Minimax Policy for Heavy-tailed Bandits

Lai Wei Vaibhav Srivastava

***Abstract*— We study the stochastic Multi-Armed Bandit (MAB) problem under worst-case regret and heavy-tailed reward distribution. We modify the minimax policy MOSS [1] for the sub-Gaussian reward distribution by using saturated empirical mean to design a new algorithm called Robust MOSS. We show that if the moment of order $1+\epsilon$ for the reward distribution exists, then the refined strategy has a worst-case regret matching the lower bound while maintaining a distribution-dependent logarithm regret.**

***Index Terms*— Heavy-tailed distribution, stochastic MAB, worst-case regret, minimax policy.**

I. Introduction

THE dilemma of exploration versus exploitation is common in scenarios involving decision-making in unknown environments. In these contexts, exploration means learning the environment while exploitation means taking empirically computed best actions. When finite time performance is concerned, i.e., scenarios in which one cannot learn indefinitely, ensuring a good balance of exploration and exploitation is the key to a good performance. MAB and its variations are prototypical models for these problems, and they are widely used in many areas such as network routing, recommendation systems and resource allocation; see [2, Chapter 1].

The stochastic MAB problem was originally proposed by Robbins [3]. In this problem, at each time, an agent chooses an arm from a set of K arms and receives the associated reward. The reward at each arm is a stationary random variable with an unknown mean. The objective is to design a policy that maximizes the expected cumulative reward or equivalently minimizes the *expected cumulative regret*, defined by the expected cumulative difference between the maximum mean reward and the reward obtained using the policy.

The worst-case regret is defined by the supremum of the expected cumulative regret computed over a class of reward distributions, e.g., sub-Gaussian distributions, or distributions with bounded support. The *minimax regret* is defined as the minimum worst-case regret, where the minimum is computed over all the policies. By construction, the worst-case regret uses minimal information about the underlying distribution and the associated regret bounds are called *distribution-free bounds*. In contrast, the standard regret bounds depend on the difference between the mean rewards from the optimal and suboptimal arms, and the corresponding bounds are referred as *distribution-dependent bounds*.

This work has been supported by NSF Award IIS-1734272.

L. Wei and V. Srivastava are with the Department of Electrical and Computer Engineering. Michigan State University, East Lansing, MI 48823 USA. e-mail: weilai1@msu.edu; e-mail: vaibhav@egr.msu.edu

In their seminal work, Lai and Robbins [4] establish that the expected cumulative regret admits an asymptotic distribution-dependent lower bound that is a logarithmic function of the time-horizon T. Here, asymptotic refers to the limit $T \to +\infty$. They also propose a general method of constructing Upper Confidence Bound (UCB) based policies that attain the lower bound asymptotically. By assuming rewards to be bounded or more generally sub-Gaussian, several subsequent works design simpler algorithms with finite time performance guarantees, e.g., the UCB1 algorithm by Auer et al. [5]. By using Kullback-Leibler(KL) divergence based upper confidence bounds, Garivier and Cappé [6] designed KL-UCB, which is proved to have efficient finite time performance as well as asymptotic optimality.

In the worst-case setting, the lower and upper bounds are distribution-free. Assuming the rewards are bounded, Audibert and Bubeck [1] establish a $\Omega(\sqrt{KT})$ lower bound on the minimax regret. They also studied a modified UCB algorithm called Minimax Optimal Strategy in the Stochastic case (MOSS) and proved that it achieves an order-optimal worst-case regret while maintaining a logarithm distribution-dependent regret. Degenne and Perchet [7] extend MOSS to an any-time version called MOSS-anytime.

The rewards being bounded or sub-Gaussian is a common assumption that gives sample mean an exponential convergence and simplifies the MAB problem. However in many applications, such as social networks [8] and financial markets [9], the rewards are heavy-tailed. For the standard stochastic MAB problem, Bubeck et al. [10] relax the sub-Gaussian assumption by only assuming the rewards to have finite moments of order $1+\epsilon$ for some $\epsilon \in (0,1]$. They present the robust UCB algorithm and show that it attains an upper bound on the cumulative regret that is within a constant factor of the distribution-depend lower bound in the heavy-tailed setting. However, the solutions provided in [10] are not able to provably achieve an order optimal worst-case regret. Specifically, the factor of optimality is a poly-logarithmic function of time-horizon.

In this paper, we study the minimax heavy tail bandit problem in which reward distributions admit moments of order $1+\epsilon$, with $\epsilon > 0$. We propose and analyze Robust MOSS algorithm to show that it achieves worst-case regret matching with the lower bound while maintaining a distribution-dependent logarithm regret. To the best of our knowledge, Robust MOSS is the first algorithm to achieve order optimal worst-case regret for heavy-tailed bandits. Our results build on techniques in [1] and [10], and augment them with new analysis based on maximal Bennett inequalities.

The remaining paper is organized as follows. We describe

the minimax heavy-tailed multiarmed bandit problem and present some background material in Section II. We present and analyze the Robust MOSS algorithm in Sections III and IV, respectively, and numerically compare it with the state of the art in Section V. We conclude in Section VI.

II. Background & Problem Description

A. Stochastic MAB Problem

In a stochastic MAB problem, an agent chooses an arm φ_t from the set of K arms $\{1,\dots,K\}$ at each time $t \in \{1,\dots,T\}$ and receives the associated reward. The reward at each arm k is drawn from an unknown distribution f_k with unknown mean μ_k. Let the maximum mean reward among all arms be μ^*. We use $\Delta_k = \mu^* - \mu_k$ to measure the suboptimality of arm k. The objective is to maximize the expected cumulative reward or equivalently to minimize the *expected cumulative regret* defined by

$$R_T := \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^* - X_{\varphi_t}\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\Delta_{\varphi_t}\right],$$

which is the difference between the expected cumulative reward obtained by selecting the arm with the maximum mean reward μ^* and selecting arms $\varphi_1,\dots,\varphi_T$.

The expected cumulative regret R_T is implicitly defined for a fixed distribution of rewards from each arm $\{f_1,\dots,f_K\}$. The worst-case regret is the expected cumulative regret for the worst possible choice of reward distributions. In particular,

$$R_T^{\text{worst}} = \sup_{\{f_1,\dots,f_K\}} R_T.$$

The regret associated with the policy that minimizes the above worst-case regret is called *minimax regret*.

B. Problem Description: Heavy-tailed Stochastic MAB

In this paper, we study the heavy-tailed stochastic MAB problem, which is the stochastic MAB problem with following assumptions.

Assumption 1: Let X be a random reward drawn from any arm $k \in \{1,\dots,K\}$. There exists a constant $u \in \mathbb{R}_{>0}$ such that $\mathbb{E}\left[|X|^{1+\epsilon}\right] \le u^{1+\epsilon}$ for some $\epsilon \in (0,1]$.

Assumption 2: Parameters T, K, u and ϵ are known.

C. MOSS Algorithm for Worst-Case Regret

We now present the MOSS algorithm proposed in [1]. The MOSS algorithm is designed for stochastic MAB problem with bounded rewards and in this paper, we extend it to design Robust MOSS algorithm for heavy-tailed bandits.

Suppose that arm k is sampled $n_k(t)$ times until time $t-1$, and $\bar{\mu}^k_{n_k(t)}$ is the associated empirical mean, then, at time t, MOSS picks the arm that maximizes the following UCB

$$g^k_{n_k(t)} = \bar{\mu}^k_{n_k(t)} + \sqrt{\frac{\max\left(\ln\left(\frac{T}{Kn_k(t)}\right),0\right)}{n_k(t)}}.$$

If the rewards from the arms have bounded support $[0,1]$, then the worst-case regret for MOSS satisfies $R_T^{\text{worst}} \le 49\sqrt{KT}$, which is order optimal [1]. Meanwhile, MOSS maintains a logarithm distribution-dependent regret bound.

D. A Lower Bound for Heavy-tailed Minimax Regret

We now present the lower bound on the minimax regret for the heavy tailed bandit problem derived in [10].

Theorem 1 ([10, Th. 2]): For any fixed time horizon T and the stochastic MAB problem under Assumptions 1 and 2 with $u = 1$,

$$R_T^{\text{worst}} \ge 0.01K^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}.$$

Remark 1: Since R_T scales with u, the lower bound for heavy tail bandit is $\Omega\big(uK^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}\big)$. This lower bound also indicates that within a finite horizon T, it is almost impossible to differentiate the optimal arm from arm k, if $\Delta_k \in O\big(u(K/T)^{\frac{\epsilon}{1+\epsilon}}\big)$. As a special case, rewards with bounded support $[0,1]$ correspond to $\epsilon = 1$ and $u = 1$. Then, the lower bound $\Omega(\sqrt{KT})$ match with the regret upper bound achieved by MOSS.

III. A Robust Minimax Policy

To deal with the heavy-tailed reward distribution, we replace the empirical mean with a saturated empirical mean. Although saturated empirical mean is a biased estimator, it has better convergence properties. We construct a novel UCB index to evaluate the arms, and at each time slot the arm with the maximum UCB index is picked.

A. Robust MOSS

In Robust MOSS, we consider a robust mean estimator called saturated empirical mean which is formally defined in the following subsection. Let $n_k(t)$ be the number of times that arm k has been selected until time $t-1$. At time t, let $\hat{\mu}^k_{n_k(t)}$ be the saturated empirical mean reward computed from the $n_k(t)$ samples at arm k. Robust MOSS initializes by selecting each arm once and subsequently, at each time t, selects the arm that maximizes the following UCB

$$g^k_{n_k(t)} = \hat{\mu}^k_{n_k(t)} + (1+\eta)c_{n_k(t)},$$

where $\eta > 0$ is an appropriate constant, $c_{n_k(t)} = u \times \left[\phi(n_k(t))\right]^{\frac{\epsilon}{1+\epsilon}}$ and

$$\phi(n) = \frac{\ln_+\left(\frac{T}{Kn}\right)}{n},$$

where $\ln_+(x) := \max(\ln x, 1)$. Note that both $\phi(n)$ and c_n are monotonically decreasing in n.

B. Saturated Empirical Mean

The robust saturated empirical mean is similar to the truncated empirical mean used in [10], which is employed to extend UCB1 to achieve logarithm distribution-dependent regret for the heavy-tailed MAB problem. Let $\{X_i\}_{i\in\{1,\dots,m\}}$ be a sequence of i.i.d. random variables with mean μ and $\mathbb{E}\left[|X_i|^{1+\epsilon}\right] \le u^{1+\epsilon}$, where $u > 0$. Pick $a > 1$ and let $h(m) = a^{\lfloor \log_a(m)\rfloor + 1}$ such that $h(m) \ge m$. Define the saturation point B_m by

$$B_m := u \times \left[\phi\big(h(m)\big)\right]^{-\frac{1}{1+\epsilon}}.$$

Then, the saturated empirical mean estimator is defined by

$$\hat{\mu}_m := \frac{1}{m}\sum_{i=1}^{m} \mathrm{sat}(X_i, B_m), \tag{1}$$

where $\mathrm{sat}(X_i, B_m) := \mathrm{sign}(X_i)\min\{|X_i|, B_m\}$.

Define $d_i := \mathrm{sat}(X_i, B_m) - \mathbb{E}[\mathrm{sat}(X_i, B_m)]$. The following lemma examines the estimator bias and provides an upper bound on the error of saturated empirical mean.

Lemma 2 (Error of saturated empirical mean): For an i.i.d. sequence of random variables $\{X_i\}_{i\in\{1,\dots,m\}}$ such that $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}\left[X_i^{1+\epsilon}\right] \le u^{1+\epsilon}$, the saturated empirical mean (1) satisfies

$$\left|\hat{\mu}_m - \mu - \frac{1}{m}\sum_{i=1}^{m} d_i\right| \le \frac{u^{1+\epsilon}}{B_m^{\epsilon}}.$$

Proof: Since $\mu = \mathbb{E}\left[X_i\left(\mathbf{1}_{\{|X_i|\le B_m\}} + \mathbf{1}_{\{|X_i|>B_m\}}\right)\right]$, the error of estimator $\hat{\mu}_m$ satisfies

$$\begin{aligned}\hat{\mu}_m - \mu =& \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{sat}(X_i, B_m) - \mu\right)\\ =& \frac{1}{m}\sum_{i=1}^{m} d_i + \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}[\mathrm{sat}(X_i, B_m)] - \mu\right),\end{aligned}$$

where the second term is the bias of $\hat{\mu}_m$. We now compute an upper bound on the bias.

$$\begin{aligned}\left|\mathbb{E}[\mathrm{sat}(X_i, B_m)] - \mu\right| &\le \mathbb{E}\left[|X_i|\,\mathbf{1}_{\{|X_i|>B_m\}}\right]\\ &\le \mathbb{E}\left[\frac{|X_i|^{1+\epsilon}}{(B_m)^{\epsilon}}\right] \le \frac{u^{1+\epsilon}}{(B_m)^{\epsilon}},\end{aligned}$$

which concludes the proof. ∎

We now establish properties of d_i.

Lemma 3 (Properties of d_i): For any $i \in \{1,\dots,m\}$, d_i satisfies (i) $|d_i| \le 2B_m$ (ii) $\mathbb{E}[d_i^2] \le u^{1+\epsilon}B_m^{1-\epsilon}$.

Proof: Property (i) follows immediately from definition of d_i, and property (ii) follows from

$$\mathbb{E}[d_i^2] \le \mathbb{E}\left[\mathrm{sat}^2(X_i, B_m)\right] \le \mathbb{E}\left[|X_i|^{1+\epsilon}B_m^{1-\epsilon}\right].$$ ∎

IV. Analysis of Robust MOSS

In this section, we analyze Robust MOSS to provide both distribution-free and distribution-dependent regret bounds.

A. Properties of Saturated Empirical Mean Estimator

To derive the concentration property of saturated empirical mean, we use a maximal Bennett type inequality as shown in Lemma 4.

Lemma 4 (Maximal Bennett's inequality [11]): Let $\{X_i\}_{i\in\{1,\dots,n\}}$ be a sequence of bounded random variables with support $[-B, B]$, where $B \ge 0$. Suppose that $\mathbb{E}[X_i|X_1,\dots,X_{i-1}] = \mu_i$ and $\mathrm{Var}[X_i|X_1,\dots,X_{i-1}] \le v$. Let $S_m = \sum_{i=1}^{m}(X_i - \mu_i)$ for any $m \in \{1,\dots,n\}$. Then, for any $\delta \ge 0$

$$\mathbb{P}\left(\exists m \in \{1,\dots,n\} : S_m \ge \delta\right) \le \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right),$$

$$\mathbb{P}\left(\exists m \in \{1,\dots,n\} : S_m \le -\delta\right) \le \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right),$$

where $\psi(x) = (1+1/x)\ln(1+x) - 1$.

Remark 2: For $x \in (0,\infty)$, function $\psi(x)$ is monotonically increasing in x.

Now, we establish an upper bound on the probability that the UCB underestimates the mean at arm k by an amount x.

Lemma 5: For any arm $k \in \{1,\dots,K\}$ and any $t \in \{K+1,\dots,T\}$ and $x > 0$, if $\eta\psi(2\eta/a) \ge 2a$, the probability of event $\left\{g^k_{n_k(t)} \le \mu_k - x\right\}$ is no greater than

$$\frac{K}{T}\frac{a}{\ln(a)}\Gamma\left(\frac{1}{\epsilon}+2\right)\left(\frac{\psi\left(2\eta/a\right)}{2a}\frac{x}{u}\right)^{-\frac{1+\epsilon}{\epsilon}}.$$

Proof: It follows from Lemma 2 that

$$\begin{aligned}&\mathbb{P}\left(g^k_{n_k(t)} \le \mu_k - x\right)\\ \le&\mathbb{P}\left(\exists m \in \{1,\dots,T\} : \hat{\mu}^k_m + (1+\eta)c_m \le \mu_k - x\right)\\ \le&\mathbb{P}\left(\exists m \in \{1,\dots,T\} : \sum_{i=1}^{m}\frac{d^k_i}{m} \le \frac{u^{1+\epsilon}}{B_m^{\epsilon}} - (1+\eta)c_m - x\right)\\ \le&\mathbb{P}\left(\exists m \in \{1,\dots,T\} : \frac{1}{m}\sum_{i=1}^{m} d^k_i \le -x - \eta c_m\right),\end{aligned}$$

where d^k_i is defined similarly to d_i for i.i.d. reward sequence at arm k and the last inequality is due to

$$\frac{u^{1+\epsilon}}{B_m^{\epsilon}} = u\left[\phi\left(h(m)\right)\right]^{\frac{\epsilon}{1+\epsilon}} \le u\left[\phi(m)\right]^{\frac{\epsilon}{1+\epsilon}} = c_m. \tag{2}$$

Recall $a > 1$. We apply a peeling argument [12, Sec 2.2] with geometric grid $a^s \le m < a^{s+1}$ over time interval $\{1,\dots,T\}$. Since c_m is monotonically decreasing with m,

$$\begin{aligned}&\mathbb{P}\left(\exists m \in \{1,\dots,T\} : \frac{1}{m}\sum_{i=1}^{m} d^k_i \le -x - \eta c_m\right)\\ \le&\sum_{s\ge 0}\mathbb{P}\left(\exists m \in [a^s, a^{s+1}) : \sum_{i=1}^{m} d^k_i \le -a^s\left(x + \eta c_{a^{s+1}}\right)\right).\end{aligned}$$

Also notice that $B_m = B_{a^s}$ for all $m \in [a^s, a^{s+1})$. Then with properties in Lemma 3, we apply Lemma 4 to get

$$\begin{aligned}&\sum_{s\ge 0}\mathbb{P}\left(\exists m \in [a^s, a^{s+1}) : \sum_{i=1}^{m} d^k_i \le -a^s\left(x + \eta c_{a^{s+1}}\right)\right)\\ \le&\sum_{s\ge 0}\exp\left(-\frac{a^s\left(x + \eta c_{a^{s+1}}\right)}{2B_{a^s}}\psi\left(\frac{2B_{a^s}\left(x + \eta c_{a^{s+1}}\right)}{au^{1+\epsilon}B_{a^s}^{1-\epsilon}}\right)\right)\\ &\left(\text{since } \psi(x) \text{ is monotonically increasing}\right)\\ \le&\sum_{s\ge 0}\exp\left(-\frac{a^s\left(x + \eta c_{a^{s+1}}\right)}{2B_{a^s}}\psi\left(\frac{2\eta B^{\epsilon}_{a^s}c_{a^{s+1}}}{au^{1+\epsilon}}\right)\right)\\ &\left(\text{plug in } c_{a^{s+1}},\ B_{a^s} \text{ and use } h(a^s) = a^{s+1}\right)\\ =&\sum_{s\ge 1}\exp\left(-a^s\left(\frac{x}{B_{a^{s-1}}} + \eta\phi(a^s)\right)\frac{\psi\left(2\eta/a\right)}{2a}\right)\\ &\left(\text{plug in } \phi(a^s) \text{ and use } \eta\psi(2\eta/a) \ge 2a, \ln_+(y) \ge \ln(y)\right)\\ \le&\sum_{s\ge 1}\exp\left(-a^s\frac{x}{B_{a^{s-1}}}\frac{\psi\left(2\eta/a\right)}{2a}\right)\frac{K}{T}a^s.\end{aligned} \tag{3}$$

Let $b = x\psi\left(2\eta/a\right)/(2au)$. Since $B_{a^{s-1}} \leq ua^{\frac{s}{1+\epsilon}}$, we have

$$\begin{aligned}
(3) \leq & \frac{K}{T}\sum_{s\geq 1} a^s \exp\left(-ba^{\frac{\epsilon s}{1+\epsilon}}\right) \\
\leq & \frac{K}{T}\int_1^{+\infty} a^y \exp\big(-ba^{\frac{(y-1)\epsilon}{1+\epsilon}}\big)dy \\
= & \frac{K}{T}a\int_0^{+\infty} a^y \exp\big(-ba^{\frac{y\epsilon}{1+\epsilon}}\big)dy \\
& \Big(\text{where we set } z = ba^{\frac{y\epsilon}{1+\epsilon}}\Big) \\
= & \frac{K}{T}\frac{a}{\ln(a)}\frac{1+\epsilon}{\epsilon}b^{-\frac{1+\epsilon}{\epsilon}}\int_b^{+\infty} z^{\frac{1+\epsilon}{\epsilon}-1}\exp\big(-z\big)dz \\
\leq & \frac{K}{T}\frac{a}{\ln(a)}\Gamma\left(\frac{1}{\epsilon}+2\right)b^{-\frac{1+\epsilon}{\epsilon}},
\end{aligned}$$

which concludes the proof. ∎

The following is a straightforward corollary of Lemma 5.

Corollary 6: For any arm $k \in \{1,\ldots,K\}$ and any $t \in \{K+1,\ldots,T\}$ and $x > 0$, if $\eta\psi(2\eta/a) \geq 2a$, the probability of event $\left\{g^k_{n_k(t)} - 2(1+\eta)c_{n_k(t)} \geq \mu_k + x\right\}$ shares the same bound in Lemma 5.

B. Distribution-free Regret Bound

The distribution-free upper bound for Robust MOSS, which is the main result for the paper, is presented in this section. We show that the algorithm achieves order optimal worst-case regret.

Theorem 7: For the heavy-tailed stochastic MAB problem with K arms and time horizon T, if η and a are selected such that $\eta\psi(2\eta/a) \geq 2a$, then Robust MOSS satisfies

$$R_T^{\text{worst}} \leq CuK^{\frac{\epsilon}{1+\epsilon}}(T/e)^{\frac{1}{1+\epsilon}} + 2uK,$$

where $C = \Gamma\left(1/\epsilon+2\right)\left[a/\left(6+3\eta\right)\right]^{\frac{1}{\epsilon}}\left[3/\psi\left(6+3\eta\right)\right]^{\frac{1+\epsilon}{\epsilon}} + \epsilon\Gamma\left(1/\epsilon+2\right)\left(6+3\eta\right)^{-\frac{1}{\epsilon}}\left[6a/\psi(2\eta/a)\right]^{\frac{1+\epsilon}{\epsilon}} a/\ln(a) + (6+3\eta)\left[e+(1+\epsilon)e^{\frac{-\epsilon}{1+\epsilon}}\right]$.

Remark 3: Parameter a and η as inputs to Robust MOSS can be selected by minimizing the leading constant C in the upper bound on the regret in Theorem 7. We have found that selecting a slightly larger than 1 and selecting smallest η that satisfies $\eta\psi(2\eta/a) \geq 2a$ yields good performance.

Proof: Since both the UCB and the regret scales with u defined in Assumption 1, to simplify the expressions, we assume $u = 1$. Also notice that Assumption 1 indicates $|\mu_k| \leq u$, so $\Delta_k \leq 2$ for any $k \in \{1,\ldots,K\}$. In the following, any terms with superscript or subscript "$*$" and "k" are with respect to the best and the k-th arm, respectively. The proof is divided into 4 steps.

Step 1: We follow a decoupling technique inspired by the proof of regret upper bound in MOSS [1]. Take the set of δ-bad arms as $\mathcal{B}_\delta$ as

$$\mathcal{B}_\delta := \{k \in \{1,\ldots,K\} \mid \Delta_k > \delta\}, \tag{4}$$

where we assign $\delta = (6+3\eta)\left(eK/T\right)^{\frac{\epsilon}{1+\epsilon}}$. Thus,

$$\begin{aligned}
R_T \leq & T\delta + \sum_{t=1}^{K}\Delta_k + \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\}\left(\Delta_{\varphi_t} - \delta\right)\right] \\
\leq & T\delta + 2K + \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\}\left(\Delta_{\varphi_t} - \delta\right)\right]. \quad (5)
\end{aligned}$$

Furthermore, we make the following decomposition

$$\begin{aligned}
& \sum_{t=K+1}^{T}\mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\}\left(\Delta_{\varphi_t} - \delta\right) \\
= & \sum_{t=K+1}^{T}\mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} \leq \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\}\left(\Delta_{\varphi_t} - \delta\right) \quad (6) \\
& + \sum_{t=K+1}^{T}\mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} > \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\}\left(\Delta_{\varphi_t} - \delta\right).
\end{aligned}$$

Notice that (6) describes regret from underestimating optimal arm $*$. For the second summand, since $g^{\varphi_t}_{n_{\varphi_t}(t)} \geq g^*_{n^*(t)}$,

$$\begin{aligned}
& \sum_{t=K+1}^{T}\mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} > \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\}\left(\Delta_{\varphi_t} - \delta\right) \\
\leq & \sum_{t=K+1}^{T}\mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^{\varphi_t}_{n_{\varphi_t}(t)} > \mu_{\varphi_t} + \frac{2\Delta_{\varphi_t}}{3}\right\}\Delta_{\varphi_t} \\
= & \sum_{k\in\mathcal{B}_\delta}\sum_{t=K+1}^{T}\mathbf{1}\left\{\varphi_t = k, g^k_{n_k(t)} > \mu_k + \frac{2\Delta_k}{3}\right\}\Delta_k, \quad (7)
\end{aligned}$$

which characterizes the regret caused by overestimating δ-bad arms.

Step 2: In this step, we bound the expectation of (6). When event $\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} \leq \mu^* - \Delta_{\varphi_t}/3\right\}$ happens, we know

$$\Delta_\varphi \leq 3\mu^* - 3g^*_{n^*(t)} \text{ and } g^*_{n^*(t)} < \mu^* - \frac{\delta}{3}.$$

Thus, we get

$$\begin{aligned}
& \mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} \leq \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\}(\Delta_{\varphi_t} - \delta) \\
\leq & \mathbf{1}\left\{g^*_{n^*(t)} < \mu^* - \frac{\delta}{3}\right\} \times \left(3\mu^* - 3g^*_{n^*(t)} - \delta\right) := Y_t
\end{aligned}$$

Since Y_t is a positive random variable, its expected value can be computed involving only its cumulative density function:

$$\begin{aligned}
\mathbb{E}\left[Y_t\right] = & \int_0^{+\infty}\mathbb{P}\left(Y_t > x\right)dx \\
\leq & \int_0^{+\infty}\mathbb{P}\left(3\mu^* - 3g^*_{n^*(t)} - \delta > x\right)dx \\
= & \int_\delta^{+\infty}\mathbb{P}\Big(\mu^* - g^*_{n^*(t)} > \frac{x}{3}\Big)dx.
\end{aligned}$$

Then we apply Lemma 5 at optimal arm $*$ to get

$$\mathbb{E}\left[Y_t\right] \leq \frac{KC_1}{T}\int_\delta^{+\infty}\frac{1}{\epsilon}x^{-\frac{1+\epsilon}{\epsilon}}dx = \frac{KC_1}{T\delta^{\frac{1}{\epsilon}}}$$

where $C_1 = \epsilon\Gamma\left(1/\epsilon+2\right)\left[6a/\psi(2\eta/a)\right]^{\frac{1+\epsilon}{\epsilon}} a/\ln(a)$. We conclude this step by

$$\mathbb{E}\left[(6)\right] \leq \sum_{t=K+1}^{T} Y_t \leq C_1 K\delta^{-\frac{1}{\epsilon}}.$$

Step 3: In this step, we bound the expectation of (7). For each arm $k \in \mathcal{B}_\delta$,

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t = k, g^k_{n_k(t)} \geq \mu_k + \frac{2\Delta_k}{3}\right\} \\
&= \sum_{t=K+1}^{T}\sum_{m=1}^{t-K} \mathbf{1}\left\{\varphi_t = k, n_k(t) = m\right\} \mathbf{1}\left\{g^k_m \geq \mu_k + \frac{2\Delta_k}{3}\right\} \\
&= \sum_{m=1}^{T-K} \mathbf{1}\left\{g^k_m \geq \mu_k + \frac{2\Delta_k}{3}\right\} \sum_{t=m+K}^{T} \mathbf{1}\left\{\varphi_t = k, n_k(t) = m\right\} \\
&\leq \sum_{m=1}^{T} \mathbf{1}\left\{g^k_m \geq \mu_k + \frac{2\Delta_k}{3}\right\} \\
&\leq \sum_{m=1}^{T} \mathbf{1}\left\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \geq \frac{2\Delta_k}{3} - (2+\eta)c_m\right\}, \qquad (8)
\end{aligned}$$

where in the last inequality we apply Lemma 2 and use the fact that $u^{1+\epsilon}/B^\epsilon_m \leq c_m$ in (2). We set

$$l_k = \left\lceil \left(\frac{6+3\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}} \ln\left(\frac{T}{K}\left(\frac{\Delta_k}{6+3\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\right\rceil.$$

With $\Delta_k \geq \delta$, we get l_k is no less than

$$\left(\frac{6+3\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}} \ln\left(\frac{T}{K}\left(\frac{\delta}{6+3\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right) = \left(\frac{6+3\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}}.$$

Furthermore, since c_m is monotonically decreasing with m, for $m \geq l_k$,

$$c_m \leq c_{l_k} \leq \left\lceil \frac{\ln_+\left(\frac{T}{K}\left(\frac{\Delta_k}{6+3\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right)}{l_k}\right\rceil^{\frac{\epsilon}{1+\epsilon}} \leq \frac{\Delta_k}{6+3\eta}. \qquad (9)$$

With this result and $l_k \geq 1$, we continue from (8) to get

$$\begin{aligned}
\mathbb{E}\left[(8)\right] \leq& l_k - 1 + \sum_{m=l_k}^{T} \mathbb{P}\left\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \geq \frac{2\Delta_k}{3} - (2+\eta)c_m\right\} \\
\leq& l_k - 1 + \sum_{m=l_k}^{T} \mathbb{P}\left\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \geq \frac{\Delta_k}{3}\right\} \qquad (10)
\end{aligned}$$

Therefore by using Lemma 4 together with statement (ii) from Lemma 3, we get

$$\begin{aligned}
&\sum_{m=l_k}^{T} \mathbb{P}\left\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \geq \frac{\Delta_k}{3}\right\} \\
&\leq \sum_{m=l_k}^{T} \exp\left(-\frac{m\Delta_k}{3B_m}\psi\left(B^\epsilon_m\Delta_k\right)\right) \\
&\leq \sum_{m=l_k}^{T} \exp\left(-\frac{m\Delta_k}{3B_m}\psi\left(6+3\eta\right)\right), \qquad (11)
\end{aligned}$$

where the last step is due to that $\psi(x)$ is monotonically increasing and $B^\epsilon_m\Delta_k \geq (6+3\eta)B^\epsilon_{m_1}c_m \geq 6+3\eta$ from (9) and (2). Since $B_m = \phi\big(h(m)\big)^{-\frac{1}{1+\epsilon}} \leq \phi(am)^{-\frac{1}{1+\epsilon}} \leq (am)^{\frac{1}{1+\epsilon}}$, we have

$$\begin{aligned}
(11) \leq& \sum_{m=1}^{T} \exp\left(-m^{\frac{\epsilon}{1+\epsilon}}a^{-\frac{1}{1+\epsilon}}\psi\left(6+3\eta\right)\frac{\Delta_k}{3}\right). \\
\leq& \int_0^{+\infty} \exp\left(-\beta y^{\frac{\epsilon}{1+\epsilon}}\right) dy \\
&\left(\text{where we set } \beta = a^{-\frac{1}{1+\epsilon}}\psi\left(6+3\eta\right)\Delta_k/3\right) \\
=& \frac{1+\epsilon}{\epsilon}\beta^{-\frac{1+\epsilon}{\epsilon}} \int_0^{+\infty} z^{\frac{1+\epsilon}{\epsilon}-1}\exp\left(-z\right) dy \\
&\left(\text{where } z = \beta y^{\frac{\epsilon}{1+\epsilon}}\right) \\
=& \Gamma\left(\frac{1}{\epsilon}+2\right)\beta^{-\frac{1+\epsilon}{\epsilon}}.
\end{aligned}$$

Plugging it into (10),

$$\mathbb{E}\left[(8)\right] \leq C_2\Delta_k^{-\frac{1+\epsilon}{\epsilon}} + C_3\Delta_k^{-\frac{1+\epsilon}{\epsilon}} \ln\left(\frac{T}{KC_3}\Delta_k^{\frac{1+\epsilon}{\epsilon}}\right)$$

where $C_2 = \Gamma\left(1/\epsilon+2\right)a^{\frac{1}{\epsilon}}\left[3/\psi\left(6+3\eta\right)\right]^{\frac{1+\epsilon}{\epsilon}}$ and $C_3 = (6+3\eta)^{\frac{1+\epsilon}{\epsilon}}$. Put it together with $\Delta_k \geq \delta$ for all $k \in \mathcal{B}_\delta$,

$$\begin{aligned}
\mathbb{E}\left[(7)\right] &\leq \sum_{k\in\mathcal{B}_\delta} C_2\Delta_k^{-\frac{1}{\epsilon}} + C_3\Delta_k^{-\frac{1}{\epsilon}} \ln\left(\frac{T}{KC_3}\Delta_k^{\frac{1+\epsilon}{\epsilon}}\right) \\
&\leq C_2K\delta^{-\frac{1}{\epsilon}} + (1+\epsilon)e^{\frac{-\epsilon}{1+\epsilon}}C_3K\delta^{-\frac{1}{\epsilon}},
\end{aligned}$$

where we use the fact that $x^{-\frac{1}{\epsilon}}\ln\left(Tx^{\frac{1+\epsilon}{\epsilon}}/\left(KC_3\right)\right)$ takes its maximum at $x = \delta\exp(\epsilon^2/(1+\epsilon))$.

Step 4: Plugging the results in step 2 and step 3 into (5),

$$R^{\text{worst}}_T \leq T\delta + \left[C_1 + C_2 + (1+\epsilon)e^{\frac{-\epsilon}{1+\epsilon}}C_3\right]K\delta^{-\frac{1}{\epsilon}} + 2K.$$

Straightforward calculation concludes the proof. ■

C. Distribution-dependent Regret Upper Bound

We now show that robust MOSS also preserves a logarithm upper bound on the distribution-dependent regret.

Theorem 8: For the heavy-tailed stochastic MAB problem with K arms and time horizon T, if $\eta\psi(2\eta/a) \geq 2a$, the regret R_T for Robust MOSS is no greater than

$$\sum_{k:\Delta_k>0}\left(\frac{u^{1+\epsilon}}{\Delta_k}\right)^{\frac{1}{\epsilon}}\left[C_1\ln\left(\frac{T}{KC_1}\left(\frac{\Delta_k}{u}\right)^{\frac{1+\epsilon}{\epsilon}}\right) + C_2K\right] + \Delta_k.$$

where $C_1 = (4+4\eta)^{\frac{1+\epsilon}{\epsilon}}$ and $C_2 = \max\left(eC_1, 2\Gamma(1/\epsilon+2)\left(8a/\psi(2\eta/a)\right)^{\frac{1+\epsilon}{\epsilon}} a/\ln(a)\right)$.

Proof: Let $\delta = (4+4\eta)\left(eK/T\right)^{\frac{\epsilon}{1+\epsilon}}$ and define $\mathcal{B}_\delta$ the same as (4). Since $\Delta_k \leq \delta$ for all $k \notin \mathcal{B}_\delta$, the regret satisfies

$$\begin{aligned}
R_T &\leq \sum_{k\notin\mathcal{B}_\delta} T\Delta_k + \sum_{t=1}^{T}\mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\}\Delta_{\varphi_t} \\
&\leq \sum_{k\notin\mathcal{B}_\delta} eK\left(\frac{4+4\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}}\Delta_k + \sum_{k\in\mathcal{B}_\delta}\sum_{t=1}^{T}\mathbf{1}\{\varphi_t = k\}\Delta_k. \qquad (12)
\end{aligned}$$

Pick arbitrary $l_k \in \mathbb{Z}_+$, thus

$$\sum_{t=1}^{T} \mathbf{1}\{\varphi_t = k\} \leq l_k + \sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t = k, n_k(t) \geq l_k\right\}$$
$$\leq l_k + \sum_{t=K+1}^{T} \mathbf{1}\left\{g^k_{n_k(t)} \geq g^*_{n^*(t)}, n_k(t) \geq l_k\right\}.$$

Observe that $g^k_{n_k(t)} \geq g^*_{n^*(t)}$ implies at least one of the following is true

$$g^*_{n^*(t)} \leq \mu^* - \Delta_k/4, \tag{13}$$
$$g^k_{n_k(t)} \geq \mu_k + \Delta_k/4 + 2(1+\eta)c_{n_k(t)}, \tag{14}$$
$$(1+\eta)c_{n_k(t)} > \Delta_k/4. \tag{15}$$

We select

$$l_k = \left\lceil \left(\frac{4+4\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}} \ln\left(\frac{T}{K}\left(\frac{\Delta_k}{4+4\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\right\rceil.$$

Similarly as (9), $n_k(t) \geq l_k$ indicates $c_{n_k(t)} \leq \Delta_k/(4+4\eta)$, so (15) is false. Then we apply Lemma 5 and Corollary 6,

$$\mathbb{P}\left\{g^k_{n_k(t)} \geq g^*_{n^*(t)}, n_k(t) \geq l_k\right\}$$
$$\leq \mathbb{P}\left((13) \text{ or } (14) \text{ is true }\right) \leq \frac{C_2' K}{T}\Delta_k^{-\frac{1+\epsilon}{\epsilon}},$$

where $C_2' = 2\Gamma\left(1/\epsilon + 2\right)\left(8a/\psi(2\eta/a)\right)^{\frac{1+\epsilon}{\epsilon}} a/\ln(a)$. Substituting it into (12), R_T is upper bounded by

$$\sum_{k \notin \mathcal{B}_\delta} \frac{eC_1 K}{\Delta_k^{\frac{1}{\epsilon}}} + \sum_{k \in \mathcal{B}_\delta}\left[\frac{C_1}{\Delta_k^{\frac{1}{\epsilon}}}\ln\left(\frac{T}{KC_1}\Delta_k^{\frac{1+\epsilon}{\epsilon}}\right) + \frac{C_2' K}{\Delta_k^{\frac{1}{\epsilon}}} + \Delta_k\right].$$

Considering the scaling factor u, the proof can be concluded with easy computation. ■

V. Numerical Illustration

In this section, we compare Robust MOSS with MOSS and Robust UCB (with truncated empirical mean or Catoni's estimator) [10] in a 3-armed heavy-tailed bandit setting. The mean rewards are $\mu_1 = -0.3$, $\mu_2 = 0$ and $\mu_3 = 0.3$ and sampling at each arm k returns a random reward equals to μ_k added by sampling noise ν, where $|\nu|$ is a generalized Pareto random variable and the sign of ν has equal probability to be positive and negative. The PDF of reward at arm k is

$$f_k(x) = \frac{1}{2\sigma}\left(1 + \frac{\xi|x-\mu_k|}{\sigma}\right)^{-\frac{1}{\xi}-1} \text{ for } x \in (-\infty, +\infty),$$

where we select $\xi = 0.33$ and $\sigma = 0.32$. Thus, for a random reward X from any arm, we know $\mathbb{E}[X^2] \leq 1$, which means $\epsilon = 1$ and $u = 1$. We select parameters $a = 1.1$ and $\eta = 2.2$ for Robust MOSS so that condition $\eta\psi(2\eta/a) \geq 2a$ is met.

Fig.1 shows the mean cumulative regret together with quantiles of cumulative regret distribution as a function of time, which are computed using 200 simulations of each policy. The simulation result shows that there is a chance MOSS loses stability in heavy-tailed MAB and suffers linear cumulative regret while other algorithms work consistently and maintain sub-linear cumulative regrets. Robust MOSS slightly outperforms Robust UCB in this specific problem.

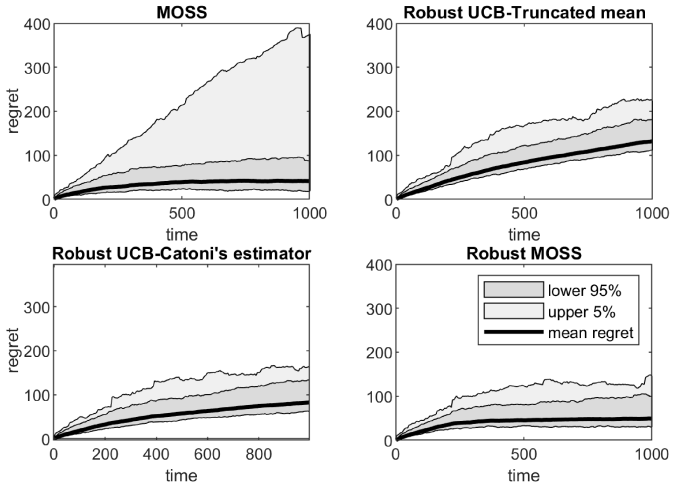

Fig. 1. Comparison of 4 algorithms in heavy-tailed MAB: On each graph, the bold curve is the mean regret while light shaded and dark shaded regions correspond respectively to upper 5% and lower 95% quantile cumulative regrets.

VI. Conclusions and Future Direction

We proposed the Robust MOSS algorithm for heavy-tailed bandit problem. We evaluate the algorithm by deriving upper bounds on the associated distribution-free and distribution-dependent regrets. Our analysis shows that Robust MOSS achieves order optimal performance in both scenarios. The saturated mean estimator centers at zero which make the algorithm not translation invariant. Exploration of translation invariant robust mean estimator in this context remains an open problem.

References

[1] J. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Conference on Learning Theory*, 2009, pp. 217–226.

[2] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[3] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

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

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[6] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *JMLR: Workshop and Conference Proceedings*, vol. 19: COLT 2011, 2011, pp. 359–376.

[7] R. Degenne and V. Perchet, "Anytime optimal algorithms in stochastic multi-armed bandits," in *International Conference on Machine Learning*, 2016, pp. 1587–1595.

[8] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, p. 47, 2002.

[9] M. Vidyasagar, "Law of large numbers, heavy-tailed distributions, and the recent financial crisis," in *Perspectives in Mathematical System Theory, Control, and Signal Processing*. Berlin, Heidelberg: Springer, 2010, pp. 285–295.

[10] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7711–7717, 2013.

[11] X. Fan, I. Grama, and Q. Liu, "Hoeffdings inequality for supermartingales," *Stochastic Processes and their Applications*, vol. 122, no. 10, pp. 3545–3559, 2012.

[12] S. Bubeck, "Bandits games and clustering foundations," Ph.D. dissertation, Université des Sciences et Technologie de Lille - Lille I, 2010.