Minimax Optimal Bandits for Heavy Tail Rewards

Kyungjae Lee, *Member, IEEE*, and Sungbin Lim, *Member, IEEE*

***Abstract*— Stochastic multiarmed bandits (stochastic MABs) are a problem of sequential decision-making with noisy rewards, where an agent sequentially chooses actions under unknown reward distributions to minimize cumulative regret. The majority of prior works on stochastic MABs assume that the reward distribution of each action has bounded supports or follows light-tailed distribution, i.e., sub-Gaussian distribution. However, in a variety of decision-making problems, the reward distributions follow a heavy-tailed distribution. In this regard, we consider stochastic MABs with heavy-tailed rewards, whose pth moment is bounded by a constant ν_p for $1 < p \leq 2$. First, we provide theoretical analysis on sub-optimality of the existing exploration methods for heavy-tailed rewards where it has been proven that existing exploration methods do not guarantee a minimax optimal regret bound. Second, to achieve the minimax optimality under heavy-tailed rewards, we propose a minimax optimal robust upper confidence bound (MR-UCB) by providing tight confidence bound of a p-robust estimator. Furthermore, we also propose a minimax optimal robust adaptively perturbed exploration (MR-APE) which is a randomized version of MR-UCB. In particular, unlike the existing robust exploration methods, both proposed methods have no dependence on ν_p. Third, we provide the gap-dependent and independent regret bounds of proposed methods and prove that both methods guarantee the minimax optimal regret bound for a heavy-tailed stochastic MAB problem. The proposed methods are the first algorithm that theoretically guarantees the minimax optimality under heavy-tailed reward settings to the best of our knowledge. Finally, we demonstrate the superiority of the proposed methods in simulation with Pareto and Fréchet noises with respect to regrets.**

***Index Terms*— Heavy-tailed noise, mini-max optimality, multi-armed bandits (MABs), regret analysis.**

I. Introduction

A STOCHASTIC multiarmed bandit (stochastic MAB) is a fundamental decision-making problem under uncertain environment. In this problem, an intelligent agent selects an action among a set of K actions and receives a noisy reward corresponding to the selected action. Then, the goal of the agent is to find an optimal action, whose expected reward is the maximum, over total rounds T. However, due to the noise in rewards, the agent needs estimation of true expected rewards. Hence, the agent should explore entire set of actions including suboptimal one to obtain accurate estimations; however, selecting suboptimal actions for exploration will make the agent lose a large amount of rewards compared to an optimal one. In this regard, the agent faces a natural dilemma between exploration and exploitation: collecting more information to estimate rewards accurately (exploration) and selecting the best action based on experiences (exploitation).

The exploration methods should carefully balance this dilemma to efficiently find an optimal action. Specifically, efficiency of exploration methods can be measured by a cumulative regret which is defined as an expected cumulative difference between the maximum rewards and the expected reward of selected actions. Hence, the smaller the regret, the more efficient the algorithm. Most exploration methods have conducted regret analysis to guarantee their efficiency. Especially, the majority of researches have assumed that the noise of rewards follows sub-Gaussian distribution whose tail probability is dominated by the tail of Gaussian distribution. Under the sub-Gaussian assumption, it is well known that, for any algorithm, the gap-independent cumulative regret cannot be lower than $\Omega(\sqrt{KT})$ [1]. Several approaches have been proposed to achieve the gap-independent lower bound $\Omega(\sqrt{KT})$, which is called a minimax optimal [2], [3], [4], [5], [6].

While many methods have been studied under sub-Gaussian noise, there still needs developing a robust exploration method to address real-world problems which are not covered by sub-Gaussian assumptions. However, few researches have investigated the stochastic MAB problem under heavy-tailed noise whose pth moment is bounded by a constant ν_p. In general, heavy-tailed noise covers wider range of noise distributions than sub-Gaussian noise. Bubeck *et al.* [7] have first addressed the heavy-tailed noise in a bandit problem by proposing a robust upper confidence bound (robust UCB) whose gap-independent regret is $O((K \ln(T)^{1-1/p} T^{1/p})$. Furthermore, Bubeck *et al.* [7] have shown that, for any algorithm, the worst case cumulative regret cannot be lower than $\Omega(K^{1-1/p} T^{1/p})$ under heavy-tailed noise assumptions. To achieve the lower bound, Lee *et al.* [8] have proposed perturbation-based explorations called APE^2 which has achieved $O(K^{1-1/p} T^{1/p} \ln(K))$ regret bound that is optimal only with respect to T but is still suboptimal with respect to K as factor $\ln(K)$. Furthermore, Wei and Srivastava [9] have proposed minimax optimal strategy by modifying the upper confidence bound of the truncated mean estimator, but the truncated mean estimator requires the prior knowledge about ν_p, which is not desirable for the bandit setting that assumes no prior knowledge about reward distributions. In this regard, we develop the minimax optimal exploration for heavy-tailed rewards without using the problem-dependent knowledge, ν_p.

Manuscript received 13 January 2022; revised 1 June 2022; accepted 22 August 2022. This work was supported in part by the Chung-Ang University Research Grants in 2021 and in part by the Institute for Information & communication Technology Planning & evaluation (IITP) Grant funded by the Korean Government (MSIT) (No. 2022-0-00612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI). *(Corresponding author: Kyungjae Lee.)*

Kyungjae Lee is with the Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea (e-mail: kyungjae.lee@ai.cau.ac.kr).

Sungbin Lim is with the Artificial Intelligence Graduate School and the Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, South Korea (e-mail: sungbin@unist.ac.kr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3203035.

Digital Object Identifier 10.1109/TNNLS.2022.3203035

In this article, we propose a true minimax optimal exploration method which can guarantee $\Theta(K^{1-1/p}T^{1/p})$ regret bound without using the prior information of ν_p. To remove the dependency of ν_p, we employ a robust estimator which is proposed in our prior work [8]. In [8], we have proposed the robust estimator, called a p-robust estimator, whose error probability decays exponentially fast while it does not depend on ν_p. More specifically, the error probability follows $O(\exp(-n^{1-1/p}\epsilon))$ where n is the number of sample and ϵ is an error bound we consider. Note that the proposed robust mean estimator has worse decaying rate than other existing robust mean estimators, which have $O(\exp(-n\epsilon^{p/(p-1)}))$, but it does not require prior information about ν_p while other estimators essentially need ν_p to guarantee the decaying rate $O(\exp(-n\epsilon^{p/(p-1)}))$. Since the p-robust estimator has worse decaying rate, naíve UCB style exploration with the p-robust estimator shows suboptimal regret bound while it can remove dependency on ν_p. Hence, to reduce the regret bound, we modify the confidence bound of p-robust estimator by borrowing a technique in MOSS [10]. Furthermore, we also extend modified upper confidence bound to the perturbation-based exploration and we derive the condition of perturbation for the minimax optimality. From the theoretical results in Lee *et al.* [8], we first prove that the unbounded perturbation, whose supporting set is unbounded, cannot achieve the minimax optimality since $\ln(K)$ factor cannot be removed. However, we also prove that we should employ a bounded perturbation to reduce the sup-optimal factor $\ln(K)$. We finally propose a randomized version of robust UCB for the minimax optimality by combining the bounded perturbation method and modified confidence bound. We believe that the proposed methods can be extended into further structured bandit problems such as [11], [12], [13], [14], [15], [16], [17], and [18]. Our contribution can be summarized as follows.

1) We analyze that robust UCB and perturbation-based exploration cannot achieve the minimax optimality. Especially, unbounded perturbation cannot remove the suboptimal factor $\ln(K)$.
2) For the minimax optimality, we propose a modified upper confidence bound and prove that its gap-independent regret bound can matches to the lower bound $\Omega(K^{1-1/p}T^{1/p})$. Hence, the modified upper confidence bound method is minimax optimal.
3) We also propose a bounded perturbation method by combining with the modified upper confidence bound and prove that its gap-independent bound also matches to the lower bound $\Omega(K^{1-1/p}T^{1/p})$. Thus, the proposed bounded perturbation method is minimax optimal.
4) For both modified upper confidence bound and bounded perturbation method, we employ the p-robust estimator in [8] that does not require ν_p as a prior knowledge. In this regard, the proposed exploration methods have no dependency on ν_p while, interestingly, they can achieve the minimax optimality.
5) We also verify the proposed methods show superior performance compared to other robust exploration methods for heavy-tailed noise.

II. Background

Consider a set of K actions, $\mathcal{A} := \{a_1, \ldots, a_K\}$, and corresponding mean rewards $\{r_{a_1}, \ldots, r_{a_K}\}$. At time $t = 1, 2, \ldots, T$, an exploration algorithm chooses an action a_t and receives a noisy reward for the selected action

$$\mathbf{R}_{t,a_t} := r_{a_t} + \epsilon_{t,a_t} \tag{1}$$

where ϵ_{t,a_t} is an identical and independently distributed zero mean random noise for each time step and each action. In multiarmed bandits (MABs), r_{a_k} is generally assumed to be unknown. Then, the goal of the exploration method is to efficiently verify an optimal action $a_\star := \arg\max_a r_a$. The performance of the exploration strategy is often measured by the cumulative regret over total round T, defined as follows:

$$\mathcal{R}_T := \sum_{t=1}^{T} r_\star - \mathbb{E}_{1:t}\big[r_{a_t}\big] = \sum_{k=1}^{K} \Delta_{a_k} \mathbb{E}\big[n_{a_k}(T)\big] \tag{2}$$

where $r_\star := \max_{a\in\mathcal{A}} r_a$ and $n_a(t)$ is the number of times selecting a over t rounds, i.e., $n_a(t) = \sum_{k=1}^{t} \mathbb{I}[a_k = a]$. Hence, the smaller $\mathcal{R}_T$, the better exploration performance.

A. Minimax Optimality Under Sub-Gaussian Noise

In stochastic MABs, many researches usually assume that each $\epsilon_{t,a}$ follows a σ_a-sub-Gaussian distribution with zero mean, that is, the following inequality holds for all $s \in \mathbb{R}$ and $a \in \mathcal{A}$:

$$\mathbb{E}\big[\exp\big(s\big(\epsilon_{t,a} - \mathbb{E}\big[\epsilon_{t,a}\big]\big)\big)\big] \le \exp\big(\sigma^2 s^2/2\big). \tag{3}$$

Under the sub-Gaussian assumption, it is well known that the gap-dependent lower bound is $\Omega(\sum_{a_k\neq a_\star} \ln(T)/\Delta_{a_k})$ and the gap-independent lower bound is $\Omega(\sqrt{KT})$, respectively, where Ω indicates a lower bound [2], [19], [20]. There exist several minimax optimal methods which guarantee matching the lower bound to solve the stochastic MABs under sub-Gaussian noise. In this article, we introduce two well-known algorithms using confidence bounds under sub-Gaussian assumptions, which is highly related to the proposed method. Auer *et al.* [19] have proposed upper confidence bound (UCB) using the confidence bound of sample mean estimators, i.e., $(2\ln(T)/n_a(t))^{1/2}$. Audibert and Bubeck [2] have analyzed that the minimax regret bound of UCB is $O((KT\ln(T))^{1/2})$ that is suboptimal. Hence, Audibert and Bubeck [2] have extended UCB to minim-max optimal strategy in stochastic MAB (MOSS) by modifying the confidence bound as $(\ln_+(T/(Kn_a(t)))/n_a(t))^{1/2}$ where $\ln_+(x) = \max(\ln(x), 0)$. From this modification, MOSS achieved the minimax optimal regret bound $\Theta(\sqrt{KT})$.

B. Minimax Optimality Under Heavy-Tailed Noise

While sub-Gaussian assumption has been well analyzed, only few methods have extended an assumption on noise to heavy-tailed noise whose pth moment is bounded, i.e.,

$$\mathbb{E}\big[|\epsilon|^p\big] \le \nu_p \tag{4}$$

where ν_p is a constant and $p \in (1, 2]$ is the maximum number of the bounded moments. For heavy-tailed

TABLE I
REGRET BOUNDS OF ALGORITHMS WITH PRIOR INFORMATION

Algorithm		Gap-Dependent Bound $O(\cdot)$	Gap-Independent Bound $\Theta(\cdot)$	Prior Info.
Robust UCB [7]		$\sum_{a\neq a_\star} \ln(T)/\Delta_a^{1/(p-1)}$	$(K\ln(T))^{1-\frac{1}{p}}T^{\frac{1}{p}}$	p and ν_p
Robust MOSS [9]		$\sum_{a\neq a_\star} \ln\left(T\Delta_a^{p/(p-1)}/K\right)/\Delta_a^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}$	
MR-UCB [This work]		$\sum_{a\neq a_\star} \ln\left(T\Delta_a^{p/(p-1)}/K\right)^{p/(p-1)}/\Delta_a^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}$	p
APE^2 (Unbounded) [8]	TYPE I	$\sum_{a\neq a_\star} \ln\left(T\Delta_a^{p/(p-1)}\right)^{p/(p-1)}/\Delta_a^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}\ln(K)$	
	TYPE II	$\sum_{a\neq a_\star} \ln(K)^{\frac{p}{p-1}}\left(T\Delta_a^{p/(p-1)}\right)^{\frac{p}{\ln(K)(p-1)}}/\Delta_a^{1/(p-1)}$		
MR-APE^2 (Bounded) [This work]		$\sum_{a\neq a_\star} \ln\left(T\Delta_a^{p/(p-1)}/K\right)^{p/(p-1)}/\Delta_a^{1/(p-1)}$	$K^{1-\frac{1}{p}}T^{\frac{1}{p}}$	

$O(\cdot)$ is an upper bound. $\Theta(\cdot)$ is a tight bound. Prior Info. indicates prior information. p indicates the maximum order of the bounded moment of rewards. ν_p is an upper bound of the p-th moment of rewards. *Unbounded* and *Bounded* indicates unbounded perturbations and bounded perturbations, respectively. TYPE I and II indicates the type of distribution of perturbations. TYPE I includes Weibull, Gamma, and generalized extreme value (GEV) distribution, TYPE II includes Pareto and Fréchet distribution

noise, it is well known that the gap-dependent lower bound is $\Omega(\sum_{a\neq a_\star}\ln(T)/\Delta_a^{1/(p-1)})$ and the gap-independent lower bound is $\Omega(K^{1-1/p}T^{1/p})$ [7]. However, most algorithms suffer from the sub-optimality in terms of the gap-independent regret bounds. Bubeck *et al.* [7] have first proposed the robust UCB using the confidence bounds of general robust estimators. Bubeck *et al.* [7] have analyzed the regret bound of the robust UCB where the gap-dependent bound is $O(\sum_{a\neq a_\star}\ln(T)/\Delta_a^{1/(p-1)}+\Delta_a)$ and gap-independent bound is $O((K\ln(T))^{1-1/p}T^{1/p})$, respectively. However, the robust UCB requires ν_p in prior to define a confidence bound of the robust estimator. Then, this condition restricts the viability of robust UCB since ν_p is generally not accessible in bandit settings. Furthermore, the upper regret bound of robust UCB has the suboptimal factor of $\ln(T)^{1-1/p}$. More precisely, Lee *et al.* [8] have proven that the lower bound of robust UCB is also $\Omega((K\ln(T))^{1-1/p}T^{1/p})$, hence, it is a tight bound. In other words, unfortunately, we cannot remove the suboptimal factor, $\ln(T)^{1-1/p}$. A similar restriction also appears in [21]. Vakili *et al.* [21] have proposed a deterministic sequencing of exploration and exploitation (DSEE) by exploring every action with a deterministic sequence. It is shown that DSEE has the gap-dependent bound $O(\ln(T))$, but, its result holds when ν_p and the minimum gap $\min_{a\in\mathcal{A}/a_\star}\Delta_a$ are known as prior information. Furthermore, in practice, DSEE often shows poor performance since the deterministic sequence cannot perform adaptive exploration. While other existing robust exploration methods have not guaranteed the minimax optimality, Wei and Srivastava [9] have recently proposed robust version of MOSS which can guarantee $\Theta(K^{1-1/p}T^{1/p})$; however, the robust MOSS has a limitation in that ν_p is an essential prior information to achieve the minimax optimality. Agrawal *et al.* [22] also have proposed KL_{inf}-UCB by adding two variants to the original UCB algorithm and proved that the problem-dependent regret bound of KL_{inf}-UCB is $O(\log(T)^{2/3})$; however, it also requires ν_p as a prior knowledge to achieve the proposed regret bound.

The dependence on ν_p is a crucial issue in a bandit problem since ν_p is problem-dependent prior information. Cesa-Bianchi *et al.* [23] have first removed in [23] only for $p=2$ by developing a robust estimator using the influence function in the Catoni's M estimator [24]. For exploration, the Boltzmann–Gumbel exploration (BGE) has been proposed. We observe one interesting fact that the robust estimator proposed in [23] has a weak tail bound, whose error probability decays slower than that of the original Catoni's M estimator [24]. However, BGE achieved gap-dependent bound $O(\sum_{a\neq a_\star}\ln(T\Delta_a^2)^2/\Delta_a+\Delta_a)$ and gap-independent bound $O(\sqrt{KT}\ln(K))$ for $p=2$. While $\ln(K)$ factor remains, BGE has a better bound than robust UCB in terms of T. Lee *et al.* [8] have extended Cesa's estimator to a p-robust estimator for $p\in(1,2]$ and have applied perturbation-based exploration inspired by BGE, which is named an adaptively perturbed exploration with a p-robust estimator (APE^2). By combining p-robust estimator and perturbation methods, Lee *et al.* [8] showed that APE^2 can achieve the regret bound of $O(K^{1-1/p}T^{1/p}\ln(K))$ which is partially optimal with respect to T but suboptimal with respect to K as the factor of $\ln(K)$.

In this article, we apply the idea of MOSS to our p-robust estimator in [8] where upper confidence bound of the p-robust estimator is modified to be tighter than the original UCB. By combining MOSS and p-robust estimator, we can enjoy both benefits of MOSS, i.e., the minimax optimality, and the p-robust estimator, i.e., independence on ν_p. Then, we also propose a randomized version of robust UCB by extending the modification of robust UCB to the perturbation-based exploration method. A comparison of existing robust exploration methods including ours can be shown in Table I. Table I shows a gap-dependent and gap-independent regret bounds and essential prior information.

III. SUB-OPTIMALITY OF EXISTING METHODS

In this section, we discuss pessimistic results about existing methods. First, we restate the sub-optimality of the robust UCBs of Bubeck *et al.* [7]. Second, we newly prove the sub-optimality of the unbounded perturbation methods in Lee *et al.* [8]. The perturbation-based exploration employs a random perturbation to encourage exploration. Hence, its cumulative regret is closely related to the distribution of random perturbation and Lee *et al.* [8] have revealed the relationship between distribution of perturbation and cumulative regret bounds. Unfortunately, from the results of Lee *et al.* [8], we prove that the perturbation-based exploration is minimax suboptimal if the random perturbation is unbounded.

A. Sub-Optimality of Robust UCBs

The robust UCB employs a class of robust estimators which satisfies the following assumption.

Assumption 1 (in [7]): Let $\{R_k\}_{k=1}^{\infty}$ be i.i.d. random variables with the finite pth moment for $p \in (1, 2]$. Let $\nu_p \geq \mathbb{E}[|R_k|^p]$ and $r = \mathbb{E}[R_k]$. Assume that, for all $\delta \in (0, 1)$ and n number of observations, there exists an estimator $\hat{r}_n(\eta, \nu_p, \delta)$ with a parameter η such that

$$\mathbb{P}\big(\hat{r}_n > r + \nu_p^{1/p}(\eta \ln(1/\delta)/n)^{1-1/p}\big) \leq \delta \tag{5}$$

and

$$\mathbb{P}\big(r > \hat{r}_n + \nu_p^{1/p}(\eta \ln(1/\delta)/n)^{1-1/p}\big) \leq \delta. \tag{6}$$

There exist several robust estimators that satisfy Assumption 1, such as truncated mean, median of mean, and Catoni's M estimator [24]. This assumption naturally provides the confidence bound of the estimator $\hat{r}_n$, hence, we can easily employ UCB-based exploration with the robust estimators in Assumption 1. However, we would like to note that the estimator in Assumption 1 essentially requires ν_p as prior information to define the estimator, which is not available under bandit setting.

Using the confidence bound in Assumption 1, we can derive a robust UCB strategy. For every step, robust UCB chooses an action based on the following strategy:

$$a_t := \arg\max_{a \in \mathcal{A}} \left\{ \hat{r}_{t-1,a} + \nu_p^{1/p} \left(\frac{\eta \ln\big(t^2\big)}{n_a(t-1)} \right)^{1-1/p} \right\} \tag{7}$$

where $\hat{r}_{t-1,a}$ is an estimator which satisfies Assumption 1 with $\delta := t^{-2}$. In our previous work, we have shown that there exists a MAB problem that makes the strategy (7) have the following lower bound of $\mathcal{R}_T$.

Theorem 1 (in [8]): There exists a K-armed stochastic bandit problem for which the regret of robust UCB has the following lower bound, for $T > \max(10, [(\nu^{(1/(p-1))}/\eta(K-1))]^2)$:

$$\mathcal{R}_T \geq \Omega\big((K \ln(T))^{1-1/p} T^{1/p}\big). \tag{8}$$

The proof can be found in [8]. Theorem 1 clearly shows that the lower regret bound of the robust UCB is $\Omega((K \ln(T))^{1-1/p} T^{1/p})$. The theorem tells that there always exists a MAB problem that causes the suboptimal regret bound for the robust UCB. Hence, the robust UCB cannot remove the suboptimal factor $\ln(T)^{1-1/p}$ from the gap-independent regret bound. Consequently, the robust UCB has two main drawbacks for a stochastic MAB. First, theorem 1 tells us the pessimistic fact that the sub-optimality of the robust UCB is caused by a fundamental issue of exploration strategy, rather than, by the lack of mathematical techniques such as employing a loose upper bound. Secondly, the estimators employed in the robust UCB usually require ν_p as a prior knowledge.

B. Sub-Optimality of Adaptively Perturbed Exploration With Unbounded Perturbation

Lee *et al.* [8] have proposed an APE^2 that can guarantee the minimax optimality with respect to T while removing dependency on ν_p. However, it still has a limitation in that its gap-independent regret bound is suboptimal with respect to K. Especially, we prove that unbounded perturbation cannot guarantee the minimax optimality in heavy-tailed MAB problems.

In APE^2, Lee *et al.* [8] have extended Catoni's M estimator by generalizing Catoni's influence function where a new influence function $\psi_p(x)$ is defined as

$$\psi_p(x) := \operatorname{sgn}(x) \ln\big(b_p|x|^p + |x| + 1\big) \tag{9}$$

where $\operatorname{sgn}(x)$ is a sign of x, $\mathbb{I}[\cdot]$ is an indicator function, and

$$b_p := \left[2\left(\frac{(2-p)}{(p-1)}\right)^{1-2/p} + \left(\frac{(2-p)}{(p-1)}\right)^{2-2/p} \right]^{-\frac{p}{2}}.$$

Using $\psi_p(x)$, Lee *et al.* [8] define a p-robust estimator and derive its confidence bounds as follows.

Theorem 2 (in [8]): Let $\{Y_k\}_{k=1}^{\infty}$ be i.i.d. random variables sampled from a heavy-tailed distribution with a finite pth moment, $\nu_p := \mathbb{E}|Y_k|^p$, for $p \in (1, 2]$. Let $y := \mathbb{E}[Y_k]$ and define an estimator as

$$\hat{Y}_n := \frac{c}{n^{1-1/p}} \cdot \sum_{k=1}^{n} \psi_p\left(\frac{Y_k}{cn^{1/p}}\right) \tag{10}$$

where $c > 0$ is an arbitrary constant. Then, for all $\delta > 0$

$$\mathbb{P}\big(\hat{Y}_n > y + c \ln\big(\exp\big(b_p \nu_p / c^p\big)/\delta\big)/n^{1-1/p}\big) \leq \delta \tag{11}$$

and

$$\mathbb{P}\big(y > \hat{Y}_n + c \ln\big(\exp\big(b_p \nu_p / c^p\big)/\delta\big)/n^{1-1/p}\big) \leq \delta. \tag{12}$$

The entire proof can be found in [8]. Compared to Assumption 1, a p-robust estimator has clear benefits in that a p-robust estimator does not depends on ν_p while robust estimators defined in Assumption 1 require ν_p as prior knowledge to guarantee the confidence bounds. This property of a p-robust estimator makes APE^2 independent on ν_p. However, a p-robust estimator has a drawback since the confidence bound of (10) is looser than Assumption 1 for a fixed δ.

By combining the estimator in (10) with a perturbation method, APE^2 selects an action based on the following decision rule:

$$a_t := \arg\max_{a \in \mathcal{A}} \big\{ \hat{r}_{t-1,a} + \beta_{t-1,a} G_{t,a} \big\} \tag{13}$$

where $\beta_{t-1,a} := c/(n_a(t-1))^{1-1/p}$, $n_a(t-1)$ is the number of times a has been selected, $G_{t,a}$ is sampled from F, and $F(g) := \mathbb{P}(G < g)$.

The lower bound of APE^2 is derived by constructing a counterexample as follows.

Theorem 3 (in [8]): Let $F(g)$ be a log-concave CDF. For $0 < c < (K-1)/(K-1+2^{p/(p-1)})$ and $T \geq (c^{1/(p-1)}(K-1)/2^{p/(p-1)})|F^{-1}(1-(1/K))|^{p/(p-1)}$, there exists a K-armed stochastic bandit problem where the regret of APE^2 is lower bounded by

$$\mathcal{R}_T \geq \Omega\big(K^{1-1/p} T^{1/p} F^{-1}(1-1/K)\big). \tag{14}$$

The proof is done by constructing the worst case bandit problem whose rewards are deterministic. When the rewards

are deterministic, no exploration is required, but, APE^2 unnecessarily explores suboptimal actions due to the perturbation. In other words, the lower bound captures the regret of APE^2 caused by useless exploration. The lower bound tells us that tail behavior of perturbation plays a crucial role in determining the effect of K on the regret bound. From the lower bound, we can derive a novel pessimistic result on APE^2 that employs unbounded perturbation for exploration.

Corollary 1: If the support of $F(g)$ is bounded, then, the lower bound of $\mathcal{R}_T$ of APE^2 becomes $\Omega(K^{1-1/p}T^{1/p})$.

Corollary 1 is induced by Theorem 3. Due to the term $F^{-1}(1-1/K)$, if G has an unbounded support, then, $F^{-1}(1-1/K)$ will grow as K increases and, thus, the lower bound of APE^2 has a suboptimal dependency on K. In other words, if G is unbounded, then, the lower bound of APE^2 cannot match $\Omega(K^{1-1/p}T^{1/p})$. From this observation, we conclude that bounded perturbation is needed to obtain the minimax optimality. Furthermore, from the observation of the sub-optimality of the robust UCB, we argue that the confidence bound of robust estimator in [7] is too loose to capture the error tightly and, thus, causes unnecessary exploration. To handle this issue, we modify the confidence bound of a p-robust estimator much tighter and extend the modified confidence bound to the perturbation method.

IV. Minimax Optimal Strategy for Heavy-Tailed Rewards

We propose two novel exploration methods to guarantee the minimax optimality under heavy-tailed noise. The first one is a minimax optimal robust upper confidence bound (MR-UCB), whose confidence bound is modified to a much tighter one, and the second one is a minimax optimal robust adaptively perturbed exploration (MR-APE), which is a randomized version of robust UCB using a bounded perturbation. The main benefit of MR-UCB and MR-APE is not only minimax optimality but also the minimal requirement of prior knowledge.

A. Minimax Optimal Robust UCB

In general, the regret bound of UCB often depends on the convergence rate of estimators. Especially, a robust estimator should satisfy two key properties to achieve efficient exploration performance. The first one is that the error probability decays exponentially fast and the second one is tight confidence bound for exploration. The main idea to design a minimax optimal exploration without dependency on ν_p is employing a p-robust estimator with tight confidence bound. The p-robust estimator satisfies exponential decaying from Theorem 2. However, if we employ the naïve confidence bound in (11) and (12), then, its minimax regret bound is suboptimal with respect to T. Hence, we propose more tight confidence bound than the naïve confidence bound. In MR-UCB, the selection rule is defined as

$$a_t := \arg\max_{a\in\mathcal{A}} \left\{\hat{r}_{t-1,a} + \beta_{t-1,a}\right\} \tag{15}$$

$$\beta_{t-1,a} := c\ln_+\left(\frac{T}{K n_a(t-1)}\right)/[n_a(t-1)]^{1-1/p} \tag{16}$$

where $\ln_+(x) := \max(\ln(x), 1)$. Similar to MOSS [2], we simply modify confidence bound from the naïve confidence bound, $O(\ln(T))$, to tighter one, $O(\ln(T/n_a(t-1)))$, that becomes tighter than $O(\ln(T))$ as the number of selecting a increases. Then, we derive the gap-dependent and gap-independent regret bounds as follows.

B. Minimax Optimal Robust Adaptively Perturbed Exploration

MR-APE is a randomized algorithm of MR-UCB. MR-APE replaces the optimism in MR-UCB with simple randomization. Instead of directly employing the confidence bound of the p-robust estimator, MR-APE is to employ a value randomly chosen between lower and upper confidence intervals using bounded perturbation within $[-1, 1]$. Then, the selection rule of MR-APE is defined as

$$a_t := \arg\max_{a\in\mathcal{A}} \left\{\hat{r}_{t-1,a} + (1+\epsilon)\beta_{t-1,a}G_{t,a}\right\} \tag{17}$$

where $G_{t,a}$ is a bounded random perturbation within $[-1, 1]$ and ϵ is an auxiliary hyperparameter. If the sampled perturbation is negative, the perturbation term can be interpreted as the lower confidence bound. Otherwise, the perturbation term is similar to the upper confidence bound. Hence, MR-APE employs both lower and upper confidence bounds for decision-making. Furthermore, if we set $G_{t,a} = 1$ and $\epsilon = 0$ almost surely, then, MR-APE is equivalent to MR-UCB. The entire algorithm is summarized in Algorithm 1.

C. Theoretical Analysis

We provide gap-dependent and gap-independent upper bounds of the cumulative regret of MR-UCB and MR-APE. First, we derive the gap-dependent regret bounds and then, extend gap-dependent bounds to the gap-independent bounds. The main idea of our proof is decomposing the event of selecting suboptimal actions into three events. Before decomposition, we assume that $r_{a_1} > r_{a_2} > r_{a_3} > \cdots > r_{a_K}$ without loss of generality. Then, let us define $Z := \min_{1<t\le T} \hat{r}_{t-1,a^*} + \beta_{t-1,a^*}$ and $z_a := r_{a^*} - \Delta_a/6$. Then, using Z and z_a, we define the event $\bar{E}_a := \{z_a \le Z\}$. Based on $\bar{E}_a$, we decompose the expected regret into three terms as follows, for any $k_0 \in [1, \ldots, K]$:

$$\mathcal{R}_T \le T\Delta_{a_{k_0}} + T\sum_{j=k_0+1}^{K} \mathbb{P}\left(\bar{E}_{a_j}^c\right)\left(\Delta_{a_j} - \Delta_{a_{j-1}}\right) \tag{18}$$

$$+ \sum_{k=k_0+1}^{K} \Delta_{a_k} \sum_{t=1}^{T} \mathbb{P}\left(\bar{E}_{a_k} \cap E_{t,a_k}\right) \tag{19}$$

where $E_{t,a} := \{a_t = a\}$ indicates the event of selecting a at time t. By computing the bound of each term, we can derive the gap-dependent and gap-independent upper bounds. We would like to note that this decomposition technique follows the proof of MOSS [2] and it generally holds without any special assumption on reward distributions. However, in [2], the remaining part for proving the minimax optimality of MOSS heavily depends on the sub-Gaussian assumption. In particular, to prove the minimax optimality, the second term

Algorithm 1 Minimax Optimal Robust Adaptively Perturbed Exploration (MR-APE)

Input: p, c, T, ϵ, and $F^{-1}(y)$
Output: $\{\hat{r}_{T,a}\}_{a\in\mathcal{A}}$
1: Initialize $\{\hat{r}_{0,a}=0, n_a(0)=0\}$ for all $a\in\mathcal{A}$ and select $a_1,\ldots,a_K$ and receive $\mathbf{R}_{1,a_1},\ldots,\mathbf{R}_{K,a_K}$ once
2: **for** $t=K+1,\ldots,T$ **do**
3: $\beta_{t-1,a}\leftarrow c\ln_+\left(\frac{T}{Kn_a(t-1)}\right)/(n_a(t-1))^{1-1/p}$ and $G_{t,a}\leftarrow F^{-1}(u)$ with $u\sim$ Uniform$(0,1)$
4: $\hat{r}_{t-1,a}\leftarrow c/(n_a(t-1))^{1-1/p}\sum_{k=1}^{t-1}\mathbb{I}[a_k=a]\psi_p(\mathbf{R}_{k,a}/(c\cdot(n_a(t-1))^{1/p}))$
5: Choose $a_t=\arg\max_{a\in\mathcal{A}}\{\hat{r}_{t-1,a}+(1+\epsilon)\beta_{t-1,a}G_{t,a}\}$, receive $\mathbf{R}_{t,a_t}$, and update $n_{a_t}(t)\leftarrow n_{a_t}(t-1)+1$
6: **end for**

is bounded using Hoeffding's maximal inequality that cannot be employed under unbounded heavy-tailed noise. Hence, to bound the second term we employ the integration bound that provides the upper bound of the summation. Consequently, we achieve the minimax optimal regret bound without using Hoeffding's maximal inequality.

1) Gap-Dependent Regret Bound of MR-UCB: Now, we provide the gap-dependent bound of each term for MR-UCB. The upper bound of the second term can be obtained as the following lemmas.

Lemma 1: For the second term of (18), MR-UCB satisfies the following inequality:

$$T\sum_{j=k_0+1}^{K}\mathbb{P}\left(\bar{E}_{a_j}^c\right)\left(\Delta_{a_j}-\Delta_{a_{j-1}}\right)\le O\left(\sum_{j=k_0+1}^{K}\frac{Kc^{\frac{p}{p-1}}}{\Delta_{a_j}^{\frac{1}{p-1}}}\right). \quad (20)$$

The entire proof of this lemma can be found in Appendix B. The main idea of to compute the upper bound of $\mathbb{P}(\bar{E}_{a_j}^c)$ is to employ the integration bound where there exist a upper bound $f(s)$ such that $\mathbb{P}(\bar{E}_{a_j}^c)=\mathbb{P}(Z<z_{a_j})\le\sum_{s=1}^{T}f(s)$ holds and the summation of $f(s)$ can be bounded by the integral of $f(s)$. The trick that bounds the summation by the integration will be generally used throughout the proof of our lemmas.

The final term of (19) can be bounded by the following lemma.

Lemma 2: For the final term of (19), MR-UCB satisfies the following inequality:

$$\sum_{k=k_0+1}^{K}\Delta_{a_k}\sum_{t=1}^{T}\mathbb{P}\left(\bar{E}_{a_k}\cap E_{t,a_k}\right) \quad (21)$$

$$\le O\left(\sum_{k=k_0+1}^{K}\frac{\max\left(3\ln\left(\frac{T}{K}\Delta_{a_k}^{\frac{p}{p-1}}\right)/2,1\right)^{\frac{p}{p-1}}}{\Delta_{a_k}^{\frac{1}{p-1}}}+\frac{c^{\frac{p}{p-1}}e^{\frac{bpv_p}{c^p}}}{\Delta_{a_k}^{\frac{1}{p-1}}}\right). \quad (22)$$

The entire proof of this lemma can be found in Appendix B. The main idea of the proof is counting the number of rounds for making confidence bound $\beta_{t-1,a}$ small enough. For small $\beta_{t-1,a}$, the final term can be bounded by the summation of the probability of estimation error. By combining two lemmas and setting $k_0=1$ whose $\Delta_1=0$ by definition, then, we can obtain the gap-dependent regret bound.

Theorem 4: Assume that $\nu_p<\infty$ and $\hat{r}_{t,k}$ is a p-robust estimator. Then, the gap-dependent regret bound of MR-UCB is

$$O\left(\sum_{a\neq a_\star}\frac{\max\left(\ln\left(\frac{T}{K}\Delta_a^{\frac{p}{p-1}}\right),1\right)^{\frac{p}{p-1}}}{\Delta_a^{\frac{1}{p-1}}}+\frac{Kc^{\frac{p}{p-1}}e^{\frac{bpv_p}{c^p}}}{\Delta_a^{\frac{1}{p-1}}}\right). \quad (23)$$

The proof is simply done by combining two lemmas and pick $k_0=1$ that makes $T\Delta_{a_{k_0}}=0$ since $r_{a_1}=r_\star$. The gap-dependent bound of MR-UCB shows the poly-logarithmic dependency on T. Compared to gap-dependent bound of robust UCB, the superiority of the gap-dependent bound can vary depending on $\{\Delta_a\}$. In general, the gap-dependent bound of MR-UCB follows $\ln(\Delta_a^{p/(p-1)}T)^{p/(p-1)}/\Delta_a^{1/(p-1)}$ while that of robust UCB follows $\ln(T)/\Delta_a^{1/(p-1)}$. Hence, if Δ_a is sufficiently large, then, $\ln(T)$ dominates $\ln(\Delta_a^{p/(p-1)})$ and this fact results in that robust UCB can have a smaller regret bound since $\ln(T)<\ln(T)^{p/(p-1)}$. On the other hand, if Δ_a is sufficiently small, then, $\ln(\Delta_a^{p/(p-1)})$ becomes a negative value for $\Delta_a\ll 1$ and, hence, it can dominantly reduce the term $\ln(\Delta_a^{p/(p-1)}T)^{p/(p-1)}$. In this regard, MR-UCB can have a smaller regret than robust UCB. From this fact, we can observe that MR-UCB is superior to robust UCB for a challenging MAB problem that has small gaps, which requires large samples to distinguish optimal action from suboptimal actions. This property makes it available that MR-UCB guarantee the optimal minimax regret bound.

2) Gap-Independent Regret Bound of MR-UCB: The gap-independent regret bounds can be derived from the similar strategies of gap-dependent bound. Now, we compute the gap-independent bounds for each term in (18) and (19).

Lemma 3: For the second term of (18), MR-UCB satisfies the following gap-independent inequality:

$$T\sum_{j=k_0+1}^{K}\mathbb{P}\left(\bar{E}_{a_j}^c\right)\left(\Delta_{a_j}-\Delta_{a_{j-1}}\right)\le O\left(c^{\frac{p}{p-1}}e^{\frac{bpv_p}{c^p}}K^{1-\frac{1}{p}}T^{\frac{1}{p}}\right). \quad (24)$$

The proof can be found in Appendix C. The main strategy of the proof is to bound the summation, $\sum_{j=k_0+1}^{K}\mathbb{P}(\bar{E}_{a_j}^c)(\Delta_{a_j}-\Delta_{a_{j-1}})$, by the integration, $\Delta-\Delta_{a_{k_0}}+\int_{\Delta}^{1}\mathbb{P}(Z<r_{a^\star}-(u/6))du$, which is borrowed from [2]. Then, the probability $\mathbb{P}(Z<r_{a^\star}-(u/6))$ can be bounded using the same technique in Lemma 1.

The third term of (19) can be bounded as follows.

Lemma 4: For the final term of (19), MR-UCB satisfies the following gap-independent inequality:

$$\sum_{k=k_0+1}^{K} \Delta_{a_k} \sum_{t=1}^{T} \mathbb{P}\big(\bar{E}_{a_k} \cap E_{t,a_k}\big) \leq O\Big(c^{\frac{p}{p-1}} K^{1-\frac{1}{p}} T^{\frac{1}{p}}\Big). \quad (25)$$

The proof can be found in Appendix C. The proof starts from Lemma 2. We pick k_0 such that $\Delta_{a_{k_0}} < \Delta < \Delta_{a_{k_0+1}}$ where $\Delta = \max(e^p, e^{-(3(p-1)/2p)})(K/T)^{1-1/p}$. Then, the gap-dependent bound in Lemma 2 is a decreasing function for $\Delta_a > \Delta$. Hence, we can replace Δ_a of Lemma 2 with Δ to get an upper bound. By combining two lemmas, we can obtain gap-independent bound of MR-UCB as following theorems.

Theorem 5: Assume that $v_p < \infty$ and $\hat{r}_{t,k}$ is a p-robust estimator. Then, the gap-independent regret bound is

$$\mathcal{R}_T \leq O\Big(c^{\frac{p}{p-1}} e^{\frac{b_p v_p}{c^p}} K^{1-\frac{1}{p}} T^{\frac{1}{p}}\Big). \quad (26)$$

From Lemmas 3 and 4, we can bound the second term of (18) and third term of (19) with $O(K^{1-1/p}T^{1/p})$. Then, the remaining part of the proof is to check the first term in (18), $T\Delta_{k_0}$, is bounded by $O(K^{1-1/p}T^{1/p})$. Fortunately, since we pick k_0 such that $\Delta_{a_{k_0}} < \Delta < \Delta_{a_{k_0+1}}$ holds, we have $T\Delta_{k_0} < T\Delta = O(K^{1-1/p}T^{1/p})$. Consequently, we can guarantee that the gap-independent regret bound of MR-UCB is $O(K^{1-1/p}T^{1/p})$ that matches the global minimax optimal regret bound for heavy-tailed MAB problems.

3) Gap-Dependent Regret Bound of MR-APE: Now, we will derive the gap-dependent regret bound of MR-APE. We can derive the regret bound of MR-APE by only proving the third term of (19) since other two terms in (18) can be bounded using the same way of MR-UCB. For the third term of (19), we first introduce $x_a := r_a + \Delta_a/3$ and $y_a := r_{a^\star} - \Delta_a/3$. Then, let us define three events, $\hat{E}_{t,a} := \{\hat{r}_{t,a} \leq x_a\}$, $\tilde{E}_{t,a} := \{\hat{r}_{t-1,a} + (1+\epsilon)\beta_{t-1,a}G_{t,a} \leq y_a\}$, and $\bar{E}_{t,a} := \{z_a \leq \hat{r}_{t-1,a^\star} + \beta_{t-1,a^\star}\}$. From the definition of three events, we have $E_{t,a} \cap \bar{E}_a \subset E_{t,a} \cap \bar{E}_{t,a}$ since $z_a \leq \min_{1<t\leq T} \hat{r}_{t-1,a^\star} + \beta_{t-1,a^\star}$ implies $z_a \leq \hat{r}_{t-1,a^\star} + \beta_{t-1,a^\star}$. Then, we decompose $E_{t,a} \cap \bar{E}_{t,a}$ into three subsets

$$E_{t,a} \cap \bar{E}_{t,a} = E_{t,a}^{(1)} \cup E_{t,a}^{(2)} \cup E_{t,a}^{(3)} \quad (27)$$

where $E_{t,a}^{(1)} = E_{t,a} \cap \bar{E}_{t,a} \cap \hat{E}_{t,a}^c$, $E_{t,a}^{(2)} = E_{t,a} \cap \bar{E}_{t,a} \cap \hat{E}_{t,a} \cap \tilde{E}_{t,a}^c$, and $E_{t,a}^{(3)} = E_{t,a} \cap \bar{E}_{t,a} \cap \hat{E}_{t,a} \cap \tilde{E}_{t,a}$. Hence, the final term of (19) can be bounded using the following inequality:

$$\mathbb{P}\big(E_{t,a} \cap \bar{E}_a\big) \leq \mathbb{P}\Big(E_{t,a}^{(1)}\Big) + \mathbb{P}\Big(E_{t,a}^{(2)}\Big) + \mathbb{P}\Big(E_{t,a}^{(3)}\Big).$$

Each term has the following meanings.

1) The first event, $E_{t,a}^{(1)}$, mainly counts the number of times that the suboptimal action a is selected due to the estimation error of $\hat{r}_{t-1,a}$. Hence, this term will be bounded by the error probability of the p-robust estimator.
2) The second event, $E_{t,a}^{(2)}$, considers the case of choosing suboptimal action due to the large perturbation, $G_{t,a}$, while its reward estimation is well concentrated. This term can be controlled by coefficient $\beta_{t-1,a}$ since this event depends on the magnitude of sampled perturbation.
3) The final event, $E_{t,a}^{(3)}$, indicates that suboptimal action was selected even though $\hat{r}_{t-1,a}$ is well estimated and the perturbation, $G_{t,a}$ is not too large. This event can happen when the estimation of the optimal reward is incorrect and the perturbation of the optimal action, $G_{t,a_\star}$, is not large enough to overcome the under-estimation.

The basic idea of deriving bounds of $E_{t,a}^{(1)}$, $E_{t,a}^{(2)}$, and $E_{t,a}^{(3)}$ is followed by Kim and Tewari [5], Lee *et al.* [8], and Cesa-Bianchi *et al.* [23]. We apply techniques in [5], [8], and [23] to our modified confidence bound. Now, we provide the gap-dependent bounds for three terms.

Lemma 5: The probabilities of $E_{t,a}^{(1)}$, $E_{t,a}^{(2)}$, and $E_{t,a}^{(3)}$ can be bounded as follows:

$$\sum_{t=1}^{T} \mathbb{P}\Big(E_{t,a}^{(1)}\Big) \leq \frac{(3c)^{\frac{p}{p-1}} \exp\Big(\frac{b_p v_p}{c^p}\Big) \Gamma\Big(\frac{2p-1}{p-1}\Big)}{\Delta_a^{\frac{p}{p-1}}} \quad (28)$$

$$\sum_{t=1}^{T} \mathbb{P}\Big(E_{t,a}^{(2)}\Big) \leq \frac{\max\Big(3(1+\epsilon)\ln\Big(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\Big), 1\Big)^{\frac{p}{p-1}}}{\Delta_{a_k}^{p/(p-1)}} \quad (29)$$

$$\sum_{t=1}^{T} \mathbb{P}\Big(E_{t,a}^{(3)}\Big) \leq M_\epsilon \frac{\max\Big(6(2+\epsilon)\ln\Big(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\Big), 1\Big)^{\frac{p}{p-1}}}{\Delta_{a_k}^{p/(p-1)}}. \quad (30)$$

The entire proofs of the lemma can be found in Appendix D. By combining all lemmas, we can bound the third term of (19). Consequently, we have the following gap-dependent regret bound of MR-APE.

Theorem 6: Assume that the pth moment of rewards is bounded by a constant $v_p < \infty$, $\hat{r}_{t,k}$ is a p-robust estimator and G is a bounded perturbation within $[-1, 1]$, and there exists a constant M_ϵ only dependent on ϵ such that $\mathbb{P}(G < (1/(1+\epsilon)))/\mathbb{P}(G > (1/(1+\epsilon))) < M_\epsilon$. Then, the gap-dependent regret bound of MR-APE is

$$\mathcal{R}_T \leq O\left(\sum_{k=1}^{K} \frac{M_\epsilon^+ \max\Big(\Big(2+\epsilon)\ln\Big(\frac{T}{K}\Delta_{a_k}^{\frac{p}{p-1}}\Big), 1\Big)^{\frac{p}{p-1}}}{\Delta_{a_k}^{\frac{1}{p-1}}} + \frac{K c^{\frac{p}{p-1}} e^{\frac{b_p v_p}{c^p}}}{\Delta_{a_k}^{\frac{1}{p-1}}}\right) \quad (31)$$

where $M_\epsilon^+ := \max(M_\epsilon, 1)$.
The proof is simply done by combining two lemmas with the proof of MR-UCB, and picking $k_0 = 1$ that makes $T\Delta_{a_{k_0}} = 0$ since $r_{a_1} = r_\star$. We can observe that the gap-dependent bound of MR-APE is the same as that of MR-UCB up to a constant.

4) Gap-Independent Regret Bound of MR-APE: Now, we can derive the gap-independent regret bound of MR-APE using the same technique of MR-UCB.

Theorem 7: Assume that the pth moment of rewards is bounded by a constant $v_p < \infty$, $\hat{r}_{t,k}$ is a p-robust estimator and G is a bounded perturbation within $[-1, 1]$, and there exists a constant M_ϵ only dependent on ϵ such that $\mathbb{P}(G < (1/(1+$

$\epsilon)))/\mathbb{P}(G > (1/(1+\epsilon))) < M_\epsilon$. Then, the gap-independent regret bound of MR-APE is

$$\mathcal{R}_T \leq O\left(\max\left(M_\epsilon^+(2+\epsilon)^{\frac{p}{p-1}}, c^{\frac{p}{p-1}}e^{\frac{b_p v_p}{c^p}}\right)K^{1-1/p}T^{1/p}\right). \quad (32)$$

The proof is omitted here and can be found in Appendix E. Similar to the gap-dependent bound, the gap-independent bound of MR-APE also has the same order of T and K as that of MR-UCB. Consequently, MR-APE also guarantee the minimax optimal regret bound.

5) Comparison Between MR-UCB and MR-APE: While MR-APE and MR-UCB have the same mini-max optimal regret bound, the main difference between MR-APE and MR-UCB comes from the gap-dependent regret bounds in Theorems 4 and 6. In Theorem 4, the logarithmic term in the gap-dependent bound of MR-UCB is independent on c and only the final term $O(Kc^{p/(p-1)}\exp(b_p v_p/c^p)/\Delta_{a_k}^{1/(p-1)})$ depends on c. In this regard, controlling c does not affect the order of T. However, in Theorem 6, MR-APE has an auxiliary controllable parameters M_ϵ and ϵ that can affect to the logarithmic term of the gap-dependent bound of MR-APE. Furthermore, we can interpret MR-APE as the unifying framework between UCB-like exploration and perturbed exploration. Intuitively speaking, from the condition of $\mathbb{P}(G < 1/(1+\epsilon))/\mathbb{P}(G > 1/(1+\epsilon)) < M_\epsilon$, most of probability mass of the perturbation is located near one, hence, MR-APE has randomness but works similar to MR-UCB. More specifically, the condition on G and M_ϵ can be rewritten as $\mathbb{P}(G > 1/(1+\epsilon))^{-1} - 1 < M_\epsilon$, then, if $\mathbb{P}(G > 1/(1+\epsilon))$ is getting smaller, the constant M_ϵ should become larger to satisfy the condition. In this case, MR-APE mainly employs the perturbation for exploration, rather than depends on the confidence bound. On the other hand, if $\mathbb{P}(G > 1/(1+\epsilon))$ is getting bigger, most of probability mass should be located near $G = 1$ to make M_ϵ small enough. Under this condition, MR-APE behaves similar to MR-UCB. Such property of MR-APE allows it to enable more adaptive exploration in practice, since we can control not only c but also M_ϵ and ϵ using the distribution of perturbation.

V. Experimental Results

A. Experimental Setup

We verify the properties of the proposed methods and compare the proposed methods to other existing methods. First, we compare two proposed methods, MR-UCB and MR-APE. Especially, we prepare various types of MR-APE that have different bounded perturbations. We separate bounded perturbations into two groups. The first group is a positive bounded perturbation whose random variable only has a positive value. The second group is a both-sided bounded perturbation whose random variable can have both positive and negative values. For the first group, we employ Bernoulli distribution and Uniform distribution with $[0, 1]$ as a bounded perturbation in MR-APE. For the second group, we employ Rademacher distribution whose value can have -1 or 1, and Uniform distribution with $[-1, 1]$. Hence, the proposed exploration scheme is tested in five different algorithms: MR-UCB, MR-APE with Bernoulli, MR-APE with Uniform$(0, 1)$, MR-APE with Rademacher, and MR-APE with Uniform$[-1, 1]$. We compare the proposed methods with existing robust exploration methods such as robust UCB [7], DSEE [21], and APE^2 with unbounded perturbation [8]. For APE^2, we utilize Fréchet and Pareto distributions as an unbounded perturbation. Hence, the comparisons are conducted with APE^2 with Fréchet and APE^2 with Pareto. Note that robust UCB and DSEE utilize the truncated mean estimator, and APE^2, MR-UCB, and MR-APE mainly utilize the p-robust estimator.

We prepare synthetic and real-world data for simulations. First, for all synthetic simulations, we synthesize a heavy-tailed MAB problem with K actions. The optimal action has 1 mean reward and $K-1$ suboptimal actions have $1-\Delta$ mean reward. Hence, Δ determines the gap between the maximum reward and other rewards. By controlling Δ, we can measure how the gap influence the regret of each exploration method. Then, we add a heavy-tailed noise to the observation of rewards. The heavy-tailed noise is created by transforming a Pareto and Fréchet random variable. Let z_t be a heavy-tailed random variable, $z_t \sim \text{Pareto}(\alpha_\epsilon, \lambda_\epsilon)$ where α_ϵ is a shape parameter and λ_ϵ is a scale parameter. Then, a noise is defined as $\epsilon_t := b_t(z_t - \mathbb{E}[z_t])$ where b_t is a Rademacher random variable that has $+1$ value with probability $1/2$ and get -1 value with probability $1/2$. From the definition, ϵ_t is a mean zero heavy-tailed noise. In simulation, we observe a noisy reward $\mathbf{R}_{t,a} := r_a + \epsilon_{t,a}$ for every step. Each simulation runs T rounds and we measure the time average regret $\mathcal{R}_t/t := \sum_{k=1}^t (r_{a_\star} - r_{a_k})/t$ for $t \in [1, T]$. Second, for real-world data, we employ cryptocurrency dataset [25] that contains daily returns of cryptocurrency from April 1, 2019 to July 1, 2021. We select ten cryptocurrency, such as Bitcoin, Ethereum, Doge, Monero, Stellar, or EOS, based on market value. Then, the goal of this simulation is to identify the most profitable currency, which is motivated by the practical scenario that an investor wants to invest a fixed budget in a cryptocurrency and get return as much as possible. For this scenario, an action is defined as buying a specific currency and the corresponding reward is defined as the daily profit. Note that it is a well-known fact that the financial data often show the inherent characteristic of heavy tails [26], [27], hence, we believe that identifying the most profitable cryptocurrency is a practical application of the proposed methods.

Consequently, we prepare four simulations. The first simulation compares the performance of exploration methods on various p and Δ with two heavy-tailed noises. The second simulation verifies the effect of increasing K for the regret bound. The third simulation measures the effect of scale hyperparameter for the performance of exploration methods. The final simulation compares the performance of exploration methods on real-world cryptocurrency dataset.

B. Performance Comparison for Various Noises, p, and Δ

We compare the performance of every exploration method. For MR-UCB, robust UCB, MR-APE, and APE^2, we optimize the hyperparameter c using a grid search. We would like to note that, for robust UCB, we modify the confidence bound in Assumption 1 by multiplying a scale parameter c since

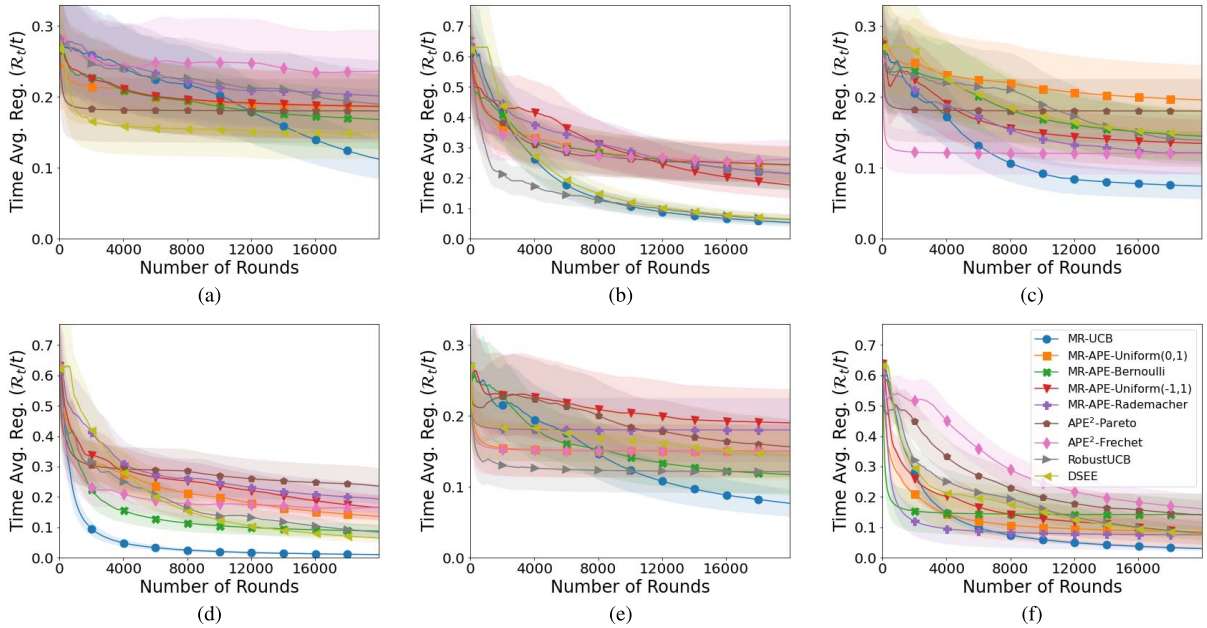

Fig. 1. Time-average regret for various p and Δ with Pareto noise. A bold line indicates the average value over ten different random seeds and shaded region indicates half-standard deviation area. All figures share the legend. (a) $p = 1.2$, $\Delta = 0.3$, $K = 10$. (b) $p = 1.2$, $\Delta = 0.7$, $K = 10$. (c) $p = 1.5$, $\Delta = 0.3$, $K = 10$. (d) $p = 1.5$, $\Delta = 0.7$, $K = 10$. (e) $p = 1.8$, $\Delta = 0.3$, $K = 10$. (f) $p = 1.8$, $\Delta = 0.7$, $K = 10$.

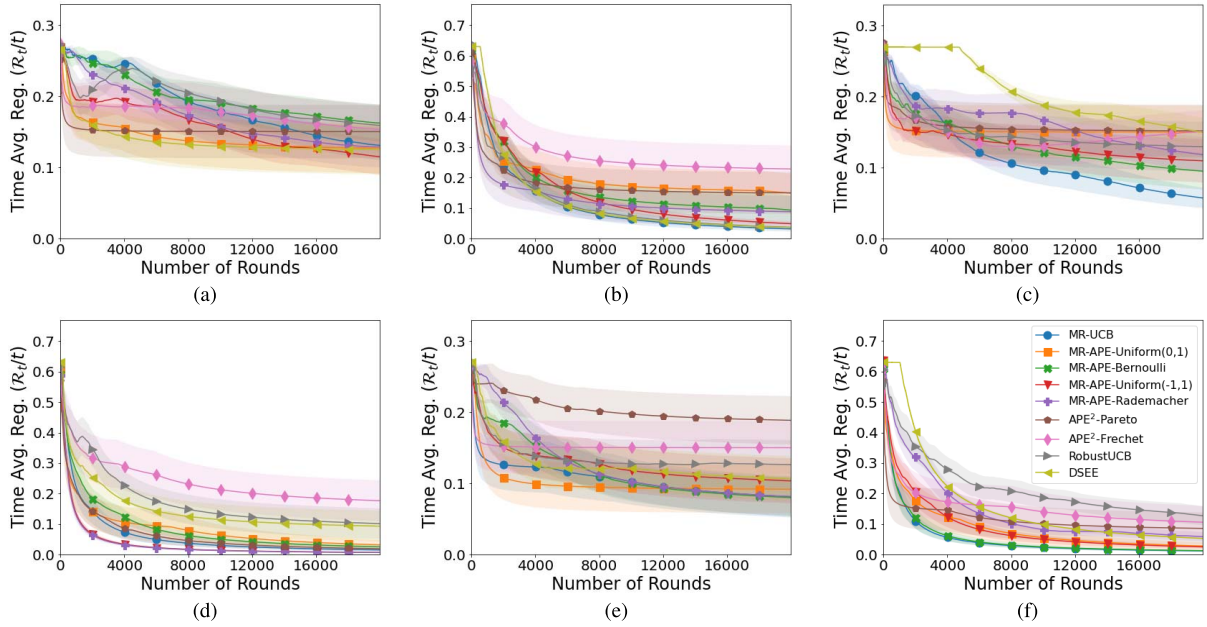

Fig. 2. Time-average regret for various p and Δ with Fréchet noise. A bold line indicates the average value over ten different random seeds and shaded region indicates half-standard deviation area. All figures share the legend. (a) $p = 1.2$, $\Delta = 0.3$, $K = 10$. (b) $p = 1.2$, $\Delta = 0.7$, $K = 10$. (c) $p = 1.5$, $\Delta = 0.3$, $K = 10$. (d) $p = 1.5$, $\Delta = 0.7$, $K = 10$. (e) $p = 1.8$, $\Delta = 0.3$, $K = 10$. (f) $p = 1.8$, $\Delta = 0.7$, $K = 10$.

the original robust UCB consistently shows poor performance even if ν_p is given. Then, c for robust UCB is also optimized using a grid search. We prepare six synthetic MAB problems by combining $\Delta = 0.3, 0.7$ and $p = 1.2, 1.5, 1.8$ for two noise types. Figs. 1 and 2 show the results of Pareto noise and Fréchet noise, respectively.

As shown in Fig. 1, first, MR-UCB consistently outperforms other exploration methods except for the case of ($p = 1.2, \Delta = 0.7$). In the MAB with ($p = 1.2, \Delta = 0.7$), MR-UCB shows comparable performance with robust UCB and MR-APE with Bernoulli. For $\Delta = 0.3$, as shown in Fig. 1(a), (c), and (e), we can observe that MR-UCB significantly outperform other methods while the performance gap between MR-UCB and other methods is marginal when $\Delta = 0.7$. Furthermore, the performance gap between MR-UCB and other methods increases as the order of the moment, p, decreases. This observation implies that MR-UCB shows more robust performance against heavy-tailed noise. As p is getting closer to 2, a robust estimator generally converges much faster than the case that p is close to 1, hence, reward estimators used by all exploration methods are concentrated with a fewer number of trials. This fact reduces the performance gap

between MR-UCB and other methods since the algorithm can distinguish the optimal action from suboptimal actions with fewer trials. However, as p is close to 1, the convergence speed of reward estimators is getting slower and requires a lot of samples to concentrate on the true mean. It hinders the convergence speed of exploration methods, however, MR-UCB outperforms other exploration methods as shown in Fig. 1(a) and (b).

As shown in Fig. 1, for MR-APE, we can observe that bounded perturbation usually shows the second-best performance for various settings under Pareto noises. In particular, MR-APE with positive perturbations generally outperforms MR-APE with both-sided perturbations. Furthermore, MR-APE with Bernoulli often shows the second best performance for various settings such as $(p = 1.2, \Delta = 0.3)$, $(p = 1.2, \Delta = 0.7)$, $(p = 1.5, \Delta = 0.7)$, and $(p = 1.8, \Delta = 0.3)$. However, we can observe that there is no clear dominance between MR-APE, APE^2, and robust UCB. From the tendency shown in Fig. 1, we can observe that the performance gap between MR-APE and other exploration methods, such as APE^{2}s, Robust UCB, and DSEE, increases as Δ decreases from 0.7 to 0.3.

For Fréchet noise setting, MR-UCB also outperforms APE^2 with unbounded perturbations and Robust UCB as shown in Fig. 2. However, unlike the Pareto noise setting, MR-APE with bounded positive perturbations shows comparable performance with MR-UCB in various problem settings, and even outperforms MR-UCB in several settings. In particular, MR-APE with Uniform$(0, 1)$ shows similar performance to MR-UCB including $(p = 1.8, \Delta = 0.7)$, $(p = 1.8, \Delta = 0.3)$, $(p = 1.5, \Delta = 0.7)$, $(p = 1.5, \Delta = 0.7)$, $(p = 1.5, \Delta = 0.7)$ and outperforms MR-UCB for $(p = 1.2, \Delta = 0.7)$. While MR-APE that randomizes MR-UCB shows inferior performance for Pareto noise settings, in Fréchet noise settings, MR-APE has advantages over MR-UCB. In summary, from the empirical results shown in Figs. 1 and 2, MR-UCB that employs modified upper confidence bound clearly outperforms other exploration methods for heavy-tailed MAB problems and MR-APEs shows comparable performance in general cases but dominates other algorithms in several special cases.

C. Performance Comparison for Varying K

In this experiment, we verify the effect of the number of actions in heavy-tailed bandits. We employ a Pareto noise setting with $p = 1.8$ and $\Delta = 0.7$. For all exploration methods, we measure the final time average regret after 20 000 rounds. The simulation is conducted varying K from $10, 30, 50, 70,$ and 100. In Fig. 3, we plot the average value over ten random seeds. For each K, we conduct the hyperparameter optimization using a grid search.

As shown in Fig. 3, all algorithms show a similar tendency that $\mathcal{R}_T/T$ increases as the number of actions increases since the number of exploring an individual action is reduced if K increases with fixed T. Hence, the plot in Fig. 3 shows the effect of K on the cumulative regret. First, the most robust algorithm against increasing the number of cation is MR-UCB. Especially, MR-UCB outperforms all other

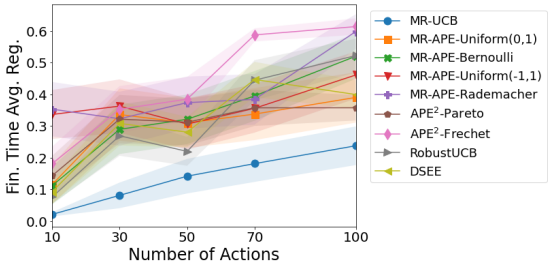

Fig. 3. Effect of number of actions. The time-average final regret $\mathcal{R}_T/T$ at the final round is plotted for different K. All regrets are measured under $p = 1.8$ and $\Delta = 0.7$ with Pareto noise distribution. The bold line is an average value of $\mathcal{R}_T/T$ over ten different random seeds and the shaded area indicates a half-standard deviation region.

exploration methods as the number of actions K increases. However, the performance of APE^2 with Fréchet is drastically getting worse as K increases while MR-APEs with bounded perturbations show a moderate performance drop. This result clearly supports the fact that using modified confidence bound helps to reduce the regret by removing the suboptimal factor $\ln(K)$ from APE^2 with unbounded perturbations. Other methods except for MR-UCB and APE^2 with Fréchet show comparable performance with each other. Interestingly, Robust UCB and DSEE show similar performance to MR-APEs such as Uniform, Bernoulli, and Rademacher perturbations. These results indicate that the regret bound of Robust UCB and DSEE has the same dependency on $K^{1/p}$ as the regret bound of MR-APEs while it has suboptimal factor $\ln(T)^{1/p}$ with respect to T. In summary, we can conclude that MR-UCB outperforms other exploration methods as the number of actions increases under heavy-tailed settings since the modified confidence bound removes the suboptimal factor of K in the minimax regret bounds of MR-UCB.

D. Effect of Hyperparameter

In this experiment, we verify the sensitivity of each exploration method with respect to the hyperparameter c. For MR-UCB, robust UCB, MR-APE, the exploration tendency depends on scale parameter c. To verify the effect of c for each algorithm, we measure the final time average regret with 50 different c values after 20 000 rounds. For this simulation, we set $K = 10$, $\Delta = 0.7$, and $T = 20\,000$ and run each algorithm with ten different random seeds. In Fig. 4, we can observe valley-shaped plots for varying hyperparameter c. In general, if c is small, then, an algorithm shows the worst regret since small c makes the algorithm rarely explore an action space. With the similar reason, if c is large, then, an exploration method also shows the worst regret since large c hinders exploitation or convergence to the optimal action. Hence, the regret is reduced at the proper range of c as shown in the valley-shaped plots in Fig. 4. For each algorithm, we would like to focus on analyzing the plateau of valley that shows sensitivity of exploration tendency with respect to hyperparameter. The wide plateau implies that the algorithm is less sensitive to hyperparameters and the proper hyperparameter can be easily found with smaller number of grid search. On the contrary, the narrow plateau indicates that the algorithm is more sensitive for hyperparameter optimizations. To visually

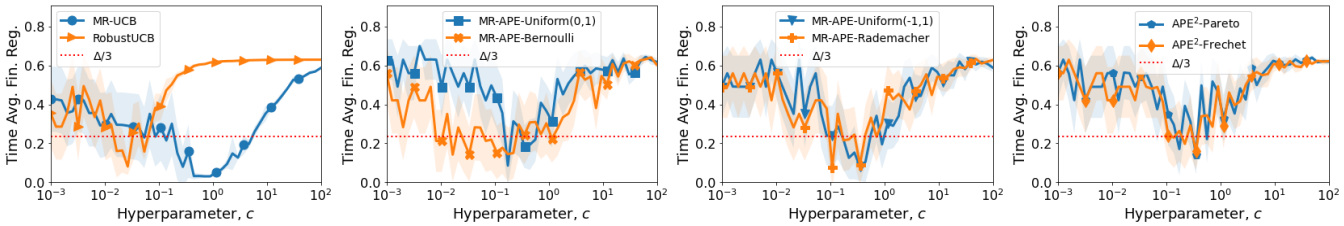

Fig. 4. Effect of hyperparameter. The time-average final regret $\mathcal{R}_T/T$ at final round is plotted for different c. All regrets are measured under $p = 1.8$ and $\Delta = 0.7$ with Pareto noise distribution. The red dotted line indicates $\mathcal{R}_T/T = \Delta/3$. The bold line is average value of $\mathcal{R}_T/T$ over ten different random seeds and shaded area indicates a half-standard deviation region. (a) MR-UCB and robust UCB. (b) MR-APE with uniform(0, 1). (c) MR-APE with uniform$(-1, 1)$. (d) Unbounded APE^2.

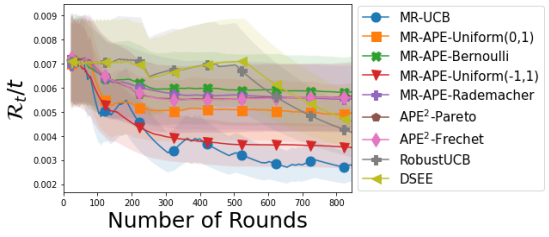

Fig. 5. Time average regret for cryptocurrency dataset. The bold line is an average value over ten different random seeds and the shaded area indicates a half-standard deviation region.

measure the range of plateau, we mark the threshold at red dotted line in Fig. 4.

In Fig. 4(a), it can be observed that MR-UCB has the wider plateau than Robust-UCB. Hence, this plot implies that MR-UCB is less sensitive than Robust-UCB. Especially, the performance of the best hyperparameter of MR-UCB is lower than that of Robust-UCB. Consequently, the result shows that MR-UCB is more robust and has a better performance than Robust-UCB with respect to hyperparameter optimization. Comparing MR-APE with Uniform$(0, 1)$ with MR-APE with Bernoulli, MR-APE with Bernoulli shows much robust performance with wider range of the plateau of valley in hyperparameters. Especially, MR-APE with Bernoulli perturbation has much wider plateau than MR-APE with Uniform perturbation. Furthermore, the plateau of MR-APE with Bernoulli is much wider and shows lower cumulative regret than the plateau of MR-UCB. This result shows that the randomization of MR-UCB such as Bernoulli perturbation has the effect of widening the plateau of valley in hyperparameter space.

In practice, finding a proper hyperparameter c is a demanding task for applying exploration methods in practical applications. Hence, algorithms that are less sensitive to hyperparameters are more suitable for practical problems since such properties reduce the cost of optimizing hyperparameters. From the experimental results of hyperparameter optimization, we can conclude that MR-UCB and MR-APE with bounded perturbations are more desirable for practical applications.

E. Performance Comparison for Cryptocurrency Dataset

In this experiment, we test all exploration methods on real-world cryptocurrency dataset [25]. Similar to other simulations, we optimize hyperparameters of each algorithm using a grid search. In Fig. 5, we plot the average value over 10 random seeds. As shown in Fig. 5, MR-UCB shows the best performance and MR-APE with Uniform$(-1, 1)$ shows the second best performance. Especially, among the set of bounded perturbations, the uniform perturbation on $(-1, 1)$ shows the best performance. Furthermore, the results show the similar tendency to the results from synthetic simulations. It is worth mentioning that MR-UCB and MR-APE with Uniform $(-1, 1)$ clearly outperform Robust-UCB, DSEE, and APE^2. Overall, with synthetic and real-world simulations, we have verified the superiority of the proposed methods.

VI. Conclusion

We have studied the minimax optimality under heavy-tailed noise assumption for stochastic MABs where the pth moment of rewards is bounded by a constant ν_p for $1 < p \leq 2$. We first investigated and found two critical drawbacks of existing robust explorations. First, existing robust exploration methods often depend on a robust mean estimator that requires prior knowledge about ν_p where ν_p is not accessible in many real-world problems. Second, we proved the sub-optimality of existing robust exploration methods for heavy-tailed rewards. Based on the analysis of the sub-optimality of existing methods, we have proposed two algorithms, MR-UCB and MR-APE, that can guarantee the minimax optimality with minimal information. Both proposed methods are independent on ν_p and this fact allows us to employ the proposed exploration methods with minimal prior knowledge compared to existing exploration methods. MR-UCB utilizes the modified confidence bounds that can provide more precise confidence bound of robust mean estimators. Then, MR-APE is the randomized version of MR-UCB that employ bounded perturbation whose scale follows the modified confidence bound in MR-UCB. Furthermore, we analyzed both gap-dependent and gap-independent regret bounds of two proposed methods and guaranteed that both proposed methods have the minimax optimal regret bounds. In simulations, we demonstrate the superiority of the proposed methods for various heavy-tailed synthetic and real-world data. Furthermore, MR-UCB clearly outperforms other algorithms as the number of actions increases under heavy-tailed noise. Consequently, we can conclude that the proposed methods have benefits over heavy-tailed MAB problems.

Appendix A
Proof of Corollary 1

Proof of Corollary 1: If the supporting interval of $F(g)$ is bounded, then, there exists some constants A and B such that, for all $y \in [0, 1]$, $A < F^{-1}(y) < B$ holds. Then, the following

inequalities also hold, $A < F^{-1}(1-1/K) < B$, due to the bounded support of $F(g)$. From this fact, we get

$$K^{1-1/p}T^{1/p}A \leq K^{1-1/p}T^{1/p}F^{-1}(1-1/K) < K^{1-1/p}T^{1/p}B.$$

This fact induces the following relation:

$$K^{1-1/p}T^{1/p}F^{-1}(1-1/K) = \Theta\big(K^{1-1/p}T^{1/p}\big).$$

Therefore, we have

$$\mathcal{R}_T \geq \Omega\big(K^{1-1/p}T^{1/p}F^{-1}(1-1/K)\big) = \Omega\big(K^{1-1/p}T^{1/p}\big).$$

□

Appendix B
Proofs for Theorem 4

Lemma 6 (First Step of Theorem 5 in [2]): Without loss of generality, assume that $r_{a_1} > r_{a_2} > \cdots > r_{a_K}$. Then, for any $k_0 \in [1, \ldots, K]$, the following bound holds:

$$\mathbb{E}[\mathcal{R}_T] \leq T\Delta_{a_{k_0}} + T\sum_{j=k_0+1}^{K}\mathbb{P}\Big(\bar{E}_{a_j}^c\Big)\big(\Delta_{a_j} - \Delta_{a_{j-1}}\big) \quad (33)$$

$$+ \sum_{k=k_0+1}^{K}\Delta_{a_k}\sum_{t=1}^{T}\mathbb{P}\big(\bar{E}_{a_k} \cap E_{t,a_k}\big). \quad (34)$$

Proof: The proof can be found in [2]. □

Proof of Lemma 1: To prove Lemma 1, we employ the concentration property of the p-robust estimator. We have

$$\sum_{j=k_0+1}^{K}\mathbb{P}\Big(\bar{E}_{a_j}^c\Big)\big(\Delta_{a_j} - \Delta_{a_{j-1}}\big) \quad (35)$$

$$= \sum_{j=k_0+1}^{K}\mathbb{P}\left(Z < r_{a^*} - \frac{\Delta_{a_j}}{6}\right)\big(\Delta_{a_j} - \Delta_{a_{j-1}}\big). \quad (36)$$

Then, we can bound the probability $\mathbb{P}(Z < r_{a^*} - (\Delta_{a_j}/6))$ as

$$\mathbb{P}\big(Z < r_{a^*} - \Delta_{a_j}/6\big) \quad (37)$$

$$= \mathbb{P}\left(\min_{1<t\leq T}\hat{r}_{t-1,a^*} + \beta_{t-1,a^*} < r_{a^*} - \frac{\Delta_{a_j}}{6}\right) \quad (38)$$

$$\leq \sum_{s=1}^{T}\mathbb{P}\left(r_{a^*} - \hat{r}_{s,a^*} > \frac{c\ln_+\left(\frac{T}{Ks}\right)}{s^{1-\frac{1}{p}}} + \frac{\Delta_{a_j}}{6}\right) \quad (39)$$

$$\leq \sum_{s=1}^{T}\exp\left(-\ln_+\left(\frac{T}{Ks}\right) - \frac{\Delta_{a_j}s^{1-\frac{1}{p}}}{6c}\right) \quad (40)$$

$$\leq \frac{K}{T}\sum_{s=1}^{T}s\exp\left(-\frac{\Delta_{a_j}s^{1-\frac{1}{p}}}{6c}\right) \quad (41)$$

$$\leq \frac{K}{T}\left(\frac{6cp}{e(p-1)\Delta_{a_j}}\right)^{\frac{p}{p-1}} + \frac{K}{T}\int_0^{\infty}xe^{-\frac{\Delta_{a_j}x^{1-\frac{1}{p}}}{6c}}dx \quad (42)$$

$$\leq \frac{K}{T}\left(\frac{6cp}{e(p-1)\Delta_{a_j}}\right)^{\frac{p}{p-1}} + \frac{K\Gamma\left(\frac{3p-1}{p-1}\right)}{2T}\left(\frac{6c}{\Delta_{a_j}}\right)^{\frac{p}{p-1}}. \quad (43)$$

Then, we can obtain a gap-dependent bound for the second term of (18) as follows:

$$T\sum_{j=k_0+1}^{K}\mathbb{P}\Big(\bar{E}_{a_j}^c\Big)\big(\Delta_{a_j} - \Delta_{a_{j-1}}\big) \quad (44)$$

$$\leq Cc^{\frac{p}{p-1}}K\sum_{j=k_0+1}^{K}\frac{\Delta_{a_j}}{\Delta_{a_j}^{\frac{p}{p-1}}} = O\left(\sum_{j=k_0+1}^{K}\frac{Kc^{\frac{p}{p-1}}}{\Delta_{a_j}^{\frac{1}{p-1}}}\right) \quad (45)$$

where C is a constant independent on c, T, K, and Δ_a. □

Proof of Lemma 2: To prove the upper bound, we first introduce the stopping time $\tau_k = \min\{t : B_{k,t} < z_{a_k}\}$ where $B_{k,t} := \hat{r}_{t-1,a_k} + c\ln_+(T/(Kn_{a_k}(t-1)))/(t-1)^{1-1/p}$. Then, we have $\{Z > z_{a_k}\} \subset \{n_{a_k,T} < \tau_k\}$ from the definition of Z and selection rule of MR-UCB. Then, $\sum_{t=1}^{T}\mathbb{P}(\bar{E}_{a_k} \cap E_{t,a_k})$ can be first bounded by $\mathbb{E}[\mathbb{I}[Z > z_{a_k}]n_{a_k,T}]$. Hence, by combining two facts, we have

$$\sum_{k=k_0+1}^{K}\Delta_{a_k}\sum_{t=1}^{T}\mathbb{P}\big(\bar{E}_{a_k} \cap E_{t,a_k}\big) \quad (46)$$

$$\leq \sum_{k=k_0+1}^{K}\Delta_{a_k}\mathbb{E}\big[\mathbb{I}\big[Z > z_{a_k}\big]n_{a_k,T}\big] \leq \sum_{k=k_0+1}^{K}\Delta_{a_k}\mathbb{E}[\tau_k]. \quad (47)$$

Then, we can compute the upper bound of the expectation of τ_k as follows:

$$\mathbb{E}[\tau_k] \leq \ell_0 + \sum_{l=\ell_0+1}^{\infty}\mathbb{P}(l < \tau_k) \quad (48)$$

$$= \ell_0 + \sum_{l=\ell_0+1}^{\infty}\mathbb{P}\big(\forall t \leq l, B_{k,t} > z_{a_k}\big) \quad (49)$$

$$\leq \ell_0 + \sum_{l=\ell_0+1}^{\infty}\mathbb{P}\left(\hat{r}_{l,a_k} - r_{a_k} > \frac{5\Delta_{a_k}}{6} - \frac{\ln_+\left(\frac{T}{Kl}\right)}{l^{1-\frac{1}{p}}}\right). \quad (50)$$

Then, we bound the expectation of τ_k by properly setting ℓ_0. Let us take ℓ_0 as follows:

$$\ell_0 = \max\left(\frac{\left[6\ln\left(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\right)\right]^{\frac{p}{p-1}}}{\left(4\Delta_{a_k}\right)^{p/(p-1)}}, \frac{1}{\Delta_{a_k}^{p/(p-1)}}\right).$$

For $l > \ell_0$, we have $l > \Delta_{a_k}^{-p/(p-1)}$ due to the definition of ℓ_0, and thus, $\ln_+((T/Kl))/l^{1-1/p} \leq 4\Delta_{a_k}/6$. Hence, the following condition holds:

$$5\Delta_{a_k}/6 - \ln_+\left(\frac{T}{Kl}\right)/l^{1-\frac{1}{p}} \geq 5\Delta_{a_k}/6 - 4\Delta_{a_k}/6. \quad (51)$$

Hence, we can bound $\mathbb{E}[\tau_k]$ as follows:

$$\mathbb{E}[\tau_k] \leq \ell_0 + \sum_{l=\ell_0+1}^{\infty}\mathbb{P}\left(\hat{r}_{l,a_k} - r_{a_k} > \frac{\Delta_{a_k}}{6}\right) \quad (52)$$

$$\leq \ell_0 + \exp\left(\frac{b_p\nu_p}{c^p}\right)\sum_{l=\ell_0+1}^{\infty}\exp\left(-\frac{l^{1-\frac{1}{p}}\Delta_{a_k}}{6c}\right) \quad (53)$$

$$\leq \ell_0 + \exp\left(\frac{b_p\nu_p}{c^p}\right)\int_0^{\infty}\exp\left(-\frac{x^{1-\frac{1}{p}}\Delta_{a_k}}{6c}\right)dx \quad (54)$$

$$\leq \max\left(\frac{\left[6\ln\left(\frac{T}{K}\Delta_{a_k}^{\frac{p}{p-1}}\right)\right]^{\frac{p}{p-1}}}{\left(4\Delta_{a_k}\right)^{\frac{p}{p-1}}}, \frac{1}{\Delta_{a_k}^{\frac{p}{p-1}}}\right) \quad (55)$$

$$+ \frac{(6c)^{\frac{p}{p-1}}e^{\frac{b_p\nu_p}{c^p}}\Gamma\left(\frac{2p-1}{p-1}\right)}{\Delta_{a_k}^{\frac{p}{p-1}}}. \quad (56)$$

Finally, using the bound of $\mathbb{E}[\tau_k]$, we get

$$\sum_{k=k_0+1}^{K} \Delta_{a_k} \mathbb{E}[\tau_k] \tag{57}$$

$$\leq O\left(\sum_{k=k_0+1}^{K} \frac{\max\left(3\ln\left(\frac{T}{K}\Delta_{a_k}^{\frac{p}{p-1}}\right)/2, 1\right)^{\frac{p}{p-1}}}{\Delta_{a_k}^{\frac{1}{p-1}}} \right. \tag{58}$$

$$\left. + \frac{c^{\frac{p}{p-1}} e^{\frac{b_p v_p}{c^p}}}{\Delta_{a_k}^{\frac{1}{p-1}}} \right). \tag{59}$$

□

Proof of Theorem 4: The proof can be done by combining Lemma 1 and 2. By combining all gap-dependent bounds and setting $k_0 = 1$, we can obtain the gap-dependent bounds as

$$O\left(\sum_{k=2}^{K} \frac{\max\left(\ln\left(\frac{T}{K}\Delta_{a_k}^{\frac{p}{p-1}}\right), 1\right)^{\frac{p}{p-1}}}{\Delta_{a_k}^{\frac{1}{p-1}}} + \frac{K c^{\frac{p}{p-1}} e^{\frac{b_p v_p}{c^p}}}{\Delta_{a_k}^{\frac{1}{p-1}}} \right). \tag{60}$$

□

APPENDIX C
PROOFS FOR THEOREM 5

Proof of Lemma 3: The proof starts from Lemma 2 as follows:

$$\sum_{k=k_0+1}^{K} \Delta_{a_k} \sum_{t=1}^{T} \mathbb{P}\left(\bar{E}_{a_k} \cap E_{t,a_k}\right) \tag{61}$$

$$\leq O\left(\sum_{k=k_0+1}^{K} \frac{\max\left(3\ln\left(\frac{T}{K}\Delta_{a_k}^{\frac{p}{p-1}}\right)/2, 1\right)^{\frac{p}{p-1}}}{\Delta_{a_k}^{\frac{1}{p-1}}} + \frac{c^{\frac{p}{p-1}} e^{\frac{b_p v_p}{c^p}}}{\Delta_{a_k}^{\frac{1}{p-1}}} \right). \tag{62}$$

Then, for all gap-independent bounds, we set k_0 such that $\Delta_{a_{k_0}} < \Delta < \Delta_{a_{k_0+1}}$ where $\Delta = \max(e^p, e^{-(3(p-1)/2p)})(K/T)^{1-1/p}$. For $\Delta_a > e^p(K/T)^{1-1/p}$ and $e^p(K/T)^{1-1/p} > e^{-(3(p-1)/2p)}(K/T)^{1-1/p}$, the upper bound in (61) is a decreasing function. Hence, replacing Δ_a with Δ makes the upper bound greater. Consequently, we have

$$\sum_{k=k_0+1}^{K} \Delta_{a_k} \mathbb{E}[\tau_k] \leq O\left(c^{\frac{p}{p-1}} e^{\frac{b_p v_p}{c^p}} K^{1-\frac{1}{p}} T^{\frac{1}{p}}\right). \tag{63}$$

□

Proof of Lemma 4: Let Δ be $(e^{(1/4)}(K/T))^{1-(1/p)}$. Let k_0 be an index of the action such that $\Delta_{a_{k_0}} \leq \Delta < \Delta_{a_{k_0+1}}$

$$\sum_{j=k_0+1}^{K} \mathbb{P}\left(\bar{E}_{a_j}^c\right)\left(\Delta_{a_j} - \Delta_{a_{j-1}}\right) \tag{64}$$

$$= \sum_{j=k_0+1}^{K} \mathbb{P}\left(Z < r_{a^\star} - \frac{\Delta_{a_j}}{6}\right)\left(\Delta_{a_j} - \Delta_{a_{j-1}}\right) \tag{65}$$

$$\leq \Delta - \Delta_{a_{k_0}} + \int_{\Delta}^{1} \mathbb{P}\left(Z < r_{a^\star} - \frac{u}{6}\right) du. \tag{66}$$

For a fixed $u \in [\Delta, 1]$, we have

$$\mathbb{P}(Z < r_{a^\star} - u/6) \tag{67}$$

$$= \mathbb{P}\left(\min_{1<t\leq T} \hat{r}_{t-1,a^\star} + \beta_{t-1,a^\star} < r_{a^\star} - u/6\right) \tag{68}$$

$$\leq \sum_{s=1}^{T} \mathbb{P}\left(\left(r_{a^\star} - \hat{r}_{s,a^\star}\right) > \frac{c\ln_+(T/(Ks))}{s^{1-\frac{1}{p}}} + \frac{u}{6}\right) \tag{69}$$

$$\leq \sum_{s=1}^{T} \exp\left(-\ln_+(T/(Ks)) - us^{1-\frac{1}{p}}/(6c)\right) \tag{70}$$

$$\leq \frac{K}{T} \sum_{s=1}^{T} s \exp\left(-us^{1-\frac{1}{p}}/(6c)\right) \tag{71}$$

$$\leq \frac{K}{T}\left(\frac{6cp}{e(p-1)u}\right)^{\frac{p}{p-1}} + \frac{K}{T}\int_0^\infty x e^{-\frac{ux^{1-\frac{1}{p}}}{6c}} dx \tag{72}$$

$$= \frac{K}{T}\left(\frac{6cp}{e(p-1)u}\right)^{\frac{p}{p-1}} + \frac{K}{2T}\Gamma\left(\frac{3p-1}{p-1}\right)\left(\frac{6c}{u}\right)^{\frac{p}{p-1}} \tag{73}$$

$$\leq C c^{\frac{p}{p-1}} \frac{K}{T} u^{-\frac{p}{p-1}} \tag{74}$$

where C is a large constant including only c and p from the above inequality. From the above inequality, we can bound the integration as follows:

$$\int_{\Delta}^{1} \mathbb{P}(Z < r_{a^\star} - u/6) du \leq C c^{\frac{p}{p-1}} \frac{K}{T} \int_{\Delta}^{1} u^{-\frac{p}{p-1}} du \tag{75}$$

$$= C c^{\frac{p}{p-1}} \frac{K}{T}\left[-(p-1)u^{-\frac{1}{p-1}}\right]_{\Delta}^{1} \leq C' c^{\frac{p}{p-1}} \frac{K}{T} \Delta^{-\frac{1}{p-1}} \tag{76}$$

where C' is a constant greater than $C(p-1)$. Finally, we get

$$T \sum_{j=k_0+1}^{K} \mathbb{P}\left(\bar{E}_{a_j}^c\right)\left(\Delta_{a_j} - \Delta_{a_{j-1}}\right) \tag{77}$$

$$\leq T\Delta + C' c^{\frac{p}{p-1}} K \Delta^{-\frac{1}{p-1}} \leq O\left(c^{\frac{p}{p-1}} K^{1-\frac{1}{p}} T^{\frac{1}{p}}\right). \tag{78}$$

□

Proof of Theorem 5: The proof can be done by combining Lemma 3 and 4. We combine all gap-independent bounds and for all gap-independent bounds, we set k_0 such that $\Delta_{a_{k_0}} < \Delta < \Delta_{a_{k_0+1}}$ where $\Delta = \max(e^p, e^{-(3(p-1)/2p)})(K/T)^{1-1/p}$. Then, $T\Delta_{k_0} < T\Delta = O(K^{1-1/p}T^{1/p})$ holds. Finally, we can obtain the gap-independent bounds

$$\mathbb{E}[\mathcal{R}_T] \leq O\left(c^{\frac{p}{p-1}} e^{\frac{b_p v_p}{c^p}} K^{1-1/p} T^{1/p}\right). \tag{79}$$

□

APPENDIX D
PROOFS FOR THEOREM 6

Proof of Lemma 5: For a fixed $a \in \mathcal{A}$, We first define a stopping time τ_k for the kth selection of a. Using τ_k, the

following bound can be derived:

$$\sum_{t=1}^{T}\mathbb{P}\Big(E_{t,a}^{(1)}\Big)=\sum_{t=1}^{T}\mathbb{P}\big(E_{t,a}\cap\bar{E}_{t,a}\cap\hat{E}_{t,a}^{c}\big) \tag{80}$$

$$\leq\sum_{k=0}^{T-1}\mathbb{P}\big(\hat{r}_{a,\tau_k}>x_a\big)\leq\exp\left(\frac{b_p\nu_p}{c^p}\right)\sum_{k=0}^{T-1}e^{-\frac{k^{1-\frac{1}{p}}\Delta_a}{3c}} \tag{81}$$

$$\leq\frac{(3c)^{\frac{p}{p-1}}\exp\Big(\frac{b_p\nu_p}{c^p}\Big)\Gamma\Big(\frac{2p-1}{p-1}\Big)}{\Delta_a^{\frac{p}{p-1}}}. \tag{82}$$

The probability of $E_{t,a}^{(2)}$ can be bounded by the probability of $\hat{E}_{\tau_k,a}\cap\tilde{E}_{\tau_k,a}^{c}$ as follows:

$$\sum_{t=1}^{T}\mathbb{P}\Big(E_{t,a}^{(2)}\Big)=\sum_{t=1}^{T}\mathbb{P}\big(E_{t,a}\cap\bar{E}_{t,a}\cap\hat{E}_{t,a}\cap\tilde{E}_{t,a}^{c}\big) \tag{83}$$

$$\leq\sum_{k=0}^{T-1}\mathbb{P}\big(\hat{E}_{\tau_k,a}\cap\tilde{E}_{\tau_k,a}^{c}\big) \tag{84}$$

$$=\sum_{k=0}^{T-1}\mathbb{P}\big(\hat{r}_{a,\tau_k}\leq x_a,\hat{r}_{\tau_k,a}+(1+\epsilon)\beta_{\tau_k,a}G_{\tau_k,a}>y_a\big) \tag{85}$$

$$\leq\sum_{k=0}^{T-1}\mathbb{P}\big(x_a+(1+\epsilon)\beta_{\tau_k,a}G_{\tau_k,a}>y_a\big) \tag{86}$$

$$\leq\ell_0+\sum_{k=\ell_0+1}^{T-1}\mathbb{P}\big(x_a+(1+\epsilon)\beta_{\tau_k,a}G_{\tau_k,a}>y_a\big). \tag{87}$$

Then, similar to Lemma 2, we properly take ℓ_0 to bound the sum of probability. Now, let us take

$$\ell_0=\max\left(3(1+\epsilon)\ln\left(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\right),1\right)^{\frac{p}{p-1}}/\Delta_{a_k}^{p/(p-1)}.$$

Then, for $l>\ell_0$, we have $l>\Delta_{a_k}^{-p/(p-1)}$, and thus

$$(1+\epsilon)\ln_+\left(\frac{T}{Kl}\right)/l^{1-1/p}\leq\Delta_{a_k}/3. \tag{88}$$

Hence, we have

$$\ell_0+\sum_{k=\ell_0+1}^{T-1}\mathbb{P}\big(x_a+(1+\epsilon)\beta_{\tau_k,a}G_{\tau_k,a}>y_a\big) \tag{89}$$

$$\leq\ell_0+\sum_{k=\ell_0+1}^{T-1}\mathbb{P}\big(G_{\tau_k,a}>1\big) \tag{90}$$

$$=\frac{\max\Big(3(1+\epsilon)\ln\Big(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\Big),1\Big)^{\frac{p}{p-1}}}{\Delta_{a_k}^{p/(p-1)}} \tag{91}$$

where $(\Delta_a/3(1+\epsilon))\beta_{\tau_k,a}^{-1}>1$ for $l>\ell_0$ and G is a random variable within $[-1,1]$ and hence, $\mathbb{P}(G>1)=0$ holds.

Finally, to prove the bound of the sum of the probability of $E_{t,a}^{(3)}$, we barrow the idea from [3] and [23]. Let $\mathcal{F}_{k,a}:=\mathbb{P}(\hat{r}_{\tau_k,a_\star}+\beta_{\tau_k,a_\star}G_{\tau_k,a_\star}>y_a)$. Using $\mathcal{F}$, we can derive the following bound:

$$\sum_{t=1}^{T}\mathbb{P}\Big(E_{t,a}^{(3)}\Big)=\sum_{t=1}^{T}\mathbb{P}\big(E_{t,a}\cap\bar{E}_{t,a}\cap\hat{E}_{t,a}\cap\tilde{E}_{t,a}\big) \tag{92}$$

$$\leq\sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-\mathcal{F}_{k,a}}{\mathcal{F}_{k,a}}\mathbb{I}\left(r_{a_\star}-\frac{\Delta_a}{6}-\beta_{\tau_k,a_\star}<\hat{r}_{\tau_k,a_\star}\right)\right] \tag{93}$$

$$\leq\sum_{k=0}^{T-1}\frac{1-\mathcal{F}_{k,a}}{\mathcal{F}_{k,a}}\leq\sum_{k=0}^{T-1}\frac{\mathbb{P}\Big(G<\frac{1}{(1+\epsilon)}-\frac{\Delta_a}{6(1+\epsilon)\beta_{\tau_k,a_\star}}\Big)}{\mathbb{P}\Big(G>\frac{1}{(1+\epsilon)}-\frac{\Delta_a}{6(1+\epsilon)\beta_{\tau_k,a_\star}}\Big)} \tag{94}$$

$$\leq\sum_{k=0}^{\ell_0}\frac{\mathbb{P}(G<1/(1+\epsilon))}{\mathbb{P}(G>1/(1+\epsilon))} \tag{95}$$

$$+\sum_{k=\ell_0+1}^{T-1}\frac{\mathbb{P}\big(G<1/(1+\epsilon)-\Delta_a/\big(6(1+\epsilon)\beta_{\tau_k,a_\star}\big)\big)}{\mathbb{P}\big(G>1/(1+\epsilon)-\Delta_a/\big(6(1+\epsilon)\beta_{\tau_k,a_\star}\big)\big)}. \tag{96}$$

Similar to the MR-UCB, let us take

$$\ell_0=\max\left(6(2+\epsilon)\ln\left(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\right),1\right)^{\frac{p}{p-1}}/\Delta_{a_k}^{p/(p-1)}.$$

For $l>\ell_0$, we have $l>\Delta_{a_k}^{-p/(p-1)}$, and thus

$$\ln_+\left(\frac{T}{Kl}\right)/l^{1-1/p}\leq\Delta_{a_k}/6/(2+\epsilon). \tag{97}$$

In other words, we have

$$1/(1+\epsilon)-\Delta_{a_k}/\big(6(1+\epsilon)\beta_{\tau_l,a_\star}\big)\leq-1. \tag{98}$$

Hence, $\mathbb{P}(G<(1/(1+\epsilon))-(\Delta_a/6(1+\epsilon)\beta_{\tau_k,a_\star}))=0$ since G is a random variable in $[-1,1]$. Finally, we get

$$\sum_{t=1}^{T}\mathbb{P}\Big(E_{t,a}^{(3)}\Big)\leq M_\epsilon\frac{\max\Big(6(2+\epsilon)\ln\Big(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\Big),1\Big)^{\frac{p}{p-1}}}{\Delta_{a_k}^{p/(p-1)}} \tag{99}$$

where $\mathbb{P}(G<1/(1+\epsilon))/\mathbb{P}(G>1/(1+\epsilon))<M_\epsilon$. □

Proof of Theorem 6: First, using Lemma 5, we can bound the final term in (19) as follows:

$$\Delta_{a_k}\sum_{t=1}^{T}\mathbb{P}\big(E_{t,a}\cap\bar{E}_a\big)=\Delta_{a_k}\mathbb{E}\big[\mathbb{I}\big[Z>z_{a_k}\big]n_{a_k,T}\big] \tag{100}$$

$$\leq\Delta_{a_k}\sum_{t=1}^{T}\mathbb{P}\big(E_{t,a}\cap\bar{E}_{t,a}\big) \tag{101}$$

$$\leq\Delta_{a_k}\sum_{t=1}^{T}\Big[\mathbb{P}\Big(E_{t,a}^{(1)}\Big)+\mathbb{P}\Big(E_{t,a}^{(2)}\Big)+\mathbb{P}\Big(E_{t,a}^{(3)}\Big)\Big] \tag{102}$$

$$\leq O\left(M_\epsilon\frac{\max\Big((2+\epsilon)\ln\Big(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\Big),1\Big)^{\frac{p}{p-1}}}{\Delta_{a_k}^{p/(p-1)}}\right. \tag{103}$$

$$\left.+\frac{(3c)^{\frac{p}{p-1}}\exp\Big(\frac{b_p\nu_p}{c^p}\Big)}{\Delta_{a_k}^{1/(p-1)}}\right). \tag{104}$$

Hence, from Lemma 6, the gap-dependent regret bound of MR-APE can be obtained as follows:

$$O\left(\sum_{k=1}^{K}M_\epsilon^{+}\frac{\max\Big((2+\epsilon)\ln\Big(\frac{T}{K}\Delta_{a_k}^{p/(p-1)}\Big),1\Big)^{\frac{p}{p-1}}}{\Delta_{a_k}^{1/(p-1)}}\right. \tag{105}$$

$$+\frac{Kc^{\frac{p}{p-1}}\exp\left(\frac{b_p v_p}{c^p}\right)}{\Delta_{a_k}^{\frac{1}{p-1}}}\Bigg) \quad (106)$$

where $M_\epsilon^+ := \max(M_\epsilon, 1)$ that combines (89) and (99). □

APPENDIX E
PROOFS FOR THEOREM 7

Proof of Theorem 7: The proof can be simply done by picking k_0 such that $\Delta_{a_{k_0}} < \Delta < \Delta_{a_{k_0+1}}$ where $\Delta = \max(e^p, e^{-(3(p-1)/2p)})(K/T)^{1-1/p}$. Then, $T\Delta_{k_0} < T\Delta = O(K^{1-1/p}T^{1/p})$ holds. Finally, we can obtain the gap-independent bounds as

$$O\left(\max\left(M_\epsilon(2+\epsilon)^{\frac{p}{p-1}}, c^{\frac{p}{p-1}}e^{\frac{b_p v_p}{c^p}}\right)K^{1-1/p}T^{1/p}\right). \quad (107)$$

□

REFERENCES

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

[2] J. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proc. 22nd Conf. Learn. Theory*, 2009, pp. 1–10.

[3] S. Agrawal and N. Goyal, "Further optimal regret bounds for Thompson sampling," in *Proc. 16th Int. Conf. Artif. Intell. Statist.*, 2013, pp. 99–107.

[4] P. Ménard and A. Garivier, "A minimax and asymptotically optimal algorithm for stochastic bandits," in *Proc. Int. Conf. Algorithmic Learn. Theory*, vol. 76, 2017, pp. 223–237.

[5] B. Kim and A. Tewari, "On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2691–2700.

[6] T. Jin, P. Xu, J. Shi, X. Xiao, and Q. Gu, "MOTS: Minimax optimal Thompson sampling," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 5074–5083.

[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] 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.*, Dec. 2020, pp. 8452–8462.

[9] L. Wei and V. Srivastava, "Minimax policy for heavy-tailed bandits," *IEEE Control Syst. Lett.*, vol. 5, no. 4, pp. 1423–1428, Oct. 2021.

[10] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Proc. 20th Int. Conf. Algorithmic Learn. Theory*, 2009, pp. 23–37.

[11] A. Rakotomamonjy, S. Koço, and L. Ralaivola, "Greedy methods, randomization approaches, and multiarm bandit algorithms for efficient sparsity-constrained optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2789–2802, Nov. 2017.

[12] O. Atan, C. Tekin, and M. van der Schaar, "Global bandits," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5798–5811, Dec. 2018.

[13] G. Ditzler, R. Polikar, and G. Rosen, "A sequential learning approach for scaling up filter-based feature subset selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2530–2544, Jun. 2018.

[14] K. Gokcesu and S. S. Kozat, "An online minimax optimal algorithm for adversarial multiarmed bandit problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5565–5580, Nov. 2018.

[15] M. M. Drugan, "Covariance matrix adaptation for multiobjective multi-armed bandits," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 8, pp. 2493–2502, Aug. 2019.

[16] M. M. Neyshabouri, K. Gokcesu, H. Gokcesu, H. Ozkan, and S. S. Kozat, "Asymptotically optimal contextual bandit algorithm using hierarchical structures," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 923–937, Mar. 2019.

[17] A. Alipour-Fanid, M. Dabaghchian, and K. Zeng, "An optimal algorithm for the stochastic bandits while knowing the near-optimal mean reward," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 5, pp. 2285–2291, May 2021.

[18] S. Yang and Y. Gao, "An optimal algorithm for the stochastic bandits while knowing the near-optimal mean reward," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2285–2291, May 2021.

[19] 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.

[20] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[21] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 759–767, Oct. 2013.

[22] S. Agrawal, S. K. Juneja, and W. M. Koolen, "Regret minimization in heavy-tailed bandits," in *Proc. Conf. Learn. Theory*, vol. 134, 2021, pp. 26–62.

[23] N. Cesa-Bianchi, C. Gentile, G. Neu, and G. Lugosi, "Boltzmann exploration done right," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6284–6293.

[24] O. Catoni, "Challenging the empirical mean and empirical variance: A deviation study," *Annales de l'IHP Probabilités et Statistiques*, vol. 48, no. 4, pp. 1148–1185, 2012.

[25] S. Rajkumar. (2021). *Cryptocurrency Historical Prices*. [Online]. Available: https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory

[26] S. T. Rachev, *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance*. Amsterdam, The Netherlands: Elsevier, 2003.

[27] H. Panahi, "Model selection test for the heavy-tailed distributions under censored samples with application in financial data," *Int. J. Financial Stud.*, vol. 4, no. 4, p. 24, Dec. 2016.

Kyungjae Lee (Member, IEEE) received the B.S. and Ph.D. degrees in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2015 and 2020, respectively.

He is currently an Assistant Professor with the Department of Artificial Intelligence, Chung-Ang University, Seoul. His current research interests include multiarmed bandit, combinatorial bandits, reinforcement learning, and its application.

Sungbin Lim (Member, IEEE) received the B.S. and Ph.D. degrees in mathematics from Korea University, Seoul, South Korea, in 2010 and 2016, respectively.

He is currently an Assistant Professor with the Artificial Intelligence Graduate School and the Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea. His current research interests include statistical machine learning, stochastic optimization, reinforcement learning, and causal inference.