_0 in Appendix C.2. Then for any $\delta \in (0, 1)$, with probability at least $1 - 3\delta$, for all $t \le T$, we have $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}_t} \le \beta_t$, where $\boldsymbol{H}_t$ is defined in Algorithm 1 and

$$\beta_t = 3\sqrt{\lambda}B + 24t^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{2\kappa}b(\log 3T)^{\frac{1-\epsilon}{2(1+\epsilon)}}(\log(2T^2/\delta))^{\frac{\epsilon}{1+\epsilon}}. \tag{3.2}$$

Proof. To derive a tight high-probability bound, we take the most advantage of the properties of Huber loss. A Chernoff bounding technique is used to bound the main error term, which requires a careful analysis of the moment generating function. See Appendix C.2 for a detailed proof. $\square$

Algorithm 2 HEAVY-OFUL

Require: Number of total rounds T, confidence level δ, regularization parameter λ, $\sigma_{\min}$, parameters for adptive Huber regression c_0, c_1, τ_0, confidence radius β_t.

1: $\kappa = d\log(1+\frac{TL^2}{d\lambda\sigma_{\min}^2})$, $\mathcal{C}_0 = \mathcal{B}_d(B)$, $\boldsymbol{H}_0 = \lambda \boldsymbol{I}$.
2: **for** $t = 1, \dots, T$ **do**
3: Observe $\mathcal{D}_t$.
4: Set $(\boldsymbol{\phi}_t, \cdot) = \operatorname{argmax}_{\boldsymbol{\phi}\in\mathcal{D}_t, \boldsymbol{\theta}\in\mathcal{C}_{t-1}} \langle \boldsymbol{\phi}, \boldsymbol{\theta}\rangle$.
5: Play $\boldsymbol{\phi}_t$ and observe R_t, ν_t.
6: Set $\sigma_t = \max\left\{\nu_t, \sigma_{\min}, \frac{\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}}{c_0}, \frac{\sqrt{LB}}{c_1^{\frac{1}{4}}(2\kappa)^{\frac{1}{4}}}\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}^{\frac{1}{2}}\right\}$.
7: Set $\tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\epsilon}{2(1+\epsilon)}}$ with $w_t = \|\boldsymbol{\phi}_t/\sigma_t\|_{\boldsymbol{H}_{t-1}^{-1}}$.
8: Update $\boldsymbol{H}_t = \boldsymbol{H}_{t-1} + \sigma_t^{-2}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top$.
9: Solve for $\boldsymbol{\theta}_t$ by Algorithm 1 and set $\mathcal{C}_t = \{\boldsymbol{\theta}\in\mathbb{R}^d | \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_{\boldsymbol{H}_t} \le \beta_t\}$.
10: **end for**

We refer to the regression process in Line 6 of Algorithm 1 as *adaptive Huber regression* in line with Sun et al. [32] to emphasize that the value of robustness parameter τ_t is chosen to adapt to data for a better trade-off between bias and robustness. Specifically, since we are in the online setting, $\boldsymbol{\phi}_t$ are dependent on $\{\boldsymbol{\phi}_s\}_{s<t}$, which is the key difference from the i.i.d. case in Sun et al. [32] where they set $\tau_t = \tau_0$, for all $t \le T$. Thus, as shown in Line 4 of Algorithm 1, inspired by Li and Sun [26], we adjust τ_t according to the importance of observations $w_t = \|\boldsymbol{\phi}_t/\sigma_t\|_{\boldsymbol{H}_{t-1}^{-1}}$, where σ_t is specified below. In the case where $\epsilon < 1$, different from Li and Sun [26], we first choose τ_t to be small for robust purposes, then gradually increase it with t to reduce the bias.

Next, we illustrate the reason for setting σ_t via Line 3 of Algorithm 1. We use $\widehat{\nu}_t \in \mathcal{F}_{t-1}$ to estimate the central moment ν_t and use moment parameter b to measure the closeness between $\widehat{\nu}_t$ and ν_t. When we choose $\widehat{\nu}_t$ as an upper bound of ν_t, b becomes a constant that equals to 1. And $\sigma_{\min}$ is a small positive constant to avoid singularity. The last two terms with respect to c_0 and c_1 are set according to the uncertainty $\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}$. In addition, setting the parameter $c_0 \le 1$ yields $w_t \le 1$, which is essential to meet the condition of elliptical potential lemma [1].

Remark 3.4. The error bound β_t in (3.2) is only related to the feature dimension d and moment parameter b. While the Bernstein-style self-normalized concentration bounds [42, 41] depend on the magnitude of ε_t, thus cannot handle heavy-tailed errors.

4 Linear Bandits

In this section, we show the algorithm HEAVY-OFUL in Algorithm 2 for heavy-tailed linear bandits in Definition 2.1. We first give a brief algorithm description, and then provide a theoretical regret analysis.

4.1 Algorithm Description

HEAVY-OFUL follows the principle of Optimism in the Face of Uncertainty (OFU) [1], and uses adaptive Huber regression in Section 3 to maintain a set $\mathcal{C}_t$ that contains the unknown coefficient $\boldsymbol{\theta}^*$ with high probability. Specifically, at the t-th round, HEAVY-OFUL estimates the expected reward of any arm $\boldsymbol{\phi}$ as $\max_{\boldsymbol{\theta}\in\mathcal{C}_{t-1}}\langle\boldsymbol{\phi},\boldsymbol{\theta}\rangle$, and selects the arm that maximizes the estimated reward. The agent then receives the reward R_t and updates the confidence set $\mathcal{C}_t$ based on the information up to round t with its center $\boldsymbol{\theta}_t$ computed by adaptive Huber regression as in Line 9 of Algorithm 2.

4.2 Regret Analysis

We next give the instance-dependent regret upper bound of HEAVY-OFUL in Theorem 4.1.

Theorem 4.1. For the heavy-tailed linear bandits in Definition 2.1, we set $c_0, c_1, \tau_0, \beta_t$ in Algorithm 2 according to Theorem 3.3 with $b = 1$. Besides, let $\lambda = d/B^2$, and $\sigma_{\min} = 1/\sqrt{T}$. Then with probability at least $1-3\delta$, the regret of HEAVY-OFUL is bounded by

$$\text{Regret}(T) = \widetilde{O}\left(dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\sum\nolimits_{t=1}^{T}\nu_t^2} + dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\right).$$

Proof. The proof uses the self-normalized concentration inequality of adaptive Huber regression and a careful analysis to bound the summation of bonuses. See Appendix D.1 for a detailed proof. □

Remark 4.2. Theorem 4.1 shows HEAVY-OFUL achieves an instance-dependent regret bound. When we assume $\nu_t, \forall t \geq 1$ have uniform upper bound ν (which can be treated as a constant), then the bound is reduced to $\widetilde{O}(dT^{\frac{1}{1+\epsilon}})$. It matches the lower bound $\Omega(dT^{\frac{1}{1+\epsilon}})$ by Shao et al. [31] up to logarithmic factors. In the deterministic scenario, where $\epsilon = 1$ and $\nu_t = 0$, for all $t \geq 1$, the bound is reduced to $\widetilde{O}(d)$. It matches the lower bound $\Omega(d)$[1] up to logarithmic factors.

5 Linear MDPs

In this section, we show the algorithm HEAVY-LSVI-UCB in Algorithm 3 for linear MDP with heavy-tailed rewards defined in Section 2.2. Let $\boldsymbol{\phi}_{k,h} := \boldsymbol{\phi}(s_{k,h}, a_{k,h})$ for short. We first give the algorithm description intuitively, then provide the computational complexity and regret bound.

5.1 Algorithm Description

HEAVY-LSVI-UCB features a novel combination of adaptive Huber regression in Section 3 and existing algorithmic frameworks for linear MDPs with bounded rewards [22, 15]. At a high level, HEAVY-LSVI-UCB employs separate estimation techniques to handle heavy-tailed rewards and transition kernels. Specifically, we utilize adaptive Huber regression proposed in Section 3 to estimate heavy-tailed rewards and weighted ridge regression [42, 15] to estimate the expected next-state value functions. Then, it follows the value iteration scheme to update the optimistic and pessimistic estimation of the optimal value function Q_h^k, V_h^k and $\check{Q}_h^k$, $\check{V}_h^k$, respectively, via a rare-switching policy as in Line 7 to 15 of Algorithm 3. We highlight the key steps of HEAVY-LSVI-UCB as follows.

Estimation for expected heavy-tailed rewards Since the expected rewards have linear structure in linear MDPs, i.e., $r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h^* \rangle$, we use adaptive Huber regression to estimate $\boldsymbol{\theta}_h^*$:

$$\boldsymbol{\theta}_{k,h} = \underset{\boldsymbol{\theta}\in\mathcal{B}_d(B)}{\text{argmin}} \frac{\lambda_R}{2}\|\boldsymbol{\theta}\|^2 + \sum_{i=1}^{k} \ell_{\tau_{i,h}}\left(\frac{R_{i,h} - \langle \boldsymbol{\phi}_{i,h}, \boldsymbol{\theta}\rangle}{\nu_{i,h}}\right), \tag{5.1}$$

where $\nu_{i,h}$ will be specified later.

Estimation for central moment of rewards By Assumption 2.5, the $(1+\epsilon)$-th central moment of rewards is linear in $\boldsymbol{\phi}$, i.e., $[\boldsymbol{\nu}_{1+\epsilon}R_h](s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\psi}_h^* \rangle$. Motivated by this, we estimate $\boldsymbol{\psi}_h^*$ by adaptive Huber regression as

$$\boldsymbol{\psi}_{k,h} = \underset{\boldsymbol{\psi}\in\mathcal{B}_d(W)}{\text{argmin}} \frac{\lambda_R}{2}\|\boldsymbol{\psi}\|^2 + \sum_{i=1}^{k} \ell_{\widetilde{\tau}_{i,h}}\left(\frac{|\varepsilon_{i,h}|^{1+\epsilon} - \langle \boldsymbol{\phi}_{i,h}, \boldsymbol{\psi}\rangle}{\nu_{i,h}}\right), \tag{5.2}$$

where W is the upper bound of $\|\boldsymbol{\psi}_h^*\|$ defined in Assumption 2.5. Since $\varepsilon_{i,h}$ is intractable, we estimate it by $\widehat{\varepsilon}_{i,h} = R_{i,h} - \langle \boldsymbol{\phi}_{i,h}, \boldsymbol{\theta}_{i,h}\rangle$, which gives $\widehat{\boldsymbol{\psi}}_{k,h}$ as

$$\widehat{\boldsymbol{\psi}}_{k,h} = \underset{\boldsymbol{\psi}\in\mathcal{B}_d(W)}{\text{argmin}} \frac{\lambda_R}{2}\|\boldsymbol{\psi}\|^2 + \sum_{i=1}^{k} \ell_{\widetilde{\tau}_{i,h}}\left(\frac{|\widehat{\varepsilon}_{i,h}|^{1+\epsilon} - \langle \boldsymbol{\phi}_{i,h}, \boldsymbol{\psi}\rangle}{\nu_{i,h}}\right). \tag{5.3}$$

The inevitable error between $\boldsymbol{\psi}_{k,h}$ and $\widehat{\boldsymbol{\psi}}_{k,h}$ can be quantified by a novel perturbation analysis of adaptive Huber regression in Appendix C.3.

[1]Consider the decision set consisting of unit bases of $\mathbb{R}^d$. Given that each arm pull can only yield information about a single coordinate, it is inevitable that d pulls are required for exploration.

Algorithm 3 HEAVY-LSVI-UCB

Require: Number of episodes K, confidence level δ, regularization parameter $\lambda_R, \lambda_V, \nu_{\min}, \sigma_{\min}$, confidence radius $\beta_{R^\epsilon}, \beta_0, \beta_R, \beta_V$.

$\kappa = d \log(1 + \frac{K}{d\lambda_R \nu_{\min}^2})$.
$\boldsymbol{\theta}_{0,h} = \widehat{\boldsymbol{w}}_{0,h} = \check{\boldsymbol{w}}_{0,h} = \mathbf{0}$, $\boldsymbol{H}_{0,h} = \lambda_R \boldsymbol{I}$, $\boldsymbol{\Sigma}_{0,h} = \lambda_V \boldsymbol{I}$, UPDATE = TRUE.
for $k = 1, \ldots, K$ **do**
  $V_{H+1}^k(\cdot) = \check{V}_{H+1}^k(\cdot) = 0$.
  for $h = H, \ldots, 1$ **do**
    Compute $\boldsymbol{\theta}_{k-1,h}$, $\widehat{\boldsymbol{w}}_{k-1,h}$ and $\check{\boldsymbol{w}}_{k-1,h}$ via (5.1) and (5.8).
    if UPDATE **then**
      $Q_h^k(\cdot,\cdot) = \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k-1,h} + \widehat{\boldsymbol{w}}_{k-1,h} \rangle + \beta_{R,k-1} \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{H}_{k-1,h}^{-1}} + \beta_V \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}$.
      $\check{Q}_h^k(\cdot,\cdot) = \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k-1,h} + \check{\boldsymbol{w}}_{k-1,h} \rangle - \beta_{R,k-1} \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{H}_{k-1,h}^{-1}} - \beta_V \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}$.
      $Q_h^k(\cdot,\cdot) = \min\{Q_h^k(\cdot,\cdot), Q_h^{k-1}(\cdot,\cdot), \mathcal{H}\}, \check{Q}_h^k(\cdot,\cdot) = \max\{\check{Q}_h^k(\cdot,\cdot), \check{Q}_h^{k-1}(\cdot,\cdot), 0\}$.
      Set $k_{\text{last}} = k$.
    else
      $Q_h^k(\cdot,\cdot) = Q_h^{k-1}(\cdot,\cdot), \check{Q}_h^k(\cdot,\cdot) = \check{Q}_h^{k-1}(\cdot,\cdot)$.
    end if
    $V_h^k(\cdot) = \max_a Q_h^k(\cdot,a)$, $\check{V}_h^k(\cdot) = \max_a \check{Q}_h^k(\cdot,a)$, $\pi_h^k(\cdot) = \operatorname{argmax}_a Q_h^k(\cdot,a)$.
  end for
  Observe initial state $s_{k,1}$.
  for $h = 1, \ldots, H$ **do**
    Take action $a_{k,h} = \pi_h^k(s_{k,h})$ and observe $R_{k,h}, s_{k,h+1}$.
    Set $\nu_{k,h}$ and $\sigma_{k,h}$ according to (5.4) and (5.9) respectively.
    Set $\tau_{k,h} = \tau_0 \frac{\sqrt{1+w_{k,h}^2}}{w_{k,h}} k^{\frac{1-\epsilon}{2(1+\epsilon)}}$, $\widetilde{\tau}_{k,h} = \widetilde{\tau}_0 \frac{\sqrt{1+w_{k,h}^2}}{w_{k,h}} k^{\frac{1-\epsilon'}{2(1+\epsilon')}}$ with $w_{k,h} = \|\boldsymbol{\phi}_{k,h}/\nu_{k,h}\|_{\boldsymbol{H}_{k,h}^{-1}}$.
    Update $\boldsymbol{H}_{k,h} = \boldsymbol{H}_{k-1,h} + \frac{1}{\nu_{k,h}^2} \boldsymbol{\phi}_{k,h} \boldsymbol{\phi}_{k,h}^\top$ and $\boldsymbol{\Sigma}_{k,h} = \boldsymbol{\Sigma}_{k-1,h} + \frac{1}{\sigma_{k,h}^2} \boldsymbol{\phi}_{k,h} \boldsymbol{\phi}_{k,h}^\top$.
  end for
  if $\exists h' \in [H]$ such that $\det(\boldsymbol{H}_{k,h'}) \geq 2\det(\boldsymbol{H}_{k_{\text{last}},h'})$ or $\det(\boldsymbol{\Sigma}_{k,h'}) \geq 2\det(\boldsymbol{\Sigma}_{k_{\text{last}},h'})$ **then**
    Set UPDATE = TRUE.
  else
    Set UPDATE = FALSE.
  end if
end for

We then set the weight $\nu_{k,h}$ for adaptive Huber regression as

$$\nu_{k,h} = \max\left\{ \widehat{\nu}_{k,h}, \nu_{\min}, \frac{\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}}{c_0}, \frac{\sqrt{\max\{B, W\}}}{c_1^{\frac{1}{4}} (2\kappa)^{\frac{1}{4}}} \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}^{\frac{1}{2}} \right\}, \tag{5.4}$$

where $\nu_{\min}$ is a small positive constant to avoid the singularity, $\widehat{\nu}_{k,h}^{1+\epsilon} = [\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h](s_{k,h}, a_{k,h}) + W_{k,h}$ is a high-probability upper bound of rewards' central moment $[\boldsymbol{\nu}_{1+\epsilon} R_h](s_{k,h}, a_{k,h})$ with

$$[\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h](s_{k,h}, a_{k,h}) = \langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{\psi}}_{k-1,h} \rangle, \tag{5.5}$$

$$W_{k,h} = \left(\beta_{R^\epsilon,k-1} + 6\mathcal{H}^\epsilon \beta_{R,k-1}\kappa\right) \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}, \tag{5.6}$$

where κ is defined in Algorithm 3, $\beta_{R^\epsilon,k} = \widetilde{O}(\sqrt{d}\nu_{R^\epsilon} k^{\frac{1-\epsilon'}{2(1+\epsilon')}}/\nu_{\min})$ and $\beta_{R,k} = \widetilde{O}(\sqrt{d} k^{\frac{1-\epsilon}{2(1+\epsilon)}})$.

Estimation for expected next-state value functions For any value function $f : \mathcal{S} \to \mathbb{R}$, we define the following notations for simplicity:

$$\boldsymbol{w}_h[f] = \int_{s\in\mathcal{S}} \boldsymbol{\mu}_h^*(s) f(s) \mathrm{d}s, \quad \widehat{\boldsymbol{w}}_{k,h}[f] = \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^k \sigma_{i,h}^{-2} \boldsymbol{\phi}_{i,h} f(s_{i,h+1}), \tag{5.7}$$

where $\sigma_{i,h}$ will be specified later. Note for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, by linear structure of transition probabilities, we have

$$[\mathbb{P}_h f](s,a) = \int_{s'\in\mathcal{S}} \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h^*(s') \rangle f(s') \mathrm{d}s' = \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h[f] \rangle.$$

In addition, for any $f, g : \mathcal{S} \to \mathbb{R}$, it holds that $\boldsymbol{w}_h[f+g] = \boldsymbol{w}_h[f] + \boldsymbol{w}_h[g]$ and $\widehat{\boldsymbol{w}}_{k,h}[f+g] = \widehat{\boldsymbol{w}}_{k,h}[f] + \widehat{\boldsymbol{w}}_{k,h}[g]$ due to the linear property of integration and ridge regression.

We remark $\widehat{\boldsymbol{w}}_{k,h}[f]$ is the estimation of $\boldsymbol{w}_h[f]$ by weighted ridge regression on $\{\boldsymbol{\phi}_{i,h}, f(s_{i,h+1})\}_{i\in[k]}$. And we estimate the coefficients $\widehat{\boldsymbol{w}}_{k,h}, \check{\boldsymbol{w}}_{k,h}, \widetilde{\boldsymbol{w}}_{k,h}$

$$\widehat{\boldsymbol{w}}_{k,h} = \widehat{\boldsymbol{w}}_{k,h}[V_{h+1}^k], \quad \check{\boldsymbol{w}}_{k,h} = \widehat{\boldsymbol{w}}_{k,h}[\check{V}_{h+1}^k], \quad \widetilde{\boldsymbol{w}}_{k,h} = \widehat{\boldsymbol{w}}_{k,h}[(V_{h+1}^k)^2], \tag{5.8}$$

where V_h^k and $\check{V}_h^k$ are optimistic and pessimistic estimation of the optimal value functions.

Estimation for variance of next-state value functions Inspired by He et al. [15], we set the weight $\sigma_{k,h}$ for weighted ridge regression in (5.7) as

$$\sigma_{k,h} = \max\left\{\widehat{\sigma}_{k,h}, \sqrt{d^3 H D_{k,h}}, \sigma_{\min}, \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}, \sqrt{d^{\frac{5}{2}} H \mathcal{H}}\, \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}^{\frac{1}{2}}\right\}, \tag{5.9}$$

where $\sigma_{\min}$ is a small constant to avoid singularity, $\widehat{\sigma}_{k,h}^2 = [\widehat{\mathbb{V}}_h V_{h+1}^k](s_{k,h}, a_{k,h}) + E_{k,h}$ with

$$\left[\widehat{\mathbb{V}}_h V_{h+1}^k\right](s_{k,h}, a_{k,h}) = \langle \boldsymbol{\phi}_{k,h}, \widetilde{\boldsymbol{w}}_{k-1,h}\rangle_{[0,\mathcal{H}^2]} - \langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h}\rangle_{[0,\mathcal{H}]}^2, \tag{5.10}$$

$$E_{k,h} = \min\left\{4\mathcal{H}\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \check{\boldsymbol{w}}_{k-1,h}\rangle + 11\mathcal{H}\beta_0 \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}, \mathcal{H}^2\right\}. \tag{5.11}$$

$$D_{k,h} = \min\left\{2\mathcal{H}\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \check{\boldsymbol{w}}_{k-1,h}\rangle + 4\mathcal{H}\beta_0 \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}, \mathcal{H}^2\right\}, \tag{5.12}$$

where $\beta_0 = \widetilde{O}(\sqrt{d^3 H \mathcal{H}^2}/\sigma_{\min})$. Here $\widehat{\sigma}_{k,h}^2$ and $D_{k,h}$ are upper bounds of $[\mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h})$ and $\max\{[\mathbb{V}_h(V_{h+1}^k - V_{h+1}^*)](s_{k,h}, a_{k,h}), [\mathbb{V}_h(V_{h+1}^* - \check{V}_{h+1}^k)](s_{k,h}, a_{k,h})\}$, respectively.

5.2 Computational Complexity

Theorem 5.1. For the linear MDPs with heavy-tailed rewards defined in Section 2.2, the computational complexity of HEAVY-LSVI-UCB is $\widetilde{O}(d^4|\mathcal{A}|H^3K + HK\mathcal{R})$. Here $\mathcal{R}$ is the cost of the optimization algorithm for solving adaptive Huber regression in (5.1). Furthermore, we can specialize $\mathcal{R}$ by adopting the Nesterov accelerated method, which gives $\mathcal{R} = \widetilde{O}(d + d^{-\frac{1-\epsilon}{2(1+\epsilon)}} H^{\frac{1-\epsilon}{2(1+\epsilon)}} K^{\frac{1+2\epsilon}{2(1+\epsilon)}})$.

Proof. See Appendix E for a detailed proof. □

Such a complexity allows us to focus on the complexity introduced by the RL algorithm rather than the optimization subroutine for solving adaptive Huber regression. Compared to that of LSVI-UCB++ [15], $\widetilde{O}(d^4|\mathcal{A}|H^3K)$, the extra term $\widetilde{O}(HK\mathcal{R})$ causes a slightly worse computational time in terms of K. This is due to the absence of a closed-form solution of adaptive Huber regression in (5.1). Thus extra optimization steps are unavoidable. Nevertheless, Nesterov accelerated method gives $\mathcal{R} = \widetilde{O}\left(K^{\frac{1+2\epsilon}{2(1+\epsilon)}}\right)$ with respect to K, which implies the computational complexity of HEAVY-LSVI-UCB is better than that of LSVI-UCB [22], $\widetilde{O}(d^2|\mathcal{A}|HK^2)$ in terms of K, thanks to the rare-switching updating policy. We conduct numerical experiments in Appendix B to further corroborate the computational efficiency of adaptive Huber regression.

5.3 Regret Bound

Theorem 5.2 (Informal)**.** For the linear MDPs with heavy-tailed rewards defined in Section 2.2, we set parameters in Algorithm 3 as follows: $\lambda_R = d/\max\{B^2, W^2\}$, $\lambda_V = 1/\mathcal{H}^2$, $\nu_{\min}$, $\sigma_{\min}$, c_0, c_1, τ_0, $\widetilde{\tau}_0$, β_{R^ϵ}, β_0, β_R, β_V in Appendix F.1. Then for any $\delta \in (0,1)$, with probability at least $1 - 16\delta$, the regret of HEAVY-LSVI-UCB is bounded by

$$\mathrm{Regret}(K) = \widetilde{O}\left(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^*K}\right),$$

where $\epsilon \in (0,1]$, $\mathcal{U}^* = \min\{\mathcal{U}_0^*, \mathcal{U}\}$, $\mathcal{V}^* = \min\{\mathcal{V}_0^*, \mathcal{V}\}$ with $\mathcal{U}_0^*$, $\mathcal{V}_0^*$ defined in Appendix F.2 and $\mathcal{H}, \mathcal{U}, \mathcal{V}$ defined in Assumption 2.7.

Proof. See Appendix F.2 for a formal version of Theorem 5.2 and its detailed proof. □

Quantities $\mathcal{U}^*$, $\mathcal{V}^*$ We make a few explanations for the quantities $\mathcal{U}^*$, $\mathcal{V}^*$. On one hand, $\mathcal{U}^*$ is upper bounded by $\mathcal{U}$, which is the upper bound of the sum of the $(1+\epsilon)$-th central moments of reward functions along a single trajectory. On the other hand, $\mathcal{U}^*$ is no more than $\mathcal{U}_0^*$, which is the sum of the $(1+\epsilon)$-th central moments with respect to the averaged occupancy measure of the first K episodes. $\mathcal{V}^*$ is defined similar to $\mathcal{U}^*$, but measures the randomness of transition probabilities.

Remark 5.3. When $\epsilon = 1$, we can show that this regret is bounded by $\widetilde{O}(d\sqrt{H\mathcal{G}^* K})$, where $\mathcal{G}^*$ is an variance-dependent quantity defined by Li and Sun [26]. Thus, our result recovers their variance-aware regret bound. See Remark F.15 in Appendix F.2 for a detailed proof.

To demonstrate the optimality of our results and establish connections with existing literature, we can specialize Theorem 5.2 to obtain the worst-case regret [22, 2, 15] and first-order regret [33].

Corollary 5.4 (Worst-case regret)**.** For the linear MDPs with heavy-tailed rewards defined in Section 2.2 and for any $\delta \in (0, 1)$, with probability at least $1 - 16\delta$, the regret of HEAVY-LSVI-UCB is bounded by

$$\widetilde{O}(dHK^{\frac{1}{1+\epsilon}} + d\sqrt{H^3 K}).$$

Proof. Notice $\mathcal{U}^*$ and $\mathcal{V}^*$ are upper bounded by $H\nu_R^2$ and $\mathcal{H}^2$ (total variance lemma in Jin et al. [21]) respectively. When $\mathcal{H} = H$, and we treat ν_R as a constant, the result follows. □

Next, we give the regret lower bound of linear MDPs with heavy-tailed rewards in Theorem 5.5, which shows our proposed HEAVY-LSVI-UCB is minimax optimal in the worst case.

Theorem 5.5. For any algorithm, there exists an H-episodic, d-dimensional linear MDP with heavy-tailed rewards such that for any K, the algorithm's regret is

$$\Omega(dHK^{\frac{1}{1+\epsilon}} + d\sqrt{H^3 K}).$$

Proof. Intuitively, the proof of Theorem 5.5 follows from a combination of the lower bound constructions for heavy-tailed linear bandits in Shao et al. [31] and linear MDPs in Zhou et al. [42]. See Appendix G for a detailed proof. □

Theorem 5.5 shows that for sufficiently large K, the reward term dominates in the regret bound. Thus, in heavy-tailed settings, the main difficulty is learning the reward functions.

Corollary 5.6 (First-order regret)**.** For the linear MDPs with heavy-tailed rewards defined in Section 2.2 and for any $\delta \in (0, 1)$, with probability at least $1 - 16\delta$, the regret of HEAVY-LSVI-UCB is bounded by

$$\widetilde{O}(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{H}V_1^* K}).$$

And when the rewards are uniformly bounded in $[0, 1]$, the result is reduced to the first-order regret bound of $\widetilde{O}(d\sqrt{H^2 V_1^* K})$.

Proof. See Section F.3 for a detailed proof. □

Our first-order regret $\widetilde{O}(d\sqrt{H^2 V_1^* K})$ is minimax optimal in the worst case since $V_1^* \le H$. And it improves the state-of-the-art result $\widetilde{O}(d\sqrt{H^3 V_1^* K})$ [26] by a factor of $\sqrt{H}$.

6 Conclusion

In this work, we propose two computationally efficient algorithms for heavy-tailed linear bandits and linear MDPs, respectively. Our proposed algorithms, termed as HEAVY-OFUL and HEAVY-LSVI-UCB, are based on a novel self-normalized concentration inequality for adaptive Huber regression, which may be of independent interest. HEAVY-OFUL and HEAVY-LSVI-UCB achieve minimax optimal and instance-dependent regret bounds scaling with the central moments. We also provide a lower bound for linear MDPs with heavy-tailed rewards to demonstrate the optimality of HEAVY-LSVI-UCB. To the best of our knowledge, we are the first to study heavy-tailed rewards in RL with function approximation and provide a new algorithm for this setting which is both statistically and computationally efficient.

Acknowledgments

Liwei Wang is supported in part by NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844, NSF IIS 2229881. Lin F. Yang is supported in part by NSF grant 2221871, and an Amazon Research Grant.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[2] Alekh Agarwal, Yujia Jin, and Tong Zhang. VOQL: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR, 2023.

[3] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

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

[5] 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.

[6] Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Nearly optimal catoni's m-estimator for infinite variance. In *International Conference on Machine Learning*, pages 1925–1944. PMLR, 2022.

[7] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[8] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[9] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[10] Peng Chen, Xinghu Jin, Xiang Li, and Lihu Xu. A generalized catoni's m-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics*, 15(2):5523–5544, 2021.

[11] Hana Choi, Carl F Mela, Santiago R Balseiro, and Adam Leary. Online display advertising markets: A literature review and future directions. *Information Systems Research*, 31(2): 556–575, 2020.

[12] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

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

[14] A Ben Hamza and Hamid Krim. Image denoising: A nonlinear robust statistical approach. *IEEE transactions on signal processing*, 49(12):3045–3054, 2001.

[15] Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

[16] Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 8971–9019, 2022.

[17] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[18] John Hull. *Risk management and financial institutions,+ Web Site*, volume 733. John Wiley & Sons, 2012.

[19] Charles Jebarajakirthy, Haroon Iqbal Maseeh, Zakir Morshed, Amit Shankar, Denni Arli, and Robin Pentecost. Mobile advertising: A systematic literature review and future research agenda. *International Journal of Consumer Studies*, 45(6):1258–1291, 2021.

[20] Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

[21] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

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

[23] Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *Advances in Neural Information Processing Systems*, 35:1060–1072, 2022.

[24] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

[25] Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 2023.

[26] Xiang Li and Qiang Sun. Variance-aware robust reinforcement learning with linear function approximation with heavy-tailed rewards. *arXiv preprint arXiv:2303.05606*, 2023.

[27] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[28] Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650. PMLR, 2016.

[29] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

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

[31] Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

[32] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[33] Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR, 2022.

[34] Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2936–2942. International Joint Conferences on Artificial Intelligence Organization, 2020.

[35] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[36] 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.

[37] Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021.

[38] Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4977–5020. PMLR, 2023.

[39] Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *arXiv preprint arXiv:2305.08841*, 2023.

[40] Han Zhong, Jiayi Huang, Lin Yang, and Liwei Wang. Breaking the moments condition barrier: No-regret algorithm for bandits with super heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 34:15710–15720, 2021.

[41] Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349, 2022.

[42] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[43] 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.

[44] Vincent Zhuang and Yanan Sui. No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 3385–3393. PMLR, 2021.

A Related Work

RL with linear function approximation In this work, we focus on the setting of linear MDPs [35, 22, 16, 2, 15, 39], where the reward functions and transition probabilities can be expressed as linear functions of some known, d-dimensional state-action features. Jin et al. [22] proposed the first computationally efficient algorithm LSVI-UCB that achieves $\widetilde{O}(\sqrt{d^3H^4K})$ worst-case regret in online learning settings. LSVI-UCB constructs upper confidence bounds for action-value functions based on least squares regression. Then Agarwal et al. [2], He et al. [15] improve it to $\widetilde{O}(d\sqrt{H^3K})$, which matches the minimax lower bound $\Omega(d\sqrt{H^3K})$ by Zhou et al. [42] up to logarithmic factors. There is another line of works that make the linear mixture MDP assumption [29, 20, 4, 42, 41, 38] where the transition kernel is a linear combination of some basis transition probability functions.

Instance-dependent regret in bandits and RL Recently, there are plenty of works that achieve instance-dependent regret in bandits and RL [24, 42, 37, 23, 41, 38, 36, 43, 33, 26]. Instance-dependent regret uses fine-grained quantities that inherently characterize the problems, and thus provides tighter guarantee than worst-case regret. These results approximately provide two different kinds of regret. The first is first-order regret, which was originally achieved by Zanette and Brunskill [36] in tabular MDPs. Then Wagenmaker et al. [33] proposed a computationally inefficient[2] algorithm FORCE in linear MDPs that achieves $\widetilde{O}(\sqrt{d^3H^3V_1^*K})$ first-order regret, where V_1^* is the optimal value function. Recently, Li and Sun [26] improved it to a computationally efficient result of $\widetilde{O}(d\sqrt{H^3V_1^*K})$. Since first-order regret depends on the optimal value function, which is non-diminishing when the MDPs are less stochastic, it is typically sub-optimal in such deterministic cases. The second is variance-aware regret, which is well-studied in linear bandits with light-tailed rewards [24, 41]. These works are based on weighted ridge regression and a Bernstein-style concentration inequality. However, hardly few works consider heavy-tailed rewards. One exception is Li and Sun [26], which improved adaptive Huber regression and provided the first variance-aware regret in the presence of finite-variance rewards. Unfortunately, little has been done in the case $\epsilon < 1$.

Heavy-tailed rewards in bandits and RL Bubeck et al. [7] made the first attempt to study heavy-tailed rewards in Multi-Armed Bandits (MAB). Since then, robust mean estimators, such as median-of-means and truncated mean, have been broadly utilized in linear bandits [28, 31, 34, 40] to achieve tight regret bounds. As far as our knowledge, the only work that considers heavy-tailed rewards where the variance could be non-existent in RL is Zhuang and Sui [44], which established a regret bound in tabular MDPs that is tight with respect to S, A, K by plugging truncated mean into UCRL2 [3] and Q-Learning [21].

Robust mean estimators and robust regression Lugosi and Mendelson [27] provided an overview of most robust mean estimators for heavy-tailed distributions, including the median-of-means estimator, truncated mean, and Catoni's M-estimator. The median-of-means estimator has limitation with minimum sample size. Truncated mean uses raw $(1+\epsilon)$-th moments, thus is sub-optimal in the deterministic case. The original version of Catoni's M-estimator in Catoni [9] requires the existence of finite variance. Then Chen et al. [10], Bhatt et al. [6] generalized it to handle heavy-tailed distributions while preserving the same order as $\epsilon \to 1$. Wagenmaker et al. [33] extended Catoni's M-estimator to general heterogeneous online regression settings with some covering arguments. Different from Catoni [9], their results scale with the second raw moments instead of variance. Sun et al. [32] imposed adaptive Huber regression to handle homogeneous offline heavy-tailed noise by utilizing Huber loss [17] as a surrogate of squared loss. Li and Sun [26] modified adaptive Huber regression to handle finite-variance noise in heterogeneous online regression settings.

B Experiments

In this section, we conduct empirical evaluations of the proposed algorithm, HEAVY-OFUL, for heavy-tailed linear bandit problems in Definition 2.1 which can be regarded as a special case of linear MDPs. Comparisons are made between MENU and TOFU [31], which give the worst-case optimal regret bound in such settings (See Table 1). To the best of our knowledge, we are the first to address

[2] Wagenmaker et al. [33] also provided a computationally efficient alternative with an extra factor of $\sqrt{d}$.

the challenge of heavy-tailed rewards in RL with function approximation, even when ϵ is less than 1. Consequently, no other algorithms in the RL literature can be readily compared to our approach (See Table 2).

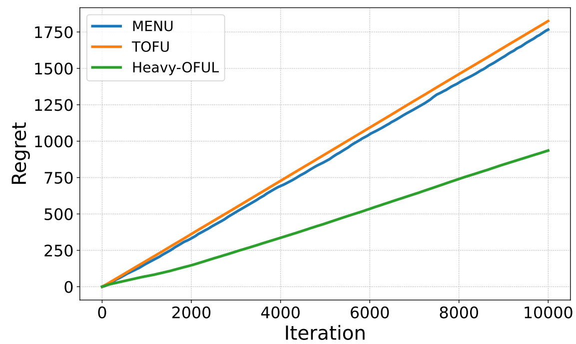

Figure 1: Comparisons of our algorithm (HEAVY-OFUL) versus MENU and TOFU in heavy-tailed linear bandits problems for 1×10^4 rounds.

We generate 5 independent paths for each algorithm and show the average cumulative regret. The experimental setup is as follows: Let the feature dimension $d = 10$. For the chosen arm $\boldsymbol{\phi}_t \in \mathcal{D}_t$, reward is $R_t = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^* \rangle + \varepsilon_t$, where $\boldsymbol{\theta}^* = \mathbf{1}_d/\sqrt{d} \in \mathbb{R}^d$ so that $\|\boldsymbol{\theta}^*\|_2 = 1$. ε_t is first sampled from a Student's t-distribution with degree of freedom $\mathrm{df} = 2$, then is multiplied by a scaling factor α such that the central moments of ε_t in each rounds are different, where $\log_{10}(\alpha) \sim \mathrm{Unif}(0, 2)$. Note the variance of ε_t does not exist and we choose $\epsilon = 0.99$. Normalization is made to ensure $L = B = 1$. As shown in Figure 1, results demonstrate the effectiveness of the proposed algorithm, which further corroborates our theoretical findings.

C Proofs for Section 3

C.1 Properties of Huber Loss

The following properties of Huber loss are important in the proofs.

Property 1. Denote $\Psi_\tau(x)$ as the derivative of Huber loss, i.e., $\Psi_\tau(x) := \ell'_\tau(x)$, where $\Psi_\tau(x)$ is Huber loss defined in Definition 3.1. Then the followings are true:

(1) $|\Psi_\tau(x)| = \min\{|x|, \tau\}$,

(2) $\Psi_\tau(x) = \tau \Psi_1(\frac{x}{\tau})$,

(3) $-\log(1 - x + |x|^{1+\epsilon}) \le \Psi_1(x) \le \log(1 + x + |x|^{1+\epsilon})$ for any $\epsilon \in (0, 1]$,

(4) $\Psi'_\tau(x) = \ell''_\tau(x) = \mathbb{1}\{|x| \le \tau\}$.

Here, (1) gives an upper bound of the derivative of Huber loss. (2) demonstrates the homogeneous property. (3) shows the similarity between the derivative of Huber loss and influence function in Catoni [9], thereby motivating us to study the moment generating function for characterizing the noise distributions (See details in the proof of Lemma C.4 in Appendix C.6). And (4) is to characterize the second derivative of Huber loss.

C.2 Proof of Theorem 3.3

Parameters for Adaptive Huber Regression. First, we set the parameters in Algorithm 1 as follows:

$$c_0 = \frac{1}{\sqrt{23 \log \frac{2T^2}{\delta}}}, \quad c_1 = \frac{(\log 3T)^{\frac{1-\epsilon}{1+\epsilon}}}{48 \left(\log \frac{2T^2}{\delta}\right)^{\frac{2}{1+\epsilon}}}, \quad \tau_0 = \frac{\sqrt{2\kappa} b (\log 3T)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\epsilon}}}.$$

Then, we give the proof of Theorem 3.3.

Proof of Theorem 3.3. Denote $z_t(\boldsymbol{\theta}) := \frac{y_t - \langle \boldsymbol{\phi}_t, \boldsymbol{\theta} \rangle}{\sigma_t}$. By the definition of b and σ_t in Line 1 and 3 of Algorithm 1, respectively,

$$\mathbb{E}[|z_t(\boldsymbol{\theta}^*)|^{1+\epsilon}|\mathcal{F}_{t-1}] = \mathbb{E}\left[\left|\frac{\varepsilon_t}{\sigma_t}\right|^{1+\epsilon}\middle|\mathcal{F}_{t-1}\right] \le b^{1+\epsilon}.$$

Notice the gradient of $L_t(\boldsymbol{\theta})$ is give by

$$\nabla L_t(\boldsymbol{\theta}) = \lambda \boldsymbol{\theta} - \sum_{s=1}^{t} \frac{\boldsymbol{\phi}_s}{\sigma_s} \Psi_{\tau_s}(z_s(\boldsymbol{\theta})). \tag{C.1}$$

Recall $\boldsymbol{\theta}_t$ is the solution of Line 6 in Algorithm 1, which is a constrained convex optimization problem. Thus it holds that $\langle \nabla L_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta} \rangle \le 0,\ \forall \boldsymbol{\theta} \in \mathcal{B}_d(B)$. Since $\|\boldsymbol{\theta}^*\| \le B$, we have

$$\langle \nabla L_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}^* \rangle \le 0. \tag{C.2}$$

By the mean value theorem for vector-valued functions, we have

$$\nabla L_t(\boldsymbol{\theta}_t) - \nabla L_t(\boldsymbol{\theta}^*) = \int_0^1 \nabla^2 L_t((1-\eta)\boldsymbol{\theta}^* + \eta \boldsymbol{\theta}_t) \mathrm{d}\eta \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*). \tag{C.3}$$

In Lemma C.1, we prove that $\nabla^2 L_t(\boldsymbol{\theta})$ approximates $\boldsymbol{H}_t$ well as long as $\boldsymbol{\theta} \in \mathcal{B}_d(B)$.

Lemma C.1. Assume $\mathbb{E}[z_t(\boldsymbol{\theta}^*)|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[|z_t(\boldsymbol{\theta}^*)|^{1+\epsilon}|\mathcal{F}_{t-1}] \le b^{1+\epsilon}$, set

$$\tau_0 \ge \frac{\sqrt{2\kappa} b (\log 3T)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\epsilon}}}, \quad c_0 \le \frac{1}{\sqrt{23 \log \frac{2T^2}{\delta}}}, \quad c_1 \le \frac{(\log 3T)^{\frac{1-\epsilon}{1+\epsilon}}}{48 \left(\log \frac{2T^2}{\delta}\right)^{\frac{2}{1+\epsilon}}}.$$

Then with probability at least $1-\delta$, for all $t \le T$ and $\boldsymbol{\theta} \in \mathcal{B}_d(B)$, we have

$$\frac{1}{3}\boldsymbol{H}_t \preceq \nabla^2 L_t(\boldsymbol{\theta}) \preceq \boldsymbol{H}_t.$$

Proof. See Appendix C.4 for a detailed proof. □

We set τ_0, c_0, c_1 to follow Lemma C.1. Lemma C.1 implies that for all $t \le T$ and $\eta \in [0,1]$, we have

$$\nabla^2 L_t((1-\eta)\boldsymbol{\theta}^* + \eta \boldsymbol{\theta}_t) \succeq \frac{1}{3}\boldsymbol{H}_t.$$

Thus, multiplying $\boldsymbol{\theta}_t - \boldsymbol{\theta}^*$ in both sides of (C.3) yields

$$\begin{aligned}
&\langle \boldsymbol{\theta}_t - \boldsymbol{\theta}^*, \nabla L_t(\boldsymbol{\theta}_t) - \nabla L_t(\boldsymbol{\theta}^*) \rangle \\
&= \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}^*, \int_0^1 \nabla^2 L_t((1-\eta)\boldsymbol{\theta}^* + \eta \boldsymbol{\theta}_t) \mathrm{d}\eta \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) \rangle \\
&\ge \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}^*, \frac{1}{3}\boldsymbol{H}_t \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) \rangle \\
&= \frac{1}{3}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}_t}^2.
\end{aligned} \tag{C.4}$$

On the other hand, by (C.2), we have

$$\langle \boldsymbol{\theta}_t - \boldsymbol{\theta}^*, \nabla L_t(\boldsymbol{\theta}_t) - \nabla L_t(\boldsymbol{\theta}^*) \rangle \le \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}^*, -\nabla L_t(\boldsymbol{\theta}^*) \rangle \le \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}_t} \cdot \|\nabla L_t(\boldsymbol{\theta}^*)\|_{\boldsymbol{H}_t^{-1}}.$$

That implies

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}_t} \le 3\|\nabla L_t(\boldsymbol{\theta}^*)\|_{\boldsymbol{H}_t^{-1}}. \tag{C.5}$$

Next we use Lemma C.2 to give a high probability upper bound of $\|\nabla L_t(\boldsymbol{\theta}^*)\|_{\boldsymbol{H}_t^{-1}}$.

Lemma C.2. Assume $\mathbb{E}[z_t(\boldsymbol{\theta}^*)|\mathcal{F}_{t-1}] = 0$, $\mathbb{E}[|z_t(\boldsymbol{\theta}^*)|^{1+\epsilon}|\mathcal{F}_{t-1}] \leq b^{1+\epsilon}$ and

$$\tau_0 \geq \frac{\sqrt{2\kappa} b (\log 3T)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\epsilon}}},$$

then with probability at least $1-2\delta, \forall t \geq 1$, we have

$$\|\nabla L_t(\boldsymbol{\theta}^*)\|_{\boldsymbol{H}_t^{-1}} \leq \sqrt{\lambda} B + \alpha_t,$$

where

$$\alpha_t = 4t^{\frac{1-\epsilon}{2(1+\epsilon)}} \left[\frac{(\sqrt{2\kappa} b)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^\epsilon} + \tau_0 \log \frac{2t^2}{\delta} \right].$$

Proof. See Appendix C.5 for a detailed proof. □

Notice the conditions of Lemma C.2 are satisfied. On the event where Lemma C.1 and C.2 hold, whose probability is at least $1-3\delta$, combining (C.5) and Lemma C.2 gives

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}_t} \leq 3\sqrt{\lambda} B + 12 t^{\frac{1-\epsilon}{2(1+\epsilon)}} \left[\frac{(\sqrt{2\kappa} b)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^\epsilon} + \tau_0 \log \frac{2t^2}{\delta} \right].$$

Choosing $\tau_0 = \frac{\sqrt{2\kappa} b (\log 3T)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\epsilon}}}$ gives the final result. □

C.3 Perturbation Analysis of Adaptive Huber Regression

Lemma C.3. Consider the setting in Definition 3.2. Let $L_t(\cdot), \boldsymbol{\theta}_t$ be defined in Theorem 3.3. Define $\widehat{L}_t(\boldsymbol{\theta}) := \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \sum_{s=1}^t \ell_{\tau_s}\left(\frac{\widehat{y}_s - \langle \boldsymbol{\phi}_s, \boldsymbol{\theta}\rangle}{\sigma_s}\right)$, where the parameters τ_s, σ_s are chosen according to Theorem 3.3. $\widehat{L}_t(\cdot)$ is the same as $L_t(\cdot)$ except for substituting y_s with $\widehat{y}_s \in \mathcal{F}_s$. And let $\widehat{\boldsymbol{\theta}}_t := \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{B}_d(B)} \widehat{L}_t(\boldsymbol{\theta})$. Suppose $|\widehat{y}_s - y_s| \leq \widehat{\beta}_s \|\boldsymbol{\phi}_s\|_{\boldsymbol{H}_s^{-1}}$ for $s \leq t$. Then, under the event where Theorem 3.3 holds, we have

$$\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\|_{\boldsymbol{H}_t} \leq 6\kappa \max_{s \leq t} \widehat{\beta}_s.$$

Proof. Notice $\boldsymbol{\theta}_t$ and $\widehat{\boldsymbol{\theta}}_t$ are solutions of convex optimization problems, we have

$$\langle \nabla L_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t \rangle \leq 0, \quad \langle \nabla \widehat{L}_t(\widehat{\boldsymbol{\theta}}_t), \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t \rangle \leq 0. \tag{C.6}$$

With a similar argument as for (C.4) in the proof of Theorem 3.3 in Appendix C.2, on the event where Theorem 3.3 holds, we have

$$\langle \boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t, \nabla L_t(\boldsymbol{\theta}) - \nabla L_t(\widehat{\boldsymbol{\theta}}_t) \rangle \geq \frac{1}{3} \|\boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t\|_{\boldsymbol{H}_t}^2.$$

Combined with (C.6), it holds that

$$\|\boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t\|_{\boldsymbol{H}_t}^2 \leq 3\langle \boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t, \nabla \widehat{L}_t(\widehat{\boldsymbol{\theta}}_t) - \nabla L_t(\widehat{\boldsymbol{\theta}}_t) \rangle \leq 3\|\boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t\|_{\boldsymbol{H}_t} \|\nabla \widehat{L}_t(\widehat{\boldsymbol{\theta}}_t) - \nabla L_t(\widehat{\boldsymbol{\theta}}_t)\|_{\boldsymbol{H}_t^{-1}}.$$

Thus

$$
\begin{aligned}
\|\boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t\|_{\boldsymbol{H}_t} &\le 3\|\nabla \widehat{L}_t(\widehat{\boldsymbol{\theta}}_t) - \nabla L_t(\widehat{\boldsymbol{\theta}}_t)\|_{\boldsymbol{H}_t^{-1}} \\
&= 3\left\|\sum_{s=1}^{t} \frac{\boldsymbol{\phi}_s}{\sigma_s}\left[\Psi_{\tau_s}\left(\frac{\widehat{y}_s - \langle \phi_s, \widehat{\boldsymbol{\theta}}_t\rangle}{\sigma_s}\right) - \Psi_{\tau_s}\left(\frac{y_s - \langle \phi_s, \widehat{\boldsymbol{\theta}}_t\rangle}{\sigma_s}\right)\right]\right\|_{\boldsymbol{H}_t^{-1}} \\
&\overset{(a)}{\le} 3\left\|\sum_{s=1}^{t} \frac{\boldsymbol{\phi}_s}{\sigma_s}\left|\frac{\widehat{y}_s - y_s}{\sigma_s}\right|\right\|_{\boldsymbol{H}_t^{-1}} \\
&\overset{(b)}{\le} 3\cdot(\max_{s\le t}\widehat{\beta}_s)\cdot\left\|\sum_{s=1}^{t} \frac{\boldsymbol{\phi}_s}{\sigma_s}\cdot\left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_s^{-1}}\right\|_{\boldsymbol{H}_t^{-1}} \\
&\le 3\cdot(\max_{s\le t}\widehat{\beta}_s)\cdot\sum_{s=1}^{t}\left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_t^{-1}}\cdot\left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_s^{-1}}
\end{aligned}
$$

where (a) is due to (1) of Property 1 and (b) uses the condition that $|\widehat{y}_s - y_s| \le \widehat{\beta}_s\,\|\boldsymbol{\phi}_s\|_{\boldsymbol{H}_s^{-1}}$. Notice $\boldsymbol{H}_t^{-1} \preceq \boldsymbol{H}_s^{-1}$ for $s \le t$, we have $\left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_t^{-1}} \le \left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_s^{-1}}$, which further implies

$$
\left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_s^{-1}}^2 = \frac{\boldsymbol{\phi}_s^\top}{\sigma_s}\left(\boldsymbol{H}_{s-1}^{-1} - \frac{\boldsymbol{H}_{s-1}^{-1}\frac{\boldsymbol{\phi}_s}{\sigma_s}\frac{\boldsymbol{\phi}_s^\top}{\sigma_s}\boldsymbol{H}_{s-1}^{-1}}{1+\frac{\boldsymbol{\phi}_s^\top}{\sigma_s}\boldsymbol{H}_{s-1}^{-1}\frac{\boldsymbol{\phi}_s^\top}{\sigma_s}}\right)\frac{\boldsymbol{\phi}_s}{\sigma_s} = w_s^2 - \frac{w_s^4}{1+w_s^2} = \frac{w_s^2}{1+w_s^2} \tag{C.7}
$$

due to Sherman-Morrison formula with $w_t = \|\boldsymbol{\phi}_t/\sigma_t\|_{\boldsymbol{H}_{t-1}^{-1}}$. Then It follows that

$$
\sum_{s=1}^{t}\left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_t^{-1}}\cdot\left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_s^{-1}} \le \sum_{s=1}^{t}\frac{w_s^2}{1+w_s^2} \le \sum_{s=1}^{t}\min\left\{1, w_s^2\right\} \le 2\kappa,
$$

where the last inequality holds due to Lemma I.6. Finally, we have

$$
\|\boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t\|_{\boldsymbol{H}_t} \le 6\kappa \max_{s\le t}\widehat{\beta}_s.
$$

□

C.4 Proof of Lemma C.1

In the following proof, we denote $z_t^* := z_t(\boldsymbol{\theta}^*) = \frac{y_t - \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^*\rangle}{\sigma_t} = \frac{\varepsilon_t}{\sigma_t}$ for short.

Proof of Lemma C.1. Using (4) of Property 1, the Hessian matrix of $L_t(\boldsymbol{\theta})$ is given by

$$
\nabla^2 L_t(\boldsymbol{\theta}) = \lambda \boldsymbol{I} + \sum_{s=1}^{t}\frac{1}{\sigma_s^2}\boldsymbol{\phi}_s\boldsymbol{\phi}_s^\top \mathbb{1}\{|z_s(\boldsymbol{\theta})| \le \tau_s\} = \boldsymbol{H}_t - \sum_{s=1}^{t}\frac{1}{\sigma_s^2}\boldsymbol{\phi}_s\boldsymbol{\phi}_s^\top \mathbb{1}\{|z_s(\boldsymbol{\theta})| > \tau_s\}. \tag{C.8}
$$

Thus

$$
\nabla^2 L_t(\boldsymbol{\theta}) \preceq \boldsymbol{H}_t
$$

holds trivially. And we only need to prove

$$
\nabla^2 L_t(\boldsymbol{\theta}) \succeq \frac{1}{3}\boldsymbol{H}_t.
$$

We can decompose $z_s(\boldsymbol{\theta})$ as

$$
z_s(\boldsymbol{\theta}) = \frac{y_s - \langle \boldsymbol{\phi}_s, \boldsymbol{\theta}\rangle}{\sigma_s} = \frac{\varepsilon_s + \langle \boldsymbol{\phi}_s, \boldsymbol{\theta}^* - \boldsymbol{\theta}\rangle}{\sigma_s} = z_s^* + \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{\theta}^* - \boldsymbol{\theta}\rangle}{\sigma_s}.
$$

Then notice

$$
\mathbb{1}\{|z_s(\boldsymbol{\theta})| > \tau_s\} \le \mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\} + \mathbb{1}\left\{\left|\frac{\langle \boldsymbol{\phi}_s, \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle}{\sigma_s}\right| > \frac{\tau_s}{2}\right\}.
$$

For all $\boldsymbol{u} \in \mathcal{B}_d(1)$, we have

$$\boldsymbol{u}^\top \nabla^2 L_t(\boldsymbol{\theta}) \boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{H}_t \boldsymbol{u} - \sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u} \rangle^2}{\sigma_s^2} \mathbb{1}\{|z_s(\boldsymbol{\theta})| > \tau_s\}$$

$$\geq \boldsymbol{u}^\top \boldsymbol{H}_t \boldsymbol{u} - \underbrace{\sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u} \rangle^2}{\sigma_s^2} \mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\}}_{\text{(i)}} - \underbrace{\sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u} \rangle^2}{\sigma_s^2} \mathbb{1}\left\{\left|\frac{\langle \boldsymbol{\phi}_s, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle}{\sigma_s}\right| > \frac{\tau_s}{2}\right\}}_{\text{(ii)}}.$$

Next we establish upper bounds for term (i) and (ii) respectively.

Term (i). Notice

$$\begin{aligned}\sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u} \rangle^2}{\sigma_s^2} \mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\} &\overset{(a)}{\leq} \sum_{s=1}^t \left\|\frac{\boldsymbol{\phi}_s}{\sigma_s}\right\|_{\boldsymbol{H}_t^{-1}}^2 \|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2 \mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\} \\ &\overset{(b)}{=} \|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2 \sum_{s=1}^t \frac{w_s^2}{1+w_s^2} \mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\} \\ &\overset{(c)}{\leq} c_0^2 \|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2 \sum_{s=1}^t \mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\},\end{aligned} \tag{C.9}$$

where (a) holds due to the Cauchy-Schwartz inequality, (b) holds with the same argument as (C.7), while (c) holds due to $\frac{w_s^2}{1+w_s^2} \leq w_s^2$ and our choice of σ_s in Line 3 of Algorithm 1 such that $\sigma_s \geq \frac{\|\boldsymbol{\phi}_s\|_{\boldsymbol{H}_{s-1}^{-1}}}{c_0}$.

Next we give an upper bound of $\sum_{s=1}^t \mathbb{1}\{|z_s^*| > \frac{\tau_s}{2}\}$. Define $X_s := \mathbb{1}\{|z_s^*| > \frac{\tau_s}{2}\}$ and $Y_s := X_s - \mathbb{E}[X_s|\mathcal{F}_{s-1}]$, which are $\mathcal{F}_s$-measurable. Lemma I.1 implies that with probability at least $1 - 2t^2/\delta$, for any fixed $t \geq 1$, we have

$$\sum_{s=1}^t Y_s \leq \sqrt{2V \log \frac{2t^2}{\delta}} + \frac{2M}{3} \log \frac{2t^2}{\delta}, \tag{C.10}$$

where $\sum_{s=1}^t \mathbb{E}[Y_s^2|\mathcal{F}_{s-1}] \leq V$ and $|Y_t| \leq M$. It follows that $|Y_s| \leq 1$ and

$$\begin{aligned}\sum_{s=1}^t \mathbb{E}[X_s|\mathcal{F}_{s-1}] &= \sum_{s=1}^t \mathbb{P}\left(|z_s^*| > \frac{\tau_s}{2}\right) \\ &\overset{(a)}{\leq} 2^{1+\epsilon} \sum_{s=1}^t \frac{b^{1+\epsilon}}{\tau_s^{1+\epsilon}} \\ &\overset{(b)}{=} \left(\frac{2b}{\tau_0}\right)^{1+\epsilon} \sum_{s=1}^t \left(\frac{w_s}{\sqrt{1+w_s^2}}\right)^{1+\epsilon} s^{-\frac{1-\epsilon}{2}},\end{aligned}$$

where (a) holds due to Markov's inequality and (b) holds due to the definition of τ_s in Line 4 of Algorithm 1. Notice

$$\sum_{s=1}^t \left(\frac{w_s}{\sqrt{1+w_s^2}}\right)^{1+\epsilon} s^{-\frac{1-\epsilon}{2}} \leq \left(\sum_{s=1}^t \frac{w_s^2}{1+w_s^2}\right)^{\frac{1+\epsilon}{2}} \left(\sum_{s=1}^t s^{-1}\right)^{\frac{1-\epsilon}{2}} \leq (2\kappa)^{\frac{1+\epsilon}{2}} (\log 3t)^{\frac{1-\epsilon}{2}}, \tag{C.11}$$

where the first inequality holds due to Hölder's inequality and the second holds due to $\frac{w_s^2}{1+w_s^2} \leq \min\{1, w_s^2\}$, Lemma I.6 and $\sum_{s=1}^t s^{-1} \leq 1 + \log t \leq \log 3t$. Thus

$$\sum_{s=1}^t \mathbb{E}[X_s|\mathcal{F}_{s-1}] \leq \left(\frac{2b}{\tau_0}\right)^{1+\epsilon} (2\kappa)^{\frac{1+\epsilon}{2}} (\log 3t)^{\frac{1-\epsilon}{2}} = \left(\frac{2\sqrt{2\kappa} b (\log 3t)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\tau_0}\right)^{1+\epsilon}.$$

Then it follows that

$$\sum_{s=1}^{t}\mathbb{E}[Y_s^2|\mathcal{F}_{s-1}] \le \sum_{s=1}^{t}\mathbb{E}[X_s^2|\mathcal{F}_{s-1}] = \sum_{s=1}^{t}\mathbb{E}[X_s|\mathcal{F}_{s-1}].$$

Setting $V = \left(\frac{2\sqrt{2\kappa}b(\log 3t)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\tau_0}\right)^{1+\epsilon}$ and $M = 1$ in (C.10), using a union bound over $t \ge 1$ and the fact that $\sum_{t=1}^{\infty}\frac{1}{2t^2} \le 1$, with probability at least $1-\delta, \forall t \ge 1$, we have

$$\begin{aligned}
&\sum_{s=1}^{t}\mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\} = \sum_{s=1}^{t}X_s \\
&\le \sum_{s=1}^{t}\mathbb{E}[X_s|\mathcal{F}_{s-1}] + \sqrt{2V\log\frac{2t^2}{\delta}} + \frac{2M}{3}\log\frac{2t^2}{\delta} \qquad \text{(C.12)}\\
&= \left(\frac{2\sqrt{2\kappa}b(\log 3t)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\tau_0}\right)^{1+\epsilon} + \sqrt{2\left(\frac{2\sqrt{2\kappa}b(\log 3t)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\tau_0}\right)^{1+\epsilon}\log\frac{2t^2}{\delta}} + \frac{2}{3}\log\frac{2t^2}{\delta}.
\end{aligned}$$

Choosing $\tau_0 \ge \frac{\sqrt{2\kappa}b(\log 3T)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\left(\log\frac{2T^2}{\delta}\right)^{\frac{1}{1+\epsilon}}}$ yields $\left(\frac{2\sqrt{2\kappa}b(\log 3t)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\tau_0}\right)^{1+\epsilon} \le 2^{1+\epsilon}\log\frac{2T^2}{\delta}$.

Then it follows from (C.9) and (C.12) that

$$\begin{aligned}
\sum_{s=1}^{t}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{u}\rangle^2}{\sigma_s^2}\mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\} &\le c_0^2\|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2\sum_{s=1}^{t}\mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\}\\
&\le c_0^2\|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2\left(2^{1+\epsilon} + \sqrt{2\cdot 2^{1+\epsilon}} + \frac{2}{3}\right)\log\frac{2T^2}{\delta}\\
&\le \frac{23}{3}c_0^2\log\frac{2T^2}{\delta}\|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2\\
&\le \frac{1}{3}\|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2,
\end{aligned}$$

where the third inequality holds due to $\epsilon \le 1$ and the last holds due to the definition of c_0.

Term (ii).

$$\begin{aligned}
\sum_{s=1}^{t}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{u}\rangle^2}{\sigma_s^2}\mathbb{1}\left\{\left|\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle}{\sigma_s}\right| > \frac{\tau_s}{2}\right\} &\le \sum_{s=1}^{t}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{u}\rangle^2}{\sigma_s^2}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle^2}{\sigma_s^2}\frac{4}{\tau_s^2}\\
&\overset{(a)}{\le} 16\sum_{s=1}^{t}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{u}\rangle^2}{\sigma_s^2}\frac{L^2B^2}{\sigma_s^2\tau_0^2}\frac{w_s^2}{1+w_s^2}s^{-\frac{1-\epsilon}{1+\epsilon}}\\
&\overset{(b)}{\le} 16\sum_{s=1}^{t}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{u}\rangle^2}{\sigma_s^2}\frac{L^2B^2w_s^2}{\sigma_s^2\cdot 2\kappa b^2}\frac{2\kappa b^2}{\tau_0^2} \qquad \text{(C.13)}\\
&\overset{(c)}{\le} 16\sum_{s=1}^{t}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{u}\rangle^2}{\sigma_s^2}c_1\frac{2\kappa b^2}{\tau_0^2}\\
&\overset{(d)}{\le} \frac{1}{3}\sum_{s=1}^{t}\frac{\langle\boldsymbol{\phi}_s,\boldsymbol{u}\rangle^2}{\sigma_s^2},
\end{aligned}$$

where (a) holds due to Cauchy-Schwartz inequality with $\|\boldsymbol{\phi}_s\| \le L$, $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \le 2B$ and the definition of τ_s in Line 4 of Algorithm 1. (b) holds due to $\frac{1}{1+w_s^2}s^{-\frac{1-\epsilon}{1+\epsilon}} \le 1$. (c) uses the definition of σ_s in Line 3 in Algorithm 1, that is $\frac{L^2B^2w_s^2}{\sigma_s^2\cdot 2\kappa b^2} \le c_1$ implied by $\sigma_s \ge \frac{\sqrt{LB}\|\boldsymbol{\phi}_s\|_{\boldsymbol{H}_{s-1}^{-1}}^{\frac{1}{2}}}{c_1^{\frac{1}{4}}(2\kappa b^2)^{\frac{1}{4}}}$. (d) holds due to our choice of τ_0 and c_1.

Putting pieces together. Combining (C.12) and (C.13), with probability at least $1-\delta$, for all $t \le T, \boldsymbol{\theta} \in \mathcal{B}_d(B)$ and $\boldsymbol{u} \in \mathcal{B}_d(1)$, we have

$$
\begin{aligned}
&\boldsymbol{u}^\top \nabla^2 L_t(\boldsymbol{\theta})\boldsymbol{u} \\
&\ge \boldsymbol{u}^\top \boldsymbol{H}_t \boldsymbol{u} - \sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u}\rangle^2}{\sigma_s^2} \mathbb{1}\left\{|z_s^*| > \frac{\tau_s}{2}\right\} - \sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u}\rangle^2}{\sigma_s^2} \mathbb{1}\left\{\left|\frac{\langle \boldsymbol{\phi}_s, \boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle}{\sigma_s}\right| > \frac{\tau_s}{2}\right\}. \\
&\ge \boldsymbol{u}^\top \boldsymbol{H}_t \boldsymbol{u} - \frac{1}{3}\boldsymbol{u}^\top \boldsymbol{H}_t \boldsymbol{u} - \frac{1}{3}\sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u}\rangle^2}{\sigma_s^2} \\
&\ge \frac{2}{3}\boldsymbol{u}^\top \boldsymbol{H}_t \boldsymbol{u} - \frac{1}{3}\left(\lambda\|\boldsymbol{u}\|^2 + \sum_{s=1}^t \frac{\langle \boldsymbol{\phi}_s, \boldsymbol{u}\rangle^2}{\sigma_s^2}\right) \\
&= \frac{1}{3}\boldsymbol{u}^\top \boldsymbol{H}_t \boldsymbol{u},
\end{aligned}
$$

which implies $\nabla^2 L_t(\boldsymbol{\theta}) \succeq \frac{1}{3}\boldsymbol{H}_t$. □

C.5 Proof of Lemma C.2

Proof of Lemma C.2. The expression of $\nabla L_t(\boldsymbol{\theta})$ in (C.1) gives

$$
\|\nabla L_t(\boldsymbol{\theta}^*)\|_{\boldsymbol{H}_t^{-1}} \le \lambda\|\boldsymbol{\theta}\|_{\boldsymbol{H}_t^{-1}} + \Bigg\|\underbrace{\sum_{s=1}^t \Psi_{\tau_s}(z_s^*)\frac{\boldsymbol{\phi}_s}{\sigma_s}}_{\boldsymbol{d}_t}\Bigg\|_{\boldsymbol{H}_t^{-1}}.
$$

We next construct upper bounds of $\lambda\|\boldsymbol{\theta}\|_{\boldsymbol{H}_t^{-1}}$ and $\|\boldsymbol{d}_t\|_{\boldsymbol{H}_t^{-1}}$, respectively.

Bound $\lambda\|\boldsymbol{\theta}\|_{\boldsymbol{H}_t^{-1}}$. Since $\boldsymbol{H}_t \succeq \lambda \boldsymbol{I}$, we have $\boldsymbol{H}_t^{-1} \preceq \frac{1}{\lambda}\boldsymbol{I}$. Thus, $\lambda\|\boldsymbol{\theta}\|_{\boldsymbol{H}_t^{-1}} \le \lambda\frac{1}{\sqrt{\lambda}}\|\boldsymbol{\theta}\| \le \sqrt{\lambda}B$.

Bound $\|\boldsymbol{d}_t\|_{\boldsymbol{H}_t^{-1}}$. We aim to decompose $\|\boldsymbol{d}_t\|^2_{\boldsymbol{H}_t^{-1}}$ into two terms and bound them separately. The fact that $\boldsymbol{H}_t = \boldsymbol{H}_{t-1} + \frac{\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top}{\sigma_t^2}$ together with the Sherman-Morrison formula implies that

$$
\boldsymbol{H}_t^{-1} = \boldsymbol{H}_{t-1}^{-1} - \frac{\boldsymbol{H}_{t-1}^{-1}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\boldsymbol{H}_{t-1}^{-1}}{\sigma_t^2(1+w_t^2)}, \quad \text{where } w_t^2 := \frac{\boldsymbol{\phi}_t^\top\boldsymbol{H}_{t-1}^{-1}\boldsymbol{\phi}_t}{\sigma_t^2} = \left\|\frac{\boldsymbol{\phi}_t}{\sigma_t}\right\|^2_{\boldsymbol{H}_{t-1}}. \tag{C.14}
$$

Clearly, w_t is $\mathcal{F}_{t-1}$-measurable and thus is predictable. By definition of $\boldsymbol{d}_t$ and (C.14),

$$
\begin{aligned}
&\|\boldsymbol{d}_t\|^2_{\boldsymbol{H}_t^{-1}} \qquad\qquad (\text{C.15}) \\
&= \left(\boldsymbol{d}_{t-1} + \Psi_{\tau_t}(z_t^*)\frac{\boldsymbol{\phi}_t}{\sigma_t}\right)^\top \boldsymbol{H}_t^{-1}\left(\boldsymbol{d}_{t-1} + \Psi_{\tau_t}(z_t^*)\frac{\boldsymbol{\phi}_t}{\sigma_t}\right) \\
&= \|\boldsymbol{d}_{t-1}\|^2_{\boldsymbol{H}_{t-1}^{-1}} - \frac{1}{1+w_t^2}\left(\frac{\boldsymbol{d}_{t-1}^\top\boldsymbol{H}_{t-1}^{-1}\boldsymbol{\phi}_t}{\sigma_t}\right)^2 + 2\Psi_{\tau_t}(z_t^*)\frac{\boldsymbol{d}_{t-1}^\top\boldsymbol{H}_t^{-1}\boldsymbol{\phi}_t}{\sigma_t} + \Psi^2_{\tau_t}(z_t^*)\frac{\boldsymbol{\phi}_t^\top\boldsymbol{H}_t^{-1}\boldsymbol{\phi}_t}{\sigma_t^2} \\
&\le \|\boldsymbol{d}_{t-1}\|^2_{\boldsymbol{H}_{t-1}^{-1}} + \underbrace{2\Psi_{\tau_t}(z_t^*)\frac{\boldsymbol{d}_{t-1}^\top\boldsymbol{H}_t^{-1}\boldsymbol{\phi}_t}{\sigma_t}}_{I_1} + \underbrace{\Psi^2_{\tau_t}(z_t^*)\frac{\boldsymbol{\phi}_t^\top\boldsymbol{H}_t^{-1}\boldsymbol{\phi}_t}{\sigma_t^2}}_{I_2}. \qquad\qquad (\text{C.16})
\end{aligned}
$$

For I_1, by (C.14), we have

$$
\begin{aligned}
I_1 &= 2\Psi_{\tau_t}(z_t^*)\frac{1}{\sigma_t}\boldsymbol{d}_{t-1}^\top\left(\boldsymbol{H}_{t-1}^{-1} - \frac{\boldsymbol{H}_{t-1}^{-1}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\boldsymbol{H}_{t-1}^{-1}}{\sigma_t^2(1+w_t^2)}\right)\boldsymbol{\phi}_t \\
&= 2\Psi_{\tau_t}(z_t^*)\frac{1}{1+w_t^2}\frac{\boldsymbol{d}_{t-1}^\top\boldsymbol{H}_{t-1}^{-1}\boldsymbol{\phi}_t}{\sigma_t}.
\end{aligned}
$$

For I_2, we have

$$\begin{aligned}
I_2 &= \Psi_{\tau_t}^2(z_t^*)\frac{\boldsymbol{\phi}_t^\top \boldsymbol{H}_t^{-1}\boldsymbol{\phi}_t}{\sigma_t^2} \\
&= \Psi_{\tau_t}^2(z_t^*)\frac{1}{\sigma_t^2}\boldsymbol{\phi}_t^\top\left(\boldsymbol{H}_{t-1}^{-1} - \frac{\boldsymbol{H}_{t-1}^{-1}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\boldsymbol{H}_{t-1}^{-1}}{\sigma_t^2(1+w_t^2)}\right)\boldsymbol{\phi}_t \\
&= \Psi_{\tau_t}^2(z_t^*)\left(w_t^2 - \frac{w_t^4}{1+w_t^2}\right) \\
&= \Psi_{\tau_t}^2(z_t^*)\frac{w_t^2}{1+w_t^2}.
\end{aligned}$$

Using the equations for I_1, I_2 and iterating (C.15), we have

$$\|\boldsymbol{d}_t\|_{\boldsymbol{H}_t^{-1}}^2 \le \underbrace{\sum_{s=1}^t \Psi_{\tau_s}(z_s^*)\frac{2}{1+w_s^2}\frac{\boldsymbol{d}_{s-1}^\top\boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s}{\sigma_s}}_{\text{(i)}} + \underbrace{\sum_{s=1}^t \Psi_{\tau_s}^2(z_s^*)\frac{w_s^2}{1+w_s^2}}_{\text{(ii)}}.$$

Next we use Lemma C.4 and Lemma C.5 to bound (i) and (ii), respectively. We remark that (i) is the leading term. In the proof of Lemma C.4, inspired by [9, 32], we make use of (3) of Property 1 to carefully quantify the moment generating function, and thus achieves a tight bound.

Lemma C.4. Assume $\mathbb{E}[z_t(\boldsymbol{\theta}^*)|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[|z_t(\boldsymbol{\theta}^*)|^{1+\epsilon}|\mathcal{F}_{t-1}] \le b^{1+\epsilon}$, let A_t define the event where $\|\boldsymbol{d}_s\|_{\boldsymbol{H}_s^{-1}} \le \alpha_s$ for $s \le t$. Then with probability at least $1-\delta, \forall t \ge 1$, we have

$$\sum_{s=1}^t \Psi_{\tau_s}(z_s^*)\frac{2\boldsymbol{d}_{s-1}^\top\boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s\,\mathbb{1}_{A_{s-1}}}{(1+w_s^2)\sigma_s} \le (\max_{s\le t}\alpha_s)2t^{\frac{1-\epsilon}{2(1+\epsilon)}}\left[\frac{(\sqrt{2}\kappa b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^\epsilon} + \tau_0\log\frac{2t^2}{\delta}\right].$$

Proof. See Appendix C.6 for a detailed proof. □

Lemma C.5. Assume $\mathbb{E}[z_t(\boldsymbol{\theta}^*)|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[|z_t(\boldsymbol{\theta}^*)|^{1+\epsilon}|\mathcal{F}_{t-1}] \le b^{1+\epsilon}$. Then with probability at least $1-\delta, \forall t \ge 1$, we have

$$\sum_{s=1}^t \Psi_{\tau_s}^2(z_s^*)\frac{w_s^2}{1+w_s^2} \le \left[t^{\frac{1-\epsilon}{2(1+\epsilon)}}\left(\sqrt{\tau_0^{1-\epsilon}(\sqrt{2}\kappa b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}} + \tau_0\sqrt{2\log\frac{2t^2}{\delta}}\right)\right]^2.$$

Proof. See Appendix C.7 for a detailed proof. □

Recall $\tau_0 \ge \frac{\sqrt{2}\kappa b(\log 3T)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\left(\log\frac{2T^2}{\delta}\right)^{\frac{1}{1+\epsilon}}}$. Thus $\sqrt{\tau_0^{1-\epsilon}(\sqrt{2}\kappa b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}} \le \tau_0\sqrt{\log\frac{2t^2}{\delta}}$. We choose

$$\alpha_t = 4t^{\frac{1-\epsilon}{2(1+\epsilon)}}\left[\frac{(\sqrt{2}\kappa b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^\epsilon} + \tau_0\log\frac{2t^2}{\delta}\right].$$

Then on the event where Lemma C.4 and C.5 hold, whose probability is at least $1-2\delta, \forall t \ge 1$, we have

$$\sum_{s=1}^t \Psi_{\tau_s}(z_s^*)\frac{2}{1+w_s^2}\frac{\boldsymbol{d}_{s-1}^\top\boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s\,\mathbb{1}_{A_{s-1}}}{\sigma_s} \le \frac{\alpha_t^2}{2}, \quad \sum_{s=1}^t \Psi_{\tau_s}^2(z_s^*)\frac{w_s^2}{1+w_s^2} \le \frac{\alpha_t^2}{2}.$$

Finally, we can conclude that all $\{A_t\}_{t\ge 0}$ is true and thus $\|\boldsymbol{d}_t\|_{\boldsymbol{H}_t^{-1}} \le \alpha_t$. □

C.6 Proof of Lemma C.4

Proof of Lemma C.4. It follows that

$$
\begin{aligned}
&\Psi_{\tau_s}(z_s^*)\frac{2}{1+w_s^2}\frac{\boldsymbol{d}_{s-1}^\top \boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s \mathbb{1}_{A_{s-1}}}{\sigma_s} \\
&= \tau_0 \Psi_1\left(\frac{z_s^*}{\tau_s}\right)\frac{2}{w_s\sqrt{1+w_s^2}} s^{\frac{1-\epsilon}{2(1+\epsilon)}} \frac{\boldsymbol{d}_{s-1}^\top \boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s \mathbb{1}_{A_{s-1}}}{\sigma_s} \\
&= (\max_{s\le t}\alpha_s) 2\tau_0 t^{\frac{1-\epsilon}{2(1+\epsilon)}} \Psi_1\left(\frac{z_s^*}{\tau_s}\right) \underbrace{\frac{1}{\sqrt{1+w_s^2}}\left(\frac{s}{t}\right)^{\frac{1-\epsilon}{2(1+\epsilon)}} \frac{\boldsymbol{d}_{s-1}^\top \boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s \mathbb{1}_{A_{s-1}}}{(\max_{s\le t}\alpha_s) w_s \sigma_s}}_{\text{denoted as } M_s},
\end{aligned}
$$

where the first equality is by (2) of Property 1 and the definition of τ_s. Note that M_s is $\mathcal{F}_{s-1}$ measurable and

$$
\begin{aligned}
|M_s| &\le \frac{|\boldsymbol{d}_{s-1}^\top \boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s| \mathbb{1}_{A_{s-1}}}{(\max_{s\le t}\alpha_s) w_s \sigma_s} \\
&\le \frac{\|\boldsymbol{d}_{s-1}\|_{\boldsymbol{H}_{s-1}^{-1}} \|\boldsymbol{\phi}_s\|_{\boldsymbol{H}_{s-1}^{-1}} \mathbb{1}_{A_{s-1}}}{(\max_{s\le t}\alpha_s) w_s \sigma_s} \\
&\le 1,
\end{aligned}
$$

where the first inequality holds due to $\frac{1}{\sqrt{1+w_s^2}}\left(\frac{s}{t}\right)^{\frac{1-\epsilon}{2(1+\epsilon)}} \le 1$, the second holds due to Cauchy-Schwartz inequality and the last holds due to the definition of $\mathbb{1}_{A_{s-1}}$.

Next we make use of (3) of Property 1 to carefully quantify the moment generating function of $\sum_{s=1}^t M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right)$ and leverage a Chernoff bounding technique to complete the proof.

It follows from (3) of Property 1 that

$$
\begin{aligned}
M_s\Psi_1\left(\frac{z_s^*}{\tau_s}\right) \le\ & M_s \mathbb{1}\{M_s \ge 0\} \log\left(1+\frac{z_s^*}{\tau_s}+\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon}\right) \\
& - M_s \mathbb{1}\{M_s < 0\} \log\left(1-\frac{z_s^*}{\tau_s}+\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon}\right).
\end{aligned}
$$

Thus

$$
\begin{aligned}
\exp\left\{M_s\Psi_1\left(\frac{z_s^*}{\tau_s}\right)\right\} &\le \left(1+\frac{z_s^*}{\tau_s}+\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon}\right)^{M_s \mathbb{1}\{M_s\ge 0\}} \left(1-\frac{z_s^*}{\tau_s}+\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon}\right)^{-M_s \mathbb{1}\{M_s<0\}} \\
&\overset{(a)}{\le} 1 + M_s\frac{z_s^*}{\tau_s} + |M_s|\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon} \\
&\overset{(b)}{\le} 1 + M_s\frac{z_s^*}{\tau_s} + \left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon},
\end{aligned}
$$

where (a) holds due to $(1+u)^v \le 1+uv, \forall u \ge -1, v\in(0,1]$ and (b) holds due to $|M_s|\le 1$. Then it follows that

$$
\begin{aligned}
\mathbb{E}\left[\exp\left\{M_s\Psi_1\left(\frac{z_s^*}{\tau_s}\right)\right\}\middle|\mathcal{F}_{s-1}\right] &\le \mathbb{E}\left[1+M_s\frac{z_s^*}{\tau_s}+\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon}\middle|\mathcal{F}_{s-1}\right] \\
&= 1 + \frac{M_s}{\tau_s}\mathbb{E}[z_s^*|\mathcal{F}_{s-1}] + \mathbb{E}\left[\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon}\middle|\mathcal{F}_{s-1}\right] \\
&\le 1 + \frac{b^{1+\epsilon}}{\tau_s^{1+\epsilon}},
\end{aligned}
$$

where the first equality holds due to $\frac{M_s}{\tau_s} \in \mathcal{F}_{s-1}$. Hence, we have

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left\{\sum_{s=1}^{t} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right)\right\}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\exp\left\{\sum_{s=1}^{t-1} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right)\right\} \mathbb{E}\left[\exp\left\{M_t \Psi_1\left(\frac{z_t^*}{\tau_t}\right)\right\} \middle| \mathcal{F}_{t-1}\right]\right] \\
&\le \left(1+\frac{b^{1+\epsilon}}{\tau_t^{1+\epsilon}}\right) \mathbb{E}\left[\exp\left\{\sum_{s=1}^{t-1} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right)\right\}\right] \le \cdots \le \prod_{s=1}^{t}\left(1+\frac{b^{1+\epsilon}}{\tau_s^{1+\epsilon}}\right) \\
&\overset{(b)}{\le} \exp\left\{b^{1+\epsilon}\sum_{s=1}^{t}\frac{1}{\tau_s^{1+\epsilon}}\right\} = \exp\left\{\left(\frac{b}{\tau_0}\right)^{1+\epsilon}\sum_{s=1}^{t}\left(\frac{w_s}{\sqrt{1+w_s^2}}\right)^{1+\epsilon} s^{-\frac{1-\epsilon}{2}}\right\} \\
&\overset{(c)}{\le} \exp\left\{\left(\frac{b}{\tau_0}\right)^{1+\epsilon}(2\kappa)^{\frac{1+\epsilon}{2}}(\log 3t)^{\frac{1-\epsilon}{2}}\right\} = \exp\left\{\frac{(\sqrt{2\kappa}b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^{1+\epsilon}}\right\},
\end{aligned}
$$

where (a) holds due to tower property of conditional expectation and $\sum_{s=1}^{t-1} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right) \in \mathcal{F}_{t-1}$, (b) holds due to $1+x \le \exp\{x\}$ and (c) holds with the same argument as (C.11).

Then, for any $t \ge 1$, a high probability upper bound of $\sum_{s=1}^{t} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right)$ can be constructed by

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{s=1}^{t} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right) \ge \frac{(\sqrt{2\kappa}b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^{1+\epsilon}} + \log\frac{2t^2}{\delta}\right) \\
&\overset{(a)}{=} \mathbb{P}\left(\exp\left\{\sum_{s=1}^{t} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right)\right\} \ge \exp\left\{\frac{(\sqrt{2\kappa}b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^{1+\epsilon}}\right\}\cdot\frac{2t^2}{\delta}\right) \\
&\overset{(b)}{\le} \frac{\mathbb{E}\left[\exp\left\{\sum_{s=1}^{t} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right)\right\}\right]}{\exp\left\{\frac{(\sqrt{2\kappa}b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^{1+\epsilon}}\right\}}\frac{\delta}{2t^2} \le \frac{\delta}{2t^2},
\end{aligned}
$$

where (a) holds due to the non-decreasing property of $\exp\{x\}$ and (b) holds due to Markov's inequality.

Finally, with a union bound over $t \ge 1$ and the fact that $\sum_{t=1}^{\infty}\frac{1}{2t^2} \le 1$, with probability at least $1-\delta, \forall t \ge 1$, we have

$$
\begin{aligned}
&\sum_{s=1}^{t} \Psi_{\tau_s}(z_s^*)\frac{2}{1+w_s^2}\frac{\boldsymbol{d}_{s-1}^\top \boldsymbol{H}_{s-1}^{-1}\boldsymbol{\phi}_s \mathbb{1}_{A_{s-1}}}{\sigma_s} \\
&= (\max_{s\le t}\alpha_s) 2\tau_0 t^{\frac{1-\epsilon}{2(1+\epsilon)}} \sum_{s=1}^{t} M_s \Psi_1\left(\frac{z_s^*}{\tau_s}\right) \\
&\le (\max_{s\le t}\alpha_s) 2\tau_0 t^{\frac{1-\epsilon}{2(1+\epsilon)}} \left[\frac{(\sqrt{2\kappa}b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^{1+\epsilon}} + \log\frac{2t^2}{\delta}\right] \\
&= (\max_{s\le t}\alpha_s) 2 t^{\frac{1-\epsilon}{2(1+\epsilon)}} \left[\frac{(\sqrt{2\kappa}b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}}{\tau_0^{\epsilon}} + \tau_0\log\frac{2t^2}{\delta}\right].
\end{aligned}
$$

□

C.7 Proof of Lemma C.5

Proof of Lemma C.5. It follows that

$$
\sum_{s=1}^{t} \Psi_{\tau_s}^2(z_s^*)\frac{w_s^2}{1+w_s^2} = \sum_{s=1}^{t} \tau_s^2 \Psi_1^2\left(\frac{z_s^*}{\tau_s}\right)\frac{w_s^2}{1+w_s^2} \le \tau_0^2 t^{\frac{1-\epsilon}{1+\epsilon}} \sum_{s=1}^{t} \Psi_1^2\left(\frac{z_s^*}{\tau_s}\right).
$$

Define $X_s := \Psi_1^2\left(\frac{z_s^*}{\tau_s}\right)$ and $Y_s = X_s - \mathbb{E}[X_s|\mathcal{F}_{s-1}]$, which are $\mathcal{F}_s$-measurable. Next we make use of Lemma I.1, with probability at least $1 - 2t^2/\delta$, for any $t \geq 1$, we have

$$\sum_{s=1}^{t} Y_s \leq \sqrt{2V \log \frac{2t^2}{\delta}} + \frac{2M}{3} \log \frac{2t^2}{\delta} \tag{C.17}$$

where $\sum_{s=1}^{t} \mathbb{E}[Y_s^2|\mathcal{F}_{s-1}] \leq V$ and $|Y_t| \leq M$. It follows that $|Y_s| \leq |X_s| + |\mathbb{E}[X_s|\mathcal{F}_{s-1}]| \leq 2$ by (1) of Property 1 and

$$\begin{aligned}
\sum_{s=1}^{t} \mathbb{E}[X_s|\mathcal{F}_{s-1}] &\overset{(a)}{\leq} \sum_{s=1}^{t} \mathbb{E}\left[\left|\frac{z_s^*}{\tau_s}\right|^{1+\epsilon} \Big|\mathcal{F}_{s-1}\right] \\
&\leq \sum_{s=1}^{t} \left(\frac{b}{\tau_s}\right)^{1+\epsilon} = \left(\frac{b}{\tau_0}\right)^{1+\epsilon} \sum_{s=1}^{t} \left(\frac{w_s}{\sqrt{1+w_s^2}}\right)^{1+\epsilon} s^{-\frac{1-\epsilon}{2}} \\
&\overset{(b)}{\leq} \left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}},
\end{aligned} \tag{C.18}$$

where (a) holds due to (1) of Property 1 and (b) holds with the same argument as (C.11).

Through conducting a similar argument as (C.18), we have

$$\sum_{s=1}^{t} \mathbb{E}[Y_s^2|\mathcal{F}_{s-1}] \leq \sum_{s=1}^{t} \mathbb{E}[X_s^2|\mathcal{F}_{s-1}] \leq \left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}}.$$

Thus, setting $V = \left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}}$ and $M = 2$ in (C.17), we have

$$\begin{aligned}
\sum_{s=1}^{t} X_s &\leq \sum_{s=1}^{t} \mathbb{E}[X_s|\mathcal{F}_{s-1}] + \sqrt{2V \log \frac{2t^2}{\delta}} + \frac{2M}{3} \log \frac{2t^2}{\delta} \\
&\leq \left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}} + \sqrt{2\left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}} \log \frac{2t^2}{\delta}} + \frac{4}{3} \log \frac{2t^2}{\delta}.
\end{aligned}$$

Finally, with a union bound over $t \geq 1$ and the fact that $\sum_{t=1}^{\infty} \frac{1}{2t^2} \leq 1$, with probability at least $1 - \delta, \forall t \geq 1$, we have

$$\begin{aligned}
&\sum_{s=1}^{t} \Psi_{\tau_s}^2(z_s^*) \frac{w_s^2}{1+w_s^2} \\
&\leq \tau_0^2 t^{\frac{1-\epsilon}{1+\epsilon}} \left[\left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}} + \sqrt{2\left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}} \log \frac{2t^2}{\delta}} + \frac{4}{3} \log \frac{2t^2}{\delta}\right] \\
&\leq \tau_0^2 t^{\frac{1-\epsilon}{1+\epsilon}} \left[\sqrt{\left(\frac{\sqrt{2\kappa}b}{\tau_0}\right)^{1+\epsilon} (\log 3t)^{\frac{1-\epsilon}{2}}} + \sqrt{2 \log \frac{2t^2}{\delta}}\right]^2 \\
&= \left[t^{\frac{1-\epsilon}{2(1+\epsilon)}} \left(\sqrt{\tau_0^{1-\epsilon}(\sqrt{2\kappa}b)^{1+\epsilon}(\log 3t)^{\frac{1-\epsilon}{2}}} + \tau_0\sqrt{2 \log \frac{2t^2}{\delta}}\right)\right]^2.
\end{aligned}$$

□

D Proofs for Section 4

D.1 Proof of Theorem 4.1

Proof of Theorem 4.1. Note that with probability at least $1-3\delta$, it holds that

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{H}_t} \le \beta_t \tag{D.1}$$

by Theorem 3.3 with $b=1$. Then we have

$$\begin{aligned}
\mathrm{Regret}(T) &= \sum_{t=1}^{T} [\langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta}^* \rangle - \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^* \rangle] \\
&\le \sum_{t=1}^{T} \left[\max_{\boldsymbol{\theta} \in \mathcal{C}_{t-1}} \langle \boldsymbol{\phi}_t^*, \boldsymbol{\theta} \rangle - \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^* \rangle \right] \\
&\overset{(a)}{\le} \sum_{t=1}^{T} \left[\max_{\boldsymbol{\theta} \in \mathcal{C}_{t-1}} \langle \boldsymbol{\phi}_t, \boldsymbol{\theta} \rangle - \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}^* \rangle \right] \\
&\overset{(b)}{\le} \sum_{t=1}^{T} \|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}} \max_{\boldsymbol{\theta} \in \mathcal{C}_{t-1}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{H}_{t-1}} \\
&\overset{(c)}{\le} \sum_{t=1}^{T} 2\beta_t \|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}} \\
&\overset{(d)}{\le} 2\beta_T \sum_{t=1}^{T} \sigma_t w_t,
\end{aligned}$$

where (a) holds due to the optimism of action $\boldsymbol{\phi}_t$, (b) holds due to Cauchy-Schwartz inequality, (c) holds due to (D.1) and (d) holds due to the fact that β_t is increasing with t.

Notice

$$\sum_{t=1}^{T} w_t^2 = \min\left\{1, w_t^2\right\} \le 2\kappa \tag{D.2}$$

by $w_s \le c_0 \le 1$ and Lemma I.6. We will use (D.2) several times in the following proof.

Next we bound the sum of bonus $\sum_{t=1}^{T} \sigma_t w_t$ separately by the value of σ_t. Recall the definition of σ_t in Algorithm 2, we decompose $[T]$ as the union of three disjoint sets $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3$ where

$$\begin{aligned}
\mathcal{J}_1 &= \{t \in [T] | \sigma_t \in \{\nu_t, \sigma_{\min}\}\}, \\
\mathcal{J}_2 &= \left\{ t \in [T] \middle| \sigma_t = \frac{\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}}{c_0} \right\}, \\
\mathcal{J}_3 &= \left\{ t \in [T] \middle| \sigma_t = \frac{\sqrt{LB}}{c_1^{\frac{1}{4}} (2\kappa)^{\frac{1}{4}}} \|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}^{\frac{1}{2}} \right\}.
\end{aligned}$$

For the summation over $\mathcal{J}_1$, we have

$$\sum_{t \in \mathcal{J}_1} \sigma_t w_t = \sum_{t \in \mathcal{J}_1} \max\{\nu_t, \sigma_{\min}\} w_t \overset{(a)}{\le} \sqrt{\sum_{t=1}^{T} (\nu_t^2 + \sigma_{\min}^2)} \sqrt{\sum_{t=1}^{T} w_t^2} \overset{(b)}{\le} \sqrt{2\kappa} \sqrt{\sum_{t=1}^{T} (\nu_t^2 + \sigma_{\min}^2)},$$

where (a) holds due to Cauchy-Schwartz inequality and (b) holds due to (D.2).

And for the summation over $\mathcal{J}_2$, we have $w_t = \left\| \frac{\boldsymbol{\phi}_t}{\sigma_t} \right\|_{\boldsymbol{H}_{t-1}^{-1}} = c_0$. Then

$$\sum_{t \in \mathcal{J}_2} \sigma_t w_t = \frac{1}{c_0^2} \sum_{t \in \mathcal{J}_2} \|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}} w_t^2 \overset{(a)}{\le} \frac{L}{c_0^2 \sqrt{\lambda}} \sum_{t=1}^{T} w_t^2 \overset{(b)}{\le} \frac{2L\kappa}{c_0^2 \sqrt{\lambda}},$$

where (a) holds due to $\boldsymbol{H}_{t-1} \succeq \lambda \boldsymbol{I}$, thus $\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}} \le \frac{\|\boldsymbol{\phi}_t\|}{\sqrt{\lambda}} \le \frac{L}{\sqrt{\lambda}}$. And (b) holds due to (D.2).

Then, for the summation over $\mathcal{J}_3$, we have $\sigma_t = \frac{LB}{\sqrt{2c_1\kappa}} w_t$. Therefore

$$\sum_{t\in\mathcal{J}_3} \sigma_t w_t = \frac{LB}{\sqrt{2c_1\kappa}} \sum_{t\in\mathcal{J}_3} w_t^2 \le \frac{LB\sqrt{2\kappa}}{\sqrt{c_1}},$$

where the inequality holds due to (D.2).

Finally, putting pieces together finishes the proof. □

E Proof of Theorem 5.1

Proof of Theorem 5.1. Recall our proposed algorithm Heavy-LSVI-UCB is detailed in Algorithm 3. First, to compute $\boldsymbol{\theta}_{k-1,h}$ in line 6, we notice the loss function in (5.1) is λ_R-strongly convex and $(\lambda_R + K/\nu_{\min}^2)$-smooth, so there are plenty of convex optimization algorithms available. For example, Nesterov accelerated method can be used. According to Bubeck et al. [8], the number of iteration of Nesterov's method is $O(\sqrt{\beta/\alpha}\log(R^2/\epsilon))$ with one derivation ($O(d)$ operations) per iteration. Here the loss function is supposed to be α-strongly convex and β-smooth. R is the maximum distance of two points and ϵ is the precision. Thus the total computational cost is $\widetilde{O}(HK\mathcal{R})$ with $\mathcal{R} = \widetilde{O}\left(d\sqrt{1 + K/(\lambda_R \nu_{\min}^2)}\right) = \widetilde{O}\left(d + d^{-\frac{1-\epsilon}{2(1+\epsilon)}} H^{\frac{1-\epsilon}{2(1+\epsilon)}} K^{\frac{1+2\epsilon}{2(1+\epsilon)}}\right)$.

Second, to evaluate the updated action-value function $Q_h^k(s,a)$ in line 10 for a given pair (s,a), we take the minimum over at most $\widetilde{O}(dH)$ action-value functions (See Lemma I.8) with $O(d^2)$ operations (Using Sherman-Morrison formula to compute $H_{k-1,h}^{-1}$ and $\Sigma_{k-1,h}^{-1}$) for each function. Thus it takes $\widetilde{O}(d^3H)$ to evaluate the updated action-value function. As a result, to compute $\widehat{w}_{k-1,h}$ in line 6, notice $\widehat{w}_{k,h} = \Sigma_{k,h}^{-1}\sum_{i=1}^k \sigma_{i,h}^{-2}\phi_{i,h}V_h^k(s_{i,h+1})$, if V_h^k remains unchanged, we only need to compute the new term $\sigma_{k,h}^{-2}\phi_{k,h}V_h^k(s_{k,h+1})$, which takes $\widetilde{O}(d^3|\mathcal{A}|H)$ computational time. Else if V_h^k is updated, we need to recalculate $\{V_h^k(s_{i,h+1})\}_{i\in[k]}$, which takes $\widetilde{O}(d^3|\mathcal{A}|HK)$ computational time. Note the number of updating episode is at most $\widetilde{O}(dH)$ and the length of each episode is H, so the total computational cost is $\widetilde{O}(d^4|\mathcal{A}|H^3K)$.

Last, to take action $a_{k,h}$ in line 19, we need to compute $\{Q_h^k(s_{k,h},a)\}_{a\in\mathcal{A}}$ and take the maximum, which takes $\widetilde{O}(d^3|\mathcal{A}|H)$ time, incurring a total cost of $\widetilde{O}(d^3|\mathcal{A}|H^2K)$. Finally, combining the total costs above gives the computational complexity of Heavy-LSVI-UCB. □

F Proof of Theorem 5.2

In this section, we give the proof sketch of Theorem 5.2. In Appendix F.1, leveraging the technique in He et al. [15], we define several events and prove them hold with high probability. In Appendix F.2, we then decompose the regret into a lower order term and summations of bonus terms with respect to reward functions and transition probabilities. Finally, we adopt a novel approach that deal with the two bonus terms separately.

F.1 High-Probability Events

Parameters for Adaptive Huber Regression in Algorithm 3. First, we set the parameters for adaptive Huber regression in Algorithm 3 as follows:

$$c_0 = \frac{1}{\sqrt{23\log\frac{2HK^2}{\delta}}}, \quad c_1 = \min\left\{\frac{(\log 3K)^{\frac{1-\epsilon}{1+\epsilon}}}{48\left(\log\frac{2HK^2}{\delta}\right)^{\frac{2}{1+\epsilon}}}, \frac{(\log 3K)^{\frac{1-\epsilon'}{1+\epsilon'}}}{48\left(\log\frac{2HK^2}{\delta}\right)^{\frac{2}{1+\epsilon'}}}\right\},$$

$$\tau_0 = \frac{\sqrt{2\kappa}(\log 3K)^{\frac{1-\epsilon}{2(1+\epsilon)}}}{\left(\log\frac{2HK^2}{\delta}\right)^{\frac{1}{1+\epsilon}}}, \quad \widetilde{\tau}_0 = \frac{\sqrt{2\kappa}\nu_{R^\epsilon}(\log 3K)^{\frac{1-\epsilon'}{2(1+\epsilon')}}}{\nu_{\min}\left(\log\frac{2HK^2}{\delta}\right)^{\frac{1}{1+\epsilon'}}}.$$

Measurability. We define filtration $\{\mathcal{G}_{k,h}\}_{(k,h)\in[K]\times[H]}$ and $\{\mathcal{F}_{k,h}\}_{(k,h)\in[K]\times[H]}$ as follows. Let $I_{k,h} := \{(i,j) : i \in [k-1], j \in [H] \text{ or } i = k, j \in [h]\}$ denote the set of index pairs up to and including the k-th episode and the h-th step. We further define $\mathcal{G}_{k,h} = \sigma\left(\bigcup_{(i,j)\in I_{k,h-1}} \{s_{i,j}, a_{i,j}, r_{i,j}\} \cup \{s_{k,h}, a_{k,h}\}\right)$ and $\mathcal{F}_{k,h} = \sigma\left(\bigcup_{(i,j)\in I_{k,h}} \{s_{i,j}, a_{i,j}, r_{i,j}\}\right)$. We make a convention that $\mathcal{G}_{k,0} = \mathcal{G}_{k-1,H+1}$ and $\mathcal{F}_{k,0} = \mathcal{F}_{k-1,H}$. Note $\mathcal{G}_{k,h} \subset \mathcal{F}_{k,h} \subset \mathcal{G}_{k,h+1}$.

We first introduce the following high-probability events:

1. We define $\mathcal{E}_{R^\epsilon}$ as the event that the following inequalities hold for all $k, h \in [K] \times [H]$,
$$\|\boldsymbol{\psi}_{k,h} - \boldsymbol{\psi}_h^*\|_{\boldsymbol{H}_{k,h}^{-1}} \le \beta_{R^\epsilon,k},$$
where $\boldsymbol{\psi}_{k,h}$ is defined in (5.2)
$$\beta_{R^\epsilon,k} = 3\sqrt{\lambda_R}W + 24k^{\frac{1-\epsilon'}{2(1+\epsilon')}} \frac{\sqrt{2\kappa}\nu_{R^\epsilon}}{\nu_{\min}} (\log 3K)^{\frac{1-\epsilon'}{2(1+\epsilon')}} \left(\log \frac{2HK^2}{\delta}\right)^{\frac{\epsilon'}{1+\epsilon'}}. \tag{F.1}$$
For simplicity, we further define $\beta_{R^\epsilon} := \beta_{R^\epsilon,K}$.
2. We define $\mathcal{E}_0$ as the event that the following inequalities hold for all $k, h \in [K] \times [H]$,
$$\begin{aligned} \left\|\widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} &\le \beta_0, \\ \left\|\check{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[\check{V}_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} &\le \beta_0, \\ \left\|\widetilde{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[\left(V_{h+1}^k\right)^2\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} &\le \mathcal{H}\beta_0, \end{aligned}$$
where $\widehat{\boldsymbol{w}}_{k-1,h}, \check{\boldsymbol{w}}_{k-1,h}, \widetilde{\boldsymbol{w}}_{k-1,h}$ are defined in (5.8) and
$$\beta_0 = 2\sqrt{d\lambda_V}\mathcal{H} + \frac{3\mathcal{H}}{\sigma_{\min}}\sqrt{d^3H\iota_0^2 + \log\frac{H}{\delta}}, \tag{F.2}$$
and
$$\begin{aligned} \iota_0 = \max\Bigg\{ &\log_2\left(1 + \frac{K}{\lambda_R\nu_{\min}^2}\right), \log_2\left(1 + \frac{K}{\lambda_V\sigma_{\min}^2}\right), \log\left(1 + \frac{8(B+L)K}{\lambda_V\mathcal{H}\sqrt{d}\sigma_{\min}^2}\right), \\ &\log\left(1 + \frac{32\beta_R^2K^2}{\sqrt{d}\lambda_R\lambda_V^2\mathcal{H}^2\sigma_{\min}^4}\right), \log\left(1 + \frac{32\beta_V^2K^2}{\sqrt{d}\lambda_V^3\mathcal{H}^2\sigma_{\min}^4}\right)\Bigg\}, L = \mathcal{H}\sqrt{\frac{dK}{\lambda_V}}. \end{aligned}$$
3. We define $\mathcal{E}_{R,k}$ as the event that the following inequalities hold for all $i, h \in [k] \times [H]$,
$$\|\boldsymbol{\theta}_{i,h} - \boldsymbol{\theta}_h^*\|_{\boldsymbol{H}_{i,h}^{-1}} \le \beta_{R,i},$$
where $\boldsymbol{\theta}_{i,h}$ is defined in (5.1) and
$$\beta_{R,i} = O\left(\sqrt{\lambda_R}B + \sqrt{d}i^{\frac{1-\epsilon}{2(1+\epsilon)}}\iota\right), \tag{F.3}$$
and
$$\iota = \max\left\{\log\left(1 + \frac{K}{d\lambda_R\nu_{\min}^2}\right), \log 3K, \log\frac{2HK^2}{\delta}\right\}.$$
For simplicity, we further define $\beta_R := \beta_{R,K} = O\left(\sqrt{\lambda_R}B + \sqrt{d}K^{\frac{1-\epsilon}{2(1+\epsilon)}}\iota\right)$ and $\mathcal{E}_R := \mathcal{E}_{R,K}$.
4. We define $\mathcal{E}_h$ as the event that the following inequalities hold for all $k \in [K], h \le h' \le H$,
$$\begin{aligned} \left\|\widehat{\boldsymbol{w}}_{k-1,h'} - \boldsymbol{w}_{h'}\left[V_{h'+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h'}} &\le \beta_V, \\ \left\|\check{\boldsymbol{w}}_{k-1,h'} - \boldsymbol{w}_{h'}\left[\check{V}_{h'+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h'}} &\le \beta_V. \end{aligned}$$
where
$$\beta_V = O\left(\sqrt{d\lambda_V}\mathcal{H} + \sqrt{d}\iota_1^2\right), \tag{F.4}$$

and

$$\iota_1 = \max\left\{\iota_0, \log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right), \log\frac{4HK^2}{\delta}, \log\left(1+\frac{4(B+L)\sqrt{d^3H}}{\sigma_{\min}}\right),\right.$$
$$\left.\log\left(1+\frac{8\sqrt{d^7}H\beta_R^2}{\lambda_R\sigma_{\min}^2}\right), \log\left(1+\frac{8\sqrt{d^7}H\beta_V^2}{\lambda_V\sigma_{\min}^2}\right)\right\}, L = \mathcal{H}\sqrt{\frac{dK}{\lambda_V}}.$$

For simplicity, we further define $\mathcal{E}_V := \mathcal{E}_1$.

Our ultimate goal is to show $\mathcal{E}_R \cap \mathcal{E}_V$ holds with high probability, which is a 'refined' event where the radius β_R, β_V are smaller than $\beta_{R^\epsilon}, \beta_0$ in the 'coarse' event $\mathcal{E}_{R^\epsilon} \cap \mathcal{E}_0$. Leveraging the technique in He et al. [15], we first prove event $\mathcal{E}_{R^\epsilon} \cap \mathcal{E}_0$ holds with high probability, then come to $\mathcal{E}_R \cap \mathcal{E}_V$.

Lemma F.1. Event $\mathcal{E}_{R^\epsilon}$ holds with probability at least $1-3\delta$.

Proof. See Appendix H.1 for a detailed proof. □

Lemma F.2. On event $\mathcal{E}_{R^\epsilon} \cap \mathcal{E}_{R,k}$, for all $h \in [H]$, we have

$$|[\widehat{\boldsymbol{\nu}}_{1+\epsilon}R_h - \boldsymbol{\nu}_{1+\epsilon}R_h](s_{k,h}, a_{k,h})| \le W_{k,h},$$

where $\widehat{\boldsymbol{\nu}}_{1+\epsilon}R_h$ and $W_{k,h}$ are defined in (5.5) and (5.6) respectively.

Proof. See Appendix H.2 for a detailed proof. □

Lemma F.3. On event $\mathcal{E}_{R^\epsilon}$, event $\mathcal{E}_R$ holds with probability at least $1-3\delta$.

Proof. See Appendix H.3 for a detailed proof. □

Lemma F.4. Event $\mathcal{E}_0$ holds with probability at least $1-3\delta$.

Proof. See Appendix H.4 for a detailed proof. □

Lemma F.5. On event $\mathcal{E}_R \cap \mathcal{E}_h$, for all $k \in [K]$, we have $\check{Q}_h^k(\cdot,\cdot) \le Q_h^*(\cdot,\cdot) \le Q_h^k(\cdot,\cdot)$. In addition, we have $\check{V}_h^k(\cdot) \le V_h^*(\cdot) \le V_h^k(\cdot)$.

Proof. See Appendix H.5 for a detailed proof. □

Lemma F.6. On event $\mathcal{E}_0 \cap \mathcal{E}_R \cap \mathcal{E}_{h+1}$, for all $k \in [K]$, we have

$$\left|\left[\widehat{\mathbb{V}}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^*\right](s_{k,h}, a_{k,h})\right| \le E_{k,h},$$

where $\widehat{\mathbb{V}}_h V_{h+1}^k$ and $E_{k,h}$ are defined in (5.10) and (5.11) respectively.

Proof. See Appendix H.6 for a detailed proof. □

Lemma F.7. On event $\mathcal{E}_0 \cap \mathcal{E}_R \cap \mathcal{E}_{h+1}$, for all $i \le k \le K$, we have

$$\max\left\{\left[\mathbb{V}_h\left(V_{h+1}^k - V_{h+1}^*\right)\right](s_{i,h}, a_{i,h}), \left[\mathbb{V}_h\left(V_{h+1}^* - \check{V}_{h+1}^k\right)\right](s_{i,h}, a_{i,h})\right\} \le D_{i,h},$$

where $D_{i,h}$ is defined in (5.12).

Proof. See Appendix H.7 for a detailed proof. □

Lemma F.8. With probability at least $1-2\delta$, on event $\mathcal{E}_0 \cap \mathcal{E}_R$, event $\mathcal{E}_V$ holds.

Proof. See Appendix H.8 for a detailed proof. □

F.2 Regret Analysis

In this section, we will prove the regret bound based on the events defined in Appendix F.1, which hold with probability at least $1-11\delta$. By the optimism of Lemma F.5, we have

$$\text{Regret}(K) = \sum_{k=1}^{K}\left[V_1^*(s_{k,1}) - V_1^{\pi^k}(s_{k,1})\right] \leq \sum_{k=1}^{K}\left[V_1^k(s_{k,1}) - V_1^{\pi^k}(s_{k,1})\right].$$

Next we bound the regret with the summations of two bonus terms, i.e. $\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}$ and $\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}$ with respect to reward functions and transition probabilities by Lemma F.9. Then we bound them separately by Lemma F.11 and Lemma F.12.

Lemma F.9. With probability at least $1-\delta$, on event $\mathcal{E}_R \cap \mathcal{E}_V$, it follows that

$$\begin{aligned}
&\sum_{k=1}^{K}\left[V_1^k(s_{k,1}) - V_1^{\pi^k}(s_{k,1})\right] \\
&\leq 6\beta_R\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 6\beta_V\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 36H\mathcal{H}\log\frac{4\lceil\log_2 HK\rceil}{\delta}.
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h},a_{k,h}) \\
&\leq 8H\beta_R\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 8H\beta_V\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 29H^2\mathcal{H}\log\frac{4\lceil\log_2 HK\rceil}{\delta}.
\end{aligned}$$

Proof. See Appendix H.9 for a detailed proof. □

Lemma F.10. With probability at least $1-\delta$, on event $\mathcal{E}_R \cap \mathcal{E}_V$, it follows that

$$\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H\left[\mathbb{P}_h(V_{h+1}^k - \check{V}_{h+1}^{k})\right](s_{k,h},a_{k,h}) \\
&\leq 8H\beta_R\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 16H\beta_V\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 29H^2\mathcal{H}\log\frac{4\lceil\log_2 HK\rceil}{\delta}.
\end{aligned}$$

Proof. See Appendix H.10 for a detailed proof. □

Lemma F.11. Set $\lambda_R = \frac{d}{\max\{B^2,W^2\}}$. Then with probability at least $1-\delta$, on event $\mathcal{E}_{R^\epsilon}\cap\mathcal{E}_R$, we have

$$\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} = \widetilde{O}\Bigg(&\sqrt{dH\mathcal{U}^*K} + \sqrt{dH^2K\nu_{\min}^2} + \left(\frac{\nu_{R^\epsilon}d^{\frac{2+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}K^{\frac{1-\epsilon'}{2(1+\epsilon')}}}{\nu_{\min}}\right)^{\frac{1}{\epsilon}} \\
&+\left(d^{\frac{4+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}\mathcal{H}^\epsilon K^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)^{\frac{1}{\epsilon}} + (\nu_R + \max\{B,W\})\sqrt{d}H\Bigg).
\end{aligned}$$

Choosing $\nu_{\min} = \widetilde{O}\left(\nu_{R^\epsilon}^{\frac{1}{1+\epsilon}}d^{\frac{1}{1+\epsilon}}H^{\frac{1-\epsilon}{2(1+\epsilon)}}K^{-\frac{(1+\epsilon)(1+\epsilon')-2}{2(1+\epsilon)(1+\epsilon')}}\right)$ further gives

$$\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} = \widetilde{O}\Bigg(&\sqrt{dH\mathcal{U}^*K} + \nu_{R^\epsilon}^{\frac{1}{1+\epsilon}}d^{\frac{3+\epsilon}{2(1+\epsilon)}}H^{\frac{3+\epsilon}{2(1+\epsilon)}}K^{\frac{1}{(1+\epsilon)(1+\epsilon')}} \\
&+\left(d^{\frac{4+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}\mathcal{H}^\epsilon K^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)^{\frac{1}{\epsilon}} + (\nu_R + \max\{B,W\})\sqrt{d}H\Bigg).
\end{aligned}$$

Proof. See Appendix H.11 for a detailed proof. □

Lemma F.12. Set $\lambda_V = \frac{1}{\mathcal{H}^2}$, let $\mathcal{A}_0$ denote the event where Lemma F.9 and Lemma F.10 hold. Then with probability at least $1 - 2\delta$, on event $\mathcal{E}_0 \cap \mathcal{E}_R \cap \mathcal{E}_V \cap \mathcal{A}_0$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|_{\Sigma_{k-1,h}^{-1}} = \widetilde{O}\Bigg(\sqrt{dH\mathcal{V}^* K} + \sqrt{dH^2 K\sigma_{\min}^2} + \frac{d^{5.5}H^{2.5}\mathcal{H}^2}{\sigma_{\min}} + \sqrt{d^4 H^3 \mathcal{H}\beta_R \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}} + d^{4.5}H^3\mathcal{H}\Bigg).$$

Choosing $\sigma_{\min} = \widetilde{O}\left(\sqrt{d^5 H^{1.5}\mathcal{H}^2} K^{-\frac{1}{4}}\right)$ further gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|_{\Sigma_{k-1,h}^{-1}} = \widetilde{O}\Bigg(\sqrt{dH\mathcal{V}^* K} + \sqrt{d^6 H^{3.5}\mathcal{H}^2} K^{\frac{1}{4}} + \sqrt{d^4 H^3 \mathcal{H}\beta_R \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}} + d^{4.5}H^3\mathcal{H}\Bigg).$$

Proof. See Appendix H.12 for a detailed proof. □

At the end of this section, we state the formal version of Theorem 5.2 and provide its proof.

Theorem F.13 (Formal version of Theorem 5.2)**.** For the time-inhomogeneous linear MDPs with heavy-tailed rewards defined in Section 2.2, we set parameters in Algorithm 3 as follows: $\lambda_R = d/\max\{B^2, W^2\}$, $\lambda_V = 1/\mathcal{H}^2$, $\nu_{\min}$ in Lemma F.11, $\sigma_{\min}$ in Lemma F.12, β_{R^ϵ}, β_0, β_R, β_V in (F.1), (F.2), (F.3), (F.4), respectively. Then for any $\delta \in (0, 1)$, with probability at least $1 - 16\delta$, the regret of HEAVY-LSVI-UCB is bounded by

$$\begin{aligned}\mathrm{Regret}(K) = \widetilde{O}\Bigg(& d\sqrt{H\mathcal{U}^*} K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^* K} + \nu_{R^\epsilon}^{\frac{1}{1+\epsilon}} d^{\frac{2+\epsilon}{1+\epsilon}} H^{\frac{3+\epsilon}{2(1+\epsilon)}} K^{\frac{2+(1-\epsilon)(1+\epsilon')}{2(1+\epsilon)(1+\epsilon')}} \\ &+ \sqrt{d^7 H^{3.5}\mathcal{H}^2} K^{\frac{1}{4}} + \left(d^{2+\epsilon} H^{\frac{1+\epsilon}{2}} \mathcal{H}^\epsilon K^{\frac{1-\epsilon}{2}}\right)^{\frac{1}{\epsilon}} \\ &+ (\nu_R + \max\{B, W\}) dHK^{\frac{1-\epsilon}{2(1+\epsilon)}} + d^5 H^3 \mathcal{H}\Bigg),\end{aligned}$$

where

$$\mathcal{U}^* = \min\{\mathcal{U}_0^*, \mathcal{U}\} \text{ with } \mathcal{U}_0^* = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H} [\boldsymbol{\nu}_{1+\epsilon} R_h]^{\frac{2}{1+\epsilon}} (s_h, a_h)\right], \tag{F.5}$$

$$\mathcal{V}^* = \min\{\mathcal{V}_0^*, \mathcal{V}\} \text{ with } \mathcal{V}_0^* = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H} [\mathbb{V}_h V_{h+1}^*](s_h, a_h)\right], \tag{F.6}$$

$$d_h^{\pi^k}(s, a) = \mathbb{P}^{\pi^k}((s_h, a_h) = (s, a) | s_0 = s_{k,1}) \tag{F.7}$$

and $\mathcal{H}, \mathcal{U}, \mathcal{V}$ are defined in Assumption 2.7.

Proof of Theorem 5.2. Based on high-probability events in Appendix F.1 and Lemmas in this section, we have

$$
\begin{aligned}
&\mathrm{Regret}(K) \\
&= \sum_{k=1}^{K}\left[V_1^*(s_{k,1}) - V_1^{\pi^k}(s_{k,1})\right] \overset{(a)}{\leq} \sum_{k=1}^{K}\left[V_1^k(s_{k,1}) - V_1^{\pi^k}(s_{k,1})\right] \\
&\overset{(b)}{\leq} 6\beta_R \sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 6\beta_V \sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 36H\mathcal{H}\log\frac{4\lceil\log_2 HK\rceil}{\delta} \\
&\overset{(c)}{\leq} \widetilde{O}\bigg(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^*K} + \nu_{R^\epsilon}^{\frac{1}{1+\epsilon}} d^{\frac{2+\epsilon}{1+\epsilon}} H^{\frac{3+\epsilon}{2(1+\epsilon)}} K^{\frac{2+(1-\epsilon)(1+\epsilon')}{2(1+\epsilon)(1+\epsilon')}} + \sqrt{d^7H^{3.5}\mathcal{H}^2}K^{\frac{1}{4}} \\
&\quad + \left(d^{2+\epsilon}H^{\frac{1+\epsilon}{2}}\mathcal{H}^\epsilon K^{\frac{1-\epsilon}{2}}\right)^{\frac{1}{\epsilon}} + (\nu_R + \max\{B,W\})dHK^{\frac{1-\epsilon}{2(1+\epsilon)}} + d^5H^3\mathcal{H}\bigg),
\end{aligned}
$$

where (a) holds due to the optimism of Lemma F.5, (b) holds due to Lemma F.9. (c) holds due to Lemma F.11, Lemma F.12 and

$$
\beta_V\sqrt{d^4H^3\mathcal{H}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}} \leq \frac{1}{2}\beta_V^2 d^4H^3\mathcal{H} + \frac{1}{2}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}.
$$

□

Remark F.14. If $\epsilon > \frac{1}{1+\epsilon'}$, then $K^{\frac{2+(1-\epsilon)(1+\epsilon')}{2(1+\epsilon)(1+\epsilon')}} < \sqrt{K}$. When the number of episodes K is sufficiently large, the regret can be simplified to $\widetilde{O}\left(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^*K}\right)$.

Remark F.15. When $\epsilon = 1$, we return to the setting in Li and Sun [26] and the regret reduces to $\widetilde{O}(d\sqrt{H\mathcal{U}^*K} + d\sqrt{H\mathcal{V}^*K})$. While their variance-aware regret bound is $\widetilde{O}(d\sqrt{H\mathcal{G}^*K})$, where $\mathcal{G}^*$ is a variance-dependent quantity defined in their work. We provide the definition of $\mathcal{G}^*$ and its relationship with $\mathcal{U}^*$, $\mathcal{V}^*$ below.

$$
\begin{aligned}
\mathcal{G}^* &= \min\left\{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H}\left[\mathbb{V}_hR_h + \mathbb{V}_hV_{h+1}^*\right](s_h,a_h)\right], \mathcal{U}+\mathcal{V}\right\} \\
&= \min\{\mathcal{U}_0^* + \mathcal{V}_0^*, \mathcal{U}+\mathcal{V}\} \\
&\geq \min\{\mathcal{U}_0^*, \mathcal{U}\} + \min\{\mathcal{V}_0^*, \mathcal{V}\} \\
&= \mathcal{U}^* + \mathcal{V}^*,
\end{aligned}
$$

where $\mathcal{U}^*$, $\mathcal{V}^*$ and $d_h^{\pi^k}(s,a)$ are defined in Theorem 5.2. Thus we have $d\sqrt{H\mathcal{U}^*K} + d\sqrt{H\mathcal{V}^*K} \leq 2d\sqrt{H(\mathcal{U}^*+\mathcal{V}^*)K} \leq 2d\sqrt{H\mathcal{G}^*K}$, which implies that we recover their result.

F.3 Proof of Corollary 5.6

Proof of Corollary 5.6. First, it holds that $\mathcal{V}^* \leq \mathcal{V} \leq \mathcal{H}V_1^*$ according to (H.19) in the proof of Lemma H.4 in Appendix H.16. Thus the first result follows. Next, to make a fair comparison with the state-of-the-art result of first-order regret [33], we assume the reward functions are uniformly bounded, i.e., $R_h(s,a) \in [0,1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$. Then $\epsilon = 1$, $\mathcal{H} = H$ and

$$
\begin{aligned}
\mathcal{U}^* \leq \mathcal{U}_0^* &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H}\left[\boldsymbol{\nu}_{1+\epsilon}R_h\right]^{\frac{2}{1+\epsilon}}(s_h,a_h)\right] \\
&\overset{(a)}{\leq} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H}r_h(s_h,a_h)\right] \overset{(b)}{\leq} V_1^*,
\end{aligned}
$$

where (a) holds due to $R_h(s,a)$ is bounded in $[0,1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$, (b) uses the optimality of V_1^*. Finally, the proof is completed by the fact that $\mathcal{V}^* \leq HV_1^*$ and $\mathcal{U}^* \leq V_1^*$. □

G Proof of Lower Bound

Proof of Theorem 5.5. The proof of Theorem 5.5 follows from a combination of the lower bound constructions for heavy-tailed linear bandits in Shao et al. [31] and linear MDPs in Zhou et al. [42]. On one hand, we construct a linear MDP with deterministic transition probabilities by concatenating H hard instances in Shao et al. [31] together. Summing the regret over the H components yields $\Omega(dHK^{\frac{1}{1+\epsilon}})$. On the other hand, Zhou et al. [42] shows the regret is at least $\Omega(d\sqrt{H^3K})$. Combining the results together gives the final lower bound. $\square$

H Omitted Proofs in Appendix F

H.1 Proof of Lemma F.1

Proof of Lemma F.1. For each $h \in [H]$, we apply Theorem 3.3 with $L = 1, B = W, T = k, \delta = \delta/H, \mathcal{F}_t = \mathcal{G}_{i,h}, y_t = |\varepsilon_{i,h}|^{1+\epsilon}, \boldsymbol{\theta}_t = \boldsymbol{\psi}_{k,h}$ and $b = \nu_{R^\epsilon}/\nu_{\min}$. We choose parameters $c_0, c_1, \widetilde{\tau}_0$ as in Theorem 3.3. Then with probability at least $1 - 3\delta$, $\|\boldsymbol{\psi}_{k,h} - \boldsymbol{\psi}_h^*\|_{\boldsymbol{H}_{k,h}^{-1}} \le \beta_{R^\epsilon,k}$, that is event $\mathcal{E}_{R^\epsilon}$ holds. $\square$

H.2 Proof of Lemma F.2

Proof of Lemma F.2. Notice that

$$
\begin{aligned}
& \left|\left[\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h - \boldsymbol{\nu}_{1+\epsilon} R_h\right](s_{k,h}, a_{k,h})\right| \\
= & \left|\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{\psi}}_{k-1,h} - \boldsymbol{\psi}_h^* \rangle\right| \\
\le & \left|\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{\psi}}_{k-1,h} - \boldsymbol{\psi}_{k-1,h} \rangle\right| + \left|\langle \boldsymbol{\phi}_{k,h}, \boldsymbol{\psi}_{k-1,h} - \boldsymbol{\psi}_h^* \rangle\right| \\
\le & \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} \|\widehat{\boldsymbol{\psi}}_{k-1,h} - \boldsymbol{\psi}_{k-1,h}\|_{\boldsymbol{H}_{k-1,h}} + \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} \|\boldsymbol{\psi}_{k-1,h} - \boldsymbol{\psi}_h^*\|_{\boldsymbol{H}_{k-1,h}},
\end{aligned}
$$

where the last inequality holds due to Cauchy-Schwartz inequality. And

$$
\|\boldsymbol{\psi}_{k-1,h} - \boldsymbol{\psi}_h^*\|_{\boldsymbol{H}_{k-1,h}} \le \beta_{R^\epsilon,k-1} \tag{H.1}
$$

since $\mathcal{E}_{R^\epsilon}$ holds.

We next give an upper bound of $\|\widehat{\boldsymbol{\psi}}_{k-1,h} - \boldsymbol{\psi}_{k-1,h}\|_{\boldsymbol{H}_{k-1,h}}$ by a novel perturbation analysis of adaptive Huber regression. For each $h \in [H]$, we apply Lemma C.3 with $t = k-1, \boldsymbol{\theta}_t = \boldsymbol{\psi}_{k-1,h}, \widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\psi}}_{k-1,h}, \mathcal{F}_s = \mathcal{G}_{i,h}, y_s = |\varepsilon_{i,h}|^{1+\epsilon}, \widehat{y}_s = |\widehat{\varepsilon}_{i,h}|^{1+\epsilon}$. Notice

$$
\begin{aligned}
\left||\widehat{\varepsilon}_{i,h}|^{1+\epsilon} - |\varepsilon_{i,h}|^{1+\epsilon}\right| &\le |\widehat{\varepsilon}_{i,h} - \varepsilon_{i,h}|^{1+\epsilon} \\
&= |\langle \boldsymbol{\phi}_{i,h}, \boldsymbol{\theta}_{i,h} - \boldsymbol{\theta}_h^* \rangle|^{1+\epsilon} \\
&\overset{(a)}{\le} \mathcal{H}^\epsilon \|\boldsymbol{\theta}_{i,h} - \boldsymbol{\theta}_h^*\|_{\boldsymbol{H}_{i,h}} \|\boldsymbol{\phi}_{i,h}\|_{\boldsymbol{H}_{i,h}^{-1}} \\
&\overset{(b)}{\le} \mathcal{H}^\epsilon \beta_{R,i-1} \|\boldsymbol{\phi}_{i,h}\|_{\boldsymbol{H}_{i,h}^{-1}},
\end{aligned}
$$

where (a) holds due to Cauchy-Schwartz inequality, and $\langle \boldsymbol{\phi}_{i,h}, \boldsymbol{\theta}_{i,h} \rangle$ together with $\langle \boldsymbol{\phi}_{i,h}, \boldsymbol{\theta}_h^* \rangle$ are bounded in $[0, \mathcal{H}]$. And (b) is due to $\mathcal{E}_{R,k-1}$ holds. Then we have $\widehat{\beta}_s = \mathcal{H}^\epsilon \beta_{R,i-1}$ in Lemma C.3, and it follows that

$$
\|\widehat{\boldsymbol{\psi}}_{k-1,h} - \boldsymbol{\psi}_{k-1,h}\|_{\boldsymbol{H}_{k-1,h}} \le 6\mathcal{H}^\epsilon \beta_{R,k-1} \kappa. \tag{H.2}
$$

Combining (H.1) with (H.2), we have

$$
\left|\left[\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h - \boldsymbol{\nu}_{1+\epsilon} R_h\right](s_{k,h}, a_{k,h})\right| \le \left(\beta_{R^\epsilon,k-1} + 6\mathcal{H}^\epsilon \beta_{R,k-1}\kappa\right) \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}.
$$

That is $\left|\left[\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h - \boldsymbol{\nu}_{1+\epsilon} R_h\right](s_{k,h}, a_{k,h})\right| \le W_{k,h}$. $\square$

H.3 Proof of Lemma F.3

Proof of Lemma F.3. For each $h \in [H]$, we use Theorem 3.3 with $L = 1, \delta = \delta/H, T = k, \mathcal{F}_t = \mathcal{G}_{i,h}, y_t = \widetilde{R}_{i,h} := R_{i,h} \mathbb{1}\{|[\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h - \boldsymbol{\nu}_{1+\epsilon} R_h](s_{i,h}, a_{i,h})| \le W_{i,h}\}$. Thus $b = 1$ by definition. Denote $\widetilde{\boldsymbol{\theta}}_{k,h}$ as the counterpart of $\boldsymbol{\theta}_{k,h}$ being the solution of adaptive Huber regression where $R_{i,h}$ is replaced by $\widetilde{R}_{i,h}$ for all $i \le k$. Then by Theorem 3.3, with probability at least $1 - 3\delta$, we have

$$\left\| \widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^* \right\|_{\boldsymbol{H}_{k,h}^{-1}} \le \beta_{R,k} \tag{H.3}$$

for all $h \in [H]$ and $k \in [K]$.

Then, we continue the proof by induction on event $\mathcal{E}_{R,k}$ over k. First, event $\mathcal{E}_{R,0}$ holds trivially. Next, we suppose $\mathcal{E}_{R,k-1}$ holds and will prove $\mathcal{E}_{R,k}$ holds. Since $\mathcal{E}_{R^\epsilon} \cap \mathcal{E}_{R,k-1}$ holds, it follows that $\widetilde{R}_{i,h} = R_{i,h}$ for all $h \in [H]$ and $i \le k$ by Lemma F.2 and the definition of $\widetilde{R}_{i,h}$. And thus $\widetilde{\boldsymbol{\theta}}_{k,h} = \boldsymbol{\theta}_{k,h}$. By (H.3), for all $h \in [H]$, we have

$$\|\boldsymbol{\theta}_{i,h} - \boldsymbol{\theta}_h^*\|_{\boldsymbol{H}_{i,h}^{-1}} \le \beta_{R,i},$$

where

$$\begin{aligned}\beta_{R,i} &= 3\sqrt{\lambda_R} B + 24 i^{\frac{1-\epsilon}{2(1+\epsilon)}} \sqrt{2\kappa} (\log 3K)^{\frac{1-\epsilon}{2(1+\epsilon)}} \left(\log \frac{2HK^2}{\delta}\right)^{\frac{\epsilon}{1+\epsilon}} \\ &= O\left(\sqrt{\lambda_R} B + \sqrt{d} i^{\frac{1-\epsilon}{2(1+\epsilon)}} \iota\right).\end{aligned}$$

That is $\mathcal{E}_{R,k}$ holds.

Finally, induction over all $k \in [K]$ completes the proof. □

H.4 Proof of Lemma F.4

We will use the self-normalized bound with a covering argument in Lemma I.5 for estimating next-state value function in the following proof frequently, which is the core technique used in Jin et al. [22], He et al. [15].

Proof of Lemma F.4. Define $\mathcal{V}^+$ as the class of optimistic value functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the parametric form given in (I.2) and $\mathcal{V}^-$ the class of pessimistic value functions with the parametric form given in (I.3). By Lemma I.8 and Lemma I.9,

$$\log \mathcal{N}(\mathcal{V}^{\pm}, \epsilon) \le \left[d \log\left(1 + \frac{4(B+L)}{\epsilon}\right) + d^2 \log\left(1 + \frac{8 d^{1/2} \beta_{R,K}^2}{\lambda_R \epsilon^2}\right)\left(1 + \frac{8 d^{1/2} \beta_V^2}{\lambda_V \epsilon^2}\right)\right] |\mathcal{K}|, \tag{H.4}$$

where $L = \mathcal{H}\sqrt{\frac{dK}{\lambda_V}}$ and $|\mathcal{K}| \le dH \log_2\left(1 + \frac{K}{\lambda_R \nu_{\min}^2}\right)\left(1 + \frac{K}{\lambda_V \sigma_{\min}^2}\right)$.

The case where $f = V_{h+1}^k$. One can find that $f \in \mathcal{V}^+$. To make use of Lemma I.5, we first specify the parameters defined therein. We have $\|f\|_\infty \le C_0 = \mathcal{H}$ and $\epsilon_1 = \frac{\lambda_V \mathcal{H}\sqrt{d}}{K} \nu_{\min}^2$. By (H.4), it follows that

$$\begin{aligned}&\log \mathcal{N}(\mathcal{V}^+, \epsilon_1) \\ &\le \left[d \log\left(1 + \frac{4(B+L)K}{\lambda_V \mathcal{H}\sqrt{d}\sigma_{\min}^2}\right) + d^2 \log\left(1 + \frac{8\beta_{R,K}^2 K^2}{\sqrt{d}\lambda_R \lambda_V^2 \mathcal{H}^2 \sigma_{\min}^4}\right)\left(1 + \frac{8\beta_V^2 K^2}{\sqrt{d}\lambda_V^3 \mathcal{H}^2 \sigma_{\min}^4}\right)\right] \\ &\qquad \cdot dH \log_2\left(1 + \frac{K}{\lambda_R \nu_{\min}^2}\right)\left(1 + \frac{K}{\lambda_V \sigma_{\min}^2}\right) \\ &\le 6 d^3 H \iota_0^2.\end{aligned}$$

By the third condition of Lemma I.5, with probability at least $1 - \frac{\delta}{H}$, $\left\|\widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k,h}} \le \beta_0$ for all $k \in [K]$. A union which finishes the proof.

The case where $f = \check{V}_{h+1}^k$. The analysis on $\check{V}_{h+1}^k$ is similar to (i).

The case where $f = \left(V_{h+1}^k\right)^2$. The analysis on $\left(V_{h+1}^k\right)^2$ is similar to (i) except for the following two changes. First, $C_0 = \mathcal{H}^2$ and $\epsilon_1' = \frac{\lambda_V \mathcal{H}^2 \sqrt{d}}{K}\sigma_{\min}^2$. Second, with $[\mathcal{V}^+]^2 = \{f^2 : f \in \mathcal{V}^+\}$, we have $[V_{h+1}^k]^2 \in [\mathcal{V}^+]^2$ and

$$\log \mathcal{N}([\mathcal{V}^+]^2, \epsilon_1') \overset{(a)}{\leq} \log \mathcal{N}(\mathcal{V}^+, \frac{\epsilon_1'}{2\mathcal{H}}) \leq \log \mathcal{N}(\mathcal{V}^+, \frac{\epsilon_1}{2}) \leq 6d^3 H \iota_0^2.$$

Here (a) uses the fact that the $\frac{\epsilon_1'}{2\mathcal{H}}$-cover of $\mathcal{V}^+$ is a ϵ_1-cover of $[\mathcal{V}^+]^2$. $\square$

H.5 Proof of Lemma F.5

Proof of Lemma F.5. We prove the optimism inequality by induction. When $h = H+1$, results hold trivially since $V_{H+1}^k = V_{H+1}^* = 0$. We assume the statement is true for $h+1$, that is $V_{h+1}^k(\cdot) \geq V_{h+1}^*(\cdot)$. Next we prove the case of h.

For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $k \in [K]$,

$$\begin{aligned}
&\langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k-1,h} + \widehat{\boldsymbol{w}}_{k-1,h}\rangle + \beta_{R,k-1}\left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} + \beta_V \left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} - Q_h^*(\cdot,\cdot) \\
&= \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k-1,h} - \boldsymbol{\theta}_h^*\rangle + \beta_{R,k-1}\left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} \\
&\qquad + \langle \boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^*\right]\rangle + \beta_V\left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} \\
&\overset{(a)}{\geq} \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k-1,h} - \boldsymbol{\theta}_h^*\rangle + \beta_{R,k-1}\left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} \\
&\qquad + \langle \boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\rangle + \beta_V\left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} \\
&\overset{(b)}{\geq} \left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{H}_{k-1,h}^{-1}}\left(-\left\|\boldsymbol{\theta}_{k-1,h} - \boldsymbol{\theta}_h^*\right\|_{\boldsymbol{H}_{k-1,h}} + \beta_{R,k-1}\right) \\
&\qquad + \left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\left(-\left\|\widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} + \beta_V\right) \\
&\overset{(c)}{\geq} 0,
\end{aligned}$$

where (a) is due to $\mathcal{E}_h$ holds, $\mathcal{E}_{h+1} \subset \mathcal{E}_h$ and thus $\left\langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{w}_h\left[V_{h+1}^k\right] - \boldsymbol{w}_h\left[V_{h+1}^*\right]\right\rangle = [\mathbb{P}_h(V_{h+1}^k - V_{h+1}^*)](\cdot,\cdot) \geq 0$ by induction. (b) is due to Cauchy-Schwartz inequality. And (c) is due to $\mathcal{E}_R \cap \mathcal{E}_h$ holds.

We assume the sequence of updating episodes $1 = k_1 \leq k_2 \leq \cdots \leq k_{N_K} \leq K$, such that the latest update episode before k is $k_{\text{last}} = k_{N_k}$. Then for all $k \in [K]$, we have

$$\begin{aligned}
&Q_{h+1}^k(\cdot,\cdot) \\
&= \min_{i \leq N_k}\left\{\langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k_i-1,h} + \widehat{\boldsymbol{w}}_{k_i-1,h}\rangle + \beta_{R,k_i-1}\left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{H}_{k_i-1,h}^{-1}} + \beta_V\left\|\boldsymbol{\phi}(\cdot,\cdot)\right\|_{\boldsymbol{\Sigma}_{k_i-1,h}^{-1}}, \mathcal{H}\right\} \\
&\geq Q_h^*(\cdot,\cdot),
\end{aligned}$$

which implies the case of h is true.

The proof for pessimism inequality is similar to the optimism. $\square$

H.6 Proof of Lemma F.6

Proof of Lemma F.6. First, $\left|\left[\widehat{\mathbb{V}}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^*\right](s_{k,h}, a_{k,h})\right| \leq \mathcal{H}^2$ since both $\left[\widehat{\mathbb{V}}_h V_{h+1}^k\right](\cdot,\cdot)$ and $\left[\mathbb{V}_h V_{h+1}^*\right](\cdot,\cdot)$ are bounded in $[0, \mathcal{H}^2]$. Then it follows that

$$\begin{aligned}
&\left|\left[\widehat{\mathbb{V}}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^*\right](s_{k,h}, a_{k,h})\right| \\
&\leq \left|\left[\widehat{\mathbb{V}}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^k\right](s_{k,h}, a_{k,h})\right| + \left|\left[\mathbb{V}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^*\right](s_{k,h}, a_{k,h})\right|.
\end{aligned}$$

We next bound the two terms in the RHS of the last inequality separately.

Bound the first term. It follows that

$$\begin{aligned}
&\left|\left[\widehat{\mathbb{V}}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^k\right](s_{k,h}, a_{k,h})\right| \\
\leq\ &\left|\langle \boldsymbol{\phi}_{k,h}, \widetilde{\boldsymbol{w}}_{k-1,h}\rangle_{[0,\mathcal{H}^2]} - \left\langle \boldsymbol{\phi}_{k,h}, \boldsymbol{w}_h\left[\left(V_{h+1}^k\right)^2\right]\right\rangle\right| \\
&\qquad + \left|\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h}\rangle_{[0,\mathcal{H}]}^2 - \left\langle \boldsymbol{\phi}_{k,h}, \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\rangle^2\right| \\
\overset{(a)}{\leq}\ &\left|\left\langle \boldsymbol{\phi}_{k,h}, \widetilde{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[\left(V_{h+1}^k\right)^2\right]\right\rangle\right| + 2\mathcal{H}\left|\left\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\rangle\right| \\
\overset{(b)}{\leq}\ &\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\left\|\widetilde{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[\left(V_{h+1}^k\right)^2\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} \\
&\qquad + 2\mathcal{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\left\|\widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} \\
\overset{(c)}{\leq}\ &3\mathcal{H}\beta_0\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}},
\end{aligned}$$

where (a) is due to both $\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h}\rangle_{[0,\mathcal{H}]}$ and $\left\langle \boldsymbol{\phi}_{k,h}, \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\rangle$ are bounded in $[0,\mathcal{H}]$. (b) holds due to Cauchy-Schwartz inequality and (c) is due to $\mathcal{E}_0$ holds.

Bound the second term. We have

$$\begin{aligned}
&\left|\left[\mathbb{V}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^*\right](s_{k,h}, a_{k,h})\right| \\
\leq\ &\left|\left[\mathbb{P}_h\left(V_{h+1}^k\right)^2 - \left(V_{h+1}^*\right)^2\right](s_{k,h}, a_{k,h})\right| + \left|\left[\mathbb{P}_h V_{h+1}^k\right]^2(s_{k,h}, a_{k,h}) - \left[\mathbb{P}_h V_{h+1}^*\right]^2(s_{k,h}, a_{k,h})\right| \\
=\ &\left|\left[\mathbb{P}_h\left(V_{h+1}^k + V_{h+1}^*\right)\left(V_{h+1}^k - V_{h+1}^*\right)\right](s_{k,h}, a_{k,h})\right| \\
&\qquad + \left|\left[\mathbb{P}_h\left(V_{h+1}^k + V_{h+1}^*\right)\right](s_{k,h}, a_{k,h})\left[\mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^*\right)\right](s_{k,h}, a_{k,h})\right| \\
\overset{(a)}{\leq}\ &4\mathcal{H}\left[\mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^*\right)\right](s_{k,h}, a_{k,h}) \\
\overset{(b)}{\leq}\ &4\mathcal{H}\left[\mathbb{P}_h\left(V_{h+1}^k - \check{V}_{h+1}^k\right)\right](s_{k,h}, a_{k,h}) \\
\overset{(c)}{\leq}\ &4\mathcal{H}\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \check{\boldsymbol{w}}_{k-1,h}\rangle \\
&\qquad + 4\mathcal{H}\left(\left\|\widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} + \left\|\check{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}}\right)\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} \\
\overset{(d)}{\leq}\ &4\mathcal{H}\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \check{\boldsymbol{w}}_{k-1,h}\rangle + 8\mathcal{H}\beta_0\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}},
\end{aligned}$$

where (a) holds due to $V_{h+1}^k(\cdot), V_{h+1}^*(\cdot)$ are both bounded in $[0,\mathcal{H}]$, $\mathcal{E}_R \cap \mathcal{E}_{h+1}$ holds and the optimism by Lemma F.5, (b) holds due to the pessimism by Lemma F.5, (c) holds due to Cauchy-Schwartz inequality and (d) is due to $\mathcal{E}_0$ holds.

Putting pieces together, we have

$$\left|\left[\widehat{\mathbb{V}}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^*\right](s_{k,h}, a_{k,h})\right| \leq 4\mathcal{H}\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \check{\boldsymbol{w}}_{k-1,h}\rangle + 11\mathcal{H}\beta_0\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}.$$

□

H.7 Proof of Lemma F.7

Proof of Lemma F.7. First, $\left[\mathbb{V}_h\left(V_{h+1}^k - V_{h+1}^*\right)\right](s_{i,h}, a_{i,h}) \le \mathcal{H}^2$ since both $V_{h+1}^k(\cdot)$ and $V_{h+1}^*(\cdot)$ are bounded in $[0, \mathcal{H}]$. Then, for any $i \le k$, we have

$$
\begin{aligned}
& \left[\mathbb{V}_h\left(V_{h+1}^k - V_{h+1}^*\right)\right](s_{i,h}, a_{i,h}) \\
\le & \left[\mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^*\right)^2\right](s_{i,h}, a_{i,h}) \\
\overset{(a)}{\le} & 2\mathcal{H}\left[\mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^*\right)\right](s_{i,h}, a_{i,h}) \\
\overset{(b)}{\le} & 2\mathcal{H}\left[\mathbb{P}_h\left(V_{h+1}^k - \check{V}_{h+1}^k\right)\right](s_{i,h}, a_{i,h}) \\
\overset{(c)}{\le} & 2\mathcal{H}\left[\mathbb{P}_h\left(V_{h+1}^i - \check{V}_{h+1}^i\right)\right](s_{i,h}, a_{i,h}) \\
\overset{(d)}{\le} & 2\mathcal{H}\langle \boldsymbol{\phi}_{i,h}, \widehat{\boldsymbol{w}}_{i-1,h} - \check{\boldsymbol{w}}_{i-1,h}\rangle + 4\mathcal{H}\beta_0 \left\|\boldsymbol{\phi}_{i,h}\right\|_{\boldsymbol{\Sigma}_{i-1,h}^{-1}},
\end{aligned}
$$

where (a) holds due to both V_{h+1}^k, V_{h+1}^* are bounded in $[0, \mathcal{H}]$, $\mathcal{E}_R \cap \mathcal{E}_{h+1}$ holds and the optimism by Lemma F.5, (b) holds due to the pessimism by Lemma F.5, (c) holds due to V_{h+1}^k is non-increasing and $\check{V}_{h+1}^k$ is non-decreasing by definition and (d) is due to Cauchy-Schwartz inequality and $\mathcal{E}_0$ holds.

The proof of inequality for $\check{V}_{h+1}^k$ is similar to the proof of V_{h+1}^k above. □

H.8 Proof of Lemma F.8

Proof of Lemma F.8. We prove Lemma F.8 by induction. First, when $h = H$, event $\mathcal{E}_H$ holds since $V_{H+1}^k(\cdot) = V_{H+1}^*(\cdot) = \check{V}_{H+1}^k(\cdot) = 0$ for all $k \in [H]$. Then, we prove with probability at least $1 - \frac{2\delta}{H}$, $\mathcal{E}_h$ holds on event $\mathcal{E}_{h+1}$. By induction over h, with probability at least $1 - 2\delta$, on event $\mathcal{E}_0 \cap \mathcal{E}_R$, event $\mathcal{E}_V = \mathcal{E}_1$ holds.

Note $\widehat{\boldsymbol{w}}_{k-1,h} = \widehat{\boldsymbol{w}}_{k-1,h}\left[V_{h+1}^k\right] = \widehat{\boldsymbol{w}}_{k-1,h}\left[V_{h+1}^*\right] + \widehat{\boldsymbol{w}}_{k-1,h}\left[V_{h+1}^k - V_{h+1}^*\right]$ and $\boldsymbol{w}_h\left[V_{h+1}^k\right] = \boldsymbol{w}_h\left[V_{h+1}^*\right] + \boldsymbol{w}_h\left[V_{h+1}^k - V_{h+1}^*\right]$. It follows that

$$
\begin{aligned}
& \left\|\widehat{\boldsymbol{w}}_{k-1,h} - \boldsymbol{w}_h\left[V_{h+1}^k\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} \\
\le & \left\|\widehat{\boldsymbol{w}}_{k-1,h}\left[V_{h+1}^*\right] - \boldsymbol{w}_h\left[V_{h+1}^*\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} + \left\|\widehat{\boldsymbol{w}}_{k-1,h}\left[V_{h+1}^k - V_{h+1}^*\right] - \boldsymbol{w}_h\left[V_{h+1}^k - V_{h+1}^*\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}}.
\end{aligned}
$$

Next, we bound the two terms in the RHS by Lemma I.5 separately.

Bound the first term. The first condition of Lemma I.5 is satisfied since V_{h+1}^* is a deterministic function. We have $C_0 = \mathcal{H}$ and $\mathcal{A}_{k,h} = \left\{\sigma_{k,h}^2 \ge \left[\mathbb{V}_h V_h^*\right](s_{k,h}, a_{k,h})\right\}$ is $\mathcal{F}_{k,h}$-measurable. Since $\mathcal{E}_0 \cap \mathcal{E}_R \cap \mathcal{E}_{h+1}$ holds, for all $k \in [K]$, $\left|\left[\widehat{\mathbb{V}}_h V_{h+1}^k - \mathbb{V}_h V_{h+1}^*\right](s_{k,h}, a_{k,h})\right| \le E_{k,h}$ by Lemma F.6. Thus, $\sigma_{k,h}^2 \ge \left[\widehat{\mathbb{V}}_h V_{h+1}^k\right](s_{k,h}, a_{k,h}) + E_{k,h} \ge \left[\mathbb{V}_h V_h^*\right](s_{k,h}, a_{k,h})$ for all $k \in [K]$, which implies $\bigcap_{k\in[K]} \mathcal{A}_{k,h}$ holds and thus $C_\sigma = 1$. By Lemma I.5, with probability at least $1 - \delta/H$, for all $k \in [K]$,

$$
\left\|\widehat{\boldsymbol{w}}_{k-1,h}\left[V_{h+1}^*\right] - \boldsymbol{w}_h\left[V_{h+1}^*\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} \le \beta_1,
$$

where

$$
\begin{aligned}
\beta_1 &= \sqrt{d\lambda_V}\mathcal{H} + 8\sqrt{d\log\left(1 + \frac{K}{\sigma_{\min}^2 d\lambda_V}\right)\log\frac{4HK^2}{\delta}} + \frac{4}{d^{2.5}H}\log\frac{4HK^2}{\delta} \\
&\le \sqrt{d\lambda_V}\mathcal{H} + 8\sqrt{d}\iota_1 + 4\iota_1.
\end{aligned}
$$

Bound the second term. The second condition of Lemma I.5 is satisfied since $V_{h+1}^k - V_{h+1}^*$ are random functions. We have $C_0 = \mathcal{H}$ and $\mathcal{A}_{k,h} = \left\{\sigma_{k,h}^2 \ge d^3 H\left[\mathbb{V}_h\left(V_h^k - V_h^*\right)\right](s_{k,h}, a_{k,h})\right\}$ is $\mathcal{F}_{k,h}$-measurable. Since $\mathcal{E}_0 \cap \mathcal{E}_R \cap \mathcal{E}_{h+1}$ holds, for all $k \in [K]$, $\left[\mathbb{V}_h\left(V_h^k - V_h^*\right)\right](s_{k,h}, a_{k,h}) \le D_{k,h}$

by Lemma F.7. Thus, by definition of $\sigma_{k,h}$, we have $\bigcap_{k\in[K]}\mathcal{A}_{k,h}$ holds and thus $C_\sigma = \frac{1}{\sqrt{d^3H}}$. And the log-covering number of the function class that covers $V_{h+1}^k - V_{h+1}^*$ can be bounded by Lemma I.9 as $\log N_0 = |\mathcal{N}(\mathcal{V}^+, \epsilon_0)| \le 6d^3H\iota_1^2$ with $\mathcal{V}^+$ defined in (I.2) and $\epsilon_0 = \min\left\{\frac{\sigma_{\min}}{\sqrt{d^3H}}, \frac{\lambda_V\mathcal{H}\sqrt{d}}{K}\sigma_{\min}^2\right\}$. By Lemma I.5, with probability at least $1-\delta/H$, for all $k\in[K]$,

$$\left\|\widehat{\boldsymbol{w}}_{k-1,h}\left[V_{h+1}^k - V_{h+1}^*\right] - \boldsymbol{w}_h\left[V_{h+1}^k - V_{h+1}^*\right]\right\|_{\boldsymbol{\Sigma}_{k-1,h}} \le \beta_2,$$

where

$$\begin{aligned}\beta_2 &= 2\sqrt{d\lambda_V}\mathcal{H} + \frac{32}{\sqrt{d^3H}}\sqrt{d\log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right)\log\frac{4N_0HK^2}{\delta}} + \frac{4}{d^{2.5}H}\log\frac{4N_0HK^2}{\delta}\\ &= O(\sqrt{d\lambda_V}\mathcal{H} + \sqrt{d}\iota_1^2).\end{aligned}$$

The proof for the pessimism is similar to that of the optimism.

Finally, putting pieces together and we have $\beta_V = \beta_1 + \beta_2 = O(\sqrt{d\lambda_V}\mathcal{H} + \sqrt{d}\iota_1^2)$.

□

H.9 Proof of Lemma F.9

Proof of Lemma F.9. For any $k\in[K]$, recall k_{last} is the latest update episode before episode k satisfying $k_{\text{last}} \le k$ and $V_h^k(\cdot) = V_h^{k_{\text{last}}}(\cdot), \forall h\in[H]$. By Lemma I.7, due to $\boldsymbol{H}_{k-1,h} \succeq \boldsymbol{H}_{k_{\text{last}}-1,h}$ and $\det(\boldsymbol{H}_{k-1,h}) \le 2\det(\boldsymbol{H}_{k_{\text{last}}-1,h})$ by the updating rule, it follows that for any $\boldsymbol{x}\in\mathbb{R}^d$,

$$\|\boldsymbol{x}\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}} \le 2\|\boldsymbol{x}\|_{\boldsymbol{H}_{k-1,h}^{-1}}. \tag{H.5}$$

Telescoping on $V_h^k(s_{k,h}) - V_h^{\pi^k}(s_{k,h})$. By definition, $Q_h^k(\cdot,\cdot) \le \langle\boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k_{\text{last}}-1,h} + \widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h}\rangle + \beta_{R,k_{\text{last}}-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}} + \beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}}$ and $Q_h^{\pi^k}(\cdot,\cdot) = \left\langle\boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_h^* + \boldsymbol{w}_h\left[V_{h+1}^{\pi^k}\right]\right\rangle$. Since $a_{k,h} = \pi_h^k(s_{k,h}) = \operatorname{argmax}_{a\in\mathcal{A}} Q_h^k(s_{k,h}, a)$, we have

$$\begin{aligned}&V_h^k(s_{k,h}) - V_h^{\pi^k}(s_{k,h})\\ &= Q_h^k(s_{k,h}, a_{k,h}) - Q_h^{\pi^k}(s_{k,h}, a_{k,h})\\ &\le \left\langle\boldsymbol{\phi}_{k,h}, (\boldsymbol{\theta}_{k_{\text{last}}-1,h} - \boldsymbol{\theta}_h^*) + \left(\widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[V_{h+1}^{\pi^k}\right]\right)\right\rangle\\ &\qquad + \beta_{R,k_{\text{last}}-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}} + \beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}}\\ &\overset{(a)}{\le} \left\langle\boldsymbol{\phi}_{k,h}, (\boldsymbol{\theta}_{k_{\text{last}}-1,h} - \boldsymbol{\theta}_h^*) + \left(\widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[V_{h+1}^{k_{\text{last}}}\right]\right) + \left(\boldsymbol{w}_h\left[V_{h+1}^k - V_{h+1}^{\pi^k}\right]\right)\right\rangle\\ &\qquad + 2\beta_{R,k-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 2\beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\\ &\overset{(b)}{\le} \left[\mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^{\pi^k}\right)\right](s_{k,h}, a_{k,h}) + 4\beta_{R,k-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\end{aligned}$$

Here (a) uses (H.5), (b) uses

$$\begin{aligned}&\left\langle\boldsymbol{\phi}_{k,h}, (\boldsymbol{\theta}_{k_{\text{last}}-1,h} - \boldsymbol{\theta}_h^*) + \left(\widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[V_{h+1}^{k_{\text{last}}}\right]\right)\right\rangle\\ &\le \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}}\|\boldsymbol{\theta}_{k_{\text{last}}-1,h} - \boldsymbol{\theta}_h^*\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}}\\ &\qquad + \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}}\left\|\widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[V_{h+1}^{k_{\text{last}}}\right]\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}}\\ &\le \beta_{R,k_{\text{last}}-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}} + \beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}}\\ &\le 2\beta_{R,k-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 2\beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\end{aligned}$$

on $\mathcal{E}_R\cap\mathcal{E}_V$ and

$$\left\langle\boldsymbol{\phi}_{k,h}, \boldsymbol{w}_h\left[V_{h+1}^k - V_{h+1}^{\pi^k}\right]\right\rangle = \left[\mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^{\pi^k}\right)\right](s_{k,h}, a_{k,h}).$$

We define $X_{k,h} := \left[\mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^{\pi^k}\right)\right](s_{k,h}, a_{k,h}) - \left[V_{h+1}^k(s_{k,h+1}) - V_{h+1}^{\pi^k}(s_{k,h+1})\right]$. Then it follows that

$$\begin{aligned}
&V_h^k(s_{k,h}) - V_h^{\pi^k}(s_{k,h}) \\
&= V_{h+1}^k(s_{k,h+1}) - V_{h+1}^{\pi^k}(s_{k,h+1}) + X_{k,h} + 4\beta_{R,k-1}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}.
\end{aligned}$$

Summing the inequality above over h' such that $h \le h' \le H$ and we have

$$V_h^k(s_{k,h}) - V_h^{\pi^k}(s_{k,h}) \le \sum_{h'=h}^{H}\left[X_{k,h'} + 4\beta_{R,k-1}\left\|\boldsymbol{\phi}_{k,h'}\right\|_{\boldsymbol{H}_{k-1,h'}^{-1}} + 4\beta_V\left\|\boldsymbol{\phi}_{k,h'}\right\|_{\boldsymbol{\Sigma}_{k-1,h'}^{-1}}\right]. \tag{H.6}$$

Therefore, setting $h = 1$ and summing (H.6) over $k \in [K]$, we have

$$\sum_{k=1}^{K}\left[V_1^k(s_{k,h}) - V_1^{\pi^k}(s_{k,h})\right] \le \sum_{k=1}^{K}\sum_{h=1}^{H}\left[X_{k,h} + 4\beta_{R,k-1}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right]. \tag{H.7}$$

Bound $\sum_{k=1}^K \sum_{h=1}^H X_{k,h}$. Notice $X_{k,h}$ is $\mathcal{F}_{k,h+1}$-measurable with $\mathbb{E}[X_{k,h}|\mathcal{F}_{k,h}] = 0$, $|X_{k,h}| \le 2\mathcal{H}$ and

$$\mathbb{E}[X_{k,h}^2|\mathcal{F}_{k,h}] \le \mathbb{E}\left[\left[V_{h+1}^k(s_{k,h+1}) - V_{h+1}^{\pi^k}(s_{k,h+1})\right]^2 \middle| \mathcal{F}_{k,h}\right] \overset{(a)}{\le} \mathcal{H}\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h}, a_{k,h}),$$

where (a) holds due to $\left|V_{h+1}^k(\cdot) - V_{h+1}^{\pi^k}(\cdot)\right| \le \mathcal{H}$ and optimism in Lemma F.5 By the variance-aware Freedman inequality in Lemma I.2, with probability at least $1 - \frac{\delta}{2}$, it follows that

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H} X_{k,h}\right| \le 3\sqrt{\iota}\sqrt{\mathcal{H}\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h}, a_{k,h}) + 10\mathcal{H}\iota} \tag{H.8}$$

where $\iota = \log\frac{4\lceil\log_2 HK\rceil}{\delta}$.

Bound $\sum_{k=1}^K \sum_{h=1}^H \left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h}, a_{k,h})$. Notice

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h}, a_{k,h}) \\
&= \sum_{k=1}^{K}\sum_{h=2}^{H}\left[V_h^k(s_{k,h}) - V_h^{\pi^k}(s_{k,h})\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} X_{k,h} \\
&\overset{(a)}{\le} \sum_{k=1}^{K}\sum_{h=2}^{H}\sum_{h'=h}^{H}\left[X_{k,h'} + 4\beta_{R,k-1}\left\|\boldsymbol{\phi}_{k,h'}\right\|_{\boldsymbol{H}_{k-1,h'}^{-1}} + 4\beta_V\left\|\boldsymbol{\phi}_{k,h'}\right\|_{\boldsymbol{\Sigma}_{k-1,h'}^{-1}}\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} X_{k,h} \\
&= \sum_{k=1}^{K}\sum_{h=2}^{H}(H-h+1)\left[X_{k,h} + 4\beta_{R,k-1}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} X_{k,h} \\
&\overset{(b)}{\le} \sum_{k=1}^{K}\sum_{h=1}^{H} X_{k,h} b_h + 4H\beta_R\sum_{k=1}^{K}\sum_{h=2}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4H\beta_V\sum_{k=1}^{K}\sum_{h=2}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}
\end{aligned}$$

where (a) uses (H.6) and (b) uses the notation

$$b_h = \begin{cases} 1 & \text{if } h = 1, \\ H - h + 2 & \text{if } 2 \le h \le H. \end{cases}$$

Since $|b_h| \le H$, by the variance-aware Freedman inequality in Lemma I.2, with probability at least $1 - \frac{\delta}{2}$, we have

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H} X_{k,h} b_h\right| \le 3H\sqrt{\iota}\sqrt{\mathcal{H}\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h}, a_{k,h}) + 10H\mathcal{H}\iota}.$$

Thus,

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h},a_{k,h}) \\
&\le 3H\sqrt{\iota}\sqrt{\mathcal{H}\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h},a_{k,h}) + 10H\mathcal{H}\iota} \\
&\qquad + 4H\beta_R\sum_{k=1}^{K}\sum_{h=2}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4H\beta_V\sum_{k=1}^{K}\sum_{h=2}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} \\
&\overset{(a)}{\le} 8H\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 8H\beta_V\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 29H^2\mathcal{H}\iota,
\end{aligned}
\tag{H.9}
$$

where (a) holds since $x \le a\sqrt{x} + b$ implies $x \le a^2 + 2b$.

Putting pieces together. Finally, plug-in (H.7), and we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\left[V_1^k(s_{k,1}) - V_1^{\pi^k}(s_{k,1})\right] \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left[X_{k,h} + 4\beta_{R,k-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right] \\
&\overset{\text{(H.8)}}{\le} 3\sqrt{\iota}\sqrt{\mathcal{H}\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})\right](s_{k,h},a_{k,h}) + 10\mathcal{H}\iota} \\
&\qquad + 4\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} \\
&\overset{\text{(H.9)}}{\le} 3\sqrt{\iota}\sqrt{\mathcal{H}\cdot\left[8H\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 8H\beta_V\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 29H^2\mathcal{H}\iota\right]} \\
&\qquad + 4\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 10\mathcal{H}\iota \\
&\le 6\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 6\beta_V\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 36H\mathcal{H}\iota
\end{aligned}
$$

where the last inequality uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $2\sqrt{ab} \le a + b$ for non-negative numbers $a, b \ge 0$. $\square$

H.10 Proof of Lemma F.10

Proof of Lemma F.10. For any $k \in [K]$, recall k_{last} is the latest update episode before episode k satisfying $k_{\text{last}} \le k$ and $V_h^k(\cdot) = V_h^{k_{\text{last}}}(\cdot), \check{V}_h^k(\cdot) = \check{V}_h^{k_{\text{last}}}(\cdot), \forall h \in [H]$.

By definition, $Q_h^k(\cdot,\cdot) \le \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k_{\text{last}}-1,h} + \widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h}\rangle + \beta_{R,k_{\text{last}}-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}} + \beta_V\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}}$ and $\check{Q}_h^k(\cdot,\cdot) \ge \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k_{\text{last}}-1,h} + \check{\boldsymbol{w}}_{k_{\text{last}}-1,h}\rangle - \beta_{R,k_{\text{last}}-1}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}} -$

$\beta_V \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}}$. Since $a_{k,h} = \pi_h^k(s_{k,h}) = \operatorname{argmax}_{a\in\mathcal{A}} Q_h^k(s_{k,h}, a)$, we have

$$
\begin{aligned}
& V_h^k(s_{k,h}) - \check{V}_h^k(s_{k,h}) \\
& \leq Q_h^k(s_{k,h}, a_{k,h}) - \check{Q}_h^k(s_{k,h}, a_{k,h}) \\
& \leq \langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \check{\boldsymbol{w}}_{k_{\text{last}}-1,h} \rangle + 2\beta_{R,k_{\text{last}}-1} \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k_{\text{last}}-1,h}^{-1}} + 2\beta_V \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}} \\
& \overset{(a)}{\leq} \left\langle \boldsymbol{\phi}_{k,h}, \left(\widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[V_{h+1}^{k_{\text{last}}}\right]\right) - \left(\check{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[\check{V}_{h+1}^{k_{\text{last}}}\right]\right)\right\rangle \\
& \qquad + \left\langle \boldsymbol{\phi}_{k,h}, \boldsymbol{w}_h\left[V_{h+1}^k - \check{V}_{h+1}^k\right]\right\rangle + 4\beta_{R,k-1} \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 4\beta_V \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} \\
& \overset{(b)}{\leq} \left[\mathbb{P}_h\left(V_{h+1}^k - \check{V}_{h+1}^k\right)\right](s_{k,h}, a_{k,h}) + 4\beta_{R,k-1} \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} + 8\beta_V \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}
\end{aligned}
$$

Here (a) uses (H.5), (b) uses

$$
\begin{aligned}
& \left\langle \boldsymbol{\phi}_{k,h}, \left(\widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[V_{h+1}^{k_{\text{last}}}\right]\right) - \left(\check{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[\check{V}_{h+1}^{k_{\text{last}}}\right]\right)\right\rangle \\
& \leq \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}} \left\|\widehat{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[V_{h+1}^{k_{\text{last}}}\right]\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}} \\
& \qquad + \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}} \left\|\check{\boldsymbol{w}}_{k_{\text{last}}-1,h} - \boldsymbol{w}_h\left[\check{V}_{h+1}^{k_{\text{last}}}\right]\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}} \\
& \leq 2\beta_V \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k_{\text{last}}-1,h}^{-1}} \\
& \leq 4\beta_V \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}
\end{aligned}
$$

on $\mathcal{E}_R \cap \mathcal{E}_V$ and

$$
\left\langle \boldsymbol{\phi}_{k,h}, \boldsymbol{w}_h\left[V_{h+1}^k - \check{V}_{h+1}^k\right]\right\rangle = \left[\mathbb{P}_h\left(V_{h+1}^k - \check{V}_{h+1}^k\right)\right](s_{k,h}, a_{k,h}).
$$

The rest of the proof is nearly the same as that in Lemma F.9 except for the constants, and we omit it to avoid repetition. □

H.11 Proof of Lemma F.11

Proof of Lemma F.11. Notice

$$
\sum_{k=1}^K \sum_{h=1}^H \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} = \sum_{k=1}^K \sum_{h=1}^H \nu_{k,h} w_{k,h}
$$

where $w_{k,h} = \left\|\boldsymbol{\phi}_{k,h}/\nu_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}}$. And we have $w_{k,h} \leq c_0 \leq 1$ by definition of $\nu_{k,h}, c_0$ and thus for all $h \in [H]$,

$$
\sum_{k=1}^K w_{k,h}^2 = \min\left\{1, w_{k,h}^2\right\} \leq 2\kappa \tag{H.10}
$$

by Lemma I.6, where κ is defined in Algorithm 3.

Next we bound the sum of bonus $\sum_{k=1}^K \sum_{h=1}^H \nu_{k,h} w_{k,h}$ separately by the value of $\nu_{k,h}$. Recall $\nu_{k,h}$ is defined in (5.4). We decompose $[K] \times [H]$ as the union of three disjoint sets $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3$ where

$$
\begin{aligned}
& \mathcal{J}_1 = \left\{(k,h) \in [K] \times [H] | \nu_{k,h} \in \{\widehat{\nu}_{k,h}, \nu_{\min}\}\right\}, \\
& \mathcal{J}_2 = \left\{(k,h) \in [K] \times [H] \middle| \nu_{k,h} = \frac{\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}}}{c_0}\right\}, \\
& \mathcal{J}_3 = \left\{(k,h) \in [K] \times [H] \middle| \nu_{k,h} = \frac{\sqrt{\max\{B, W\}}}{c_1^{\frac{1}{4}} (2\kappa)^{\frac{1}{4}}} \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}}^{\frac{1}{2}}\right\}.
\end{aligned}
$$

For the summation over $\mathcal{J}_1$, we have

$$\begin{aligned}\sum_{(k,h)\in\mathcal{J}_1} \nu_{k,h} w_{k,h} &= \sum_{(k,h)\in\mathcal{J}_1} \max\left\{\widehat{\nu}_{k,h}, \nu_{\min}\right\} w_{k,h} \\ &\stackrel{(a)}{\leq} \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}(\widehat{\nu}_{k,h}^2 + \nu_{\min}^2)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} w_{k,h}^2} \\ &\stackrel{(b)}{\leq} \sqrt{2H\kappa}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\nu}_{k,h}^2 + HK\nu_{\min}^2},\end{aligned} \tag{H.11}$$

where (a) holds due to Cauchy-Schwartz inequality and (b) holds due to (H.10).

We provide a upper bound for $\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\nu}_{k,h}^2$ in Lemma H.1.

Lemma H.1. With probability at least $1-\delta$, on event $\mathcal{E}_{R^\epsilon}\cap\mathcal{E}_R$, we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\nu}_{k,h}^2 = O\Bigg(&\mathcal{U}^*K + (\beta_{R^\epsilon} + \mathcal{H}^\epsilon\beta_R\kappa)^{\frac{2}{1+\epsilon}}\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}\right)^{\frac{2}{1+\epsilon}} \\ &+ H\nu_R^2\log\frac{2\lceil\log_2 K\rceil}{\delta}\Bigg).\end{aligned}$$

Proof. See Appendix H.13 for a detailed proof. □

Next, for the summation over $\mathcal{J}_2$, we have $\left\|\frac{\boldsymbol{\phi}_{k,h}}{\nu_{k,h}}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} = c_0$. Then

$$\sum_{(k,h)\in\mathcal{J}_2}\nu_{k,h}w_{k,h} = \frac{1}{c_0^2}\sum_{(k,h)\in\mathcal{J}_2}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} w_{k,h}^2 \stackrel{(a)}{\leq} \frac{1}{c_0^2\sqrt{\lambda_R}}\sum_{k=1}^{K}\sum_{h=1}^{H} w_{k,h}^2 \stackrel{(b)}{\leq} \frac{2H\kappa}{c_0^2\sqrt{\lambda_R}}, \tag{H.12}$$

where (a) holds due to $\boldsymbol{H}_{k-1,h}\succeq\lambda_R\boldsymbol{I}$, thus $\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} \leq \frac{\|\boldsymbol{\phi}_{k,h}\|}{\sqrt{\lambda_R}} \leq \frac{1}{\sqrt{\lambda_R}}$. And (b) holds due to (H.10).

Then, for the summation over $\mathcal{J}_3$, we have $\nu_{k,h} = \frac{\max\{B,W\}}{\sqrt{2c_1\kappa}}w_{k,h}$. Therefore

$$\sum_{(k,h)\in\mathcal{J}_3}\nu_{k,h}w_{k,h} = \frac{\max\{B,W\}}{\sqrt{2c_1\kappa}}\sum_{(k,h)\in\mathcal{J}_3} w_{k,h}^2 \leq \frac{\max\{B,W\}H\sqrt{2\kappa}}{\sqrt{c_1}}, \tag{H.13}$$

where the inequality holds due to (H.10).

Finally, combining (H.11), (H.12), (H.13) and Lemma H.1, with probability at least $1-\delta$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} = O\left(\sqrt{H\kappa}\left(\beta_{R^\epsilon} + \mathcal{H}^\epsilon\beta_R\kappa\right)^{\frac{1}{1+\epsilon}}\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}\right)^{\frac{1}{1+\epsilon}} + C\right)$$

where

$$C = \sqrt{H\kappa}\sqrt{\mathcal{U}^*K + HK\nu_{\min}^2 + H\nu_R^2\log\frac{2\lceil\log_2 K\rceil}{\delta}} + \frac{H\kappa}{c_0^2\sqrt{\lambda_R}} + \frac{\max\{B,W\}H\sqrt{\kappa}}{\sqrt{c_1}}.$$

Using the inequality that $x\leq ax^{\frac{1}{1+\epsilon}} + b$ implies $x\leq a^{\frac{1+\epsilon}{\epsilon}} + \frac{1+\epsilon}{\epsilon}b$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} = O\left((H\kappa)^{\frac{1+\epsilon}{2\epsilon}}\left(\beta_{R^\epsilon} + \mathcal{H}^\epsilon\beta_R\kappa\right)^{\frac{1}{\epsilon}} + C\right).$$

Next, we simplify the expression above by hiding logarithmic terms. Notice $\kappa = \widetilde{O}(d)$. Setting $\lambda_R = \frac{d}{\max\{B^2, W^2\}}$, we have $\beta_{R^\epsilon} = \widetilde{O}\left(\sqrt{d}\frac{\nu_{R^\epsilon}}{\nu_{\min}}K^{\frac{1-\epsilon'}{2(1+\epsilon')}}\right)$ and $\beta_R = \widetilde{O}\left(\sqrt{d}K^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)$. Therefore,

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} &= O(C) + \widetilde{O}\left(\left(d^{\frac{1+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}\beta_{R^\epsilon}\right)^{\frac{1}{\epsilon}} + \left(d^{\frac{3+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}\mathcal{H}^\epsilon\beta_R\right)^{\frac{1}{\epsilon}}\right) \\
&= O(C) + \widetilde{O}\left(\left(\frac{\nu_{R^\epsilon}d^{\frac{2+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}K^{\frac{1-\epsilon'}{2(1+\epsilon')}}}{\nu_{\min}}\right)^{\frac{1}{\epsilon}} + \left(d^{\frac{4+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}\mathcal{H}^\epsilon K^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)^{\frac{1}{\epsilon}}\right).
\end{aligned}
$$

We then simplify C as

$$
C = \widetilde{O}\left(\sqrt{dH\mathcal{U}^*K} + \sqrt{dH^2K\nu_{\min}^2} + (\nu_R + \max\{B, W\})\sqrt{d}H\right).
$$

Finally, we have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} = \widetilde{O}\Bigg(&\sqrt{dH\mathcal{U}^*K} + \sqrt{dH^2K\nu_{\min}^2} + \left(\frac{\nu_{R^\epsilon}d^{\frac{2+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}K^{\frac{1-\epsilon'}{2(1+\epsilon')}}}{\nu_{\min}}\right)^{\frac{1}{\epsilon}} \\
&+ \left(d^{\frac{4+\epsilon}{2}}H^{\frac{1+\epsilon}{2}}\mathcal{H}^\epsilon K^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)^{\frac{1}{\epsilon}} + (\nu_R + \max\{B, W\})\sqrt{d}H\Bigg).
\end{aligned}
$$

□

H.12 Proof of Lemma F.12

Proof of Lemma F.12. Denote $b_{k,h} = \|\phi_{k,h}/\sigma_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}}$, then

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} = \sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}b_{k,h}.
$$

And we have $b_{k,h} \le 1$ by definition of $\sigma_{k,h}$ and thus for all $h \in [H]$,

$$
\sum_{k=1}^{K} b_{k,h}^2 = \min\left\{1, b_{k,h}^2\right\} \le 2\kappa_V \tag{H.14}
$$

by Lemma I.6, where $\kappa_V := d\log\left(1 + \frac{K}{d\lambda_V\sigma_{\min}^2}\right)$.

Next we bound the sum of bonus $\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}b_{k,h}$ separately by the value of $\sigma_{k,h}$. Recall $\sigma_{k,h}$ is defined in (5.9). We decompose $[K]\times[H]$ as the union of three disjoint sets $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3$ where

$$
\begin{aligned}
\mathcal{J}_1 &= \left\{(k,h) \in [K]\times[H] \,\middle|\, \sigma_{k,h} \in \left\{\widehat{\sigma}_{k,h}, \sqrt{d^3HD_{k,h}}, \sigma_{\min}\right\}\right\}, \\
\mathcal{J}_2 &= \left\{(k,h) \in [K]\times[H] \,\middle|\, \sigma_{k,h} = \|\phi_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right\}, \\
\mathcal{J}_3 &= \left\{(k,h) \in [K]\times[H] \,\middle|\, \sigma_{k,h} = \sqrt{d^{\frac{5}{2}}H\mathcal{H}}\,\|\phi_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}^{\frac{1}{2}}\right\}.
\end{aligned}
$$

For the summation over $\mathcal{J}_1$, we have

$$
\begin{aligned}
\sum_{(k,h)\in\mathcal{J}_1}\sigma_{k,h}b_{k,h} &= \sum_{(k,h)\in\mathcal{J}_1}\max\left\{\widehat{\sigma}_{k,h}, \sqrt{d^3HD_{k,h}}, \sigma_{\min}\right\}\cdot b_{k,h} \\
&\overset{(a)}{\le} \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}(\widehat{\sigma}_{k,h}^2 + d^3HD_{k,h} + \sigma_{\min}^2)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}b_{k,h}^2} \\
&\overset{(b)}{\le} \sqrt{2H\kappa_V}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2 + d^3H\sum_{k=1}^{K}\sum_{h=1}^{H}D_{k,h} + HK\sigma_{\min}^2},
\end{aligned} \tag{H.15}
$$

where (a) holds due to Cauchy-Schwartz inequality and (b) holds due to (H.14).

We provide upper bounds for $\sum_{k=1}^K \sum_{h=1}^H \widehat{\sigma}_{k,h}^2$ and $\sum_{k=1}^K \sum_{h=1}^H D_{k,h}$ in Lemma H.2 and Lemma H.3 respectively.

Lemma H.2. With probability at least $1-2\delta$, on event $\mathcal{E}_0 \cap \mathcal{E}_R \cap \mathcal{E}_V \cap \mathcal{A}_0$, we have

$$\sum_{k=1}^K \sum_{h=1}^H \widehat{\sigma}_{k,h}^2 = O\bigg(\mathcal{V}^* K + (\beta_0 + H\beta_V)\mathcal{H} \sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + H\mathcal{H}\beta_R \sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + H^2\mathcal{H}^2 \log \frac{4\lceil \log_2 HK \rceil}{\delta}\bigg).$$

Proof. See Appendix H.14 for a detailed proof. $\square$

Lemma H.3. On event $\mathcal{E}_0 \cap \mathcal{A}_0$, we have

$$\sum_{k=1}^K \sum_{h=1}^H D_{k,h} = O\bigg((\beta_0 + H\beta_V)\mathcal{H} \sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + H\mathcal{H}\beta_R \sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + H^2\mathcal{H}^2 \log \frac{4\lceil \log_2 HK \rceil}{\delta}\bigg).$$

Proof. See Appendix H.15 for a detailed proof. $\square$

Next, for the summation over $\mathcal{J}_2$, we have $\left\| \frac{\boldsymbol{\phi}_{k,h}}{\sigma_{k,h}} \right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} = 1$. Then

$$\sum_{(k,h)\in\mathcal{J}_2} \sigma_{k,h} b_{k,h} = \sum_{(k,h)\in\mathcal{J}_2} \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} b_{k,h}^2 \overset{(a)}{\le} \frac{1}{\sqrt{\lambda_V}} \sum_{k=1}^K \sum_{h=1}^H b_{k,h}^2 \overset{(b)}{\le} \frac{2H\kappa_V}{\sqrt{\lambda_V}}, \tag{H.16}$$

where (a) holds due to $\boldsymbol{\Sigma}_{k-1,h} \succeq \lambda_V \boldsymbol{I}$, thus $\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} \le \frac{\|\boldsymbol{\phi}_{k,h}\|}{\sqrt{\lambda_V}} \le \frac{1}{\sqrt{\lambda_V}}$. And (b) holds due to (H.14).

Then, for the summation over $\mathcal{J}_3$, we have $\sigma_{k,h} = d^{2.5}H\mathcal{H}b_{k,h}$. Therefore

$$\sum_{(k,h)\in\mathcal{J}_3} \sigma_{k,h} b_{k,h} = d^{2.5}H\mathcal{H} \sum_{(k,h)\in\mathcal{J}_3} b_{k,h}^2 \le 2d^{2.5}H^2\mathcal{H}\kappa_V, \tag{H.17}$$

where the inequality holds due to (H.14). $\square$

Finally, combining (H.15), (H.16), (H.17) and Lemma H.2, with probability at least $1-2\delta$, we have

$$\sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} = O\left(\sqrt{H\kappa_V}\sqrt{(\beta_0 + H\beta_V)d^3 H\mathcal{H} \sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}} + C\right)$$

where

$$C = \sqrt{H\kappa_V}\sqrt{\mathcal{V}^* K + HK\sigma_{\min}^2 + d^3H^2\mathcal{H}\beta_R \sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}} + d^3H^3\mathcal{H}^2 \log \frac{4\lceil \log_2 HK \rceil}{\delta}} + \frac{H\kappa_V}{\sqrt{\lambda_V}} + d^{2.5}H^2\mathcal{H}\kappa_V.$$

Using the inequality that $x \le a\sqrt{x} + b$ implies $x \le a^2 + 2b$, we have

$$\sum_{k=1}^K \sum_{h=1}^H \|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} = O\left((\beta_0 + H\beta_V)d^3H^2\mathcal{H}\kappa_V + C\right).$$

Next, we simplify the expression above by hiding logarithmic terms. Notice $\kappa_V = \widetilde{O}(d)$. Setting $\lambda_V = \frac{1}{\mathcal{H}^2}$, we have $\beta_0 = \widetilde{O}\left(\frac{\sqrt{d^3 H \mathcal{H}^2}}{\sigma_{\min}}\right)$ and $\beta_V = \widetilde{O}(\sqrt{d})$. Therefore,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} &= O(C) + \widetilde{O}\left(d^4 H^2 \mathcal{H}\beta_0 + d^4 H^3 \mathcal{H}\beta_V\right) \\ &= O(C) + \widetilde{O}\left(\frac{d^{5.5} H^{2.5}\mathcal{H}^2}{\sigma_{\min}} + d^{4.5} H^3 \mathcal{H}\right).\end{aligned}$$

We then simplify C as

$$C = \widetilde{O}\left(\sqrt{dH\mathcal{V}^* K} + \sqrt{dH^2 K \sigma_{\min}^2} + \sqrt{d^4 H^3 \mathcal{H}\beta_R \sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}} + d^{2.5} H^2 \mathcal{H}\right).$$

Finally, we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} = \widetilde{O}\Bigg(&\sqrt{dH\mathcal{V}^* K} + \sqrt{dH^2 K \sigma_{\min}^2} + \frac{d^{5.5} H^{2.5}\mathcal{H}^2}{\sigma_{\min}} \\ &+ \sqrt{d^4 H^3 \mathcal{H}\beta_R \sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}} + d^{4.5} H^3 \mathcal{H}\Bigg).\end{aligned}$$

H.13 Proof of Lemma H.1

Proof of Lemma H.1. On event $\mathcal{E}_{R^\epsilon} \cap \mathcal{E}_R$, by Lemma F.2, for all $(k,h) \in [K]\times[H]$, we have $\left|\left[\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h - \boldsymbol{\nu}_{1+\epsilon} R_h\right](s_{k,h}, a_{k,h})\right| \le W_{k,h}$. Thus

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\nu}_{k,h}^2 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\left[\widehat{\boldsymbol{\nu}}_{1+\epsilon} R_h\right](s_{k,h}, a_{k,h}) + W_{k,h}\right)^{\frac{2}{1+\epsilon}} \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\left[\boldsymbol{\nu}_{1+\epsilon} R_h\right](s_{k,h}, a_{k,h}) + 2W_{k,h}\right)^{\frac{2}{1+\epsilon}} \\ &\overset{(a)}{\le} 2\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\boldsymbol{\nu}_{1+\epsilon} R_h\right]^{\frac{2}{1+\epsilon}}(s_{k,h}, a_{k,h}) + 2\sum_{k=1}^{K}\sum_{h=1}^{H} W_{k,h}^{\frac{2}{1+\epsilon}},\end{aligned}$$

where (a) holds due to Jensen's inequality: for any non-negative number a, b, we have $(a+b)^{1+\xi} \le 2^\xi a + 2^\xi b \le 2a + 2b$ for $\xi \in [0,1]$. We next bound the two terms in the RHS separately.

For the first term. On one hand, we have $\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\boldsymbol{\nu}_{1+\epsilon} R_h\right]^{\frac{2}{1+\epsilon}}(s_{k,h}, a_{k,h}) \le \mathcal{U}K$ by Assumption 2.7.

On the other hand, we denote $X_k = \sum_{h=1}^{H}\left[\boldsymbol{\nu}_{1+\epsilon} R_h\right]^{\frac{2}{1+\epsilon}}(s_{k,h}, a_{k,h})$. Use the notation $\mathcal{F}_k := \mathcal{F}_{H,k}$ as the σ-field generated by all the random variables up to k-th episode for short. Then $X_k \ge 0$ is $\mathcal{F}_k$-measurable with $|X_k| \le H\nu_R^2$ and $\mathrm{Var}[X_k|\mathcal{F}_{k-1}] \le \mathbb{E}[X_k^2|\mathcal{F}_{k-1}] \le H\nu_R^2 \mathbb{E}[X_k|\mathcal{F}_{k-1}]$. With probability at least $1-\delta$, the variance-aware Freedman inequality in Lemma I.2 gives

$$\begin{aligned}\sum_{k=1}^{K} X_k &\le \sum_{k=1}^{K}\mathbb{E}[X_k|\mathcal{F}_{k-1}] + 3\sqrt{H\nu_R^2 \sum_{k=1}^{K}\mathbb{E}[X_k|\mathcal{F}_{k-1}]\log\frac{2\lceil\log_2 K\rceil}{\delta}} + 5H\nu_R^2\log\frac{2\lceil\log_2 K\rceil}{\delta} \\ &\le 3\sum_{k=1}^{K}\mathbb{E}[X_k|\mathcal{F}_{k-1}] + 7H\nu_R^2\log\frac{2\lceil\log_2 K\rceil}{\delta}.\end{aligned}$$

We relate $\mathbb{E}[X_k|\mathcal{F}_{k-1}]$ to policy π^k as

$$\mathbb{E}[X_k|\mathcal{F}_{k-1}] = \mathbb{E}\left[\sum_{h=1}^{H}\left[\boldsymbol{\nu}_{1+\epsilon} R_h\right]^{\frac{2}{1+\epsilon}}(s_{k,h}, a_{k,h})|\mathcal{F}_{k-1}\right] = \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H}\left[\boldsymbol{\nu}_{1+\epsilon} R_h\right]^{\frac{2}{1+\epsilon}}(s_h, a_h)\right],$$

where $d_h^{\pi^k}(s,a)$ is defined in (F.7). Then we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}[\boldsymbol{\nu}_{1+\epsilon}R_h]^{\frac{2}{1+\epsilon}}(s_{k,h},a_{k,h}) \\
&\leq 3\sum_{k=1}^{K}\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H}[\boldsymbol{\nu}_{1+\epsilon}R_h]^{\frac{2}{1+\epsilon}}(s_h,a_h)\right]+7H\nu_R^2\log\frac{2\lceil\log_2 K\rceil}{\delta} \\
&= 3\mathcal{U}_0^*K+7H\nu_R^2\log\frac{2\lceil\log_2 K\rceil}{\delta},
\end{aligned}$$

where $\mathcal{U}_0^*$ is defined in (F.5).

For the second term. Recall $W_{k,h}$ is defined in (5.6). We have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}W_{k,h}^{\frac{2}{1+\epsilon}} &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left[(\beta_{R^\epsilon,k-1}+6\mathcal{H}^\epsilon\beta_{R,k-1}\kappa)\,\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}\right]^{\frac{2}{1+\epsilon}} \\
&\leq (\beta_{R^\epsilon}+6\mathcal{H}^\epsilon\beta_R\kappa)^{\frac{2}{1+\epsilon}}\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}\right)^{\frac{2}{1+\epsilon}},
\end{aligned}$$

where the inequality holds due to Jensen's inequality.

Finally, putting pieces together and using $\mathcal{U}^*=\min\{\mathcal{U}_0^*,\mathcal{U}\}$ complete the proof. □

H.14 Proof of Lemma H.2

Proof of Lemma H.2. On event $\mathcal{E}_0\cap\mathcal{E}_R\cap\mathcal{E}_V$, by Lemma F.6, for all $(k,h)\in[K]\times[H]$, we have $\left|\left[\widehat{\mathbb{V}}_hV_{h+1}^k-\mathbb{V}_hV_{h+1}^*\right](s_{k,h},a_{k,h})\right|\leq E_{k,h}$. Thus

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\left[\widehat{\mathbb{V}}_hV_{h+1}^k\right](s_{k,h},a_{k,h})+E_{k,h}\right) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{V}_hV_{h+1}^*\right](s_{k,h},a_{k,h})+2\sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}.
\end{aligned}$$

We next bound the two terms in the RHS separately.

For the first term. On one hand, we denote $X_k=\sum_{h=1}^{H}[\mathbb{V}_hV_{h+1}^*](s_{k,h},a_{k,h})$. Use the notation $\mathcal{F}_k:=\mathcal{F}_{k,h}$ as the σ-field generated by all the random variables up to k-th episodes for short. Then $X_k\geq 0$ is $\mathcal{F}_k$-measurable with $|X_k|\leq H\mathcal{H}^2$ and $\mathrm{Var}[X_k|\mathcal{F}_{k-1}]\leq\mathbb{E}[X_k^2|\mathcal{F}_{k-1}]\leq H\mathcal{H}^2\mathbb{E}[X_k|\mathcal{F}_{k-1}]$. With probability at least $1-\delta$, the variance-aware Freedman inequality in Lemma I.2 gives

$$\begin{aligned}
\sum_{k=1}^{K}X_k &\leq \sum_{k=1}^{K}\mathbb{E}[X_k|\mathcal{F}_{k-1}]+3\sqrt{H\mathcal{H}^2\sum_{k=1}^{K}\mathbb{E}[X_k|\mathcal{F}_{k-1}]\log\frac{2\lceil\log_2 K\rceil}{\delta}}+5H\mathcal{H}^2\log\frac{2\lceil\log_2 K\rceil}{\delta} \\
&\leq 3\sum_{k=1}^{K}\mathbb{E}[X_k|\mathcal{F}_{k-1}]+7H\mathcal{H}^2\log\frac{2\lceil\log_2 K\rceil}{\delta}.
\end{aligned}$$

We relate $\mathbb{E}[X_k|\mathcal{F}_{k-1}]$ to π^k as

$$\mathbb{E}[X_k|\mathcal{F}_{k-1}]=\mathbb{E}\left[\sum_{h=1}^{H}\left[\mathbb{V}_hV_{h+1}^*\right](s_{k,h},a_{k,h})|\mathcal{F}_{k-1}\right]=\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H}\left[\mathbb{V}_hV_{h+1}^*\right](s_h,a_h)\right],$$

where $d_h^{\pi^k}(s,a)$ is defined in (F.7). Then we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{V}_h V_{h+1}^*\right](s_{k,h},a_{k,h})\\
&\leq 3\sum_{k=1}^{K}\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[\sum_{h=1}^{H}\left[\mathbb{V}_h V_{h+1}^*\right](s_h,a_h)\right]+7H\mathcal{H}^2\log\frac{2\lceil\log_2 K\rceil}{\delta}\\
&=3\mathcal{V}_0^* K+7H\mathcal{H}^2\log\frac{2\lceil\log_2 K\rceil}{\delta},
\end{aligned}
$$

where $\mathcal{V}_0^*$ is defined in (F.6).

We restate the total variance lemma (Lemma C.5 in Jin et al. [21]) here for completeness.

Lemma H.4 (Total variance lemma)**.** With probability at least $1-\delta$, we have

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{h+1}^{\pi^k}](s_{k,h},a_{k,h})\leq 2\mathcal{V}K+2H\mathcal{H}^2\log\frac{1}{\delta}.
$$

Proof. See Appendix H.16 for a detailed proof. □

On the other hand, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{V}_h V_{h+1}^*\right](s_{k,h},a_{k,h})\\
&=\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{V}_h V_{h+1}^{\pi^k}\right](s_{k,h},a_{k,h})+\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{V}_h V_{h+1}^*-\mathbb{V}_h V_{h+1}^{\pi^k}\right](s_{k,h},a_{k,h})\\
&\overset{(a)}{\leq} 2\mathcal{V}K+2H\mathcal{H}^2\log\frac{1}{\delta}+2\mathcal{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}_h\left(V_{h+1}^k-\check{V}_{h+1}^k\right)\right]\\
&\overset{(b)}{\leq} 2\mathcal{V}K+2H\mathcal{H}^2\log\frac{1}{\delta}+16H\mathcal{H}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}+16H\mathcal{H}\beta_V\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\\
&\qquad+58H^2\mathcal{H}^2\log\frac{4\lceil\log_2 HK\rceil}{\delta}\\
&\leq 2\mathcal{V}K+16H\mathcal{H}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}+16H\mathcal{H}\beta_V\sum_{k=1}^{K}\sum_{h=1}^{H}\|\boldsymbol{\phi}_{k,h}\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\\
&\qquad+60H^2\mathcal{H}^2\log\frac{4\lceil\log_2 HK\rceil}{\delta},
\end{aligned}
$$

where (a) holds with probability at least $1-\delta$ due to Lemma H.4 and (H.18). (b) is due to $\mathcal{A}_0$ holds.

$$
\begin{aligned}
&\left[\mathbb{V}_h V_{h+1}^*-\mathbb{V}_h V_{h+1}^{\pi^k}\right](s_{k,h},a_{k,h})\\
&=[\mathbb{P}_h V_{h+1}^*]^2(s_{k,h},a_{k,h})-[\mathbb{P}_h V_{h+1}^{\pi^k}]^2(s_{k,h},a_{k,h})\\
&\qquad-\left([\mathbb{P}_h V_{h+1}^*]^2(s_{k,h},a_{k,h})-[\mathbb{P}_h V_{h+1}^{\pi^k}]^2(s_{k,h},a_{k,h})\right)\\
&\overset{(a)}{\leq}[\mathbb{P}_h V_{h+1}^*]^2(s_{k,h},a_{k,h})-\mathbb{P}_h[V_{h+1}^{\pi^k}]^2(s_{k,h},a_{k,h})\\
&\overset{(b)}{\leq}2\mathcal{H}\left[\mathbb{P}_h(V_{h+1}^*-V_{h+1}^{\pi^k})\right](s_{k,h},a_{k,h})\\
&\overset{(c)}{\leq}2\mathcal{H}\left[\mathbb{P}_h(V_{h+1}^k-V_{h+1}^{\pi^k})\right](s_{k,h},a_{k,h}),
\end{aligned}
\tag{H.18}
$$

where (a) holds due to $V_{h+1}^*(\cdot)\geq V_{h+1}^{\pi^k}(\cdot)$. (b) holds due to $V_{h+1}^*(\cdot)$ and $V_{h+1}^{\pi^k}$ are bounded in $[0,\mathcal{H}]$. And (c) holds due to Lemma F.5.

For the second term. Recall $E_{k,h}$ is defined in (5.11). We have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} E_{k,h} &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left[4\mathcal{H}\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \check{\boldsymbol{w}}_{k-1,h}\rangle + 11\mathcal{H}\beta_0 \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right] \\
&\overset{(a)}{\leq} \sum_{k=1}^{K}\sum_{h=1}^{H}\left[4\mathcal{H}\left[\mathbb{P}_h\left(V_{h+1}^k - \check{V}_{h+1}^k\right)\right](s_{k,h}, a_{k,h}) + 22\mathcal{H}\beta_0 \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right] \\
&\overset{(b)}{\leq} (22\beta_0 + 48H\beta_V)\mathcal{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 32H\mathcal{H}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} \\
&\qquad + 116H^2\mathcal{H}^2\log\frac{4\lceil\log_2 HK\rceil}{\delta} \\
&= O\Bigg((\beta_0 + H\beta_V)\mathcal{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + H\mathcal{H}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} \\
&\qquad + H^2\mathcal{H}^2\log\frac{4\lceil\log_2 HK\rceil}{\delta}\Bigg),
\end{aligned}
$$

where (a) is due to $\mathcal{E}_0$ holds and (b) is due to $\mathcal{A}_0$ holds.

Finally, putting pieces together and using $\mathcal{V}^* = \min\{\mathcal{V}_0^*, \mathcal{V}\}$ completes the proof. □

H.15 Proof of Lemma H.3

Proof of Lemma H.3. Recall $D_{k,h}$ is defined in (5.12). We have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} D_{k,h} &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left[2\mathcal{H}\langle \boldsymbol{\phi}_{k,h}, \widehat{\boldsymbol{w}}_{k-1,h} - \check{\boldsymbol{w}}_{k-1,h}\rangle + 4\mathcal{H}\beta_0 \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right] \\
&\overset{(a)}{\leq} \sum_{k=1}^{K}\sum_{h=1}^{H}\left[2\mathcal{H}\left[\mathbb{P}_h\left(V_{h+1}^k - \check{V}_{h+1}^k\right)\right](s_{k,h}, a_{k,h}) + 8\mathcal{H}\beta_0 \left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}\right] \\
&\overset{(b)}{\leq} (8\beta_0 + 24H\beta_V)\mathcal{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + 16H\mathcal{H}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} \\
&\qquad + 58H^2\mathcal{H}^2\log\frac{4\lceil\log_2 HK\rceil}{\delta} \\
&= O\Bigg((\beta_0 + H\beta_V)\mathcal{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}} + H\mathcal{H}\beta_R\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_{k,h}\right\|_{\boldsymbol{H}_{k-1,h}^{-1}} \\
&\qquad + H^2\mathcal{H}^2\log\frac{4\lceil\log_2 HK\rceil}{\delta}\Bigg),
\end{aligned}
$$

where (a) is due to $\mathcal{E}_0$ holds and (b) is due to $\mathcal{A}_0$ holds. □

H.16 Proof of Lemma H.4

Proof of Lemma H.4. The proof leverages the technique in Lemma C.5 of Jin et al. [21]. We define the filtration $\mathcal{F}_k$ as the σ-field generated by all the random variables over the first k episodes. And let $X_k = \sum_{h=1}^{H}[\mathbb{V}_h V_{h+1}^{\pi^k}](s_{k,h}, a_{k,h})$. It follows that π^k is $\mathcal{F}_{k-1}$-measurable, X_k is $\mathcal{F}_k$-measurable, and $|X_k| \le H\mathcal{H}^2$. Let $\mathbb{E}_k(\cdot) := \mathbb{E}[\cdot|\mathcal{F}_k]$ for short. We have

$$\begin{aligned}
\mathcal{V} &\overset{(a)}{\ge} \mathrm{Var}\left(\sum_{h=1}^{H} r_h(s_{k,h}, a_{k,h})\right) \\
&= \mathbb{E}_{k-1}\left[\sum_{h=1}^{H} r_h(s_{k,h}, a_{k,h}) - V_1^{\pi^k}(s_{k,1})\right]^2 \\
&\overset{(b)}{=} \mathbb{E}_{k-1}\left[\sum_{h=1}^{H}\Big(r_h(s_{k,h}, a_{k,h}) + V_{h+1}^{\pi^k}(s_{k,h+1}) - V_h^{\pi^k}(s_{k,h})\Big)\right]^2 \\
&\overset{(c)}{=} \sum_{h=1}^{H} \mathbb{E}_{k-1}\left[r_h(s_{k,h}, a_{k,h}) + V_{h+1}^{\pi^k}(s_{k,h+1}) - V_h^{\pi^k}(s_{k,h})\right]^2 \\
&= \sum_{h=1}^{H} \mathbb{E}_{k-1}\left[V_{h+1}^{\pi^k}(s_{k,h+1}) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_{k,h}, a_{k,h})\right]^2 \\
&= \mathbb{E}_{k-1}\sum_{h=1}^{H} [\mathbb{V}_h V_{h+1}^{\pi^k}](s_{k,h}, a_{k,h}) \\
&= \mathbb{E}[X_k|\mathcal{F}_{k-1}]
\end{aligned}$$

where (a) holds due to the definition of $\mathcal{V}$, (b) holds due to $V_{H+1}^{\pi^k}(\cdot) = 0$ and (c) holds due to Markov property of MDPs.

We also note that

$$\mathrm{Var}\left(\sum_{h=1}^{H} r_h(s_{k,h}, a_{k,h})\right) \le \mathcal{H}\mathbb{E}\left[\sum_{h=1}^{H} r_h(s_{k,h}, a_{k,h})\right] \le \mathcal{H}V_1^*,$$

where the second inequality holds due to the optimality of V_1^*. Thus, we actually have

$$\mathcal{V} \le \mathcal{H}V_1^*, \tag{H.19}$$

which shows the relationship between our result with the first-order regret.

Then $\mathrm{Var}[X_k|\mathcal{F}_{k-1}] \le H\mathcal{H}^2\mathbb{E}[X_k|\mathcal{F}_{k-1}]$, and we have

$$\sum_{k=1}^{K} \mathrm{Var}[X_k|\mathcal{F}_{k-1}] \le H\mathcal{H}^2 \sum_{k=1}^{K} \mathbb{E}[X_k|\mathcal{F}_{k-1}] \le H\mathcal{H}^2\mathcal{V}K.$$

By the Freedman inequality in Lemma I.1, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} [\mathbb{V}_h V_{h+1}^{\pi^k}](s_{k,h}, a_{k,h}) = \sum_{k=1}^{K} X_k &\le \sum_{k=1}^{K} \mathbb{E}[X_k|\mathcal{F}_{k-1}] + \sqrt{2H\mathcal{H}^2\mathcal{V}K\log\frac{1}{\delta}} + \frac{2}{3}H\mathcal{H}^2\log\frac{1}{\delta} \\
&\le \mathcal{V}K + 2\sqrt{H\mathcal{H}^2\mathcal{V}K\log\frac{1}{\delta}} + \frac{2}{3}H\mathcal{H}^2\log\frac{1}{\delta} \\
&\le 2\mathcal{V}K + 2H\mathcal{H}^2\log\frac{1}{\delta}.
\end{aligned}$$

□

I Auxiliary Lemmas

Lemma I.1 (Freedman inequality [13])**.** Let $\{X_t\}_{t\in[T]}$ be a stochastic process that adapts to the filtration $\mathcal{F}_t$ so that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$, $|X_t| \le M$ and $\sum_{t=1}^{T} \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \le V$ where $M > 0$ and $V > 0$ are positive constants. Then with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T} X_t \le \sqrt{2V\ln\frac{1}{\delta}} + \frac{2M}{3}\ln\frac{1}{\delta}.$$

Lemma I.2 (Variance-aware Freedman inequality, Theorem 5 in Li et al. [25])**.** Let $\{X_t\}_{t\in[T]}$ be a stochastic process that adapts to the filtration $\mathcal{F}_t$ so that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$, $|X_t| \le M$ and $\sum_{t=1}^T \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \le V^2$ where $M > 0$ and $V > 0$ are positive constants. Then with probability at least $1 - \delta$, we have

$$\left|\sum_{t=1}^T X_t\right| \le 3\sqrt{\sum_{t=1}^T \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \cdot \log\frac{2K}{\delta}} + 5M\log\frac{2K}{\delta},$$

where $K = 1 + \lceil 2\log_2 \frac{V}{M}\rceil$.

Proof of Lemma I.2. By Theorem 5 in Li et al. [25], we have for any positive integer $K \ge 1$,

$$\mathbb{P}\left(\left|\sum_{t=1}^T X_t\right| \le \sqrt{8\max\left\{\sum_{t=1}^T \mathbb{E}[X_t^2|\mathcal{F}_{t-1}], \frac{V^2}{2^K}\right\}\cdot \ln\frac{2K}{\delta}} + \frac{4M}{3}\ln\frac{2K}{\delta}\right) \ge 1 - \delta.$$

By setting $K = 1 + \lceil 2\log_2 \frac{V}{M}\rceil$, we have $\frac{V^2}{2^K} \le M^2$. Using $\max\{a, b\} \le a + b$, $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b \ge 0$ and $\ln\frac{2K}{\delta} \ge 1$, we complete the proof. □

The following two lemmas are the counterpart lemmas of Theorem 3.3 under light-tail assumption.

Lemma I.3 (Bernstein inequality for self-normalized martingales, Lemma 4.1 in He et al. [15])**.** Let $\{\mathcal{G}_t\}_{t\ge 0}$ be a filtration and $\{\boldsymbol{x}_t, \eta_t\}_{t\ge 0}$ be a stochastic process so that $\boldsymbol{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. If $\|\boldsymbol{x}_t\| \le L$ and $\{\eta_t\}_{t\ge 1}$ satisfies that $\mathbb{E}[\eta_t|\mathcal{G}_t] = 0$, $\mathbb{E}[\eta_t^2|\mathcal{G}_t] \le \sigma^2$ and $\left|\eta_t \min\left\{1, \|\boldsymbol{x}_t\|_{\boldsymbol{Z}_{t-1}^{-1}}\right\}\right| \le M$ for all $t \ge 1$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have for all $t \ge 1$,

$$\left\|\sum_{i=1}^t \boldsymbol{x}_i\eta_i\right\|_{\boldsymbol{Z}_t^{-1}} \le 8\sigma\sqrt{d\log\left(1 + \frac{tL^2}{d\lambda}\right)\log\frac{4t^2}{\delta}} + 4M\log\frac{4t^2}{\delta},$$

where $\boldsymbol{Z}_t = \lambda\boldsymbol{I} + \sum_{i=1}^t \boldsymbol{x}_i\boldsymbol{x}_i^\top$ for $t \ge 1$ and $\boldsymbol{Z}_0 = \lambda\boldsymbol{I}$.

Lemma I.4 (Hoeffding inequality for self-normalized martingales, Theorem 1 in Abbasi-Yadkori et al. [1])**.** Let $\{\mathcal{G}_t\}_{t\ge 0}$ be a filtration and $\{\boldsymbol{x}_t, \eta_t\}_{t\ge 0}$ be a stochastic process so that $\boldsymbol{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable.. If $\|\boldsymbol{x}_t\| \le L$ and $\{\eta_t\}_{t\ge 1}$ satisfies that $\mathbb{E}[\eta_t|\mathcal{G}_t] = 0$ and $|\eta_t| \le M$ for all $t \ge 1$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have for all $t \ge 1$,

$$\left\|\sum_{i=1}^t \boldsymbol{x}_i\eta_i\right\|_{\boldsymbol{Z}_t^{-1}} \le M\sqrt{d\log\left(1 + \frac{tL^2}{d\lambda}\right) + \log\frac{1}{\delta}},$$

where $\boldsymbol{Z}_t = \lambda\boldsymbol{I} + \sum_{i=1}^t \boldsymbol{x}_i\boldsymbol{x}_i^\top$ for $t \ge 1$ and $\boldsymbol{Z}_0 = \lambda\boldsymbol{I}$.

The next lemma is a general result of the concentration bounds for estimated value functions, which is the core technique used in Jin et al. [22], Zhou and Gu [41], He et al. [15]. We include the proof here for completeness.

Lemma I.5. Fix any $h \in [H]$. Consider a specific value function $f(\cdot)$ which satisfies

(1) $\sup_{s\in\mathcal{S}} |f(s)| \le C_0$;

(2) $f \in \mathcal{V}$ where $\mathcal{V}$ is a class of functions with $\mathcal{N}(\mathcal{V}, \epsilon)$ the ϵ-covering number of $\mathcal{V}$ with respective to the distance $\mathrm{dist}(f, f') := \sup_{s\in\mathcal{S}} |f(s) - f'(s)|$.

We assume there exists a deterministic $C_\sigma > 0$ and $\mathcal{A}_{k,h}$ (which is $\mathcal{F}_{k,h}$-measurable) such that $\mathcal{A}_{k,h} \subseteq \left\{\sigma_{k,h}^2 \ge [\mathbb{V}_h f](s_{k,h}, a_{k,h})/C_\sigma^2\right\}$ for all $k \in [K]$. Let $\boldsymbol{w}_h[f], \widehat{\boldsymbol{w}}_{k,h}[f]$ be defined in (5.7), (5.7) and $\sigma_{k,h}, \boldsymbol{\Sigma}_{k,h}$ be defined in our algorithm. Under any of the following conditions, with probability at least $1 - \delta/H$, it follows for all $k \in [K]$,

$$\|\widehat{\boldsymbol{w}}_{k,h}[f] - \boldsymbol{w}_h[f]\|_{\boldsymbol{\Sigma}_{k,h}} \le \beta. \tag{I.1}$$

(i) If $f(\cdot)$ is a deterministic function and $\bigcap_{k\in[K]}\mathcal{A}_{k,h}$ is true, (I.1) holds with

$$\beta=\sqrt{d\lambda_V}C_0+8C_\sigma\sqrt{d\log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right)\log\frac{4HK^2}{\delta}}+\frac{4C_0}{d^{2.5}H\mathcal{H}}\log\frac{4HK^2}{\delta}.$$

(ii) If $f(\cdot)$ is a random function and $\bigcap_{k\in[K]}\mathcal{A}_{k,h}$ is true, (I.1) holds with

$$\beta=2\sqrt{d\lambda_V}C_0+32C_\sigma\sqrt{d\log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right)\log\frac{4HK^2N_0}{\delta}}+\frac{4C_0}{d^{2.5}H\mathcal{H}}\log\frac{4HK^2N_0}{\delta}$$

where $N_0=|\mathcal{N}(\mathcal{V},\epsilon_0)|$ and $\epsilon_0=\min\left\{C_\sigma\sigma_{\min},\frac{\lambda_V C_0\sqrt{d}}{K}\sigma_{\min}^2\right\}$.

(iii) If $f(\cdot)$ is a random function, (I.1) holds with

$$\beta=2\sqrt{d\lambda_V}C_0+\frac{C_0}{\sigma_{\min}}\sqrt{d\log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right)+\log\frac{N_1}{\delta}}.$$

where $N_1=|\mathcal{N}(\mathcal{V},\epsilon_1)|$ and $\epsilon_1=\frac{\lambda_V C_0\sqrt{d}}{K}\sigma_{\min}^2$.

Proof. By definition of $\boldsymbol{w}_h[f]$,

$$\begin{aligned}
\widehat{\boldsymbol{w}}_{k,h}[f]&=\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}f(s_{i,h+1})\\
&=\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[\boldsymbol{\phi}_{i,h}^\top\boldsymbol{w}_h[f]+f(s_{i,h+1})-[\mathbb{P}_hf](s_{i,h},a_{i,h})\right]\\
&=\boldsymbol{w}_h[f]-\lambda_V\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{w}_h[f]+\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-[\mathbb{P}_hf](s_{i,h},a_{i,h})\right].
\end{aligned}$$

Then it follows that

$$\begin{aligned}
&\|\widehat{\boldsymbol{w}}_{k,h}[f]-\boldsymbol{w}_h[f]\|_{\boldsymbol{\Sigma}_{k,h}}\\
&\le\lambda_V\left\|\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{w}_h[f]\right\|_{\boldsymbol{\Sigma}_{k,h}}+\left\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-[\mathbb{P}_hf](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}}\\
&=\lambda_V\|\boldsymbol{w}_h[f]\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}+\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-[\mathbb{P}_hf](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}\\
&\le\sqrt{d\lambda_V}C_0+\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-[\mathbb{P}_hf](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}},
\end{aligned}$$

where the last inequality holds due to $\|\boldsymbol{w}_h[f]\|\le\sqrt{d}C_0$.

(i) Assume $f(\cdot)$ is a deterministic function.

We set $\mathcal{G}_i=\mathcal{F}_{i,h}$, $\boldsymbol{x}_i=\sigma_{i,h}^{-1}\boldsymbol{\phi}_{i,h}$, $\eta_i=\sigma_{i,h}^{-1}\left[f(s_{i,h+1})-[\mathbb{P}_hf](s_{i,h},a_{i,h})\right]$ and $\boldsymbol{Z}_k=\lambda_V\boldsymbol{I}+\sum_{i=1}^k\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\boldsymbol{\phi}_{i,h}^\top=\boldsymbol{H}_{k,h}$. Since $f(\cdot)$ is deterministic, $\mathcal{A}_{k,h}\in\mathcal{F}_{k,h}$. Clearly $\boldsymbol{x}_i\in\mathcal{G}_i,\mathbb{E}[\eta_i|\mathcal{G}_i]=0$ and $\mathbb{E}[\eta_i^2|\mathcal{G}_i]\le C_\sigma^2$. We also have $\|\boldsymbol{x}_i\|\le\sigma_{\min}^{-1},|\eta_i|\le C_0\sigma_{i,h}^{-1}$ and $\|\boldsymbol{x}_i\|_{\boldsymbol{Z}_{i-1}}=w_{i,h}$. As a result, $|\eta_i|\min\left\{1,\|\boldsymbol{x}_i\|_{\boldsymbol{Z}_{i-1}}\right\}\le C_0\frac{w_{i,h}}{\sigma_{i,h}}\le\frac{C_0}{d^{2.5}H\mathcal{H}}$ where the last inequality uses $\sigma_{i,h}^2\ge\mathcal{H}d^{2.5}H\|\boldsymbol{\phi}_{i,h}\|_{\boldsymbol{H}_{k-1,h}^{-1}}$. By Lemma I.3, it follows that with probability

$1-\frac{\delta}{H}$, for all $k\in[K]$,

$$
\begin{aligned}
\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-[\mathbb{P}_h f](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} &= \left\|\sum_{i=1}^{k}\boldsymbol{x}_i\eta_i\right\|_{\boldsymbol{Z}_k^{-1}} \\
&\le 8C_\sigma\sqrt{d\log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right)\log\frac{4HK^2}{\delta}}+\frac{4C_0}{d^{2.5}H\mathcal{H}}\log\frac{4HK^2}{\delta}.
\end{aligned}
$$

Finally, on the event $\bigcap_{k\in[K]}\mathcal{A}_{k,h}$, we will have all the indicator functions equal to one.

(ii) If $f(\cdot)$ is a random function, covering arguments are used to handle the possible correlation between $f(\cdot)$ and history data.

Denote the ϵ_0-net of $\mathcal{V}$ by $\mathcal{N}(\mathcal{V},\epsilon_0)$ where $\epsilon_0=\min\left\{C_\sigma\sigma_{\min},\frac{\lambda_V C_0\sqrt{d}}{K}\sigma_{\min}^2\right\}$. Hence, for any $f\in\mathcal{V}$, there exists $\bar{f}\in\mathcal{N}(\mathcal{V},\epsilon_0)$ such that $\|\bar{f}-f\|_\infty=\sup_{s\in\mathcal{S}}|f(s)-\bar{f}(s)|\le\epsilon_0$. Then,

$$
\begin{aligned}
&\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-[\mathbb{P}_h f](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\le \underbrace{\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[\bar{f}(s_{i,h+1})-[\mathbb{P}_h \bar{f}](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}}_{\text{(i)}} \\
&\qquad+\underbrace{\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-\bar{f}(s_{i,h+1})-[\mathbb{P}_h(f-\bar{f})](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}}_{\text{(ii)}}.
\end{aligned}
$$

For term (ii), due to $\|\boldsymbol{\phi}_{i,h}\|\le 1$ and $|f(s_{i,h+1})-\bar{f}(s_{i,h+1})-[\mathbb{P}_h(f-\bar{f})](s_{i,h},a_{i,h})|\le\epsilon_0$, we have

$$
\begin{aligned}
&\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[f(s_{i,h+1})-\bar{f}(s_{i,h+1})-[\mathbb{P}_h(f-\bar{f})](s_{i,h},a_{i,h})\right]\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\le \frac{K\epsilon_0}{\sigma_{\min}^2\sqrt{\lambda_V}}\le\sqrt{d\lambda_V}C_0.
\end{aligned}
$$

We next bound term (i) as follows. On the event $\mathcal{A}_{k,h}$, by definition of ϵ_0,

$$
\begin{aligned}
[\mathbb{V}_h\bar{f}](s_{k,h},a_{k,h}) &\le 2[\mathbb{V}_h f](s_{k,h},a_{k,h})+2[\mathbb{V}_h(\bar{f}-f)](s_{k,h},a_{k,h}) \\
&\le 2C_\sigma^2\sigma_{k,h}^2+2\epsilon_0^2\le 4C_\sigma^2\sigma_{k,h}^2.
\end{aligned}
$$

For any fixed $f'\in\mathcal{V}$, we set $\mathcal{G}_i=\mathcal{F}_{i,h}$, $\boldsymbol{x}_i=\sigma_{i,h}^{-1}\boldsymbol{\phi}_{i,h}$, $\eta_i=\sigma_{i,h}^{-1}\left[\bar{f}(s_{i,h+1})-[\mathbb{P}_h\bar{f}](s_{i,h},a_{i,h})\right]\mathbb{1}_{\mathcal{A}_{k,h}}$ and $\boldsymbol{Z}_k=\lambda_V\boldsymbol{I}+\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\boldsymbol{\phi}_{i,h}^\top=\boldsymbol{H}_{k,h}$. Moreover, due to the choice of $\sigma_{i,h}$, it follows that $|\eta_i|\min\left\{1,\|\boldsymbol{x}_i\|_{\boldsymbol{Z}_{i-1}}\right\}\le C_0\frac{w_{i,h}}{\sigma_{i,h}}\le\frac{C_0}{d^{2.5}H\mathcal{H}}$. By Lemma I.3, it follows that with probability $1-\frac{\delta}{H}$, for all $k\in[K]$,

$$
\begin{aligned}
&\sup_{f'\in\mathcal{N}(\mathcal{V},\epsilon_0)}\left\|\sum_{i=1}^{k}\sigma_{i,h}^{-2}\boldsymbol{\phi}_{i,h}\left[\bar{f}(s_{i,h+1})-[\mathbb{P}_h\bar{f}](s_{i,h},a_{i,h})\right]\mathbb{1}_{\mathcal{A}_{k,h}}\right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\le 32C_\sigma\sqrt{d\log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right)\log\frac{4HK^2N_0}{\delta}}+\frac{4C_0}{d^{2.5}H\mathcal{H}}\log\frac{4HK^2N_0}{\delta}.
\end{aligned}
$$

where $N_0=|\mathcal{N}(\mathcal{V},\epsilon_0)|$.

As a result, on the event $\bigcap_{k\in[K]}\mathcal{A}_{k,h}$, all the indicator functions equal to one, which completes the proof.

(iii) The proof is almost similar to the second item except that we use Lemma I.4 to analyze the term (I). Noticing we also have $|\eta_i| = |\sigma_{i,h}^{-1}\left[\bar{f}(s_{i,h+1}) - [\mathbb{P}_h\bar{f}](s_{i,h},a_{i,h})\right]| \le \frac{C_0}{\sigma_{\min}}$. By Lemma I.4, it follows that with probability $1-\frac{\delta}{H}$, for all $k \in [K]$,

$$\sup_{f' \in \mathcal{N}(\mathcal{V},\epsilon_0)} \left\| \sum_{i=1}^{k} \sigma_{i,h}^{-2}\phi_{i,h}\left[\bar{f}(s_{i,h+1}) - [\mathbb{P}_h\bar{f}](s_{i,h},a_{i,h})\right] \right\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}$$
$$\le \frac{C_0}{\sigma_{\min}}\sqrt{d\log\left(1+\frac{K}{\sigma_{\min}^2 d\lambda_V}\right)+\log\frac{N_1}{\delta}},$$

where $N_1 := |\mathcal{N}(\mathcal{V},\varepsilon_1)|$.

□

Lemma I.6 (Lemma 11 in Abbasi-Yadkori et al. [1])**.** Let $\{\boldsymbol{x}_t\}_{t\in[T]} \subset \mathbb{R}^d$ and assume $\|\boldsymbol{x}_t\| \le L$ for all $t \in [T]$. Set $\boldsymbol{\Sigma}_t = \sum_{s=1}^{t}\boldsymbol{x}_t\boldsymbol{x}_t^\top + \lambda\boldsymbol{I}$. Then it follows that

$$\sum_{t=1}^{T}\min\left\{1, \|\boldsymbol{x}_t\|^2_{\boldsymbol{\Sigma}_{t-1}^{-1}}\right\} \le 2d\log\left(1+\frac{TL^2}{d\lambda}\right).$$

Lemma I.7 (Lemma 12 in Abbasi-Yadkori et al. [1])**.** Suppose $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying that $\boldsymbol{A} \succeq \boldsymbol{B}$, then for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\|\boldsymbol{x}\|_{\boldsymbol{B}^{-1}} \le \|\boldsymbol{x}\|_{\boldsymbol{A}^{-1}}\sqrt{\frac{\det(\boldsymbol{A})}{\det(\boldsymbol{B})}}.$$

The following lemmas are concerning function class and covering number, which are also used in He et al. [15]. Let $\mathcal{K} = \{k_1, k_2, \cdots\}$ denote the set of episodes where the algorithm updates the value function in Algorithm 3. For a fixed number of episodes K, we have $|\mathcal{K}| \le K$ trivially. Lemma I.8 shows $|\mathcal{K}|$ is only logarithmically related to K due to the mechanism of rare-switching value function updates.

Lemma I.8.

$$|\mathcal{K}| \le dH\log_2\left(1+\frac{K}{\lambda_R\nu_{\min}^2}\right)\left(1+\frac{K}{\lambda_V\sigma_{\min}^2}\right).$$

Proof. According to the updating policy, for each episode k_i, there exists a stage $h' \in [H]$ such that $\det(\boldsymbol{H}_{k_i-1,h'}) \ge 2\det(\boldsymbol{H}_{k_{i-1}-1,h'})$ or $\det(\boldsymbol{\Sigma}_{k_i-1,h'}) \ge 2\det(\boldsymbol{\Sigma}_{k_{i-1}-1,h'})$. Since we always have $\boldsymbol{H}_{k_i-1,h} \succeq \boldsymbol{H}_{k_{i-1}-1,h}$ and $\boldsymbol{\Sigma}_{k_i-1,h} \succeq \boldsymbol{\Sigma}_{k_{i-1}-1,h}$ for all $h \in [H]$, it then follows that

$$\prod_{h\in[H]}\det(\boldsymbol{H}_{k_i-1,h})\prod_{h\in[H]}\det(\boldsymbol{\Sigma}_{k_i-1,h}) \ge 2\prod_{h\in[H]}\det(\boldsymbol{H}_{k_{i-1}-1,h})\prod_{h\in[H]}\det(\boldsymbol{\Sigma}_{k_{i-1}-1,h}).$$

By induction, it follows that

$$\begin{aligned}\prod_{h\in[H]}\det(\boldsymbol{H}_{h,k_{|\mathcal{K}|}-1})\prod_{h\in[H]}\det(\boldsymbol{\Sigma}_{h,k_{|\mathcal{K}|}-1}) &\ge 2^{|\mathcal{K}|}\prod_{h\in[H]}\det(\boldsymbol{H}_{k_1-1,h})\prod_{h\in[H]}\det(\boldsymbol{\Sigma}_{k_1-1,h})\\ &\ge 2^{|\mathcal{K}|}\prod_{h\in[H]}\det(\lambda_R\boldsymbol{I})\prod_{h\in[H]}\det(\lambda_V\boldsymbol{I})\\ &= 2^{|\mathcal{K}|}(\lambda_R\lambda_V)^{dH}.\end{aligned}$$

On the other hand, due to $\boldsymbol{H}_{h,k_{|\mathcal{K}|}-1} \preceq \boldsymbol{H}_{h,K}$, the determinant $\det(\boldsymbol{H}_{h,k_{|\mathcal{K}|}-1})$ is upper bounded by

$$\prod_{h\in[H]}\det(\boldsymbol{H}_{h,k_{|\mathcal{K}|}-1}) \le \prod_{h\in[H]}\det(\boldsymbol{H}_{h,K}) \le \left(\lambda_R+\frac{K}{\nu_{\min}^2}\right)^{dH}.$$

A similar result holds for $\prod_{h\in[H]}\det(\boldsymbol{\Sigma}_{h,k_{|\mathcal{K}|}-1})$. Finally, we have

$$|\mathcal{K}| \le dH\log_2\left(1+\frac{K}{\lambda_R\nu_{\min}^2}\right)\left(1+\frac{K}{\lambda_V\sigma_{\min}^2}\right).$$

The proof is completed. □

The optimistic value function $V_h^k(\cdot) = \min_{k_i \le k} \max_a Q_h^{k_i}(\cdot, a)$ belongs to the function class $\mathcal{V}^+$

$$\mathcal{V}^+ = \Big\{ f | f(\cdot) = \max_{a \in \mathcal{A}} \min_{i \le |\mathcal{K}|} \min \Big\{ \langle \boldsymbol{\phi}(\cdot, a), \boldsymbol{\theta}_i + \boldsymbol{w}_i \rangle + \beta_R \|\boldsymbol{\phi}(\cdot, a)\|_{\boldsymbol{H}_i^{-1}} + \beta_V \|\boldsymbol{\phi}(\cdot, a)\|_{\boldsymbol{\Sigma}_i^{-1}}, \mathcal{H} \Big\}, \\ \|\boldsymbol{\theta}_i\| \le B, \|\boldsymbol{w}_i\| \le L, \boldsymbol{H}_i \succeq \lambda_R \boldsymbol{I}, \boldsymbol{\Sigma}_i \succeq \lambda_V \boldsymbol{I} \Big\}. \tag{I.2}$$

while the pessimistic value function $\check{V}_h^k(\cdot) = \max_{k_i \le k} \max_a \check{Q}_h^{k_i}(\cdot, a)$ belong to the function class $\mathcal{V}^-$,

$$\mathcal{V}^- = \Big\{ f | f(\cdot) = \max_{a \in \mathcal{A}} \max_{i \le |\mathcal{K}|} \max \Big\{ \langle \boldsymbol{\phi}(\cdot, a), \boldsymbol{\theta}_i + \boldsymbol{w}_i \rangle - \beta_R \|\boldsymbol{\phi}(\cdot, a)\|_{\boldsymbol{H}_i^{-1}} - \beta_V \|\boldsymbol{\phi}(\cdot, a)\|_{\boldsymbol{\Sigma}_i^{-1}}, 0 \Big\}, \\ \|\boldsymbol{\theta}_i\| \le B, \|\boldsymbol{w}_i\| \le L, \boldsymbol{H}_i \succeq \lambda_R \boldsymbol{I}, \boldsymbol{\Sigma}_i \succeq \lambda_V \boldsymbol{I} \Big\}. \tag{I.3}$$

Here B and $L = \mathcal{H}\sqrt{\frac{dK}{\lambda_V}}$ are uniform upper bounds of $\boldsymbol{\theta}_i$ and $\boldsymbol{w}_i$ respectively (See Lemma E.2 of He et al. [15] for its proof).

Lemma I.9 gives the covering number of function class $\mathcal{V}^+, \mathcal{V}^-$ and the squared version.

Lemma I.9 (Covering number of value functions)**.** Let $\mathcal{V}^\pm$ denote the class of optimistic or pessimistic value functions with definition in (I.2) and (I.3) respectively. And denote the squared version of $\mathcal{V}$ as $[\mathcal{V}]^2 := \{f^2 | f \in \mathcal{V}\}$. Let $\mathcal{N}(\mathcal{V}, \epsilon)$ be the ϵ-covering number of $\mathcal{V}$ with respective to the distance $\mathrm{dist}(f, f') := \sup_{s \in \mathcal{S}} |f(s) - f'(s)|$. Then,

$$\log \mathcal{N}(\mathcal{V}^\pm, \epsilon) \le \left[d \log \left(1 + \frac{4(B+L)}{\epsilon} \right) + d^2 \log \left(1 + \frac{8 d^{1/2} \beta_{R,K}^2}{\lambda_R \epsilon^2} \right) \left(1 + \frac{8 d^{1/2} \beta_V^2}{\lambda_V \epsilon^2} \right) \right] |\mathcal{K}|,$$

$$\log \mathcal{N}([\mathcal{V}^+]^2, \epsilon) \le \left[d \log \left(1 + \frac{8H(B+L)}{\epsilon} \right) + d^2 \log \left(1 + \frac{32 d^{1/2} H^2 \beta_{R,K}^2}{\lambda_R \epsilon^2} \right) \left(1 + \frac{32 d^{1/2} H^2 \beta_V^2}{\lambda_V \epsilon^2} \right) \right] |\mathcal{K}|,$$

where $|\mathcal{K}|$ is the number of episodes where the algorithm updates the value function in Algorithm 3.

Proof. The proof is nearly identical to Lemma E.6, E.7, E.8 in He et al. [15] except for the following differences. First, the weight in dot product is $\boldsymbol{\theta}_i + \boldsymbol{w}_i$ instead of $\boldsymbol{w}_i$, so L is replaced by $B + L$ in the results. Second, since we maintain two different matrices $\boldsymbol{H}_i$ and $\boldsymbol{\Sigma}_i$, the term with respect to the extra matrix is added accordingly. $\square$