

Received 17 June 2024, accepted 30 June 2024, date of publication 8 July 2024, date of current version 14 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3424828

## THEORY

# Heavy-Tailed Reinforcement Learning With Penalized Robust Estimator

**HYEON-JUN PARK AND KYUNGJAE LEE** 

Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Kyungjae Lee (kyungjae.lee@ai.cau.ac.kr)

This work was supported in part by the Institute of Information Communications Technology Planning Evaluation (IITP) through the Korean Government [Ministry of Science and ICT (MSIT)], Artificial Intelligence Graduate School Program (Chung-Ang University) under Grant 2021-0-01341; and in part by the Chung-Ang University Young Scientist Scholarship in 2022.

**ABSTRACT** We consider finite-horizon episodic reinforcement learning (RL) under heavy-tailed noises, where the  $p$ -th moment is bounded for any  $p \in (1, 2]$ . In this setting, existing RL algorithms are limited by their requirement for prior knowledge about the bounded moment order of the noise distribution. This requirement hinders their practical application, as such prior information is rarely available in real-world scenarios. Our proposed method eliminates the need for this prior knowledge, enabling implementation in a wider range of scenarios. We introduce two RL algorithms,  $p$ -Heavy-UCRL and  $p$ -Heavy-Q-learning, designed for model-based and model-free RL settings, respectively. Without the need for prior knowledge, these algorithms demonstrate robustness to heavy-tailed noise and achieve nearly optimal regret bounds, up to logarithmic terms, with the same dependencies on dominating terms as existing algorithms. Finally, we show that our proposed algorithms have empirically comparable performance to existing algorithms in synthetic tabular scenario.


**INDEX TERMS** Reinforcement learning, heavy-tailed noise, regret analysis.

## I. INTRODUCTION

Reinforcement Learning (RL) [16] has emerged as a critical paradigm in the training of intelligent agents, enabling them to make optimal decisions through interactions with their environment. This approach has been successfully applied across various domains, including soft robotics [21], portfolio management [27], and autonomous driving [17]. To accurately reflect the inherent randomness present in real-world applications, RL approaches typically employ noise assumptions. Traditionally, the RL framework assumes noise with bounded characteristics or sub-Gaussian distributions, which has been extensively examined in the RL literature [6], [8], [11], [13], [14], [20]. However, real-world scenarios often present complexities that challenge these conventional noise assumptions, extending beyond the scope of sub-Gaussian noise. Examples include the fields of finance [4], meteorology [9], and network communication [3], where noise characteristics can be more complex. In response to

these challenges, the research landscape of RL has progressively expanded to encompass a broader spectrum of noise assumptions [25]. A notable advancement in this direction is the incorporation of heavy-tailed noise distributions into the RL framework. These heavy-tailed distributions are adept at modeling the presence of rare but problematic events, which can substantially impede the learning process of an agent.

Specifically, researchers in the field of multi-armed bandits (MABs), which represent the simplest RL problem, have investigated various settings characterized by heavy-tailed noise distributions [7], [10], [15], [18], [19], [22], [23], [24], [26]. The objective of these bandit algorithms is to identify an optimal sampling strategy in the presence of heavy-tailed noise, primarily leveraging the optimism in the face of uncertainty (OFU) framework [2]. Algorithms based on OFU utilize the confidence interval of a mean estimator of the rewards to determine the sampling strategy. Consequently, a fundamental approach to managing heavy-tailed noise entails establishing concentration inequalities of estimator with exponentially decaying error bounds, similar to the Azuma-Hoeffding or Bernstein inequalities [1] commonly

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaojie Su .

applied in sub-Gaussian or bounded noise distribution settings. This approach necessitates the development of a robust estimator capable of accurately estimating the true reward, even in the presence of heavy-tailed noise.

Interestingly, [7] proposed such estimators having exponential decaying error bound on reward mean estimation, and the estimators have been widely used to deal with heavy-tailed rewards in bandits studies [10], [15], [18], [19], [23]. However, these estimators necessitate prior knowledge of the bounded moment order of the heavy-tailed noise distribution to be defined. In real-world scenarios, it is often unrealistic to expect such prior knowledge about the noise distribution to be available. This limitation poses constraints on the practical application of algorithms developed with the robust estimators for handling heavy-tailed noise. In RL, [25] developed robust algorithms to handle heavy-tailed noise in finite-horizon episodic Markov Decision Processes (MDPs), utilizing the robust estimators provided in [7]. As a result, the proposed RL algorithms are constrained in situations where the prior information about the noise distribution is not known in advance.

In this paper, we introduce two algorithms that enhance the practicality of existing algorithms by eliminating the need for prior knowledge of the noise distribution. The proposed algorithms maintain the same theoretical guarantee as previous algorithms up to logarithmic terms, without requiring prior knowledge of the moment order of the heavy-tailed noise distribution. This improvement is crucial for real-world applications where such prior information is often unavailable. Specifically, we utilize the  $p$ -robust estimator provided by [22], which can be defined without prior knowledge, instead of the estimators proposed by [7] that require such knowledge. By using this estimator, we introduce two RL algorithms,  $p$ -Heavy-UCRL2 and  $p$ -Heavy-Q-learning, which are designed for model-based and model-free RL settings, respectively. Since the confidence interval of the  $p$ -robust estimator shares the same exponential convergence rate as the robust estimators used in [25], up to logarithmic terms, the regret bound of the proposed algorithms is also comparable to that of the existing algorithms. Finally, we show that the experimental performance of our algorithms is almost equivalent to that of the existing algorithms in synthetic tabular setting.

## II. RELATED WORK

In reinforcement learning (RL), bounded noise or sub-Gaussian noise assumptions have long been investigated [6], [8], [11], [13], [14], [20], and more recently, this assumption has been extended to include heavy-tailed noise settings.

In the context of multi-armed bandits (MABs), which can be viewed as the stateless RL, several works have aimed to develop learning algorithms under heavy-tailed noises. Reference [7] first considered the stochastic MAB problem in the heavy-tailed noise setting. Specifically, they introduced robust mean estimators such as truncation and median-of-means and employed them to analyze a robust

bandit algorithm, referred to as RobustUCB, which is an adaptation of the classical UCB algorithm [2]. Subsequently, the truncation and median-of-means estimators have found broad application across various scenarios to address heavy-tailed noises, such as linear bandits [10], [15], [23], Lipschitz bandits [18], and Bayesian optimization [19]. Yet, using the truncation and median-of-means estimators is hindered by a fundamental limitation: the prerequisite knowledge of moment bounds pertaining to heavy-tailed noise distributions. This constraint arises from the definitions associated with these estimators, which impose constraints on the practical utilization of the truncation and median-of-means estimators. Interestingly, [22] tackled this limitation by introducing the  $p$ -robust estimator. Unlike its counterparts, this estimator operates independently of prior knowledge while having a slightly looser confidence bound compared to the truncation and median-of-means estimators. In addition, [22] showed that there is a finite-armed stochastic bandit problem for which the Robust-UCB has an unavoidable sub-optimal factor  $\ln(T)^{1-1/p}$  where  $p$  is a bounded moment order of the heavy-tailed noise distribution and proposed a perturbation-based algorithm  $APE^2$  that achieves minimax optimal regret bound  $\tilde{O}(T^{1/p})$  in terms of  $T$ .

As aforementioned, MABs under heavy-tailed noises have been studied and improved over the last decade. However, existing studies in the RL field have mainly focused on light-tailed noise settings (e.g., sub-Gaussian). Here, we will briefly review only the works closely related to our results. [6] introduced the well-known model-based RL algorithm, UCRL2, with a total regret bound  $\tilde{O}(DS\sqrt{AT})$  where  $D$  is a diameter of the given MDP,  $S$  is states, and  $A$  represents actions. In addition, [6] showed that for any RL algorithms, we can choose an MDP in which the lower bound of the algorithm matches  $\Omega(\sqrt{DSAT})$ . Following this work, [12] proposed the Bayesian RL algorithm, PSRL, which combines UCRL2 with Thompson sampling and proved that PSRL can achieve the Bayesian regret bound  $\tilde{O}(D\sqrt{SAT})$ , which is an improvement over UCRL2 by a factor of  $O(\sqrt{H})$ . In model-free RL algorithms, [14] adopted Q-learning with Hoeffding and Bernstein-style bonus terms in a finite-horizon episodic MDP setting, which attains nearly optimal regret bound in terms of  $T$  under some optimistic initialization of value functions. While there have been advancements, these algorithms are still limited by the sub-Gaussian noise assumption.

Notably, [25] first studied RL under heavy-tailed noises. Specifically, they proposed two RL algorithms, Heavy-UCRL2 and Heavy-Q-learning, which are designed in model-based and model-free setups, respectively. However, like most of the works on MABs under heavy-tailed noises, [25] utilized the truncation estimator which requires prior knowledge about the bounded moment order of heavy-tailed noises. Thus, Heavy-UCRL2 and Heavy-Q-learning assume the prior knowledge is known in advance, which hampers the practical application of the algorithms. In this work, we will relax this constraint on Heavy-UCRL2 and

Heavy-Q-learning. More specifically, we propose robust RL algorithms,  $p$ -Heavy-UCRL2 and  $p$ -Heavy-Q-learning, by leveraging the  $p$ -robust estimator. Especially, the proposed algorithms do not need any prior knowledge and enjoy the same regret bound of Heavy-UCRL2 and Heavy-Q-learning in terms of  $H$ ,  $S$ ,  $A$ , and  $T$  up to logarithmic terms.

### III. PROBLEM FORMULATION

Consider the problem of reinforcement learning (RL) whose goal is to find optimal sequential decisions to maximize cumulative reward over given rounds. RL can be formalized using the concept of an Markov decision process (MDP)  $\mathcal{M} = \mathcal{M}(\mathcal{S}, \mathcal{A}, p, r)$ , where  $\mathcal{S}$  is a finite state space,  $\mathcal{A}$  is a finite action space,  $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  is the transition model, and the stochastic reward function  $r : \mathcal{S} \times \mathcal{A} \rightarrow [r_{\min}, r_{\max}]$ . We denote the  $S$  and  $A$  denotes the cardinality of  $\mathcal{S}$  and  $\mathcal{A}$ , respectively. In RL, the reward and transition functions may not be fully known to the agent. Then, the agent interacts with the environment over a sequence of discrete time steps to learn about reward and transition models and to optimize its behavior. In each time step  $t \in [T]$ , an agent in state  $s_t \in \mathcal{S}$  selects an action  $a_t \in \mathcal{A}$  and receives a reward  $r_t$ , which is an independent and identically distributed random variable sampled from the reward distribution  $R(s_t, a_t)$ . After receiving the reward, the agent transitions to the next state  $s' \in \mathcal{S}$  according to the transition model  $p(s'|s, a)$ , where  $\sum_{s' \in \mathcal{S}} p_{s,a}(s') = 1$ . Following [25], we assume that the reward distribution can be a heavy-tailed distribution with finite moment order of  $p$  where  $p \in (1, 2]$ . Furthermore, we make assumption on MDPs as follows:

*Definition 1 (Diameter of MDP [6]):* Consider the stochastic process defined by a stationary policy  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  operating on an MDP  $\mathcal{M}$  with initial state  $s$ . Let  $T(s'|\mathcal{M}, \pi, s)$  be the random variable for the first time step in which state  $s'$  is reached in this process. Then the diameter of  $\mathcal{M}$  is defined as

$$D(\mathcal{M}) := \max_{s \neq s'} \min_{\pi: \mathcal{S} \rightarrow \mathcal{A}} \mathbb{E}[T(s'|\mathcal{M}, \pi, s)]. \quad (1)$$

An MDP  $\mathcal{M}$  is called a *communicating MDP* if and only if it has a finite diameter. In addition, we assume finite-horizon episodic MDPs where the agent interacts with the MDP over  $T = HK$  total rounds, where  $H$  is the horizon length and  $K$  is the number of episodes. For a policy  $\pi$ , the  $Q$  value function is defined as  $Q_h^\pi(s, a) := \mathbb{E}[\sum_{h'=h}^H r_{h'}|s_h = s, a_h = a, \pi]$ . Then, the value function is  $V_h^\pi(s) = \mathbb{E}[\sum_{h'=h}^H r_{h'}|s_h = s, \pi]$ . For any  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$ , the optimal  $Q$ -value and value functions are denoted as  $Q_h^*(s, a) := \max_\pi Q_h^\pi(s, a)$  and  $V_h^*(s) := \max_\pi V_h^\pi(s, a)$ . For simplicity, we set  $V_{H+1}(s) = 0$  and  $Q_{H+1}(s, a) = 0$ , respectively. Then, the goal of the agent is to minimize the *cumulative regret*,

$$\mathcal{R}_T := \sum_{t=1}^T V_1^*(s_1) - V_1^\pi(s_1), \quad (2)$$

where  $\pi$  is the chosen policy by the learner.

### IV. METHOD

In this section, we introduce two algorithms tailored for heavy-tailed noise setting. The first algorithm,  $p$ -Heavy-UCRL2, is an adaptation of UCRL2 [6], specifically designed to handle heavy-tailed noises. The second algorithm,  $p$ -Heavy-Q-learning, is a robust modification of Q-learning algorithm [14] intended for heavy-tailed noises as well. Both algorithms leverage the  $p$ -robust estimator proposed by [22]. The formal definition of the  $p$ -robust estimator is as follows:

*Definition 2 (p-Robust Estimator [22]):* Let  $\{Y_k\}_{k=1}^\infty$  be independent and identically distributed (i.i.d.) random variables sampled from a heavy-tailed distribution with a finite  $p$ -th moment,  $v_p := \mathbb{E}|Y_k|^p$ , for any  $p \in (1, 2]$ . Let  $y := \mathbb{E}[Y_k]$  and define an estimator as follows

$$\hat{Y}_n := \frac{c}{n^{1-\frac{1}{p}}} \sum_{k=1}^n \psi_p(Y_k/(cn^{1/p})) \quad (3)$$

where  $c > 0$  is a constant. Then, for all  $\epsilon > 0$ , we have

$$\mathbb{P}(\hat{Y}_n > y + c \ln(\exp(b_p v_p)/c^p)/\delta)/n^{1-\frac{1}{p}} \leq \delta \quad (4)$$

and

$$\mathbb{P}(y > \hat{Y}_n + c \ln(\exp(b_p v_p)/c^p)/\delta)/n^{1-\frac{1}{p}} \leq \delta. \quad (5)$$

---

#### Algorithm 1 $p$ -Heavy-UCRL2

---

- input**  $\delta \in (0, 1)$ ,  $\mathcal{S}$ ,  $\mathcal{A}$ ,  $p$ -robust estimator  $\hat{r}$ , parameter  $c$
- 1:  $t \leftarrow 1$ , initial state  $s_1$
  - 2: **for** episode  $k = 1, 2, \dots, K$  **do**
  - 3:  $t_k \leftarrow t$ ,  $N_k(s, a) \leftarrow 0$
  - 4: For all  $(s, a) \in \mathcal{S} \times \mathcal{A}$  initialize  $v_k(s, a)$  to 0, where  $v_k(s, a)$  is a visitation of state-action pair  $(s, a)$  in the current episode
  - 5: Set  $\mathcal{M}_k$  be the set of all MDPs with states  $\mathcal{S}$  and actions  $\mathcal{A}$  with transitions  $\hat{p}_k(\cdot|s, a)$  and rewards  $\hat{r}(s, a)$  satisfying the following inequalities:

$$|\tilde{r}(s, a) - \hat{r}_k(s, a)| \leq \frac{7c \log(2SA t_k/\delta)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \quad (6)$$

$$|\tilde{p}(\cdot|s, a) - \hat{p}_k(\cdot|s, a)|_1 \leq \sqrt{\frac{14S \log(2A t_k/\delta)}{\max\{1, N_k(s, a)\}}} \quad (7)$$

- 6: Obtain policy  $\tilde{\pi}_k$  by using extended value iteration and select action according to  $\tilde{\pi}_k$  until  $v_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_k(s_t, \tilde{\pi}_k(s_t))\}$
  - 7: **end for**
- 

Note that the  $p$ -robust estimator consists of an influence function defined as

$$\psi_p := \ln(b_p|x|^p + x + 1)\mathbb{I}[x \geq 0] - \ln(b_p|x|^p - x + 1)\mathbb{I}[x < 0] \quad (8)$$

where  $b_p := \left[2\left(\frac{2-p}{p-1}\right)^{1-\frac{2}{p}} + \left(\frac{2-p}{p-1}\right)^{2-\frac{2}{p}}\right]^{-\frac{2}{p}}$ . Intuitively, the influence function behaves linearly around a neighborhood

**Algorithm 2**  $p$ -Heavy- $Q$  Learning

---

**Require:**  $\delta \in (0, 1), \mathcal{S}, \mathcal{A}$ , parameter  $c$

- 1: Initialize  $Q_h \leftarrow Hr_{max}$  and  $N_h(s, a) \leftarrow 0$  for all  $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$
- 2: **for** episode  $k = 1, 2, \dots, K$  **do**
- 3:   **for** step  $h = 1, 2, \dots, H$  **do**
- 4:     Select action  $a_{k,h} \leftarrow \arg \max_{a'} Q_{h,k}(x_h, a_h)$
- 5:      $t \leftarrow N_h(s_h, a_h) + 1$
- 6:      $b'_t \leftarrow b_t + 2H(c \ln(2SAT)/\delta)/t^{1-\frac{1}{p}}$
- 7:      $\alpha_t \leftarrow \frac{H+1}{H+t}$
- 8:      $Q_h \leftarrow (1 - \alpha_t)Q_h(x_h, a_h) + \alpha_t(ct^{\frac{1}{p}}\psi_p(r_h(s_h, a_h)/ct^{\frac{1}{p}}) + V_{h+1} + b'_t)$
- 9:   **end for**
- 10: **end for**

---

of 0, however, when  $x$  becomes large enough, the influence function has a form similar to the logarithmic function. With this property, rewards contaminated by heavy-tailed noises can be regularized while preserving meaningful information. Furthermore, it is important to note that the definition of  $p$ -robust estimator eliminates the need for any prior knowledge about the bounded moment of heavy-tailed noises. This feature facilitates the development of robust reinforcement learning algorithms without requiring prior information on noise moment order which is often not given in practice. By leveraging the  $p$ -robust estimator, we propose  $p$ -Heavy-UCRL2 as detailed in Algorithm 1.

**A. COMPARISON WITH UCRL2 [6]**

The first algorithm,  $p$ -Heavy-UCRL2 (Algorithm 1) is a variant of the UCRL2 algorithm [6] with the  $p$ -robust estimator. In the vanilla UCRL2 algorithm, a learner constructs a plausible set of MDPs at the beginning of each episode. Then, the learner chooses an optimistic MDP in the set of plausible MDPs and estimates an optimistic policy corresponding to the chosen MDP. In particular, the rewards  $\hat{r}_k(s, a)$  and transition probabilities  $\hat{p}_k(s, a)$  are estimated considering given  $k$  episodes, and a set of plausible MDPs is defined from the estimates. Thus, the estimation step of mean rewards and transition probabilities directly affects the quality of plausible MDPs set and the selection of an optimistic MDP. In particular, the vanilla UCRL2 guarantees the estimations of reward and transition probability by using Azuma-Hoeffding inequality which has an exponential convergence rate but only applicable under the sub-Gaussian noise assumption. However, under the assumption of heavy-tailed noises, Azuma-Hoeffding inequality cannot be utilized. Thus,  $p$ -Heavy-UCRL2 estimates mean rewards by using the  $p$ -robust estimator which provides a similar convergence rate to that of Azuma-Hoeffding inequality under heavy-tailed noises. In other words, the high probability confidence bound of mean rewards is guaranteed by the confidence interval of  $p$ -robust estimator.

**B. COMPARISON WITH HEAVY-UCRL2 [25]**

Similar to ours, [25] proposed a model-based RL algorithm, referred to as Heavy-UCRL2, designed to handle heavy-tailed noises. Specifically, they revised the vanilla UCRL2 algorithm with a truncation estimator [7] capable of adapting to heavy-tailed noises. Consequently, mean rewards are estimated using the confidence interval of the truncation estimator. However, unlike the  $p$ -robust estimator, the truncation estimator assumes that the bounded moment order of heavy-tailed noises is known in advance, necessitating prior knowledge on bounded moment order for Heavy-UCRL2. This constrains the practical application of Heavy-UCRL2. In Section V, we will prove that Heavy-UCRL2 and  $p$ -Heavy-UCRL2 have the same regret bounds up to logarithmic terms of  $H$ ,  $S$ , and  $T$ , even though  $p$ -Heavy-UCRL2 does not require prior knowledge.

**C. COMPARISON WITH Q-LEARNING [14]**

Algorithm 2 is an adaptation of the Q-learning algorithm [14] with the  $p$ -robust estimator. In finite-horizon episodic MDP, [14] showed that if the  $Q$  value function is initialized with  $H$  and the exploration bonus term is given as  $b_t := c\sqrt{H^3\iota}/t$  where  $\iota := \log(SAT/\delta)$  and the Q-learning algorithm can achieve a nearly optimal regret bound. The update rule of the  $Q$  value functions of vanilla Q-learning is as follows:

$$Q_h(s, a) \leftarrow (1 - \alpha_t)Q_h(s, a) + \alpha_t[r_h(s, a) + V_{h+1}(s') + b_t], \quad (9)$$

where  $\alpha_t := \frac{H+1}{H+t}$  is a learning rate,  $(s, a) \in \mathcal{S} \times \mathcal{A}$  represents a state and action pair, and  $s'$  is the next state. The performance of this algorithm hinges crucially on the bonus term  $b_t$ , which enables the Q-learning algorithm to attain an optimal regret bound  $O(\sqrt{H^4SAT\iota})$  with respect to  $T$ . However, this bonus term was derived from the usage of Azuma-Hoeffding inequality, and thus when using this bonus term in the heavy-tailed setting, one cannot guarantee the performance of the algorithm. Instead of direct application of reward mean to update the  $Q$  value function, we first regularize contaminated rewards by using the  $p$ -robust estimator. The update rule of our algorithm is as follows:

$$Q_h \leftarrow (1 - \alpha_t)Q_h(x_h, a_h) + \alpha_t \left( ct^{\frac{1}{p}}\psi_p \left( \frac{r_h(s_h, a_h)}{ct^{\frac{1}{p}}} \right) + V_{h+1} + b'_t \right). \quad (10)$$

Here, the contamination on rewards is alleviated by the  $p$ -robust estimator. In addition, the bonus term is different from the original bonus term  $b_t$ , which is defined as  $b'_t = b_t + 2H(c \ln(2SAT)/\delta)/t^{1-\frac{1}{p}}$  where  $c$  is some constant, and  $\delta \in (0, 1)$  is a confidence parameter. This bonus term is derived from the confidence interval of the  $p$ -robust estimator and enables us to explore suitably even in the heavy-tailed noise setting.



**D. COMPARISON WITH HEAVY-Q-LEARNING [25]**

In model-free RL under heavy-tailed noises, [25] proposed a robust variant of Q-learning algorithm [14] by using the truncation estimator. Especially, Heavy-Q-learning truncates rewards contaminated by heavy-tailed noises, i.e.,  $\hat{r}_t = r_t \mathbb{I}\{|r_t| \leq B_t\}$ , where  $\mathbb{I}$  is indicator function,  $B_t := (v_p t / \log(2SAT/\delta))$  is a truncation threshold at time step  $t \in [T]$  and  $v_p$  is the bounded moment order of heavy-tailed noise. Accordingly, the Bellman update rule of Heavy-Q-learning is defined as

$$Q_h(s, a) \leftarrow (1 - \alpha_t)Q_h(s, a) + \alpha_t(\hat{r}_h(s, a) + V_{h+1}(s') + b_t''), \tag{11}$$

where  $b_t'' := b_t + 8Hv_p^{\frac{1}{p}} \left(\frac{\log(2SAT/\delta)}{t}\right)^{1-\frac{1}{p}}$ . These simple modifications induce the robustness of Q-learning. However, as in the case of Heavy-UCRL2, Heavy-Q-learning still requires prior knowledge.

**V. REGRET ANALYSIS**

In this section, we present theoretical results of  $p$ -Heavy-UCRL2 and  $p$ -Heavy-Q-learning. In the analysis of  $p$ -Heavy-UCRL2, we establish various regret bounds, including a total regret bound over  $T$  rounds (Theorem 1), a per-step regret bound (Corollary 1), an instance-dependent regret bound (Theorem 2), and a regret bound for changing MDPs (Theorem 3). For  $p$ -Heavy-Q-learning, we establish regret bounds using Hoeffding and Bernstein-style bonus terms. Significantly, both  $p$ -Heavy-UCRL2 and  $p$ -Heavy-Q-learning exhibit identical polynomial dependencies on  $H$ ,  $S$ , and  $T$  terms as the existing regret bounds in [25], which match the lower bound in RL under heavy-tailed noise [25]. It is worth noting that unlike Heavy-UCRL2 and Heavy-Q-learning, these achievements are made without necessitating prior knowledge of the bounded moment order of heavy-tailed noise distributions. Now, we start with the first result, the total regret bound of  $p$ -Heavy-UCRL2.

*Theorem 1 (Total Regret Bound of  $p$ -Heavy-UCRL2):* Let  $R_\Delta := r_{\max} - r_{\min}$ . Then, we have the following total regret upper bound of  $p$ -Heavy-UCRL2 with probability at least  $1 - \delta$ ,

$$20R_\Delta DS \sqrt{AT \log\left(\frac{T}{\delta}\right) + (2C_p + 1)(7c) \log\left(\frac{2SAT}{\delta}\right)} (SAT)^{\frac{1}{p}}, \tag{12}$$

where  $C_p$  is some constant.

The proof is deferred to Appendix A. Within the proof, the regret is bounded by three primary error terms which are induced by the extended value iteration, estimation of transition probabilities, and rewards, respectively. Specifically, heavy-tailed noises only impact the errors in reward mean estimation, while the other terms remain unaffected. Thus, the first term of (12) which is unaffected by heavy-tailed noise,  $20R_\Delta DS \sqrt{AT \log(T/\delta)}$ , retains the same dependencies on  $D$ ,  $S$ ,  $A$ , and  $T$  as the vanilla UCRL2 [6]. In contrast, the second

term of (12) represents an extra regret component associated with heavy-tailed noise, which is derived from the confidence interval of the  $p$ -robust estimator. More precisely, to bound the reward mean estimation error in constructing a plausible MDP set,  $p$ -Heavy-UCRL2 leverages the confidence interval of the  $p$ -robust estimator, which results in an additional regret bound (i.e., the second term of (12)). Note that this penalty on regret matches the lower bound  $\tilde{\Omega}((SAT)^{1/p})$  [25], which implies the extra term of regret bound is inevitable.

Similar to ours, Heavy-UCRL2 utilizes the truncation estimator to bound reward mean estimation error in the heavy-tailed setting. The regret term of Heavy-UCRL2 corresponding to the second term of (12) shows a better dependency on  $\log(2SAT/\delta)$  by an order of  $1/p$  where  $p \in (1, 2]$  is a bounded moment order of heavy-tailed noise distribution. The difference in these regret bounds arises from the difference in the confidence intervals of the  $p$ -robust estimator and the truncation estimator. However, the increase in the order of the regret bound for  $p$ -Heavy-UCRL2 is limited to the logarithmic factor  $\log(2SAT/\delta)$ , meaning that the polynomial dependencies on the dominating terms, including  $H$ ,  $S$ ,  $A$ , and  $T$ , remain consistent with the regret bound of Heavy-UCRL2. We will show that the experimental performance of both algorithms is almost identical as detailed in Section VI.

*Corollary 1 (Average Per-Step Regret):* The average per-step regret of  $p$ -Heavy-UCRL2 is at most  $\lambda$ , with probability at least  $1 - \delta$ , for any

$$T \geq \max \left\{ \left( 4^2 20^2 \frac{R_\Delta^2 D^2 S^2 A}{\lambda^2} \log\left(\frac{40R_\Delta DSA}{\delta\lambda}\right) \right), \alpha \log\left(\frac{2SA}{\delta}\right) + 2\alpha \log\left(\frac{\alpha}{\delta}\right) \right\}. \tag{13}$$

where  $\alpha = (1/\lambda)^{\frac{p}{p-1}} (4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} (SA)^{\frac{1}{p-1}}$ .

The proof is deferred to Appendix B. From Theorem 1, we can directly derive Corollary 1 which provides probably approximately correct (PAC) bound of  $p$ -Heavy-UCRL2. Compared to Heavy-UCRL2, the order of  $7c$  has increased from 1 to  $\frac{p}{p-1}$ , where  $1 < \frac{p}{p-1}$  for any  $p \in (1, 2]$ . This implies that we need slightly more samples to guarantee the regret that is smaller than  $\lambda$ .

*Theorem 2 (Instance-Dependent Regret):* For any initial state  $s \in \mathcal{S}$ , any  $T \leq 1$ , any  $\lambda > 0$ , and some constant  $C_p$  with probability at least  $1 - 3\delta$ , the regret bound of  $p$ -Heavy-UCRL2 is

$$(4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} \left( \log\left(\frac{2SAT}{\delta}\right) \right)^{\frac{p}{p-1}} \left( \frac{SA}{\lambda} \right)^{\frac{1}{p-1}} + \lambda T \tag{14}$$

Let  $g := \rho^*(M) - \max_{s \in \mathcal{S}} \max_{\pi: \mathcal{S} \rightarrow \mathcal{A}} \{\rho(M, \pi, s) : \rho(M, \pi, s) > \rho^*(M)\}$  be the average reward gap between of the best and second-best policies. Then, the expected regret

of  $p$ -Heavy-UCRL2 (with parameter  $\delta := \frac{1}{3T}$ ) is bounded as

$$\begin{aligned} & \mathbb{E}[\Delta(M, s, T)] \\ & \leq (4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} \left( \log \left( \frac{2SAT}{\delta} \right) \right)^{\frac{p}{p-1}} \left( \frac{2SA}{g} \right)^{\frac{1}{p-1}} \\ & \quad + \sum_{s,a} \left[ 1 + \log_2 \left( \max_{\pi: \pi(s)=a} T_\pi \right) \right] \max_{\pi: \pi(s)=a} T_\pi. \quad (15) \end{aligned}$$

The proof is deferred to Appendix C. This theorem shows that  $p$ -Heavy-UCRL2 can achieve the logarithmic expected regret in terms of instance-dependent term, that is, the gap between the average reward of the best and second-best policies. Compared to Heavy-UCRL2, the dependency on  $\log(2SAT/\delta)$  is increased from 1 to  $\frac{p}{p-1}$ , while the polynomial dependencies on the other terms remain unchanged. Thus we can attain the same instance-dependent regret bound to that of Heavy-UCRL2 up to logarithmic terms without requiring prior knowledge.

**Theorem 3 (Regret for Changing MDP):** Let  $p \in (1, 2]$ . Restarting  $p$ -Heavy-UCRL2 with parameter  $\frac{\delta}{2}$  at each steps  $\left\lceil \frac{i(2p-1)/(p-1)}{\ell^{(p)/(p-1)}} \right\rceil$  for  $i = 1, 2, 3, \dots$ , with probability at least  $1 - \delta$ , the regret of  $p$ -Heavy-UCRL2 is upper bounded by

$$R_\Delta \ell^{\frac{p-1}{2p-1}} T^{\frac{p}{2p-1}} (SA)^{\frac{1}{p}}. \quad (16)$$

The proof is deferred to Appendix D. This theorem shows that the regret bound of  $p$ -Heavy-UCRL for changing MDP, which is the same as that of Heavy-UCRL2 up to logarithmic terms. Furthermore, if the noise distribution has a finite variance (i.e.,  $p = 2$ ), the regret bound for changing MDP becomes  $R_\Delta \ell^{1/3} T^{2/3} (SA)^{1/2}$  that recovers the same dependencies on  $\ell, T, S$ , and  $A$  compared with vanilla UCRL2 under sub-Gaussian noise setting.

**Theorem 4 (Regret Bound of  $p$ -Heavy-Q-Learning with Hoeffding-Style Bonus):** The total regret of  $p$ -Heavy-Q-learning with Hoeffding-style bonus over  $T$  rounds is as follows,

$$O \left( r_{\max} H^2 \sqrt{SAT} \iota + \iota H^{2-\frac{1}{p}} T^{\frac{1}{p}} (SA)^{1-\frac{1}{p}} \right). \quad (17)$$

The proof is deferred to Appendix E. This theorem presents the total regret bound of  $p$ -Heavy-Q-learning over  $T$  rounds with Hoeffding-style bonus term. The first term of (17) is not affected by heavy-tailed noise, while the second term represents an additional component induced by heavy-tailed noise. In comparison with Heavy-Q-learning, the order of  $\iota = \log(SAT/\delta)$  has increased from  $1 - 1/p$  to 1. However, this increase is confined to logarithmic terms, similar to the previous  $p$ -Heavy-UCRL2 case, while the orders regarding  $H, S, A$ , and  $T$  remain unchanged.

**Theorem 5 (Regret Bound of  $p$ -Heavy-Q-Learning With Bernstein-Style Bonus:)** The total regret of  $p$ -Heavy-Q-learning with Bernstein-style bonus term over  $T$

rounds is as follows,

$$\begin{aligned} & O \left( \sqrt{H^3 r_{\max}^3 SAT} \iota + H^{2-\frac{1}{p}} \iota (SA)^{1-\frac{1}{p}} T^{\frac{1}{p}} + \sqrt{H^9 r_{\max}^2 \nu^{\frac{1}{p}} (SA) \iota^3} \right. \\ & \quad \left. + \sqrt{H^{\frac{4p-3}{p-1}} r_{\max} (SA) \iota^2} + H^{\frac{3p-2}{p-1}} \sqrt{(SA)^3 \iota^4 (p-1)} \right). \quad (18) \end{aligned}$$

The proof of Theorem 5 is deferred to Appendix F. This theorem presents the regret bound for  $p$ -Heavy-Q-learning with a Bernstein-style bonus term. Unlike the Hoeffding bonus term case, which uses the Azuma-Hoeffding inequality, Bernstein's inequality is utilized to bound the MDP's transition probability estimation error, leading to an additional regret due to the gap between true variance and estimation variance. Furthermore, the second term of (18) represents the regret induced by heavy-tailed noise, where, akin to previous cases, the order of  $\iota$  has increased from  $1 - 1/p$  to 1. Note that in a sub-Gaussian noise setting, using the Bernstein's inequality to bound transition probabilities estimation error tightens the regret bound by a factor of  $O(\sqrt{H})$  [14]. However, in a heavy-tailed noise setting, the reward estimation error dominates the regret bound, and thus, the overall regret bound remains the same as that of Hoeffding-style bonus term.

**Theorem 6 (Lower bound in heavy-tailed setting [25]):**

For any fixed  $T$  and algorithm, there exists a communicating MDP  $M$  with diameter  $D$  such that the expected regret of the algorithm is  $\Omega((SA)^{1-\frac{1}{p}} T^{\frac{1}{p}})$ . In the finite-horizon episodic setting, there exists a MDP such that the expected regret is  $\Omega(H(SA)^{1-\frac{1}{p}} T^{\frac{1}{p}})$ .

[25] proved that the lower bound of finite-horizon episodic MDP setting under heavy-tailed noise. Specifically, when the noise has only a finite variance ( $p = 2$ ), this lower bound recovers the lower bound under sub-Gaussian noises  $\Omega(\sqrt{SAT})$ . We note that the regret bound of  $p$ -Heavy-Q-learning aligns with this regret bound in both cases of Hoeffding and Bernstein-style bonus terms, up to logarithmic terms.

## VI. EXPERIMENTS

In this section, we present the experimental performance of  $p$ -Heavy-UCRL2 and  $p$ -Heavy-Q-learning on synthetic tabular MDP such as SixArms [5] and DoubleChain. The main comparison group of algorithms includes Heavy-UCRL2 and Heavy-Q-learning, both designed for the heavy-tailed noise setting. In experimental results, the proposed algorithms show similar or even better performance while both of them do not require prior information on the bounded moment order of heavy-tailed noise distribution. Furthermore, we compared our algorithms with UCRL, Q-learning, and PSRL to demonstrate the challenges faced by reinforcement learning algorithms developed under the sub-Gaussian noise assumption when applied in heavy-tailed scenarios. To generate heavy-tailed noise distribution, we make a Weibull distribution which has a scale  $\alpha$  and shape  $k$  parameters. Specifically, we set  $\alpha = 1$  and  $k = 1.6$ , the

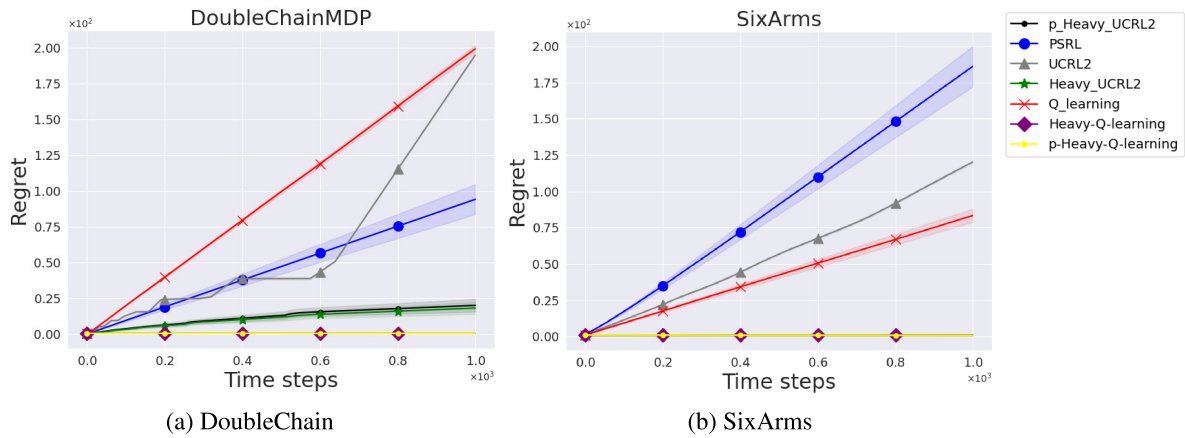


FIGURE 1. Experimental results on synthetic tabular MDPs. Fig. (a) for DoubleChain and Fig. (b) for SixArms.

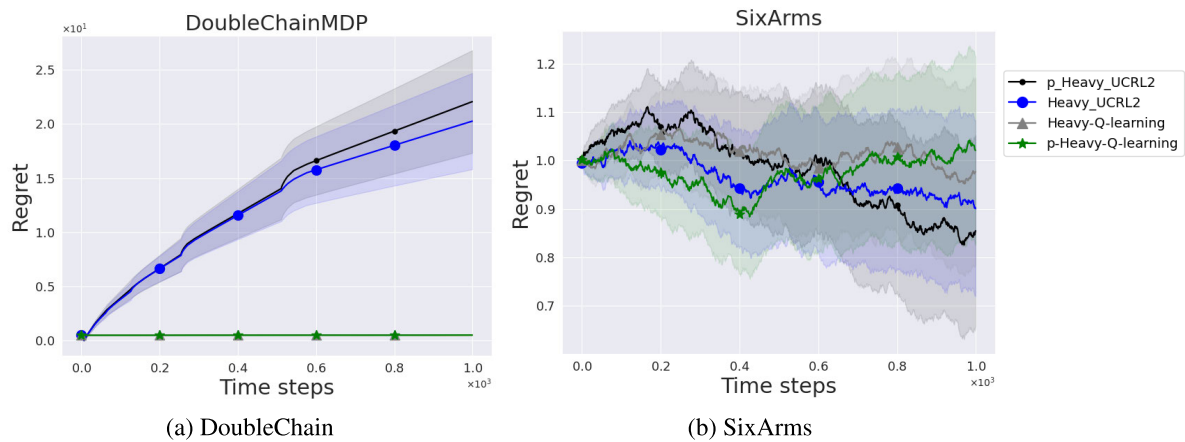


FIGURE 2. Experimental results on synthetic tabular MDPs only associated with RL algorithms under heavy-tailed noises. Fig. (a) for DoubleChain and Fig. (b) for SixArms.

mean is zero, and the noise moment order  $p = 1.2$ . We begin with the formal construction of synthetic tabular MDPs.

The following synthetic MDPs are adopted to evaluate the performance of RL algorithms in [25]. DoubleChain MDP is a combination of two  $\ell$ -length Riverswim-style MDPs where an agent has two actions, either going right or left. In this MDP, the leftmost state has a negative or zero reward and the rightmost states of each Riverswim-style sub-MDP have the highest rewards. Thus the agent tries to go right-side states to maximize cumulative rewards. In our simulation, we set the length  $\ell = 3$ , and thus the number of states  $S = |S|$  is 7 including an initial state  $s_1$ . We assume that the rewards of the rightmost states  $s_\ell$  and  $s_{2\ell}$  follow a normal distribution  $\mathcal{N}(0.5, (0.1)^2)$ . Furthermore, the rewards on states  $s_2, \dots, s_{\ell-1}$  and  $s_{\ell+1}, \dots, s_{2\ell}$  follow  $\mathcal{N}(0.1, (0.1)^2)$  and  $\mathcal{N}(-0.5, (0.1)^2)$ , respectively. The SixArms MDP consists of seven states. In an initial state  $s_0$ , an agent can choose one of six actions,  $a_1, \dots, a_6$  each of which transitions the agent to the corresponding state  $s_1, \dots, s_6$  with probability  $p_i$  for  $i \in \{1, \dots, 6\}$ . After the transition to the next state  $s_i$ ,  $\forall i \in \{1, \dots, 6\}$ , if the agent selects action  $a_{i-1}$  then the agent

returns to the current state with probability 1; otherwise, if the agent takes any other actions, the agent goes to the initial state  $s_0$ . The reward on the initial state and the other states are set to normal distribution  $\mathcal{N}(1.2, (0.1)^2)$  and  $\mathcal{N}(1 + (0.2)s_{i-1}, 0.1)$ , respectively. Note that the algorithm receives a noisy reward contaminated by heavy-tailed noises.

### A. RESULTS ON TABULAR MDP

Fig. 1 presents the experimental results of RL algorithms on DoubleChain and SixArms MDPs. As shown in both Fig. 1a and 1b, the algorithms designed for sub-Gaussian noise settings, namely UCRL2, Q-learning, and PSRL, exhibit poor performance under heavy-tailed noise conditions. These findings align with theoretical results, as the guarantees for sub-Gaussian RL algorithms are heavily dependent on concentration inequalities that hold only in sub-Gaussian noise settings. Conversely, RL algorithms tailored for heavy-tailed noise demonstrate superior performance. Specifically, the model-based RL algorithms  $p$ -Heavy-UCRL2 and Heavy-UCRL outperform the sub-

Gaussian RL algorithms by significant margins and produce nearly identical results to each other. Additionally, the model-free RL algorithms  $p$ -Heavy-Q-learning and Heavy-Q-learning show almost zero regret, with minimal differences between them. These outcomes support our theoretical results, indicating that the increase in the regret bounds is restricted to logarithmic terms. Fig. 2 shows the experimental results exclusively for RL algorithms designed for heavy-tailed noise. Consistently, there are no significant performance differences between these algorithms, despite our methods not requiring additional information about the moment order of the heavy-tailed noise. The performances of all algorithms for heavy-tailed noises are similar, however,  $p$ -Heavy-UCRL2 and  $p$ -Heavy-Q-learning have a strength in that they do not require prior knowledge of bounded moment order of heavy-tailed noise while Heavy-UCRL and Heavy-Q-learning do.

## VII. CONCLUSION

In this paper, we proposed reinforcement learning (RL) algorithms,  $p$ -Heavy-UCRL2 and  $p$ -Heavy-Q-learning, for finite-horizon episodic Markov decision processes (MDPs) under heavy-tailed noise. Unlike existing algorithms, our methods do not require any prior knowledge about the bounded moment order of the heavy-tailed noise distribution, making them applicable to a broader range of scenarios. This advantage arises from the nature of the  $p$ -robust estimator, which is defined without the need for prior knowledge. We demonstrated that our algorithms achieve the same regret bounds as existing algorithms, nearly optimal up to logarithmic terms. Finally, our experimental results validate the theoretical guarantees, showing favorable performance with existing algorithms.

## APPENDIX A

### PROOF OF THEOREM 1

We briefly outline the proof of Theorem 1. The total regret over  $T$  rounds is bounded as (20) with a probability of at least  $1 - \frac{\delta}{12T^{5/4}}$ . This regret bound can be decomposed into two major terms: the regret when the true MDP is within the plausible MDP set and the regret when it is not. By utilizing Lemma 1, we can determine the regret bound for the case when the true MDP is not included in the plausible MDP set. The regret when the true MDP is within the plausible MDP set can be further decomposed into transition and reward estimation errors. To bound the transition estimation error, we use technical Lemma 2. The reward estimation error is bounded using the confidence interval of the  $p$ -robust estimator (Definition 2). Finally, by summing the upper bounds of the regret for both cases when the true MDP is within the plausible set and when it is not—the proof is completed.

*Proof:* The proof is an adaptation of the proof of Theorem 2 in [6] and the entire proof consists of four subsections. Note that conditioned on  $N(s, a)$ , the rewards at each time step  $t \in [T]$ ,  $r_t$ , are independent. Thus by using the

confidence interval of the  $p$ -Robust estimator the estimation error for rewards under heavy-tailed noise is bounded as follows,

$$\mathbb{P}\left\{\sum_{t=1}^T r_t \leq \sum_{s,a} N(s, a)\bar{r}(s, a) - C_T \mid (N(s, a))_{s,a}\right\} \leq \left(\frac{\delta}{8T}\right)^{5/4} \leq \frac{\delta}{12T^{5/4}}, \quad (19)$$

where  $C_T := \frac{5}{4} \ln(\exp(b_p v_p / c^p) (8T/\delta))$ ,  $\delta \in (0, 1]$  is a confidence parameter, and  $N(s, a)$  is a state-action counts. Letting  $\Delta_k$  be the regret incurred by an arbitrary episode  $k \in [K]$ , the total regret  $\Delta$  can be bounded by

$$\Delta(s_1, T) \leq \sum_{k=1}^K \Delta_k + C_T \quad (20)$$

with probability at least  $1 - \frac{\delta}{12T^{5/4}}$ . In particular, the total regret  $\Delta(s_1, T)$  can be decomposed into regret that occurs when the true MDP is not in the plausible set and regret that occurs when it is included in the plausible set.

### A. REGRET FOR THE TRUE MDP NOT BEING IN THE PLAUSIBLE SET

In this section, we provide the regret when the estimated plausible set does not encompass the true MDP. The proof directly follows from the following lemma.

*Lemma 1 (Lemma 17 in [6]):* For any  $t \geq 1$ , the probability that the true MDP  $M$  is not contained in the set of plausible MDPs  $\mathcal{M}(t)$  at time  $t$  is at most  $\frac{\delta}{15t^6}$ , that is

$$\mathbb{P}\{M \notin \mathcal{M}(t)\} < \frac{\delta}{15t^6}. \quad (21)$$

Then, by this lemma and the same argument in Section IV-B in [6], we have, with probability at least  $1 - \frac{\delta}{12T^{5/4}}$ ,

$$\sum_{k=1}^m \Delta_k \mathbb{I}_{M \notin \mathcal{M}_k} \leq R_\Delta \sqrt{T} \quad (22)$$

where  $\mathbb{I}$  is an indicator function and  $R_\Delta := r_{\max} - r_{\min}$ .

### B. REGRET FOR THE TRUE MDP BEING IN THE PLAUSIBLE SET

Now we turn to bound the regret in each episode  $k \in [K]$ , assuming that the plausible MDP set contains the true MDP. We first start by showing that the condition for stopping value iteration of  $p$ -Heavy-UCRL2 is identical to that of UCRL2. Specifically, the stopping criterion of extended value iteration is as follows:

*Theorem 7 (Modification of Theorem 7 in [6]):* Let  $\mathcal{M}$  be the set of MDPs with state space  $\mathcal{S}$ , action space  $\mathcal{A}$ , transition probabilities  $\tilde{p}(\cdot|s, a)$  and mean rewards  $\tilde{r}(s, a)$  that satisfy inequalities (6) and (7).  $u_i(s)$  denote the state value of  $s$  at iteration  $i$ . Then, if  $\mathcal{M}$  contains at least one



communicating MDP, extended value iteration converges. Furthermore, stopping extended value iteration when

$$\max_{s \in \mathcal{S}} \{u_{i+1}(s) - u_i(s)\} - \min_{s \in \mathcal{S}} \{u_{i+1}(s) - u_i(s)\} < \varepsilon \quad (23)$$

the greedy policy with respect to  $u_i$  is  $\varepsilon$ -optimal.

From Theorem 7, the regret in episode  $k$  can be bounded as

$$\Delta_k \leq \sum_{s,a} v_k(s, a)(\rho^* - \bar{r}(s, a)) \quad (24)$$

$$\leq \sum_{s,a} v_k(s, a)(\bar{\rho}_k - \bar{r}(s, a)) + \sum_{s,a} \frac{v_k(s, a)}{\sqrt{t_k}} \quad (25)$$

$$= \sum_{s,a} v_k(s, a)(\bar{p}_k - \bar{r}_k(s, a)) + \sum_{s,a} v_k(s, a)(\bar{r}_k(s, a) - \bar{r}(s, a)) + \sum_{s,a} \frac{v_k(s, a)}{\sqrt{t_k}} \quad (26)$$

$$= \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{I})\mathbf{u}_i + \sum_{s,a} v_k(s, a)(\bar{r}_k(s, a) - \bar{r}(s, a)) + 2 \sum_{s,a} \frac{v_k(s, a)}{\sqrt{t_k}} \quad (27)$$

where  $\tilde{\mathbf{P}}_k := (\tilde{p}_k(s'|s, \tilde{\pi}_k(s)))_{s,s'}$  the transition probabilities of  $\tilde{\pi}_k$  on  $\tilde{M}_k$ . For the second term, we have the following:

$$\sum_{s,a} v_k(s, a)(\bar{r}_k(s, a) - \bar{r}(s, a)) \leq \sum_{s,a} v_k(s, a)(|\bar{r}_k(s, a) - \hat{r}_k(s, a)| + |\bar{r}_k(s, a) - \hat{r}_k(s, a)|) \quad (28)$$

$$\leq \sum_{s,a} v_k(s, a) \frac{2 \cdot 7 \log(2SAT_k/\delta)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \quad (29)$$

where the inequality (29) follows from the confidence interval of  $p$ -robust estimator and the assumption that the true MDP is in the plausible set of MDPs. Since  $\max\{1, N_k(s, a)\} \leq t_k \leq T$  holds, we obtain that

$$\Delta_k \leq \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{I})\mathbf{u}_i + 14c \log\left(\frac{2SAT}{\delta}\right) \times \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} + 2 \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}}. \quad (30)$$

For the first term of the above inequality, we have

$$\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{I})\mathbf{w}_k = \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\mathbf{w}_k + \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\mathbf{w}_k. \quad (31)$$

Since the true MDP is in the plausible set  $\mathcal{M}_k$ , we can leverage the following bound

$$\|\tilde{p}(\cdot|s, a) - \hat{p}_k(\cdot|s, a)\|_1 \leq \sqrt{\frac{14S \log(2A_k/\delta)}{\max\{1, N_k(s, a)\}}}. \quad (32)$$

By using the inequality (32), we can bound  $\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\mathbf{w}_k$ . Note that this bound is identical to that of the vanilla UCRL2. In other words, we can bound transition error in the same manner in [6] since heavy-tailed noise does not affect this

error term. From the argument in subsection 4.3.2 in [6], we have the following inequality:

$$\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\mathbf{w}_k \leq D \sqrt{14S \log\left(\frac{2AT}{\delta}\right)} \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \quad (33)$$

In addition, the second term of the inequality (31) can be bounded as

$$\sum_{k=1}^m \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\mathbf{w}_k \leq R_\Delta \left( D \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)} + DSA \log_2\left(\frac{8T}{SA}\right) \right) \quad (34)$$

with probability at least  $1 - \frac{\delta}{12T^{5/4}}$ .

Combining the derived results and summing up the per-episode regret  $\Delta_k$  overall episodes  $[K]$  with  $M \in \mathcal{M}_k$ , the total regret when the true MDP is in the plausible set of MDPs is as follows:

$$\begin{aligned} & \sum_{k=1}^m \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} \\ & \leq \sum_{k=1}^m \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\mathbf{w}_k \mathbb{1}_{M \in \mathcal{M}_k} + \sum_{k=1}^m \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\mathbf{w}_k \mathbb{1}_{M \in \mathcal{M}_k} \\ & \quad + 14c \log\left(\frac{2SAT}{\delta}\right) \sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \\ & \quad + 2 \sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \quad (35) \\ & \leq R_\Delta D \left( \sqrt{14S \log\left(\frac{2AT}{\delta}\right)} + 2 \right) (\sqrt{2} + 1) \sqrt{SAT} \\ & \quad + R_\Delta \left( D \sqrt{\frac{5}{2} T \log\left(\frac{8T}{\delta}\right)} + DSA \log_2\left(\frac{8T}{SA}\right) \right) \\ & \quad + 14c \log\left(\frac{2SAT}{\delta}\right) \sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \quad (36) \end{aligned}$$

To bound  $\sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}}$ , we introduce the following lemma.

*Lemma 2 [25]: For any sequence of numbers  $z_1, z_2, \dots, z_n$  with  $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$ , there exists some constant  $C_p$  such that*

$$\sum_{k=1}^n \frac{z_k}{Z_{k-1}^{1-\frac{1}{p}}} \leq C_p Z_n^{1/p} \quad (37)$$

*Proof:* The proof can be found in [25].  $\square$

By using lemma 2 and Jensen's inequality, we have

$$\sum_{s,a} \sum_k \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \leq C_p \sum_{s,a} N_{s,a}^{1/p} \leq C_p (SAT)^{1/p} \quad (38)$$

where  $C_p$  is some constant. Finally, we have

$$\begin{aligned} & \sum_{k=1}^m \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} \\ & \leq \sum_{k=1}^m \mathbf{v}_k (\tilde{\mathbf{P}}_k - \mathbf{P}_k) \mathbf{w}_k \mathbb{1}_{M \in \mathcal{M}_k} + \sum_{k=1}^m \mathbf{v}_k (\mathbf{P}_k - \mathbf{I}) \mathbf{w}_k \mathbb{1}_{M \in \mathcal{M}_k} \\ & \quad + 14c \log \left( \frac{2SAT}{\delta} \right) \sum_{k=1}^m \sum_{s,a} \frac{v_k(s,a)}{\max\{1, N_k(s,a)\}^{1-\frac{1}{p}}} \\ & \quad + 2 \sum_{k=1}^m \sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \quad (39) \\ & \leq R_\Delta D \left( \sqrt{14S \log \left( \frac{2AT}{\delta} \right) + 2} \right) (\sqrt{2} + 1) \sqrt{SAT} \\ & \quad + R_\Delta \left( D \sqrt{\frac{5}{2} T \log \left( \frac{8T}{\delta} \right)} + DSA \log_2 \left( \frac{8T}{SA} \right) \right) \\ & \quad + 14c \log \left( \frac{2SAT}{\delta} \right) C_p (SAT)^{1/p}. \quad (40) \end{aligned}$$

### C. COMBINING ALL RESULTS

Now, we have the following total regret of  $p$ -Heavy-UCRL2

$$\begin{aligned} \Delta(s_1, T) & \leq \sum_{k=1}^m \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} + \sum_{k=1}^m \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} + C_T \quad (41) \\ & \leq C_T + R_\Delta \sqrt{T} \\ & \quad + R_\Delta \left( D \sqrt{\frac{5}{2} T \log \left( \frac{8T}{\delta} \right)} + DSA \log_2 \left( \frac{8T}{SA} \right) \right) \\ & \quad + R_\Delta D \left( \sqrt{14S \log \left( \frac{2AT}{\delta} \right) + 2} \right) (\sqrt{2} + 1) \sqrt{SAT} \\ & \quad + 2C_p \left( 7c \log \left( \frac{2SAT}{\delta} \right) \right) (SAT)^{1/p} \quad (42) \end{aligned}$$

where  $C_T := \frac{5}{4} \log \left( \exp(b_p v_p / c^p)^{\frac{4}{5}} \cdot \left( \frac{8T}{\delta} \right) \right)$  the confidence interval of  $p$ -robust estimator. The heavy-tailed terms are bounded by

$$\begin{aligned} & \frac{5}{4} \log \left( e^{\left( \frac{b_p v_p}{c^p} \right)^{\frac{4}{5}} \left( \frac{8T}{\delta} \right)} \right) + 2C_p \left( 7c \log \left( \frac{2SAT}{\delta} \right) \right) (SAT)^{\frac{1}{p}} \\ & \leq (2C_p + 1) \left( 7c \log \left( \frac{2SAT}{\delta} \right) \right) (SAT)^{\frac{1}{p}} \quad (43) \end{aligned}$$

$$= B_p \log \left( \frac{2SAT}{\delta} \right) (SAT)^{\frac{1}{p}} \quad (44)$$

where  $B_p := (2C_p + 1)(7c)$ . Combining all these, we can write the total regret as follows:

$$\begin{aligned} \Delta(s_1, T) & \leq R_\Delta \sqrt{T} + R_\Delta \left( D \sqrt{\frac{5}{2} T \log \left( \frac{8T}{\delta} \right)} + DSA \log_2 \left( \frac{8T}{SA} \right) \right) \end{aligned}$$

$$\begin{aligned} & + R_\Delta D \left( \sqrt{14S \log \left( \frac{2AT}{\delta} \right) + 2} \right) (\sqrt{2} + 1) \sqrt{SAT} \\ & + B_p \log \left( \frac{2SAT}{\delta} \right) (SAT)^{\frac{1}{p}}. \quad (45) \end{aligned}$$

Expanding the RHS of the above inequality gives

$$\begin{aligned} \Delta(s_1, T) & \leq R_\Delta \sqrt{T} + R_\Delta D \sqrt{\frac{5}{2} T \log \left( \frac{8T}{\delta} \right)} \\ & \quad + R_\Delta DSA \log_2 \left( \frac{8T}{SA} \right) + R_\Delta D \sqrt{14S \log (2AT\delta)} \sqrt{2} \sqrt{SAT} \\ & \quad + R_\Delta D \sqrt{14S \log \left( \frac{2AT}{\delta} \right)} \sqrt{SAT} \\ & \quad + R_\Delta D 2\sqrt{2} \sqrt{SAT} + R_\Delta 2\sqrt{SAT} + B_p \log \left( \frac{2SAT}{\delta} \right) (SAT)^{\frac{1}{p}}. \quad (46) \end{aligned}$$

We can rewrite this inequality as

$$\begin{aligned} \Delta(s_1, T) & \leq R_\Delta DS \sqrt{AT} \left( \frac{1}{\sqrt{A}} + \sqrt{\frac{1}{A} \cdot \frac{5}{2} \log \left( \frac{8T}{\delta} \right)} \right) \\ & \quad + (\sqrt{2} + 1) \sqrt{14 \log \left( \frac{2AT}{\delta} \right) + \sqrt{8} + 2} \\ & \quad + R_\Delta DSA \log_2 \left( \frac{8T}{SA} \right) + B_p \log \left( \frac{2SAT}{\delta} \right) (SAT)^{\frac{1}{p}}. \quad (47) \end{aligned}$$

As similar to [6], assume that  $A \geq 2$ . For  $1 \leq T \leq 20^2 A \log \left( \frac{T}{\delta} \right)$ , we have  $\Delta(s_1, T) \leq 20 \sqrt{AT \log \left( \frac{T}{\delta} \right)}$  trivially. Since  $T > 34A \log \left( \frac{T}{\delta} \right)$ , we have  $A < \frac{1}{34 \log(T/\delta)} \cdot \sqrt{AT \log(T/\delta)}$  and also  $\log_2(8T) < 2 \cdot \log(T)$ . Then we can obtain

$$\begin{aligned} R_\Delta DSA \log_2 \left( \frac{8T}{SA} \right) & < 2R_\Delta DSA \log \left( \frac{T}{SA} \right) \quad (48) \\ & < \frac{2R_\Delta DS}{34 \log(T/\delta)} \log(T/SA) \sqrt{AT \log(T/\delta)} \\ & = \frac{2}{34} R_\Delta DS \sqrt{AT \log(T/\delta)}. \quad (49) \end{aligned}$$

Further,  $T > 34 \cdot A \log \left( \frac{T}{\delta} \right)$  also implies  $\log \left( \frac{2AT}{\delta} \right) \leq 2 \cdot \log \left( \frac{T}{\delta} \right)$  and  $\log \left( \frac{8T}{\delta} \right) \leq 2 \cdot \log \left( \frac{T}{\delta} \right)$ . Thus, we have that for any  $T > 1$ , with probability at least,

$$\begin{aligned} \Delta(s_1, T) & \leq R_\Delta DS \sqrt{AT \log(T/\delta)} \\ & \quad \times \left( \frac{1}{\sqrt{2}} + \sqrt{\frac{5}{2}} + (\sqrt{2} + 1) \sqrt{28} + \sqrt{8} + 2 + \frac{2}{34} \right) \\ & \quad + B'_\varepsilon \log \left( \frac{2SAT}{\delta} \right) (SAT)^{\frac{1}{p}}. \quad (51) \end{aligned}$$

Therefore we have

$$\begin{aligned} & \Delta(s_1, T) \\ & \leq 20 \cdot R_{\Delta} D S \sqrt{AT \log \left( \frac{T}{\delta} \right)} + B_p \left( \log \left( \frac{2SAT}{\delta} \right) \right) (SAT)^{\frac{1}{p}} \end{aligned} \quad (52)$$

as desired.  $\square$

**APPENDIX B  
PROOF OF COROLLARY 1**

*Proof:* To complete the proof, we need to show that there exists some  $T_0$  such that the per-step regret can be bounded by  $\lambda$  when  $T_0 \leq T$ . From Theorem 1, we have that

$$\frac{20R_{\Delta} D S \sqrt{AT \log(T/\delta)}}{T} + \frac{B_p(\log(2SAT/\delta))(SAT)^{\frac{1}{p}}}{T} < \lambda \quad (53)$$

Here,  $B_p$  is constant defined as Theorem 1. Following [6], we find  $T_0$  satisfying the first and second terms of the above inequality are bounded by  $\frac{\lambda}{2}$ , respectively. The first term is derived from result in [6]. For the second term, we have

$$\frac{B_p(\log(2SAT/\delta))(SAT)^{\frac{1}{p}}}{T} < \frac{\lambda}{2} \quad (54)$$

$$\Rightarrow 2(2C_p + 1)(7c)(\log(2SAT/\delta))(SAT)^{\frac{1}{p}} < \lambda T \quad (55)$$

$$\Rightarrow (4C_p + 2)(7c)(\log(2SAT/\delta))(SA)^{\frac{1}{p}} \frac{1}{\lambda} < T^{1-\frac{1}{p}} \quad (56)$$

$$\begin{aligned} & \Rightarrow \left( \frac{1}{\lambda} \right)^{\frac{p}{p-1}} (4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} (\log \left( \frac{2SAT}{\delta} \right))^{\frac{p}{p-1}} (SA)^{\frac{1}{p-1}} \\ & < T \end{aligned} \quad (57)$$

$$\Rightarrow \left( \frac{1}{\lambda} \right)^{\frac{p}{p-1}} (4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} (\log \left( \frac{2SAT}{\delta} \right))^{\frac{p}{p-1}} (SA)^{\frac{1}{p-1}} < T \quad (58)$$

Let  $\alpha = (1/\lambda)^{\frac{p}{p-1}} (4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} (SA)^{\frac{1}{p-1}}$ . Then, we have

$$\alpha \log \left( \frac{2SAT}{\delta} \right) < T \Rightarrow \alpha \log \left( \frac{2SA}{\delta} \right) + \alpha \log \left( \frac{T}{\delta} \right) < T \quad (59)$$

Setting  $X = 2\alpha \log \left( \frac{\alpha}{\delta} \right)$  gives

$$X = 2\alpha \log \left( \frac{\alpha}{\delta} \right) \quad (60)$$

$$= \alpha \log \left( \frac{\alpha}{\delta} \cdot \frac{\alpha}{\delta} \right) \quad (61)$$

$$> \alpha \log \left( \frac{\alpha}{\delta} \cdot 2 \log \left( \frac{\alpha}{\delta} \right) \right) \quad (62)$$

$$= \alpha \log \left( \frac{X}{\delta} \right) \quad (63)$$

Note that inequality (62) uses the fact  $x > 2 \log(x)$  for all  $x > 0$ . Therefore we can conclude that  $2\alpha \log \left( \frac{\alpha}{\delta} \right) > \alpha \log \left( \frac{T}{\delta} \right)$ , from which the proof is completed.  $\square$

**APPENDIX C  
PROOF OF THEOREM 2**

This section presents the proof of Theorem 2 which represents the logarithmic upper bound on the expected regret of  $p$ -Heavy-UCRL2. Lemma 3 plays a role similar to Lemma 2 in the proof of Theorem 1.

*Lemma 3 [25]:*

$$\sum_{k \in K_{\lambda}} \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \leq C_{\varepsilon} (L_{\varepsilon} SA)^{\frac{1}{p}} \quad (64)$$

*Proof:* The proof can be found in [25].  $\square$

Following [6], we make the definition of  $\varepsilon$ -bad episode.

*Definition 3:* An episode  $k$  is  $\varepsilon$ -bad if its average regret is larger than  $\varepsilon$ , where the average regret of length  $\ell_k$  is  $\frac{\Delta_k}{\ell_k}$  with  $\Delta_k = \sum_{t=t_k}^{t_{k+1}-1} (\rho^* - r_t)$ .

Now, we introduce Lemma 4 which provides an upper bound on the total number of rounds in  $\lambda$ -bad episodes where the average regret exceeds  $\lambda$ . Using this lemma, we can upper bound the incurred regret in the  $\lambda$ -bad episodes.

*Lemma 4:* Let  $L_{\lambda}$  be the number of steps taken by  $p$ -Heavy-UCRL2 in  $\lambda$ -bad episodes up to time step  $T$ . Then for any initial state  $s \in \mathcal{S}$ , for any  $T$  and  $\lambda > 0$ , we have

$$L_{\lambda} \leq \left( \frac{4C_p + 2}{\lambda} \right)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} \left( \log \left( \frac{2SAT}{\delta} \right) \right)^{\frac{p}{p-1}} (SA)^{\frac{1}{p-1}} \quad (65)$$

*Proof:* First, let us define  $K_{\lambda}$  and  $J_{\lambda}$  be the sets of the indices of the  $\lambda$ -bad episodes and time steps in those episodes, respectively. Then, by confidence interval of the  $p$ -robust estimator, we have the following with probability at least  $1 - \delta$ ,

$$\begin{aligned} \sum_{k \in K_{\varepsilon}} \sum_{t=t_k}^{t_{k+1}-1} r_t & \geq \sum_{k \in K_{\varepsilon}} \sum_{s,a} v_k(s, a) \bar{r}(s, a) \\ & \quad - \frac{5}{4} \ln \left( \exp(b_p v_p / c^p)^{5/4} \left( \frac{8L_{\varepsilon}}{\delta} \right) \right). \end{aligned} \quad (66)$$

By combining the fact that  $\mathbb{P} \{ \sum_{k \in K_{\varepsilon}} \Delta_k \mathbb{1}_{M \notin M_k} > 0 \} \leq \delta$  and above inequality, we have

$$\Delta_p(s, T) \leq C_p + \sum_{k \in K_p} \Delta_k \mathbb{1}_{M \in M_k} \quad (67)$$

where  $\Delta_{\lambda}$  is regret of  $\lambda$ -bad episodes and  $C_p := \frac{5}{4} \ln \left( \exp(b_p v_p / c^p)^{5/4} \left( \frac{8L_{\lambda}}{\delta} \right) \right)$ . By combining these inequalities, we have

$$\begin{aligned} \Delta_k & \leq v_k(\tilde{P}_k - I)w_k \\ & \quad + 2 \cdot 7c \log \left( \frac{2SAT}{\delta} \right) \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \\ & \quad + 2 \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \end{aligned} \quad (68)$$

$$\begin{aligned} v_k(\tilde{P}_k - I)w_k & = v_k(\tilde{P}_k - P_k)w_k + v_k(P_k - I)w_k \\ v_k(\tilde{P}_k - P_k)w_k & \end{aligned} \quad (69)$$

$$\leq R_{\Delta} D \sqrt{14S \log \left( \frac{2AT}{\delta} \right)} \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \quad (70)$$

By using (69) and (70), we can obtain the following inequality:

$$\begin{aligned} \Delta_k &\leq v_k(\tilde{P}_k - I)w_k \\ &+ 2 \cdot 7c \log \left( \frac{2SAT}{\delta} \right) \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \\ &+ 2 \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \end{aligned} \quad (71)$$

$$\begin{aligned} &= v_k(\tilde{P}_k - P_k)w_k + v_k(P_k - I)w_k \\ &+ 2 \cdot 7c \log \left( \frac{2SAT}{\delta} \right) \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \\ &+ 2 \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \end{aligned} \quad (72)$$

$$\begin{aligned} &\leq v_k(P_k - I)w_k + 2 \cdot 7c \log \left( \frac{2SAT}{\delta} \right) \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \\ &+ \left( R_{\Delta} D \sqrt{14S \log(2AT/\delta)} + 2 \right) \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \end{aligned} \quad (73)$$

In the same manner as Appendix D in [6], we have

$$\sum_{k \in K_{\lambda}} \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \leq (\sqrt{2} + 1) \sqrt{L_{\varepsilon} SA} \quad (74)$$

And also, by lemma 3, we have

$$\sum_{k \in K_{\lambda}} \sum_{s,a} \frac{v_k(s, a)}{\max\{1, N_k(s, a)\}^{1-\frac{1}{p}}} \leq C_p (L_{\varepsilon} SA)^{\frac{1}{p}} \quad (75)$$

Then, from inequalities (67), (73), (74), and (75) it follows that with probability at least  $1 - 2\delta$ ,

$$\begin{aligned} \Delta_{\varepsilon}(s, T) &\leq C_T + (R_{\Delta} D \sqrt{14S \log(2AT/\delta)} + 2)(\sqrt{2} + 1)(\sqrt{L_{\varepsilon} SA}) \\ &+ 2 \cdot 7c \log \left( \frac{2SAT}{\delta} \right) C_p (L_{\varepsilon} SA)^{\frac{1}{p}} \\ &+ \sum_{k \in K_p} v_k(P_k - I)w_k \mathbb{1}_{M \in \mathcal{M}_k} \end{aligned} \quad (76)$$

where  $C_T := \frac{5}{4} \ln(\exp(b_p v_p / c^p)^{4/5}) \left( \frac{8L_p}{\delta} \right)$ . Now, we need to bound  $\sum_{k \in K_{\varepsilon}} v_k(P_k - I)w_k \mathbb{1}_{M \in \mathcal{M}_k}$ . For this, we use an argument similar to the one applied to obtain inequality (34).

$$\begin{aligned} &\sum_{k \in K_{\varepsilon}} v_k(P_k - I)w_k \mathbb{1}_{M \in \mathcal{M}_k} \\ &\leq R_{\Delta} \left( 2D \cdot \sqrt{L_{\varepsilon} \log \left( \frac{T}{\delta} \right)} + DSA \log_2 \left( \frac{8T}{SA} \right) \right) \end{aligned} \quad (77)$$

with probability at least  $1 - \delta$ . Combining inequalities (76) and (77) yields:

$$\Delta'_{\varepsilon}(s, T)$$

$$\begin{aligned} &\leq \frac{5}{4} \ln \left( \exp(b_p v_p / c^p)^{4/5} \right) \left( \frac{8L_{\varepsilon}}{\delta} \right) \\ &+ (R_{\Delta} D \sqrt{14S \log(2AT/\delta)} + 2)(\sqrt{2} + 1)(\sqrt{L_{\varepsilon} SA}) \\ &+ 2 \cdot 7c \log \left( \frac{2SAT}{\delta} \right) C_p (L_{\varepsilon} SA)^{\frac{1}{p}} \\ &+ R_{\Delta} \left( 2D \cdot \sqrt{L_{\varepsilon} \log \left( \frac{T}{\delta} \right)} + DSA \log_2 \left( \frac{8T}{SA} \right) \right) \end{aligned} \quad (78)$$

with probability at least  $1 - 3\delta$ . We can simplify the above inequality (78) in a similar way to the proof of Theorem 1.

$$\begin{aligned} \Delta_{\lambda}(s, T) &\leq 20DS \sqrt{L_{\lambda} A \log \left( \frac{T}{\delta} \right)} + B_p \left( \log \left( \frac{2SAT}{\delta} \right) \right) (L_{\lambda} SA)^{\frac{1}{p}} \end{aligned} \quad (79)$$

Here,  $(2C_p + 1)(7c)$  is denoted by  $B_p$ . Also, by the condition of  $T$ ,

$$\Delta'_{\lambda}(s, T) \leq 2B_p \left( \log \left( \frac{2SAT}{\delta} \right) \right) (L_{\lambda} SA)^{\frac{1}{p}} \quad (80)$$

Especially, by using condition on  $T$  and the fact that  $\lambda L_{\lambda} \leq \Delta'_{\lambda}(s, T)$ , we have

$$L_{\lambda}(T) \leq \frac{1}{\lambda} \cdot \Delta_{\lambda}(s, T) \quad (81)$$

$$\leq \frac{1}{\lambda} \cdot 2 \cdot B_p \left( \log \left( \frac{2SAT}{\delta} \right) \right) (L_{\lambda} SA)^{\frac{1}{p}} \quad (82)$$

$$= \frac{1}{\lambda} \cdot 2 \cdot (2C_p + 1) \cdot (7c) \cdot \left( \log \left( \frac{2SAT}{\delta} \right) \right) (L_{\lambda} SA)^{\frac{1}{p}} \quad (83)$$

Thus we have

$$L_{\lambda}^{1-\frac{1}{p}} \leq \left( \frac{4C_p + 2}{\lambda} \right) \cdot (7c) \cdot \left( \log \left( \frac{2SAT}{\delta} \right) \right) (SA)^{\frac{1}{p}} \quad (84)$$

$$\begin{aligned} \Rightarrow L_{\lambda} &\leq \left( \frac{4C_p + 2}{\lambda} \right)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} \\ &\cdot \left( \log \left( \frac{2SAT}{\delta} \right) \right)^{\frac{p}{p-1}} (SA)^{\frac{1}{p-1}} \end{aligned} \quad (85)$$

as desired.  $\square$

*Proof:* Substituting  $L_{\lambda}$  in inequality (79) with upper bound on  $L_{\lambda}$  in inequality (85), we have

$$\begin{aligned} \Delta'_{\lambda} &\leq 2B_p \left( \log \left( \frac{2SAT}{\delta} \right) \right) (L_{\lambda} SA)^{\frac{1}{p}} \end{aligned} \quad (86)$$

$$\begin{aligned} &\leq 2B_p \left( \log \left( \frac{2SAT}{\delta} \right) \right) (SA)^{\frac{1}{p}} \left( \frac{4C_p + 2}{\lambda} \right)^{\frac{1}{p-1}} (7c)^{\frac{1}{p-1}} \\ &\times \left( \log \left( \frac{2SAT}{\delta} \right) \right)^{\frac{1}{p-1}} (SA)^{\frac{1}{p(p-1)}} \end{aligned} \quad (87)$$

$$= 2B_p \left( \log \left( \frac{2SAT}{\delta} \right) \right)^{\frac{p}{p-1}} (SA)^{\frac{1}{p-1}} \left( \frac{4C_p + 2}{\lambda} \right)^{\frac{1}{p-1}} (7c)^{\frac{1}{p-1}} \quad (88)$$



$$= (4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} \left( \log \left( \frac{2SAT}{\delta} \right) \right)^{\frac{p}{p-1}} \left( \frac{SA}{\lambda} \right)^{\frac{1}{p-1}} \quad (89)$$

with probability at least  $1 - 3\delta$ . Since the regret incurred in non  $\lambda$ -bad episodes is smaller than  $\lambda T$ , the first statement of theorem is completed. For the second part, note that the expected regret in  $\frac{\delta}{2}$ -bad episodes is upper bounded by  $(4C_p + 2)^{\frac{p}{p-1}} (7c)^{\frac{p}{p-1}} \cdot \left( \log \left( \frac{2SAT}{\delta} \right) \right)^{\frac{p}{p-1}} \left( \frac{2SA}{\delta} \right)^{\frac{1}{p-1}} + 1$ . The remaining part of the proof follows from the proof of Theorem 4 in [6].  $\square$

**APPENDIX D  
PROOF OF THEOREM 3**

*Proof:* Since we assume that horizon  $T$  is unknown, we use an alternative approach for restarting:  $p$ -Heavy-UCRL2' restarts  $p$ -Heavy-UCRL2 with confidence parameter  $\frac{\delta}{\ell^2}$  at steps  $\tau_i = \left\lceil \frac{\ell^{(1+2\varepsilon)/\varepsilon}}{\ell^{(1+\varepsilon)/\varepsilon}} \right\rceil$  for  $i = 1, 2, 3, \dots$ , dividing the algorithm into  $n$  stages. Let  $n$  be the largest natural number such that  $\left\lceil \frac{n^{(1+2\varepsilon)/\varepsilon}}{\ell^{(1+\varepsilon)/\varepsilon}} \right\rceil \leq T$ , that is,  $n$  is the number of restarts up to step  $T$ . Then, by the definition of  $\tau_i$ , we can verify  $\frac{n^{(1+2\varepsilon)/\varepsilon}}{\ell^{(1+\varepsilon)/\varepsilon}} \leq \tau_n \leq T \leq \tau_{n+1} - 1 < \frac{(n+1)^{(1+2\varepsilon)/\varepsilon}}{\ell^{(1+\varepsilon)/\varepsilon}}$  holds and consequently we have the following inequality

$$\ell^{\frac{(1+\varepsilon)}{(1+2\varepsilon)} T^{\frac{\varepsilon}{(1+2\varepsilon)}}} - 1 \leq n \leq \ell^{\frac{(1+\varepsilon)}{(1+2\varepsilon)} T^{\frac{\varepsilon}{(1+2\varepsilon)}}} \quad (90)$$

The regret  $\Delta_c$  incurred in the  $\ell$  stages in which the MDP is restarted is bounded by multiplication of  $R_\Delta$  and total number of steps in these stages. The number of steps is maximized in the last  $\ell$  stages, which is  $T_\ell$  time steps. Note that time steps  $T_\ell$  contain at most  $\tau_{n+1} - 1 - \tau_{n-\ell+1}$  steps. Then

$$T_\ell \leq \tau_{n+1} - 1 - \tau_{n-\ell+1} \quad (91)$$

$$\leq \frac{(n+1)^{(1+2\varepsilon)/\varepsilon}}{\ell^{(1+\varepsilon)/\varepsilon}} - \frac{(n+1-\ell)^{(1+2\varepsilon)/\varepsilon}}{\ell^{(1+\varepsilon)/\varepsilon}} - \frac{1}{\ell^{(1+\varepsilon)/\varepsilon}} \quad (92)$$

$$\leq \frac{1}{\ell^{(1+\varepsilon)/\varepsilon}} \cdot \left( (n+1)^{(1+2\varepsilon)/\varepsilon} - (n+1-\ell)^{(1+2\varepsilon)/\varepsilon} \right) \quad (93)$$

$$= \frac{1}{\ell^{(1+\varepsilon)/\varepsilon}} \sum_{k=0}^{\infty} \binom{1+2\varepsilon}{k} n^{\frac{1+2\varepsilon}{\varepsilon} - k} \cdot (1 - (1-\ell)^k) \quad (94)$$

$$\leq \frac{1+2\varepsilon}{\varepsilon} \cdot n^{\frac{1+2\varepsilon}{\varepsilon}} \cdot \ell^{-\frac{1}{\varepsilon}} \quad (95)$$

$$\leq \frac{1+2\varepsilon}{\varepsilon} \cdot \ell^{\frac{\varepsilon}{1+2\varepsilon}} \cdot T^{\frac{1+\varepsilon}{1+2\varepsilon}} \quad (96)$$

where we used the generalized binomial theorem. Thus  $\Delta_c$  is bounded as follow,

$$\Delta_c \leq R_\Delta \cdot \ell \cdot T_\ell = R_\Delta \cdot \frac{(1+2\varepsilon)}{\varepsilon} \cdot \ell^{\frac{(1+\varepsilon)}{(1+2\varepsilon)}} \cdot T^{\frac{(1+\varepsilon)}{(1+2\varepsilon)}} \quad (97)$$

For the case that the MDP does not change between the steps  $\tau_i$  and  $\min\{T, \tau_{i+1}\}$ , the regret  $\Delta(s_{\tau_i}, T_i)$  for these  $T_i := \min\{T, \tau_{i+1}\} - \tau_i$  steps is bounded by applying Theorem 1.

With the confidence parameter  $\frac{\delta}{\ell^2}$ , we have

$$\Delta(s_{\tau_i}, T_i) \leq 2B_p \left( \log \left( \frac{\ell^2 2SAT_i}{\delta} \right) \right) (SAT_i)^{\frac{1}{p}} \quad (98)$$

$$\leq 2B_p \left( 3 \log \left( \frac{2SAT}{\delta} \right) \right) (SAT_i)^{\frac{1}{p}} \quad (99)$$

with probability at least  $1 - \frac{\delta}{4\ell^2 T_i^{5/4}}$ , where  $B_p := 7c(2C_p + 1)$ . Summing over all stages  $i = 1, \dots, n$ , the total regret  $\Delta_f$  is bounded by

$$\Delta_f = \sum_{i=1}^n \Delta(s_{\tau_i}, T_i) \quad (100)$$

$$\leq \sum_{i=1}^n 2B_p \left( 3 \log \left( \frac{2SAT}{\delta} \right) \right) (SAT_i)^{\frac{1}{p}} \quad (101)$$

$$\leq 2B_p \left( 3n^{\frac{1}{p}} \log \left( \frac{2SAT}{\delta} \right) \right) (SAT)^{\frac{1}{p}} \quad (102)$$

$$\leq 2B_p \left( 3n \log \left( \frac{2SAT}{\delta} \right) \right) (SAT)^{\frac{1}{p}} \quad (103)$$

$$\leq 2B_p \cdot \ell^{\frac{(1+\varepsilon)}{(1+2\varepsilon)}} \cdot T^{\frac{1+\varepsilon}{(1+2\varepsilon)}} \left( 3 \log \left( \frac{2SAT}{\delta} \right) \right) \cdot (SA)^{\frac{1}{p}} \quad (104)$$

with probability at least  $1 - \sum_{i=1}^n \frac{\delta}{4\ell^2 T_i^{5/4}}$ . Here, the inequality (102) is due to Jensen's inequality  $\sum_{i=1}^n T_i^{\frac{1}{p}} \leq (nT)^{\frac{1}{p}}$ . The remaining part is similar to the proof of Theorem 6 in [6].  $\square$

**APPENDIX E  
PROOF OF THEOREM 4**

We begin with Lemma 5, which is used to prove Theorem 4. Lemma 5 demonstrates that for any episode  $k \in [K]$ ,  $Q_h^k$  serves as an upper bound for the optimal  $Q$ -value function  $Q_h^*$ . In the proof of Theorem 4, Lemma 5 is utilized to bound the total regret of  $p$ -Heavy-Q-learning with Hoeffding-style bonus. This approach is similar to the scheme used in the proof of Theorem 1 in [14].

*Lemma 5:* Let us define  $b_t = r_{\max} c \sqrt{\frac{H^3 t}{t}}$ , and let  $\gamma_t := 2H(c \ln(2SAT)/\delta)/t^{1-\frac{1}{p}}$ . Then, for any  $\delta > 0$  and for any  $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ ,  $(Q_h^k - Q_h^*)$  satisfies the following bound with probability at least  $1 - \delta$ ,

$$\begin{aligned} 0 &\leq (Q_h^k - Q_h^*) \\ &\leq \alpha_t^0 H r_{\max} + \sum_{i=1}^t \alpha_t^i (V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i})) + 3b_t + \gamma_t. \end{aligned} \quad (105)$$

*Proof:* We have the following identity for  $Q_h^*$

$$\begin{aligned} Q_h^*(s, a) &= \alpha_t^0 Q_h^*(s, a) \\ &+ \sum_{i=1}^t \alpha_t^i \left[ \bar{r}_h(s, a) + (\mathbb{P}_h - \hat{\mathbb{P}}_h^{k_i}) V_{h+1}^*(s, a) + V_{h+1}^*(s, a) \right]. \end{aligned} \quad (106)$$

Then we can obtain following upper bound

$$\begin{aligned}
& (Q_h^k - Q_h^*(s, a)) \\
& \leq \alpha_t^0 (Hr_{\max} - Q_h^*(s, a)) \\
& + \sum_{i=1}^t \alpha_t^i \left[ (V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i})) + [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](s, a) + b_i \right] \\
& + \sum_{i=1}^t \alpha_t^i (\hat{r}(s, a) - \bar{r}(s, a)) \tag{107}
\end{aligned}$$

$$\begin{aligned}
& \leq \alpha_t^0 (Hr_{\max} - Q_h^*(s, a)) \\
& + \sum_{i=1}^t \alpha_t^i \left[ (V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i})) + [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](s, a) + b_i \right] \\
& + 2H \left( (c \ln(2SAT)/\delta) / t^{1-\frac{1}{p}} \right) \tag{108}
\end{aligned}$$

$$\begin{aligned}
& = \alpha_t^0 (Hr_{\max} - Q_h^*(s, a)) \\
& + \sum_{i=1}^t \alpha_t^i \left[ (V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i})) + \sum_{i=1}^t \alpha_t^i [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](s, a) \right. \\
& \left. + \sum_{i=1}^t \alpha_t^i b_i + 2H \left( (c \ln(2SAT)/\delta) / t^{1-\frac{1}{p}} \right) \right] \tag{109}
\end{aligned}$$

$$\begin{aligned}
& \leq \alpha_t^0 (Hr_{\max} - Q_h^*(s, a)) + \sum_{i=1}^t \alpha_t^i \left[ (V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i})) \right] \\
& + cr_{\max} \sqrt{\frac{H^3 \iota}{t}} + \sum_{i=1}^t \alpha_t^i b_i + 2H \left( (c \ln(2SAT)/\delta) / t^{1-\frac{1}{p}} \right) \tag{110}
\end{aligned}$$

with probability at least  $1 - \delta$ . We now denote  $\gamma_t$  as  $2H \left( (c \ln(2SAT)/\delta) / t^{1-\frac{1}{p}} \right)$ . Then by extending above inequality we have

$$\begin{aligned}
& \alpha_t^0 (Hr_{\max} - Q_h^*(s, a)) + \sum_{i=1}^t [(V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i}))] \\
& + \sum_{i=1}^t \alpha_t^i b_i + cr_{\max} \sqrt{\frac{H^3 \iota}{t}} + \gamma_t \\
& \leq \alpha_t^0 (Hr_{\max} - Q_h^*(s, a)) + \sum_{i=1}^t [(V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i}))] \\
& + 2cr_{\max} \sqrt{\frac{H^3 \iota}{t}} + cr_{\max} \sqrt{\frac{H^3 \iota}{t}} + \gamma_t \tag{111}
\end{aligned}$$

$$\begin{aligned}
& = \alpha_t^0 (Hr_{\max} - Q_h^*(s, a)) + \sum_{i=1}^t [(V_{h+1}^{k_i} - V_{h+1}^*(s_{h+1}^{k_i}))] + 3b_t + \gamma_t \tag{112}
\end{aligned}$$

This completes the proof.  $\square$

Now we present the proof of Theorem 4.

*Proof:* The proof is a modification of the proof of Theorem 1 in [14]. By applying lemma 5 and following the same argument in proof of Theorem 1 in [14], the additional part is a regret due to the heavy-tailed noise. Since the other

parts are same, we only introduce the extra regret incurred by heavy-tailed noise  $\sum_{h=1}^H \sum_{k=1}^K \gamma_{n_h^k}$ .

$$\sum_{k=1}^K \gamma_k = \sum_{s,a} \sum_{n=1}^{N_h^K(s,a)} 2H \left( (c \ln(2SAT)/\delta) / n^{1-\frac{1}{p}} \right) \tag{113}$$

$$= 2H(c \ln(2SAT)/\delta) \sum_{s,a} \sum_{n=1}^{N_h^K} n^{\frac{1}{p}-1} \tag{114}$$

$$\leq 2H(c \ln(2SAT)/\delta) K \left( \frac{K}{SA} \right)^{\frac{1}{p}-1} \tag{115}$$

$$\leq O(\iota HK^{\frac{1}{p}} (SA)^{1-\frac{1}{p}}) \tag{116}$$

where the first inequality holds since  $\sum_{s,a} \sum_{n=1}^{N_h^K(s,a)}$  is maximized when  $N_h^K(s, a) = \frac{K}{SA}$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . Considering the above result and following the same argument in the proof of Lemma 4.2 in [14], the total regret can be bounded as  $\sum_{k=1}^K \delta_1^k \leq O(r_{\max} \sqrt{H^4 SAT} \iota + \iota H^2 K^{\frac{1}{p}} (SA)^{1-\frac{1}{p}})$  with probability at least  $1 - 2\delta$ . Rescaling  $\delta$  to  $\delta/2$  completes the proof.  $\square$

## APPENDIX F

### PROOF OF THEOREM 5

The proof is an adaptation of the proof of Theorem 5 in [14]. We claim that with the following bonus term, we can derive Theorem 5 by using a similar argument in [14]. Unlike the Azuma-Hoeffding inequality, Bernstein's inequality includes the true variance of the optimal value function in the bound. However, since we do not know the true variance, we need to estimate it, which introduces an additional variance estimation error (Lemma 9). We first introduce the definition of the empirical variance term which is employed in the proofs as follows:

$$W_t(x, a, h) := \frac{1}{t} \sum_{i=1}^t \left[ V_{h+1}^{k_i}(x_{h+1}^{k_i}) - \frac{1}{t} \sum_{j=1}^t V_{h+1}^{k_j}(x_{h+1}^{k_j}) \right]^2 \tag{117}$$

where the state-action pair  $(x, a)$  was taken at step  $h$  for  $t$  times with  $k_1, \dots, k_t$  episodes, respectively. Additionally, the Bernstein-style bonus term for some constants  $c_1$  and  $c_2$  is defined as:

$$\begin{aligned}
\beta_t' := \min \left\{ \sqrt{\frac{Hr_{\max} \iota}{t} (W_t(s, a, h) + H)} + \frac{H^{\frac{3p-1}{p-1}} \iota \sqrt{SA(p-1)}}{t} \right. \\
\left. + \frac{H^2 \iota \sqrt{r_{\max}^3}}{t} + \frac{\iota \sqrt{H^7 SA r_{\max}}}{t}, c_2 r_{\max} \sqrt{\frac{H^3 \iota}{t}} \right\} \\
+ 2H \iota / t^{1-\frac{1}{p}} \tag{118}
\end{aligned}$$

where  $\iota = \log(2SAT)/\delta$  and accordingly, we have

$$\begin{aligned}
b_1(x, a, h) & := \frac{\beta_1(x, a, h)}{2}, \\
b_t(x, a, h) & := \frac{\beta_t(x, a, h) - (1 - \alpha_t) \beta_{t-1}(x, a, h)}{2\alpha_t}. \tag{119}
\end{aligned}$$

As in the Hoeffding-style Q-learning case, the following lemma still holds by using the update rule of Algorithm 2 and the Bellman optimality equation.

*Lemma 6 (Recursion on Q [14]):* For any  $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$  and episode  $k \in [K]$ , let  $t = N_h^k(x, a)$  and suppose that  $(x, a)$  was previously taken at step  $h$  of episodes  $k_1, \dots, k_t < k$ , then

$$\begin{aligned} (Q_h^k - Q_h^*)(x, a) &= \alpha_t^0(Hr_{max} - Q_h^*(s, a)) \\ &+ \sum_{i=1}^t \alpha_t^i \left[ (V_{h+1}^{k_i} - V_{h+1}^*)(x_{h+1}^{k_i}) \right. \\ &+ \left. [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](x, a) + b_i(x, a, h) \right] \\ &+ \sum_{i=1}^t \alpha_t^i (\hat{r}_h(s, a) - \bar{r}(s, a)) \end{aligned} \quad (120)$$

From Lemma 6, we can directly derive Lemma 7.

*Lemma 7:* There exists absolute constants  $c_2$  such that if  $\beta_t'(x, a, h) \leq c_1 r_{max} \sqrt{\frac{H^3 \iota}{t}} + \gamma_t$ , then, with probability at least  $1 - \delta$ , the following holds

$$\begin{aligned} \forall (x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K], \\ (V_h^k - V_h^*)(x_h^k) \leq \alpha_t^0 Hr_{max} + \sum_{i=1}^t (V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) + \beta_t' \end{aligned} \quad (121)$$

where  $t = N_h^k(s, a)$ ,  $\gamma_t := 2H(c \ln(2SAT)/\delta)/t^{1-\frac{1}{p}}$  and  $k_1, \dots, k_t < k$  are the episodes in which  $(x, a)$  was taken at step  $h$ .

The main challenge of the proof is to bound the term  $\sum_{i=1}^t \alpha_t^i [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](x, a)$  in Lemma 6. Specifically, since we do not know the  $V_{h+1}^*$ , we need to substitute  $V_{h+1}^*$  with its estimate  $W_t$ . Therefore, it is important to show how close  $V_{h+1}^*$  and  $W_t$  are, and the Lemma 8 is a key ingredient in proving this. Lemma 8 is a variant of Lemma C.7 in [14] with additional terms due to heavy-tailed noise, and is used as a component in the proof of Lemma 9. Additionally, Lemma 9 is used to address the estimation error between true variance and sample variance.

*Lemma 8 (Technical Lemma for Lemma 9):* Suppose Lemma 7 holds. For any  $h \in [H]$ , let  $\phi_h^k := (V_h^k - V_h^*)(x_h^k)$ , and let  $w = (w_1, \dots, w_k)$  be a nonnegative weight vector. Then, we have

$$\begin{aligned} \sum_{k=1}^K w_k \phi_h^k &\leq O(r_{max} \sqrt{H^5 \iota} \cdot (SA \|w\|_\infty + \sqrt{SA \|w\|_1 \|w\|_\infty})) \\ &+ H^2 SA \|w\|_\infty + H^2 \iota (SA \|w\|_\infty)^{1-\frac{1}{p}} (\|w\|_1)^{\frac{1}{p}}. \end{aligned} \quad (122)$$

*Proof:* From the optimistic choose of bonus term, we have  $V_h^k(x_h^k) \leq \max_{a' \in \mathcal{A}} Q_h^k(x_h^k, a') = Q_h^k(x_h^k, a_h^k)$ . Then, by using the Bellman optimality equation and Lemma 6, the following holds.

$$\phi_h^k = (V_h^k - V_h^*)(x_h^k) \leq (Q_h^k - Q_h^*)(x_h^k, a_h^k) \quad (123)$$

$$\leq \alpha_t^0 Hr_{max} + \sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k_i} + 3b_t + \gamma_t \quad (124)$$

where  $\gamma_t := 2H(c \ln(2SAT)/\delta)/t^{1-\frac{1}{p}}$ . In [14], the weighted sum of  $\phi_h^k$  is upper bounded by  $O(SA \|w\|_\infty \sqrt{H^5 \iota} + \sqrt{SA \|w\|_1 \|w\|_\infty} H^{\frac{3}{2}} \iota)$ , where  $\iota := \log(SAT/\delta)$ . In our proof, since we are dealing with heavy-tailed noise, an additional regret term occurs:

$$\sum_{k=1}^K w_k \gamma_k = \sum_{k=1}^K w_k 2H \left( c \ln(SAT)/\delta \right) \cdot (n_h^k)^{\frac{1}{p}-1} \quad (125)$$

$$\leq O(H \iota) \cdot \sum_{s,a} \sum_{i=1}^{N_h^k(s,a)} w_{k(i)} \left( \frac{1}{n_h^k} \right)^{1-\frac{1}{p}} \quad (126)$$

where  $w = (w_1, \dots, w_k)$  is a weight vector. Let us define  $d = \lfloor \frac{\|w\|_1}{SA \|w\|_\infty} \rfloor$ . Since  $\|w\|_1 = \sum_{x,a} \sum_{i=1}^{N_h^k(x,a)} w_{k_i}(x, a)$ , we can continue to write above inequality as following:

$$\begin{aligned} O(H \iota) \left( \|w\|_1 + \sum_{s,a} \sum_{i=1}^d \|w\|_\infty \left( \frac{1}{i} \right)^{1-\frac{1}{p}} \right) \\ = O(H \iota) \left( \sum_{s,a} \sum_{i=1}^{N_h^k(s,a)} w_{k_i}(x,a) + \sum_{s,a} \sum_{i=1}^d \|w\|_\infty \left( \frac{1}{i} \right)^{1-\frac{1}{p}} \right) \end{aligned} \quad (127)$$

$$\leq O(H \iota) \sum_{s,a} \|w\|_\infty \left( 1 + \sum_{i=1}^d \left( \frac{1}{i} \right)^{1-\frac{1}{p}} \right) \quad (128)$$

$$\leq O \left( H \iota \left( SA \|w\|_\infty + (SA \|w\|_\infty)^{1-\frac{1}{p}} (\|w\|_1)^{\frac{1}{p}} \right) \right) \quad (129)$$

Then, the weighted summation can be rewritten as

$$\begin{aligned} \sum_{k=1}^K w_k \phi_h^k \\ \leq Hr_{max} SA \|w\|_\infty + \sum_{k=1}^K w'_k \phi_{h+1}^{k'} + O(SA \|w\|_\infty \\ + \sqrt{SA \|w\|_1 \|w\|_\infty} \cdot \sqrt{H^3 \iota} \\ + O \left( H \iota \cdot \left( SA \|w\|_\infty + (SA \|w\|_\infty)^{1-\frac{1}{p}} (\|w\|_1)^{\frac{1}{p}} \right) \right)) \end{aligned} \quad (130)$$

By recursion this for  $h, h+1, \dots, H$ , we have

$$\begin{aligned} O \left( SA \|w\|_\infty r_{max} \sqrt{H^5 \iota} + r_{max} \sqrt{SA \|w\|_1 \|w\|_\infty} H^{\frac{5}{2}} \iota \right. \\ \left. + H^2 \iota SA \|w\|_\infty + H^2 \iota (SA \|w\|_\infty)^{1-\frac{1}{p}} (\|w\|_1)^{\frac{1}{p}} \right). \end{aligned} \quad (131)$$

□

Now, we introduce the estimation error between the empirical variance and true variance of the optimal value function.

*Lemma 9 (Variance Estimation Error):* There exists an absolute constant  $c > 0$  such that for any  $\delta \in (0, 1)$  and  $k \in [K]$ , with probability at least  $1 - \delta/K$ , if

$$\phi_h^k = (V_h^k - V_h^*)(x_h^k) \leq (Q_h^k - Q_h^*)(x_h^k, a_h^k) \leq \alpha_t^0 \phi_{h+1}^{k_i} + \beta_t' \quad (132)$$

holds and  $(Q_h^{k'} - Q_h^*)(x, a) \geq 0$  for all  $k' < k$ , then for all  $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ , we have

$$\begin{aligned} & \left| \mathbb{V}_h V_{h+1}^*(x, a) - W_t(x, a, h) \right| \\ & \leq O\left(H^2 r_{\max}^2 \sqrt{\frac{l}{t}} + r_{\max} \sqrt{H^7 l} \left(\frac{SA}{t} + \sqrt{\frac{SA}{t}}\right) + \frac{H^3 SA}{t} + H^3 \left(\frac{SA}{t}\right)^{1-\frac{1}{p}}\right) \end{aligned} \quad (133)$$

where  $t = N_h^k(x, a)$ .

*Proof:* Let us define  $k_i = \min(\{k \in [K] \mid k > k_{i-1} \text{ and } (x_h^k, a_h^k = (x, a))\} \cup \{K + 1\})$  for any  $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$  with  $k_0 = 0$ . Then the difference between true variance of value function  $\mathbb{V}_h V_{h+1}^*(x, a)$  and its estimate  $W_t(x, a, h)$  can be decomposed as following:

$$P_1 : [\mathbb{V}_h V_{h+1}^*](x, a) = \mathbb{E}_{x' \sim \mathbb{P}(\cdot | x, a)} [V_{h+1}^*(x')] - [\mathbb{P}_h V_{h+1}^*](x, a)^2 \quad (134)$$

$$P_2 : \frac{1}{t} \sum_{i=1}^t [V_{h+1}^*(x_{h+1}^{k_i} - [\mathbb{P}_h V_{h+1}^*](x, a))]^2 \quad (135)$$

$$P_3 : \frac{1}{t} \sum_{i=1}^t [V_{h+1}^* - \frac{1}{t} \sum_{j=1}^t V_{h+1}^*(x_{h+1}^{k_j})]^2 \quad (136)$$

$$P_4 : W_t(x, a, h) = \frac{1}{t} \sum_{i=1}^t [V_{h+1}^{k_i}(x_{h+1}^{k_i}) - \frac{1}{t} \sum_{j=1}^t V_{h+1}^{k_j}(x_{h+1}^{k_j})]^2 \quad (137)$$

Thus the upper bound of  $|\mathbb{V}V_{h+1}^* - W_t|$  becomes the summation of  $|P_1 - P_2|$ ,  $|P_2 - P_3|$ , and  $|P_3 - P_4|$  by triangle inequality. In addition, [14] proved that by using Azuma-Hoeffding inequality,  $|P_1 - P_2|$  and  $|P_2 - P_3|$  are bounded by  $cH^2 r_{\max}^2 \sqrt{l/t}$ . To bound  $|P_3 - P_4|$ , we apply Lemma 8 with a weight vector  $w$  such that  $w_{k_i} = \frac{1}{t}$  for all  $i \in [t]$ , but  $w_{k'} = 0$  for all  $k' \notin \{k_1, \dots, k_t\}$ , which means  $\|w\|_1 = 1$  and  $\|w\|_\infty = \frac{1}{t}$ . Then we have

$$\begin{aligned} & |P_3 - P_4| \\ & \leq \frac{4H}{t} \sum_{i=1}^t (V_{h+1}^{k_i}(x_{h+1}^{k_i}) - V_{h+1}^*(x_{h+1}^{k_i})) \end{aligned} \quad (138)$$

$$\begin{aligned} & \leq O\left(H \cdot (r_{\max} \sqrt{H^5 l} (SA \|w\|_\infty + \sqrt{SA \|w\|_1} \|w\|_\infty) + H^2 l SA \|w\|_\infty + H^2 l (SA \|w\|_\infty)^{1-\frac{1}{p}} (\|w\|_1)^{\frac{1}{p}})\right) \end{aligned} \quad (139)$$

$$\begin{aligned} & = O\left(H \left(r_{\max} \sqrt{H^5 l} \left(\frac{SA}{t} + \sqrt{\frac{SA}{t}}\right) + \frac{H^2 l SA}{t} H^2 \left(\frac{SA}{t}\right)^{1-\frac{1}{p}}\right)\right) \end{aligned} \quad (140)$$

$$\begin{aligned} & = O\left(r_{\max} \sqrt{H^7 l} \left(\frac{SA}{t} + \sqrt{\frac{SA}{t}}\right) + \frac{H^3 l SA}{t} + H^3 \left(\frac{SA}{t}\right)^{1-\frac{1}{p}}\right) \end{aligned} \quad (141)$$

Hence, the gap between the empirical variance and the actual variance for UCB-Bernstein is bounded by

$$\begin{aligned} & O\left(H^2 r_{\max}^2 \sqrt{\frac{l}{t}} + r_{\max} \sqrt{H^7 l} \left(\frac{SA}{t} + \sqrt{\frac{SA}{t}}\right) + \frac{H^3 l SA}{t} + H^3 \left(\frac{SA}{t}\right)^{1-\frac{1}{p}}\right). \end{aligned} \quad (142)$$

□

Lemma 10 shows the bound on total variance over  $K$  episodes. This lemma is applied to prove Lemma 11.

*Lemma 10 (Bound on Total Variance, Lemma C.5 in [14]):* There exists an absolute constant  $c$ , such that with probability at least  $1 - \delta$ ,

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}_h V_{h+1}^{\pi_k}(x_h^k, a_h^k) \leq c r_{\max}^2 (HT + H^3 l). \quad (143)$$

*Lemma 11 (Bound on  $Q_h^k - Q_h^*$ ):* For any  $\delta \in (0, 1)$ , there exists an absolute  $c_1, c_2 > 0$  such that under the choice of  $\beta_t(x, a, h)$  in the equality (118) with probability at least  $1 - 2\delta$ , the following holds simultaneously for all  $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ :

$$\begin{aligned} & (Q_h^k - Q_h^*)(x, a) \\ & \leq \alpha_t^0 H r_{\max} + \sum_{i=1}^t \alpha_t^i (V_{h+1}^{k_i} - V_{h+1}^*(x_{h+1}^{k_i})) + \beta_t' \end{aligned} \quad (144)$$

where  $t = N_h^k(s, a)$  and  $k_1, \dots, k_t < k$  are the episodes in which  $(x, a)$  was taken at step  $h$ .

*Proof:* The proof is an adaptation of the proof of Lemma C.4 in [14]. Reference [14] proved that the following holds

$$\begin{aligned} & \left| \sum_{i=1}^t \alpha_t^i \mathbb{I}[k_i \leq k] \cdot [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) V_{h+1}^*](x, a) \right| \\ & \leq O(1) \cdot \left[ \sqrt{\frac{H}{\tau} [\mathbb{V}_h V_{h+1}^*](x, a) l} + \frac{H^2}{\tau} r_{\max} l \right] \end{aligned} \quad (145)$$

with probability at least  $1 - \frac{\delta}{(SAT)}$ . By using Lemma 10, the inequality (145) can be written as follows,

$$\begin{aligned} & \left| \sum_{i=1}^t [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) V_{h+1}^*](x, a) \right| \\ & \leq O\left(\left[\sqrt{\frac{H}{t} [\mathbb{V}_h V_{h+1}^*](x, a) l} + \frac{H^2}{t} r_{\max} l\right]\right) \end{aligned} \quad (146)$$

$$\begin{aligned} & \leq O\left(\left\{\frac{H}{t} \cdot \left(W_t(s, a, h) r_{\max} + H r_{\max} + H^2 r_{\max}^2 \sqrt{\frac{l}{t}} + H^3 l \left(\frac{SA}{t} + \left(\frac{SA}{t}\right)^{1-\frac{1}{p}}\right)\right)\right\}^{1/2} + \frac{l \sqrt{H^7 SA} r_{\max}}{t}\right) \end{aligned} \quad (147)$$



$$\leq O\left(\left\{\frac{H}{t}\left(W_t(x, a, h)r_{\max} + Hr_{\max} + \left(\frac{1}{p}\right)\left(H + (p-1)\frac{H^{\frac{3p-1}{p-1}}SA_t}{t}\right)\right)\right\}^{1/2} + \frac{H^2\iota\sqrt{r_{\max}^3}}{t} + \frac{\iota\sqrt{H^7SAr_{\max}}}{t}\right) \quad (148)$$

$$\leq O\left(\sqrt{\frac{Hr_{\max}t}{t}(W_t(s, a, h) + H)} + \frac{H^{\frac{3p-1}{p-1}}\iota\sqrt{SA(p-1)}}{t} + \frac{H^2\iota\sqrt{r_{\max}^3}}{t} + \frac{\iota\sqrt{H^7SAr_{\max}}}{t}\right) \quad (149)$$

$$\leq \beta'_t \quad (150)$$

where the weighted AM-GM inequality is used. Finally, applying the above inequality to Lemma 6, we have for all  $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ ,

$$0 \leq (Q_h^k - Q_h^*)(x, a) - \alpha_t^0(Hr_{\max} - Q_h^*(s, a)) - \sum_{i=1}^t \alpha_t^i [(V_{h+1}^{k_i} - V_{h+1}^*)(x_{h+1}^{k_i})] - \sum_{i=1}^t \alpha_t^i (\hat{r}_h(x, a) - \bar{r}(x, a)) - \sum_{i=1}^t \alpha_t^i b_i \quad (151)$$

$$\leq \left| \sum_{i=1}^t \alpha_t^i [(\mathbb{P}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](x, a) \right| \quad (152)$$

$$\leq O(1) \cdot \left[ \sqrt{\frac{Hr_{\max}t}{t}(W_t(s, a, h) + H)} + \frac{H^{\frac{3p-1}{p-1}}\iota\sqrt{SA(p-1)}}{t} + \frac{H^2\iota\sqrt{r_{\max}^3}}{t} + \frac{\iota\sqrt{H^7SAr_{\max}}}{t} \right] \quad (153)$$

which completes the proof.  $\square$

Now we turn to present the proof of Theorem 5.

*Proof:* As in the proof of Hoeffding-style bonus term, by using Lemma 11, the total regret can be bounded by

$$\sum_{k=1}^K \delta_h^k \leq O(r_{\max}\sqrt{\iota SAH^4}) + \sum_{h'=h}^H \sum_{k=1}^K (\beta_{n_{h'}}^k (s_{h'}^k, a_{h'}^k, h') + \xi_{h+1}^k) \quad (154)$$

where  $\xi_{h+1}^k := [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)(V_{h+1}^* - V_{h+1}^k)](s, a)$ . Here,  $\sum_{h'=h}^H \sum_{k=1}^K \xi_{h=h'}^k \leq O(Hr_{\max}\sqrt{t})$  holds with probability  $1 - \delta$  by Azuma-Hoeffding inequality. Then the remaining part is  $\sum_{k=1}^K \sum_{h=1}^H \beta_{n_h}^k$ . From the definition of the bonus term  $\beta'_t$ , we have

$$\sum_{k=1}^K \sum_{h=1}^H \beta_{n_h}^k \leq \sum_{k=1}^K \sum_{h=1}^H O\left(\sqrt{\frac{Hr_{\max}t}{t}(W_t(s, a, h) + H)}\right)$$

$$+ \frac{H^{\frac{2p-1}{p-1}}\iota\sqrt{SA(p-1)}}{t} + \frac{H^2\iota\sqrt{r_{\max}^3}}{t} + \frac{\iota\sqrt{H^7SAr_{\max}}}{t} + \sum_{k=1}^K \sum_{h=1}^H (Ht)/t^{1-\frac{1}{p}}. \quad (155)$$

Then using the same argument as in the regret analysis of Theorem 4, we have

$$\sum_{k=1}^K \gamma = \sum_{s,a} \sum_{n=1}^{N_h^K(s,a)} 2H \left( c \ln(2SAT)/\delta \right) / n^{1-\frac{1}{p}} \quad (156)$$

$$\leq O(HK^{\frac{1}{p}}(SA)^{1-\frac{1}{p}}) = O(T^{\frac{1}{p}}H^{1-\frac{1}{p}}(SA)^{1-\frac{1}{p}}) \quad (157)$$

The upper bound of remaining terms in (155) can be obtained from steps in the proof of Theorem 5 in [25], which completes the proof.  $\square$

### APPENDIX G TECHNICAL LEMMA

Lemma 12 [14]:

$$\alpha_t = \frac{H+1}{H+t}, \quad \alpha_t^0 = \prod_{j=1}^t (1 - \alpha_j) \alpha_t^i = \alpha_i \prod_{j=i+1}^t (1 - \alpha_j) \quad (158)$$

Then, the following properties hold for  $\alpha_t^i$ :

- 1)  $\frac{1}{\sqrt{t}} \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$  for every  $t \geq 1$ .
- 2)  $\max_{i \in [t]} \alpha_t^i \leq \frac{2H}{t}$  and  $\sum_{i=1}^t (\alpha_t^i)^2 \leq \frac{2H}{t}$  for every  $t \geq 1$ .
- 3)  $\sum_{i=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$  for every  $i \geq 1$ .

Lemma 13 (Lemma 5 in [14]): For all  $\alpha_t^i$ , we have the following:

$$t^{-\frac{p-1}{p}} \leq \sum_{i=1}^t \alpha_t^i i^{-\frac{p-1}{p}} \leq 2t^{-\frac{p-1}{p}} \quad (159)$$

### REFERENCES

- [1] D. A. Freedman, "On tail probabilities for martingales," in *The Annals of Probability*. USA: Institute of Mathematical Statistics, 1975, pp. 100–118.
- [2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.
- [3] O. Cappe, E. Moulines, J.-C. Pesquet, A. P. Petropulu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Process. Mag.*, vol. 19, no. 3, pp. 14–27, May 2002.
- [4] S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Cham, Switzerland: Springer, 2007.
- [5] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for Markov decision processes," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1309–1331, Dec. 2008.
- [6] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, pp. 1563–1600, Aug. 2010.
- [7] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7711–7717, Nov. 2013.
- [8] I. Osband, D. Russo, and B. Van Roy, "Efficient reinforcement learning via posterior sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1–11.

- [9] C. Frankenberg, A. K. Thorpe, D. R. Thompson, G. Hulley, E. A. Kort, N. Vance, J. Borchardt, T. Krings, K. Gerilowski, C. Sweeney, S. Conley, B. D. Bue, A. D. Aubrey, S. Hook, and R. O. Green, "Airborne methane remote measurements reveal heavy-tail flux distribution in four corners region," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 35, pp. 9734–9739, Aug. 2016.
- [10] A. M. Medina and S. Yang, "No-regret algorithms for heavy-tailed linear bandits," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1642–1650.
- [11] I. Osband and B. Van Roy, "On lower bounds for regret in reinforcement learning," 2016, *arXiv:1608.02732*.
- [12] S. Agrawal and R. Jia, "Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–16.
- [13] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 263–272.
- [14] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
- [15] B. Xue, G. Wang, Y. Wang, and L. Zhang, "Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1–10.
- [16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
- [17] Y. Choi, K. Lee, and S. Oh, "Distributional deep reinforcement learning with a mixture of Gaussians," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9791–9797.
- [18] S. Lu, G. Wang, Y. Hu, and L. Zhang, "Optimal algorithms for Lipschitz bandits with heavy-tailed rewards," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4154–4163.
- [19] S. Ray Chowdhury and A. Gopalan, "Bayesian optimization under heavy-tailed payoffs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–6.
- [20] H. Bourel, O. Maillard, and M. S. Talebi, "Tightening exploration in upper confidence reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1056–1066.
- [21] K. Lee, S. Kim, S. Lim, S. Choi, M. Hong, J. I. Kim, Y.-L. Park, and S. Oh, "Generalized Tsallis entropy reinforcement learning and its application to soft mobile robots," *Robotics, Sci. Syst.*, vol. 16, pp. 1–10, Jul. 2020.
- [22] K. Lee, H. Yang, S. Lim, and S. Oh, "Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 8452–8462.
- [23] B. Xue, G. Wang, Y. Wang, and L. Zhang, "Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs," 2020, *arXiv:2004.13465*.
- [24] S. Agrawal, S. K. Juneja, and W. M. Koolen, "Regret minimization in heavy-tailed bandits," in *Proc. Conf. Learn. Theory*, 2021, pp. 26–62.
- [25] V. Zhuang and Y. Sui, "No-regret reinforcement learning with heavy-tailed rewards," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3385–3393.
- [26] K. Lee and S. Lim, "Minimax optimal bandits for heavy tail rewards," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 1–15, Sep. 2022.
- [27] H. Yang, H. Park, and K. Lee, "A selective portfolio management algorithm with off-policy reinforcement learning using Dirichlet distribution," *Axioms*, vol. 11, no. 12, p. 664, Nov. 2022.



**HYEON-JUN PARK** received the B.S. degree in mathematics from Chungnam National University, in 2021. He is currently pursuing the Ph.D. degree in artificial intelligence with Chung-Ang University, Seoul, South Korea. His current research interests include multi-armed bandit, reinforcement learning, and its application.



**KYUNGJAE LEE** received the B.S. and Ph.D. degrees in electrical and computer engineering from Seoul National University, in 2015 and 2020, respectively. He is currently an Assistant Professor with the Department of Artificial Intelligence, Chung-Ang University, Seoul, South Korea. His current research interests include multi-armed bandit, combinatorial bandits, reinforcement learning, and its application.

...