Bandit algorithms: Letting go of logarithmic regret for statistical robustness

Kumar Ashutosh
IIT Bombay

Jayakrishnan Nair
IIT Bombay

Anmol Kagrecha
IIT Bombay

Krishna Jagannathan
IIT Madras

Abstract

We study regret minimization in a stochastic multi-armed bandit setting, and establish a fundamental trade-off between the regret suffered under an algorithm, and its statistical robustness. Considering broad classes of underlying arms' distributions, we show that bandit learning algorithms with logarithmic regret are *always inconsistent* and that *consistent* learning algorithms always suffer a super-logarithmic regret. This result highlights the inevitable statistical fragility of all 'logarithmic regret' bandit algorithms available in the literature – for instance, if a UCB algorithm designed for σ-subGaussian distributions is used in a subGaussian setting with a mismatched variance parameter, the learning performance could be inconsistent. Next, we show a positive result: statistically robust and consistent learning performance is attainable if we allow the regret to be *slightly worse* than logarithmic. Specifically, we propose three classes of distribution oblivious algorithms that achieve an asymptotic regret that is arbitrarily close to logarithmic.

1 Introduction

The stochastic multi-armed bandit (MAB) problem seeks to select the best among an available basket of options (a.k.a., arms), each characterized by an unknown probability distribution. Classically, these probability distributions represent rewards, and the best arm is defined as the one associated with the largest average reward. The learning algorithm, which chooses (a.k.a., pulls) one arm per decision epoch, identifies the best arm via experimentation—each pull of an arm yields one sample from the underlying reward distribution. One classical performance metric is *regret*, which evaluates an algorithm based on how often it pulls sub-optimal arms.

The standard approach towards algorithm design for regret minimization is as follows. First, it is assumed that the arm reward distributions belong to a specific (semi-)parametric class—for example, the class of bounded distributions with support contained in $[0, b]$, or the class of σ-subGaussians. Next, algorithms are proposed for such specific parametric distribution classes, often making explicit use of the parameters (such as b or σ) corresponding to the parametric distribution class. Finally, logarithmic regret guarantees are proved for such algorithms, by utilizing exponential concentration inequalities (such as Hoeffding's inequality or subGaussian concentration) for that parametric distribution class.

For distribution classes such as σ-subGaussians, a logarithmic regret guarantee may not be so surprising, because such distributions enjoy exponential concentration bounds. On the other hand, when dealing with heavy-tailed arms' distributions, it is not clear that a logarithmic regret is achievable. This is because heavy-tailed distributions (such as Pareto) are characterized by a high degree of variability, and their empirical mean estimators do not enjoy exponential concentration in the sample size. Somewhat surprisingly, a logarithmic regret guarantee was shown to be attainable in Bubeck et al. (2013) using a truncated mean estimator, for distributions satisfying a bounded moment condition. While this approach Bubeck et al. (2013) can handle heavy-tailed as well as light-tailed distributions, the algorithm still needs to know the moment bounds.

As such, a logarithmic regret guarantee has been shown to hold in a broad range of stochastic bandit settings. At this point, it is perhaps not an exaggeration to suggest that a logarithmic regret is regarded as a 'default performance expectation' from 'good' stochastic bandit learning algorithms. The present pa-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

per challenges this perceived sanctity of logarithmic regret, in the context of low-regret learning of stochastic MABs. We show that bandit algorithms that enjoy a logarithmic regret guarantee cannot be statistically robust.

Our contributions: We make two key contributions in this paper.

First, we show that bandit algorithms that enjoy a logarithmic regret guarantee are fundamentally fragile from a statistical standpoint. Equivalently, we show that statistically robust algorithms necessarily incur super-logarithmic regret. Here, an algorithm is said to be statistically robust if it exhibits *consistency*, i.e., the regret scales slower than any power-law, over a suitably broad class of MAB instances.

For example, consider an algorithm with logarithmic regret designed for σ-subGaussian arms. When this algorithm is used in a 'mismatched' bandit instance, say with σ'-subGaussian arms ($\sigma' > \sigma$), the learning performance can be *inconsistent.* That is, the regret suffered by the algorithm in the mismatched instance could have a power-law scaling in the time horizon. This is of practical concern, since the parameters that define the space of arms' distributions (usually in the form of support/moment bounds) are often themselves estimated from limited data samples, and are therefore prone to errors.

Our second contribution is a positive result: we show that statistically robust learning is achievable if we are willing to tolerate a 'slightly-worse-than-logarithmic' regret in the time horizon. Specifically, we propose three classes of algorithms that (i) are *distribution oblivious* (i.e., they require no prior information about the arm distribution parameters), and (ii) incur a regret that is slightly super-logarithmic. The first algorithm class offers this guarantee over subexponential (a.k.a., light-tailed) instances. The latter two are designed to work robustly for general distribution instances (excepting some pathological ones).

In all three algorithms, the asymptotic regret guarantee is controlled by a certain slow-growing scaling function that is used to define confidence bounds. A more slow-growing scaling function makes the regret asymptotically closer to logarithmic, but at the expense of a potential degradation in performance for shorter horizons. Furthermore, the regret for shorter horizon-lengths can be improved by incorporating (noisy) prior information about the reward distributions into the scaling function, without compromising on statistical robustness.

Related literature: There is a vast literature on the regret minimization for the stochastic MAB problem; we refer the reader to the textbook treatments Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2018). However, to the best of our knowledge, the issue of statistical robustness and its connection to logarithmic regret has not been explored before.

We are aware of only two other works that address statistical robustness in the context of bandit algorithms, both of which consider the fixed budget pure exploration setting. For the best arm identification problem, statistically robust algorithms have been demonstrated recently in Kagrecha et al. (2019). For thresholding bandit problem, the algorithm proposed in Locatelli et al. (2016) is *distribution-free*, i.e., the algorithm does not require knowledge of the σ parameter defining the space of σ-subGaussian rewards. Finally, we note that Salomon et al. (2013) prove regret lower bounds under a relaxation of the standard notion of consistency.

The remainder of this paper is organized as follows. We introduce some preliminaries and define the MAB formulation in Section 2. The trade-off between statistical robustness and logarithmic regret is established in Section 3. Our statistically robust algorithms and their performance guarantees are presented in Section 4, and we report the results of some numerical experiments in Section 5. An appendix, containing proofs omitted due to space constraints, as well as other details, are uploaded separately as the 'supplementary material' document.

2 Model and Preliminaries

In this section, we introduce some preliminaries and formally define the MAB formulation.

2.1 Preliminaries

We begin by introducing the classes of reward distributions we will work with in this paper.

- $\mathcal{B}([a,b])$ denotes the set of bounded distributions with support contained in $[a,b]$. The set of all bounded distributions is denoted by $\mathcal{B}$.
- We use $\mathcal{SG}(\sigma)$, for $\sigma > 0$, to denote σ-subGaussian distributions, and $\mathcal{SG}$ to denote all subGaussian distributions.
- We denote $\mathcal{SE}(v,\alpha)$, for $v, \alpha > 0$, to denote the following class of subexponential distributions:

$$\left\{ F : \int e^{\lambda(x-\mu(F))} dF(x) \leq e^{\frac{v^2\lambda^2}{2}} \ \forall\ |\lambda| < \frac{1}{\alpha} \right\}$$

  where $\mu(F)$ denotes the mean of F. The class of all subexponential distributions is denoted by $\mathcal{SE}$.

Distributions in $\mathcal{SE}$ are also commonly referred to as *light-tailed*, and those not in $\mathcal{SE}$ are called *heavy-tailed* (see Foss et al. (2011)).

- For $\epsilon, B > 0$, let $\mathcal{G}(\epsilon, B)$ denote the set of distributions whose $(1+\epsilon)^{th}$ absolute moment is upper bounded by B, i.e.,

$$\mathcal{G}(\epsilon, B) = \left\{ F : \int |x|^{1+\epsilon} dF(x) \leq B \right\}.$$

In the MAB literature, $\mathcal{G}(\epsilon, B)$ is often used as the class of reward distributions in order to allow for heavy-tailed rewards (see, for example, Bubeck et al. (2013); Yu et al. (2018)). Finally, the union of the sets $\mathcal{G}(\epsilon, B)$ over $\epsilon, B > 0$ is denoted by $\mathcal{G}$:

$$\mathcal{G} = \left\{ F : \int |x|^{1+\epsilon} dF(x) < \infty \text{ for some } \epsilon > 0 \right\}.$$

$\mathcal{G}$ is the most general space of reward distributions one can work with in the context of the MAB problem—it contains all light-tailed distributions and most heavy-tailed distributions of interest.

Note that $\mathcal{B} \subset \mathcal{SG} \subset \mathcal{SE} \subset \mathcal{G}$. We also recall the Kullback-Leibler divergence (or relative entropy) between distributions F and F':

$$D(F, F') = \int \log \left(\frac{dF(x)}{dF'(x)} \right) dF(x),$$

where F is absolutely continuous with respect to F'.

Much of the vast literature on MAB problems assumes that the reward distributions lie in specific parametric subsets of $\mathcal{B}$, $\mathcal{SG}$, $\mathcal{SE}$, or $\mathcal{G}$; for example $\mathcal{B}([0,1])$, $\mathcal{SG}(1)$, $\mathcal{G}(1, B)$ etc. Further, the parameter(s) corresponding to these subsets are 'baked' into the algorithms. While this approach guarantees strong performance over the parametric distribution subset under consideration (logarithmic regret, in the classical regret minimization framework), it is highly fragile to uncertainty in these parameters. Indeed, as we demonstrate in Section 3, any algorithm that enjoys logarithmic regret for a parametric subset of a distribution class *must be inconsistent* over the entire distribution class—specifically, when there is a parameter mismatch, the regret suffered could have a power-law scaling in the time horizon. In Section 4, we propose bandit algorithms that are statistically robust but incur (slightly) superlogarithmic regret.

2.2 Problem formulation

Consider a multi-armed bandit (MAB) problem with k arms. Let $\mathcal{M}$ be a distribution class (such as $\mathcal{B}, \mathcal{SG}$ etc.) An instance $\nu = (\nu_i,\ 1 \leq i \leq k)$ of the MAB problem is defined as an element of $\mathcal{M}^k$, where $\nu_i \in \mathcal{M}$ is the distribution corresponding to arm i. Let μ_i denote the mean reward associated with arm i, i.e., μ_i is the expected value of a random variable distributed according to ν_i. An *optimal* arm is an arm that maximizes the mean reward, i.e., one whose mean reward equals $\mu^* = \max_{1 \leq i \leq k} \mu_i$. The *sub-optimality gap* associated with arm i is defined as $\Delta_i := \mu^* - \mu_i$.

In this paper, our goal is to minimize *regret*. Formally, under the a policy (a.k.a., algorithm) π, let $T_i(n)$ denote the number of times i^{th} arm has been pulled after n rounds. The regret $R_n(\pi, \nu)$ associated with the policy π after n rounds is defined as

$$R_n(\pi, \nu) = \sum_{i=1}^{n} \Delta_i \mathbb{E}\left[T_i(n)\right].$$

An algorithm is said to be *consistent* over $\mathcal{M}^k$ if, for all instances $\nu \in \mathcal{M}^k$, the regret satisfies $R_n(\pi, \nu) = o(n^a)$ for all $a > 0$ (see Lattimore and Szepesvári (2018)). For example, an algorithm that guarantees polylogarithmic regret over *all* instances in $\mathcal{M}^k$ is consistent over $\mathcal{M}^k$. On the other hand, if an algorithm suffers $O(n^a)$ regret for some $a > 0$ and *some* instance in $\mathcal{M}^k$, then the algorithm is inconsistent over $\mathcal{M}^k$.

3 Impossibility of logarithmic regret for statistically robust algorithms

In this section, we shed light on a fundamental conflict between logarithmic regret and statistical robustness. Recall that in classical MAB formulations, it is assumed that arm reward distributions lie in, say $\mathcal{B}([0, b])$ or $\mathcal{SG}(\sigma)$. In such cases, algorithms that exploit this parametric information (i.e., the value of b in the former case and the value of σ in the latter) are known that achieve $O(\log(n))$ regret, where n denotes the horizon. The celebrated UCB family of algorithms is a classic example Lattimore and Szepesvári (2018). In this section, we ask the question: Are these algorithms robust with respect to the parametric information 'baked' into them? Our main result of this section answers this question in the negative. Specifically, we show that statistically robust algorithms (i.e., algorithms that maintain consistency over an entire class of distributions) necessarily incur super-logarithmic regret. In other words, algorithms that enjoy a logarithmic regret guarantee over a particular parametric sub-class of reward distributions are *not* statistically robust.

Theorem 1. *Let $\mathcal{M} \in \{\mathcal{B}, \mathcal{SG}, \mathcal{SE}, \mathcal{G}\}$. For any algorithm π that is consistent over $\mathcal{M}^k$, and for any instance $\nu \in \mathcal{M}^k$ with at least one sub-optimal arm (i.e.,*

$\Delta_i > 0$ *for some* i*),*

$$\lim_{n\to\infty} \frac{R_n(\pi,\nu)}{\log(n)} = \infty.$$

The proof of Theorem 1 is provided in Appendix A. The crux of the argument is as follows. Given an MAB instance $\nu \in \mathcal{M}^k$, the expected number of pulls $\mathbb{E}[T_i(n)]$ of any suboptimal arm i over a horizon of n pulls, under any algorithm that is consistent over $\mathcal{M}^k$, is lower bounded as

$$\liminf_{n\to\infty} \frac{\mathbb{E}[T_i(n)]}{\log(n)} \geq \frac{1}{d_i},$$

where $d_i = \inf_{\nu_i' \in \mathcal{M}}\{D(\nu_i,\nu_i') : \mu(\nu_i') > \mu^*(\nu)\}$ (see (Lattimore and Szepesvári, 2018, chap. 16)). Informally, d_i is the smallest perturbation of ν_i in relative entropy sense that would make arm i optimal. The proof of Theorem 1 follows by showing that when $\mathcal{M}$ is $\mathcal{B}$, $\mathcal{SG}$, $\mathcal{SE}$ or $\mathcal{G}$, we have $d_i = 0$ for all suboptimal arms of any instance. In other words, given any distribution $\eta \in \mathcal{M}$, there exists another distribution $\eta' \in \mathcal{M}$ such that $\mu(\eta')$ is arbitrarily large, even while $D(\eta,\eta')$ is arbitrarily small.

Theorem 1 highlights that classical bandit algorithms are not robust with respect to uncertainty in support/moment bounds. For example, consider any algorithm π that guarantees logarithmic regret over $\mathcal{SG}(1)$ (for example, the algorithms presented in Chapters 7–9 in Lattimore and Szepesvári (2018)). Theorem 1 implies that all such algorithms are *inconsistent* over $\mathcal{SG}$. This reveals an inherent fragility of such algorithms—while they might guarantee good performance over the specific parametric sub-class of reward distributions they are designed for, they are not robust to uncertainty with respect to the parameters that specify the distribution class.

Having shown that robust algorithms cannot achieve logarithmic regret, in the following section, we present statistically robust algorithms for $\mathcal{SE}$, and $\mathcal{G}$. (Of course, an algorithm that is robust over $\mathcal{SE}$ is also robust over $\mathcal{B}$ and $\mathcal{SG}$). Specifically, these algorithms attain a regret that is *slightly* super-logarithmic, while remaining consistent over $\mathcal{SE}$ and $\mathcal{G}$ respectively.

4 Statistically robust algorithms

In this section, we demonstrate how statistical robustness can be achieved by allowing for *slightly superlogarithmic* regret. In particular, we propose algorithms that are *distribution oblivious*, i.e., they do not require any prior information about the arm distributions in the form of support/moment/tail bounds. By suitably choosing a certain scaling function that parameterizes the algorithms, the associated regret can be made arbitrarily close to logarithmic (in the time horizon). However, this is not an entirely 'free lunch'—tuning the scaling function for stronger asymptotic regret guarantees can affect the regret for a moderate horizon values. Interestingly though, this trade-off between asymptotic and short-horizon performance can be tempered by incorporating (noisy) prior information about support/moment bounds on the arm distributions into the scaling functions, while maintaining statistical robustness.

We propose three distribution oblivious algorithms for robust regret minimization in this section. The first, which we call *Robust Upper Confidence Bound algorithm for Light-Tailed instances* (R-UCB-LT) algorithm is suitable for subexponential (light-tailed) instances. (An instance is said to be light-tailed if all arm distributions are light-tailed). It uses the empirical average as an estimator for the mean reward, and uses a confidence bound that that is a suitably (and robustly) scaled version of the typical non-oblivious confidence bounds in UCB algorithms.

Next, to deal with the most general class $\mathcal{G}$ of reward distributions, we propose two algorithms. The first is called R-UCB-TEA, where the qualifier TEA stands for the *Truncated Empirical Average* estimator used by the algorithm. Empirical averages, which provide good estimates of the mean for light-tailed arms, can deviate significantly from the true mean for heavy-tailed arms. To control the 'high variability' in the sample values, a truncated mean estimator is typically used; see for example, Bubeck et al. (2013); Yu et al. (2018). The truncation parameter in R-UCB-TEA is scaled with time suitably to provide statistical robustness. The second algorithm, which we call R-UCB-MoM, uses a *Median of Means* (MoM) estimator for the mean of each arm. MoM estimators, as the name suggests, partition the data into disjoint bins, and compute the median of the (empirical average) estimators corresponding to the different bins. This approach has been shown to provide favourable concentration properties for highly variable data samples (Bubeck et al., 2013); R-UCB-MoM uses a number of bins that is scaled logarithmically with time, in conjunction with a robustly scaled confidence interval. Desirably, all our algorithms are *any-time*, and have provable regret guarantees.

Before we describe the algorithms, we define the following class of functions which serve as scaling functions for both algorithms.

Definition 1. *A function* $f : \mathbb{N} \to (0,\infty)$ *is said to be* slow-growing *if* $f(t+1) \geq f(t)\ \forall\ t \in \mathbb{N}$,

$$\lim_{t\to\infty} f(t) = \infty, \text{ and } \lim_{t\to\infty} \frac{f(t)}{t^a} = 0\ \forall\ a > 0.$$

Algorithm 1 R-UCB-LT

Input k arms, slow-growing scaling function f

for $t = 1$ to k **do**

Pull arm with index $i = t$ and observe reward R_t

Update $\hat{\mu}(i, u_i) \leftarrow R_t$, $u_i \leftarrow 1$

end for

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

Calculate the upper confidence bound as

$$\mathcal{U}(i, u_i, t) = \hat{\mu}(i, u_i) + \underbrace{\sqrt{\frac{f(t)\log(t)}{u_i}}}_{\mathcal{W}(u_i, t)}$$

Pull arm i maximizing $\mathcal{U}(i, u_i, t)$ and observe reward R_t

Update empirical average $\hat{\mu}(i, u_i)$ and $u_i \leftarrow u_i + 1$

end for

4.1 Robust Upper Confidence Bound algorithm for Light-Tailed instances

The R-UCB-LT algorithm is presented in Algorithm 1. The only structural difference between R-UCB-LT and the classical UCB algorithm is in the definition of the upper confidence bound—under R-UCB-LT, the confidence width $\mathcal{W}(u_i, t)$ for arm i at time t, where u_i denotes the number of pulls of arm i prior to time t, is scaled by a slow-growing function f. This simple scaling provides statistical robustness over light-tailed instances, as established in Theorem 2 below. We prove the consistency of R-UCB-LT over all subexponential instances, albeit with superlogarithmic regret. We also provide stronger guarantees for subGaussian instances.

Theorem 2. *Consider the algorithm R-UCB-LT with a specified slow-growing scaling function f. For an instance $\nu \in \mathcal{SE}(v, \alpha)^k$, there exists threshold $t^{\mathcal{SE}}_{min}(v, \alpha)$ such that for $t > t^{\mathcal{SE}}_{min}(v, \alpha)$, the regret under R-UCB-LT satisfies*

$$R_t(\nu) \leq \sum_{i:\Delta_i > 0} \left(f(t)\log(t)\ \hat{c} + 4\Delta_i\right). \tag{1}$$

where $\hat{c} = \max\left\{\frac{4}{\Delta_i}, \Delta_i \left(\frac{\alpha}{v^2}\right)^2\right\}$, an instance-dependent constant.

For an instance $\nu \in \mathcal{SG}(\sigma)^k$, there exists a threshold $t^{\mathcal{SG}}_{min}(\sigma)$ such that for $t > t^{\mathcal{SG}}_{min}(\sigma)$, the regret under R-UCB-LT satisfies

$$R_t(\nu) \leq \sum_{i:\Delta_i > 0} \left(\frac{4f(t)\log(t)}{\Delta_i} + 4\Delta_i\right). \tag{2}$$

The key take-aways from Theorem 2 are as follows.

- R-UCB-LT is clearly consistent over $\mathcal{SE}^k$, but the regret guarantee is super-logarithmic, as demanded by Theorem 1.

- R-UCB-LT is distribution oblivious in the sense that it does not need the parameters v, α in the implementation. However, the stated regret guarantee holds for t greater than an instance-dependent threshold $t^{\mathcal{SE}}_{min}(v, \alpha)$—this is because the confidence width needs to be large enough for certain concentration properties to hold. Explicit characterization of the threshold $t^{\mathcal{SE}}_{min}(v, \alpha)$, along with (weaker) regret bounds for t less than this threshold, are provided in Appendix B.

- Choosing f to be 'slower' growing leads to better asymptotic regret guarantees, but also increases the threshold t_{min}. This implies a trade-off between asymptotic and short-horizon performance in a purely oblivious setting. However, (noisy) prior information about the class of arm distributions can be incorporated into the choice of scaling function f to dilute this trade-off. For example, if it is believed that the arm distributions are σ-subGaussian, then one may set $f(t) = 8\sigma^2 + h(t)$, where $h(\cdot)$ is slow-growing; this choice of motivated by the observation that for the well known (non-robust) α-UCB algorithm Bubeck and Cesa-Bianchi (2012), f would be replaced by $2\alpha\sigma^2$, $\alpha > 1$ for σ-subGaussian arms. This choice would make $t^{\mathcal{SG}}_{min}$ small if the arms are σ'-subGaussian, where $\sigma' \approx \sigma$, while still providing statistical robustness to the reliability of this prior information; see proof below. We also illustrate this phenomenon in our numerical experiments in Section 5.

- Stronger performance guarantees are possible for the subclass $\mathcal{SG}^k$. Indeed, given that $\mathcal{SG}(\sigma) \subset \mathcal{SE}(\sigma, \alpha)$ for all $\alpha > 0$, the guarantee (2) is stronger than (1) for $\nu \in \mathcal{SG}(\sigma)^k$.

Proof of Theorem 2

The proof is structurally similar to the proof of the UCB regret bound in Bubeck et al. (2013). We consider the cases $\nu \in \mathcal{SG}^k$ and $\nu \in \mathcal{SE}^k$ separately.

Case 1 $\nu \in \mathcal{SE}^k$

We define the following three events for any suboptimal arm i.

$$\begin{aligned} E_1 &: \quad \mathcal{U}(i^*, T_{i^*}(t-1), t) \leq \mu^* \\ E_2 &: \quad \hat{\mu}(i, T_i(t-1)) > \mu_i + \mathcal{W}(T_i(t-1), t) \\ E_3 &: \quad \Delta_i < 2\mathcal{W}(T_i(t-1), t) \end{aligned}$$

where $T_i(t)$ denotes the number of times i^{th} arm is pulled till time instant t. The three events can be interpreted as follows. Event E_1 occurs when the upper confidence bound corresponding to the optimal arm is less than its actual mean. Event E_2 corresponds to the case when the mean estimator of a sub-optimal arm is much larger than its actual mean. Finally, the event E_3 corresponds to the case when the confidence window of arm i is large. We now prove that one of these events must occur if sub-optimal arm i is chosen at time instant t. Denote I_t as the arm chosen at time t.

Claim If $I_t = i$, then one of E_1, E_2 or E_3 is true.

To justify this claim, we assume all the three events to be false and then show a contradiction. We have,

$$\begin{aligned}\mathcal{U}(i^*, T_{i^*}(t-1), t) &> \mu^* \\ &= \mu_i + \Delta_i \\ &\geq \mu_i + 2\mathcal{W}(T_i(t-1), t) \\ &\geq \hat{\mu}(i, T_i(t-1)) + \mathcal{W}(T_i(t-1), t) \\ &= \mathcal{U}(i, T_i(t-1), t)\end{aligned}$$

which is a contradiction since $I_t \neq i^*$.

Next, we derive a concentration inequality corresponding to the confidence bound used in R-UCB-LT. This inequality will be used to upper bound the probability of events E_1 and E_2. By our choice of algorithm

$$\hat{\mu}(i,u) = \frac{1}{u}\sum_{j=1}^{u} X_j \ ; \quad \mathcal{W}(u,t) = \sqrt{\frac{f(t)\log(t)}{u}}$$

We assume the underlying distribution to be $\nu \in \mathcal{SE}(v,\alpha)$. For any confidence width $\mathcal{W}$, we have the following concentration inequality (see Equation (2.18) in Wainwright (2019)):

$$\mathbb{P}\left(\frac{1}{u}\sum_{j=1}^{u} X_j - \mu \geq \mathcal{W}\right) \leq e^{-\frac{u\mathcal{W}}{2}\min\left\{\frac{\mathcal{W}}{v^2}, \frac{1}{\alpha}\right\}}$$

We are interested only in small values of the confidence window $\mathcal{W}$, and hence the first term in the minimum expression is of interest to us. For the first term to be less than the second term, we have the inequality $\mathcal{W} \leq \frac{v^2}{\alpha}$. Putting the value of confidence window $\mathcal{W}(u,t)$ in this inequality, we get, $u \geq f(t)\log(t)\left(\frac{\alpha}{v^2}\right)^2$. Denote the minimum u satisfying this inequality as u_0. Hence for all $u > u_0$ we have,

$$\mathbb{P}\left(\hat{\mu}(i^*,u) + \mathcal{W}(u,t) > \mu^*\right) \leq \exp\left(\frac{-f(t)\log(t)}{2v^2}\right).$$

Since $f(t)$ is a sub-linearly growing function, for all time $t > t_0$, we are guaranteed to have $f(t) > 8v^2$, where $t_0 = f^{-1}(8v^2)$. Substituting this inequality in the above expression yields,

$$\mathbb{P}\left(\hat{\mu}(i^*,u) + \mathcal{W}(u,t) > \mu^*\right) \leq \exp\left(-4\log(t)\right) = t^{-4}$$

for all time instances $t > t_0$ and $u > u_0$. In addition, u_0 is an increasing function with number of rounds t. This inequality is useful in establishing an upper bound on the probability of events E_1 and E_2. Next we have, by union bound over u,

$$\mathbb{P}(E_1) \leq \mathbb{P}(\exists u \in [t] : \mathcal{U}(i^*,u,t) \leq \mu^*) \leq t.t^{-4} = t^{-3}.$$

Similarly, $\mathbb{P}(E_2) \leq t^{-3}$. Let u_i' denote the maximum value of $T_i(t-1)$ for which event E_3 is true. Consequently, for all $t > u_i'$ and $u > u_0$, if $I_t = i$, then at least one of the event E_1, E_2 is true. Finally, we choose $u_i = \max(u_i', u_0, t_0)$ since we wish to apply the above concentration inequality for all time instances $t > u_i$. Now, for any sub-optimal arm i,

$$\begin{aligned}\mathbb{E}[T_i(t)] &= \mathbb{E}\left[\sum_{s=1}^{t} \mathbb{1}\{I_t = i\}\right] \\ &\leq u_i + \mathbb{E}\left[\sum_{s=u_i+1}^{t} \mathbb{1}\{I_t = i\}\right] \\ &= u_i + \mathbb{E}\left[\sum_{s=u_i+1}^{t} \mathbb{1}\{I_t = i, E_1 \text{ true or } E_2 \text{ true}\}\right] \\ &\leq u_i + \sum_{s=u_i+1}^{t} \mathbb{P}(E_1 \cup E_2) \\ &\leq u_i + \sum_{s=u_i+1}^{t} \frac{2}{s^3} \leq u_i + 4.\end{aligned}$$

Evaluating the value of u_i, we get

$$u_i = \max\left\{\frac{4f(t)\log(t)}{\Delta_i^2}, f(t)\log(t)\left(\frac{\alpha}{v^2}\right)^2, t_0\right\}$$

However, we observe that t_0 is a constant and thus the first two terms (u_i', u_0) will be more than t_0 after a time instance, say t_1. Hence $\forall t > t_{min}^{\mathcal{SE}}(\nu)$,

$$\mathbb{E}[T_i(t)] \leq \max\left\{\frac{4f(t)\log(t)}{\Delta_i^2}, f(t)\log(t)\left(\frac{\alpha}{v^2}\right)^2\right\} + 4$$

where the instance dependent $t_{min}^{\mathcal{SE}}(\nu) = \max(t_0, t_1)$.

Thus, we get the regret upper bound as

$$R_t(\nu) \leq \sum_{i:\Delta_i > 0} \left(f(t)\log(t)\ \hat{c} + 4\Delta_i\right)$$

$\forall t > t_{min}^{\mathcal{SE}}(\nu)$, where $\hat{c} = \max\left\{\frac{4}{\Delta_i}, \Delta_i\left(\frac{\alpha}{v^2}\right)^2\right\}$.

Case 2 $\nu \in \mathcal{SG}^k$

Algorithm 2 R-UCB-TEA

Input k arms, slow-growing scaling function f taking values in $(1, \infty)$

Initialize $\mathcal{R}_i = \{\ \}$, $u_i = 0$ for all arm i
for $t = 1$ to k **do**
  pull arm with index $i = t$ and observe reward r
  Append r to $\mathcal{R}_i$ and update $u_i \leftarrow u_i + 1$
end for
for $t = k+1, k+2, \ldots$ **do**
  Calculate the upper confidence bound as

$$\mathcal{U}(i, u_i, t) = \underbrace{\frac{1}{u_i} \sum_{X \in \mathcal{R}_i} X \mathbb{1}_{\{|X| \le f(t)\}}}_{\hat{\mu}(i, u_i, t)} + \underbrace{\frac{1}{\log(f(t))} + \frac{16 f(t) \log(t)}{u_i}}_{\mathcal{W}(u_i, t)}$$

  Pull arm i maximizing $\mathcal{U}(i, u_i, t)$ and observe reward R_t
  Append R_t to $\mathcal{R}_i$ and update $u_i \leftarrow u_i + 1$
end for

We observe that, $\mathcal{SG}$ is a special case of $\mathcal{SE}$ with $\alpha \to 0$. And hence, the regret expression can be obtained as

$$R_t(\nu) \le \sum_{i:\Delta_i > 0} \left(\frac{4 f(t) \log(t)}{\Delta_i} + 4\Delta_i \right) \quad \forall t > t^{\mathcal{SG}}_{min}(\nu)$$

where the instance dependent threshold $t^{\mathcal{SG}}_{min}$ is $\max(t_0, t_1)$; t_0 and t_1 same as the previous case. □

4.2 Robust Upper Confidence Bound algorithm for arbitrary instances using Truncated Empirical Average estimators

The R-UCB-LT algorithm discussed above is robust to parametric uncertainties, and guarantees 'slightly-worse-than-logarithmic' regret for any light-tailed bandit instance. However, one could argue that R-UCB-LT is still not *truly* robust—after all, how can we be certain in a practical scenario that there are no heavy-tailed arms involved? From the viewpoint of applications such as financial portfolios and insurance, heavy-tailed distributions are ubiquitously used in modelling. Therefore there is a compelling case for handling heavy-tailed as well as light-tailed arms' distributions within a common, statistically robust framework.

In this section, we propose a truly robust algorithm for the most general setting, i.e., for bandit instances in $\mathcal{G}^k$. We recall that the class $\mathcal{G}$ demands only the boundedness of the $(1+\epsilon)$-moment for some $\epsilon > 0$. This is only mildly more demanding than the finiteness of the mean,[1] which is necessary for the MAB problem to be well-posed.

Once the restriction to light-tailed reward distributions is removed, more sophisticated estimators than empirical averages are required; this is because empirical averages are highly sensitive to (relatively frequent) outliers in heavy-tailed data. One such approach is to use truncation-based estimators (see, for example, Bubeck et al. (2013)), which offer lower variability at the expense of a (controllable) bias. The R-UCB-TEA algorithm, stated formally as Algorithm 2, uses a truncation-based estimator in conjunction with a robust scaling of the confidence bound. Note that the same scaling function f is used for both truncation as well for scaling the confidence bound.

R-UCB-TEA provides the following performance guarantee over instances in $\mathcal{G}^k$. To the best of our knowledge, this is the first time a single algorithm has been shown to provide provable regret guarantees in such generality.

Theorem 3. *Consider the algorithm R-UCB-TEA with a specified slow-growing scaling function f taking values in $(1, \infty)$. For an instance $\nu \in \mathcal{G}(\epsilon, B)^k$, there exists a threshold $t_{min}(\epsilon, B)$ such that for $t > t_{min}(\epsilon, B)$, the regret under R-UCB-TEA satisfies*

$$R_t(\nu) \le \sum_{i:\Delta_i > 0} \left(\frac{32 f(t) \log(t)}{1 - \frac{2}{\Delta_i \log(f(t))}} + 4\Delta_i \right).$$

The performance guarantee of R-UCB-TEA is structurally similar to that for R-UCB-LT: The algorithm is consistent, with a super-logarithmic regret that is dictated by the growth of the scaling function f. Moreover, while slowing the growth of f improves the asymptotic regret guarantee, it causes t_{min} to increase, potentially compromising the performance for shorter horizons. As before, prior information on, say, moment bounds satisfied by the arm distributions can be incorporated into the design of f. For example, if it is believed that $\nu \in \mathcal{G}(\epsilon, B)$, a natural choice of f would be $f(t) = c + h(t)$, where $h(\cdot)$ is a slow-growing function, and $c > 1$ is the smallest constant satisfying: $\log(x) \le x^{\epsilon}/3B$ for all $x \ge c$; this choice would make t_{min} close to zero for instances in $\mathcal{G}(\epsilon', B')^k$, for $\epsilon' \approx \epsilon$, $B' \approx B$ (see Appendix C). The proof of Theorem 3 is similar to the proof of Theorem 2 and is provided in Appendix C.

[1] Distributions with finite mean that do not belong to $\mathcal{G}$ are quite pathological, and are of little practical interest.

Algorithm 3 R-UCB-MoM

Input k arms, slow-growing scaling function f, slow-decaying function g

Initialize $\mathcal{R}_i = \{\ \}$, $u_i = 0$ for all arm i
for $t = 1$ to k **do**
  pull arm with index $i = t$ and observe reward R_t
  Append R_t to $\mathcal{R}_i$ and update $u_i \leftarrow u_i + 1$
end for
for $t = k+1, k+2, \ldots$ **do**
  Calculate mean estimator $\hat{\mu}(i, u, t)$ using algorithm 4 with input $\mathcal{R}_i$, u_i and t
  Calculate the upper confidence bound as

$$\mathcal{U}(i, u_i, t) = \hat{\mu}(i, u, t) + \underbrace{f(t)\left(\frac{32\log(t)}{u}\right)^{g(t)}}_{\mathcal{W}(u_i, t)}$$

  Pull arm i maximizing $\mathcal{U}(i, u_i, t)$ and observe reward R_t
  Append R_t to $\mathcal{R}_i$ and update $u_i \leftarrow u_i + 1$
end for

Algorithm 4 Function to Calculate Median of Means (MoM)

Input $\mathcal{R}$, u, t.
if $u > 32\log(t)$ **then**:
  Take $q = \lceil 32\log(t) \rceil$ and $N = \lfloor \frac{u}{q} \rfloor$
  Compute $\hat{\mu}_l = \frac{1}{N}\sum_{m=1}^{N} R_{\{(l-1)N+m\}}$ for $l = 1, 2, \ldots, q$
  return median$(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_q)$
else:
  return median$(\mathcal{R})$
end if

4.3 Robust Upper Confidence Bound algorithm for arbitrary instances using Median of Means (MoM) estimator

Next, we present another statistically robust algorithm over $\mathcal{G}^k$. In place of the truncation-based estimator used by R-UCB-TEA, this algorithm, called R-UCB-MoM, uses a *median of means* (MoM) estimator (see Bubeck et al. (2013)). The MoM estimator works as follows. The samples (corresponding to each arm) are first divided into q bins each having an equal number of samples. The empirical mean is calculated for each of the bins, and the MoM estimator is the median of the q intra-bin estimates. Note that median computation provides robustness to extreme values (outliers) that are inherent in heavy-tailed data samples.

In addition to slow-growing scaling functions (see Definition 1), R-UCB-MoM uses another class of scaling functions, which is defined below.

Definition 2. *A function $g : \mathbb{N} \to (0, \infty)$ is said to be* slow-decaying *if*

$$g(t+1) \leq g(t)\ \forall\ t \in \mathbb{N},$$

$$\lim_{t\to\infty} g(t) = 0, \quad \lim_{t\to\infty} \frac{g(t)}{t^a} = 0\ \forall\ a > 0.$$

R-UCB-MoM, stated formally as Algorithm 3, provides the following regret guarantee over instances in $\mathcal{G}^k$.

Theorem 4. *Consider the algorithm R-UCB-MoM with a specified slow-growing scaling function f and slow-decaying scaling function g. For an instance $\nu \in \mathcal{G}(\epsilon, B)^k$, there exists a threshold $t_{min}(\epsilon, B)$ such that for $t > t_{min}(\epsilon, B)$, the regret under R-UCB-MoM satisfies*

$$R_t(\nu) \leq \sum_{i:\Delta_i > 0} \left(\Delta_i \left(\frac{2f(t)}{\Delta_i} \right)^{\frac{1}{g(t)}} 32\log(t) + 4\Delta_i \right).$$

The proof of Theorem 4 is given in Appendix D. It may not be immediately apparent from the statement of Theorem 4 how super-logarithmic the regret bound is. As we show in the following corollary, the regret bound can be made arbitrarily close to logarithmic by suitably choosing the scaling functions f and g.

Corollary 1. *For any slow-growing function $\Phi(t)$, there exists a slow-growing function $f(t)$, a slow-decaying function $g(t)$, and t_{min} such that for all i,*

$$\left(\frac{2f(t)}{\Delta_i} \right)^{\frac{1}{g(t)}} \leq \Phi(t)$$

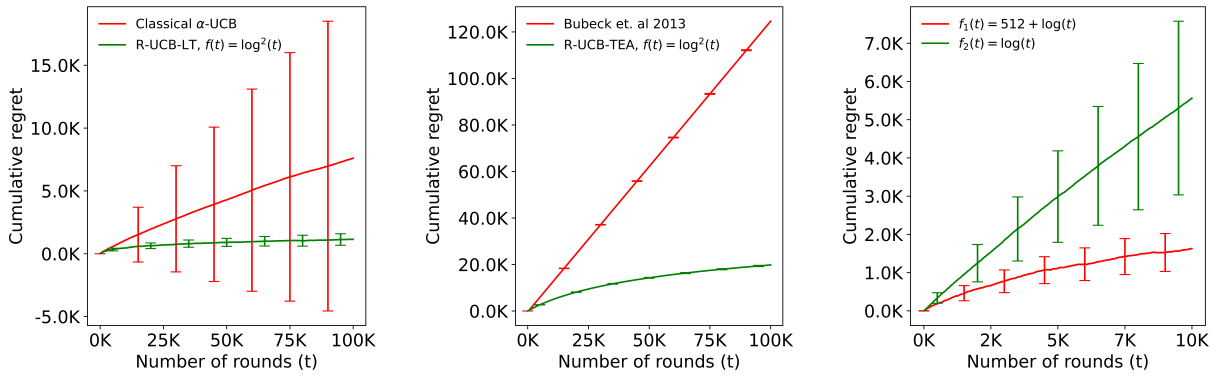
for $t > t_{min}$.

Basically, f and g can be defined in terms of the function Φ; the details can be found in Appendix D.1. As with R-UCB-TEA, configuring the scaling functions for a stronger asymptotic regret guarantee would increase the value of t_{min}, potentially compromising performance for shorter horizons. Also as before, (noisy) prior information about the arm distributions can be used to dilute this tradeoff; the details are omitted due to space constraints.

5 Experimental Analysis

In this section, we present numerical results to illustrate the performance of the algorithms presented in Section 4. Due to space constraints, we only consider R-UCB-LT and R-UCB-TEA here.

In the first experiment, we compare the proposed algorithms with standard algorithms which assume the



(a) R-UCB-LT: Comparison with standard algorithms

(b) R-UCB-TEA: Comparison with standard algorithms

(c) Using prior information to improve short-horizon regret

knowledge of distribution parameters. As per Theorem 1, we established that a statistically robust algorithm cannot have logarithmic regret. We demonstrate that when there is a parameter misestimation, the regret of the standard algorithms can be considerable. Here, we compare R-UCB-LT with standard $\alpha-$UCB in Bubeck and Cesa-Bianchi (2012) and R-UCB-TEA with heavy-tailed algorithm presented in Bubeck et al. (2013). The chosen instance for the first case is as follows: two arms both distributed as Gaussian $\mathcal{N}(\mu, \sigma)$ with parameters $(1, 1)$ and $(2, 3)$. This choice of parameters is arbitrary and a similar trend was observed in trials with other Gaussian instances. For this light-tailed instance, we compare the performance of R-UCB-LT with α-UCB, where α is 1.1 and the algorithm misestimates the instance as being 1-subGaussian. The plot of mean and standard deviation of regret, averaged over 200 instances, is given in Figure 1a. Next, for the heavy-tail case the chosen instance is as follows: two arms both distributed as Pareto Type I with (scale, shape) as (4,5) and (5,5). The specialized algorithm in Bubeck et al. (2013) assumes (incorrectly) that $\mathbb{E}[X^5] < 50$ (such moment misestimations are very common with heavy-tailed data). The obtained regret, averaged over 200 instances, is given in Figure 1b. Note that the choice of $f(t) = \log^2(t)$ is also arbitrary and an even lower regret can be obtained using slower scaling functions, for example, $\log(t)$ or $\log(\log(t))$.

In the second experiment, we demonstrate how choosing $f(t)$ based on (noisy) prior information can decrease regret over short horizons. The chosen instance for this experiment is as follows: two arms both distributed as Gaussian $\mathcal{N}(\mu, \sigma)$ with parameters $(0, 1)$ and $(1, 10)$. Now, suppose we have the (noisy) prior information that the arms are σ-subGaussian with $\sigma \approx 8$. As stated in Section 4, we incorporate this prior information into the design of $f(t)$ by choosing $f_1(t) = 512 + \log(t)$. We compare the cumulative regret for this choice with that corresponding to a completely oblivious choice of $f(t)$, i.e., $f_2(t) = \log(t)$. The experiment is repeated 200 times and obtained mean and standard deviation of regret is shown in Figure 1c. We can see that $f_1(t)$, i.e., the scaling function chosen based on the prior information, incurs lower regret. This trend in cumulative regret can be reasoned as follows. The algorithm using scaling function $f_2(t)$ uses smaller confidence widths, which results in greater susceptibility to the noise in the arm rewards. In conclusion, if noisy prior information about the possible arm distributions is available, this can be incorporated into the choice of the scaling function to improve short-horizon performance, while retaining statistical robustness.

6 Concluding remarks

In this paper, we demonstrated the fundamental trade-off between logarithmic regret and statistical robustness in stochastic MABs. We also proposed robust algorithms that incur slightly super-logarithmic regret. It would be interesting to explore similar trade-offs between statistical robustness and performance in other bandit settings, including thresholding bandits Locatelli et al. (2016), linear bandits Rusmevichientong and Tsitsiklis (2010) and combinatorial bandits Chen et al. (2013).

More broadly, we hope that this paper spawns further work on statistically robust online learning algorithms. We have focussed on one of the simplest learning paradigms (regret minimization in MABs), where a logarithmic regret emerged as a robustly unattainable performance barrier. Other fundamental performance barriers of statistically robust learning await discovery, in more challenging settings such as Markovian bandits and Markov Decision Processes.

References

Agrawal, S., Juneja, S., and Glynn, P. (2020). Optimal δ-correct best-arm selection for heavy-tailed distributions. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 61–110, USA. PMLR.

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

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159.

Foss, S., Korshunov, D., Zachary, S., et al. (2011). *An introduction to heavy-tailed and subexponential distributions*. Springer.

Kagrecha, A., Nair, J., and Jagannathan, K. (2019). Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems*, pages 11272–11281.

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. *preprint*, page 28.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

Salomon, A., Audibert, J.-Y., and Alaoui, I. E. (2013). Lower bounds and selectivity of weak-consistent policies in stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 14(Jan):187–207.

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. (2012). PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Yu, X., Shao, H., Lyu, M. R., and King, I. (2018). Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *UAI*.

A Appendix for Section 3 - Impossibility of logarithmic regret for statistically robust algorithms

This section is devoted to the proof of Theorem 1. The proof is based on the following characterization of instance-dependent lower bounds from Lattimore and Szepesvári (2018) (see Theorem 16.2):

Theorem 5. *For any algorithm π that is consistent over $\mathcal{M}^k$, and instance $\nu \in \mathcal{M}^k$,*

$$\liminf_{n\to\infty} \frac{R_n(\pi,\nu)}{\log(n)} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{d_i(\nu_i,\mu^*,\mathcal{M})},$$

where $d_i(\nu_i,\mu^*,\mathcal{M}) := \inf_{\nu_i'\in\mathcal{M}}\{D(\nu_i,\nu_i'):\ \mu(\nu_i')>\mu^*\}$.

The proof of Theorem 1 therefore follows from the following lemma, which shows that $d_i(\nu_i,\mu^*,\mathcal{M})=0$ for all suboptimal arms of any instance ν when $\mathcal{M}$ is $\mathcal{B}$, $\mathcal{SG}$, $\mathcal{SE}$, or $\mathcal{G}$.

Lemma 1. *Fix $\mathcal{M}\in\{\mathcal{B},\mathcal{SG},\mathcal{SE},\mathcal{G}\}$. For any distribution $F\in\mathcal{M}$, and for any $a>0$ and $b>\mu(F)$, there exists distribution $F'\in\mathcal{M}$ such that*

$$D(F,F')\leq a \quad \text{and} \quad \mu(F')\geq b.$$

Proof. We consider the following two cases.

Case 1: $\mathcal{M}\in\{\mathcal{SG},\mathcal{SE},\mathcal{G}\}$

If the distribution F is unbounded from above (i.e., $\bar{F}(y)>0$ for all $y\in\mathbb{R}$), then the claim follows from Lemma 1 in Agrawal et al. (2020). The idea there is to construct a new distribution F' such that for a chosen y, the CDF on the left side is decreased by a factor of e^{-a} with respect to F, and rest of the mass is pushed on the right side of y. Crucially, under this perturbation, F' remains in $\mathcal{M}$, since on both sides of y only a constant is being multiplied, thus keeping the functional form of the distribution same. The KL-divergence $D(F,F')$ is always less than a independent of the choice of y. However, the mean of F' can be made arbitrary large by choosing a suitably large value of y.

On the other hand, if F is bounded from above, then the argument below (for the case $\mathcal{M}=\mathcal{B}$) can be applied to construct F' that is also bounded from above, but satisfies the conditions required. (Specifically, the boundedness of the lower end-point of the support is not required for this argument.)

Case 2: $\mathcal{M}=\mathcal{B}$

We construct a new bounded distribution F' such that the CDF of F' is e^{-a} times the CDF of F over its support. The rest of the probability mass is uniformly distributed starting from the right end-point of the support to an arbitrary point v'.

Suppose that the support of F is contained within $[u,v]$. Define the CDF of distribution F' as follows, for $\gamma\in(0,1)$ and $v'>v$.

$$F'(x) = (1-\gamma)F(x) \qquad \forall\ x\leq v$$
$$F'(x) = 1+\gamma\frac{x-v'}{v'-v} \qquad \forall\ x\in(v,v']$$

Now,

$$D(F,F') = \int_u^v \log\left(\frac{dF(x)}{dF'(x)}\right)dF(x) = -\log(1-\gamma).$$

Choosing $\gamma=1-e^{-a}$ yields $D(F,F')=a$. Turning now to the mean of F',

$$\begin{aligned}\mu(F') &= \int_u^{v'} x dF'(x) = (1-\gamma)\mu(F) + \int_v^{v'} x\frac{\gamma}{v'-v}dx \\ &= (1-\gamma)\mu(F) + \frac{\gamma}{2}(v'+v)\end{aligned}$$

Clearly, $\mu(F')$ can be made arbitrarily large by choosing a suitably large v'.

□

B Regret bounds when $t < t_{min}$

We discuss a weaker regret bound for time instances less than the threshold time t_{min}. In the proof of theorem 2, we use a slow-growing scaling function to make the inequality oblivious to its parameters. However, we are also interested in obtaining a regret bound for $t < t_{min}$. We have,

$$\mathbb{P}\left(\hat{\mu}(i^*, u) + \mathcal{W}(u, t) > \mu^*\right) \leq \exp\left(-\hat{c} f(t) \log(t)\right)$$

where

$$\hat{c} = \begin{cases} \frac{2}{(b-a)^2}, & \text{if } \nu \in \mathcal{B}^k \\ \frac{1}{2\sigma^2}, & \text{if } \nu \in \mathcal{SG}^k \\ \frac{1}{2v^2}, & \text{if } \nu \in \mathcal{SE}^k \end{cases}$$

Substituting this weaker concentration bound in the above proof of regret bound we get,

$$\mathbb{E}[T_i(t)] \leq u_i + \sum_{s=u_i+1}^{t} t^{1-\hat{c} f(t) \log(t)}$$

as the expected number of times a sub-optimal arm is pulled. The above expression for $\mathbb{E}[T_i(t)]$ still yields a *sub-linear* upper bound, though weaker than before.

C Proof of Theorem 3 - Regret Upper Bound for R-UCB-G

We prove Theorem 3 in this section. This proof is similar to proof of Theorem 2 given in Section 4.1.

Proof. We define the following three events for any sub-optimal arm i.

$$\begin{aligned} E_1 &: \quad \mathcal{U}(i^*, T_{i^*}(t-1), t) \leq \mu^* \\ E_2 &: \quad \hat{\mu}(i, T_i(t-1), t) > \mu_i + \mathcal{W}(T_i(t-1), t) \\ E_3 &: \quad \Delta_i < 2\mathcal{W}(T_i(t-1), t) \end{aligned}$$

where $T_i(t)$ denotes the number of times i^{th} arm is pulled till time instant t. The three events can be interpreted as follows. Event E_1 occurs when the upper confidence bound corresponding to the optimal arm is less than its actual mean. Event E_2 corresponds to the case when the mean estimator of a sub-optimal arm is much more than its actual mean. As we shall see, both E_1 and E_2 are low-probability event and its probability can be upper bounded. Finally, event E_3 corresponds to the case when the confidence window of arm i is large. We now prove that one of these event must be true when a sub-optimal arm is chosen at time instant t. Denote I_t as the arm chosen at time t.

Claim If $I_t = i$, then one of E_1, E_2 or E_3 is true.

To justify this claim, we assume all the three events to be false and then show a contradiction.

We have,

$$\begin{aligned} \mathcal{U}(i^*, T_{i^*}(t-1), t) &> \mu^* \\ &= \mu_i + \Delta_i \\ &\geq \mu_i + 2\mathcal{W}(T_i(t-1), t) \\ &\geq \hat{\mu}(i, T_i(t-1), t) + \mathcal{W}(T_i(t-1), t) \\ &= \mathcal{U}(i, T_i(t-1), t) \end{aligned}$$

which is a contradiction since $I_t \neq i^*$.

Now, by our choice of algorithm

$$\hat{\mu}(i,u,t) = \frac{1}{u}\sum_{j=1}^{u} X_j \mathbb{1}_{\{|X_j| \le f(t)\}}$$

We attempt to establish a distribution oblivious concentration inequality with mean estimator chosen as $\hat{\mu}(i,u,t)$. We draw inspiration from already established non-oblivious concentration inequality based on this mean estimator (see Lemma 1 in Bubeck et al. (2013), Lemma 1 in Yu et al. (2018) which uses results from Seldin et al. (2012)).

We assume the underlying instance to be in $G(\epsilon, B)^k$. For a truncation parameter $f(t)$, we have, with a probability at least $1 - t^{-4}$

$$\begin{aligned}\mu - \hat{\mu}(i,u,t) &\le \frac{B}{f(t)^\epsilon} + \frac{1}{u}\left(2f(t)\log(2t^4) + u\frac{B}{2f(t)^\epsilon}\right) \\ &\le \frac{3B}{2f(t)^\epsilon} + \frac{16f(t)\log(t)}{u}\end{aligned}$$

The second term on the RHS comes from Equation (8) in Yu et al. (2018) and the first term is from Equation (7) in the same paper. Now, the only non-obliviousness is due to the first term. We observe that, for all $t > t_0$, $3B\log(f(t)) < 2f(t)^\epsilon$. There always exists t_0 such that this is true, since, left hand side is a sub-linear term, while right hand side is not.

For all $t > t_0$, with a probability at least $1 - t^{-4}$

$$\mu - \hat{\mu}(i,u,t) \le \frac{1}{\log(f(t))} + \frac{16f(t)\log(t)}{u}$$

$$\Rightarrow \mathbb{P}\left(\mu - \hat{\mu}(i,u,t) \ge \mathcal{W}(u,t)\right) \le t^{-4}$$

This expression establishes a distribution oblivious inequality for a general (even heavy-tailed) random variables. This inequality is valid for all time instances $t > t_0$, where t_0 is a distribution dependent constant parameter.

This inequality is useful in establishing an upper bound on the probability of events E_1 and E_2, similar to case 1 in the proof given in Section 4.1. We have,

$$\mathbb{P}(E_1) \le \mathbb{P}(\exists u \in [t] : \mathcal{U}(i^*,u,t) \le \mu^*) \le t.t^{-4} = t^{-3} \text{ by union bound over } u$$

Similarly, $\mathbb{P}(E_2) < t^{-3}$.

Now, we proceed to obtain regret upper bound similar to case 1 in the proof given in Section 4.1. We define u_i' as the maximum value of $T_i(t-1)$ for which event E_3 is true. Also, we wish to apply concentration bound for all time instants $t > u_i$. Consequently, we choose $u_i = \max(u_i', t_0)$.

Similar to the previous case, we get,

$$\mathbb{E}[T_i(t)] \le u_i + 4$$

The value of u_i can be evaluated from the inequality given in event E_3 and the choice of $\mathcal{W}(u,t)$. We get,

$$u_i = \max\left\{\frac{32f(t)\log(t)}{\Delta_i - \frac{2}{\log(f(t))}}, t_0\right\}$$

.

However, the above calculated value of u_i' is valid only when

$$\Delta_i - \frac{2}{\log(f(t))} > 0$$

Let t_1 denote the minimum value of t satisfying the equation above. Moreover, we observe that t_0 is a constant and thus the first term in the expression of u_i will be more than t_0 after a time instance, say t_2. Hence,

$$\mathbb{E}[T_i(t)] \leq \frac{32f(t)\log(t)}{\Delta_i - \frac{2}{\log(f(t))}} \quad \forall t > t_{min}(\nu)$$

where the instance dependent threshold $t_{min} = \max(t_0, t_1, t_2)$.

Thus, we get the regret upper bound as

$$\boxed{R_t(\nu) \leq \sum_{i:\Delta_i>0} \left(\frac{32f(t)\log(t)}{1 - \frac{2}{\Delta_i \log(f(t))}} + 4\Delta_i \right) \quad \forall t > t_{min}(\nu)}$$

□

D Proof of Theorem 4 – Regret Upper Bound for R-UCB-MoM

The proof of Theorem 4 is similar to the proof of theorem 2 presented in Section 4.1.

Proof. We define the following three events for any sub-optimal arm i.

$$\begin{aligned} E_1 &: \quad \mathcal{U}(i^*, T_{i^*}(t-1), t) \leq \mu^* \\ E_2 &: \quad \hat{\mu}(i, T_i(t-1), t) > \mu_i + \mathcal{W}(T_i(t-1), t) \\ E_3 &: \quad \Delta_i < 2\mathcal{W}(T_i(t-1), t) \end{aligned}$$

where $T_i(t)$ denotes the number of times i^{th} arm is pulled till time instant t. The three events can be interpreted as follows. Event E_1 occurs when the upper confidence bound corresponding to the optimal arm is less than its actual mean. Event E_2 corresponds to the case when the mean estimator of a sub-optimal arm is much more than its actual mean. As we shall see, both E_1 and E_2 are low-probability event and its probability can be upper bounded. Finally, event E_3 corresponds to the case when the confidence window of arm i is large. We now prove that one of these event must be true when a sub-optimal arm is chosen at time instant t. Denote I_t as the arm chosen at time t.

Claim If $I_t = i$, then one of E_1, E_2 or E_3 is true.

To justify this claim, we assume all the three events to be false and then show a contradiction.

We have,

$$\begin{aligned} \mathcal{U}(i^*, T_{i^*}(t-1), t) &> \mu^* \\ &= \mu_i + \Delta_i \\ &\geq \mu_i + 2\mathcal{W}(T_i(t-1), t) \\ &\geq \hat{\mu}(i, T_i(t-1), t) + \mathcal{W}(T_i(t-1), t) \\ &= \mathcal{U}(i, T_i(t-1), t) \end{aligned}$$

which is a contradiction since $I_t \neq i^*$.

Now, by our choice of algorithm $\hat{\mu}(i, u, t)$ is the *median of means* estimator. In this mean estimator, we first divide the samples into q bins, and compute the average of all the bins. Each bin will have $N = \lceil \frac{u}{q} \rceil$ samples.

We return the median of these q bins as the mean estimator. We attempt to establish a distribution oblivious concentration inequality for this mean estimator. Formally, this estimator is defined as

$$\hat{\mu}(i,u,t) = \text{median}(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_q) \quad \text{where } q = \lceil 32\log(t) \rceil \quad \text{and} \quad \hat{\mu}_l = \frac{1}{N}\sum_{m=1}^{N} X_{\{(l-1)N+m\}}$$

The choice of $q = \lceil 32\log(t) \rceil$ is useful in establishing the required concentration inequality. This requirement comes from the fact that we need at least $N = 1$ samples per bin. Further, we assume that for all arms, $u > 32\log(t)$. Hence, the inequality that we now propose is valid only for $u > 32\log(t)$.

We define a bernoulli random variable $Y_l = \mathbb{1}\{\hat{\mu}_l > \mu + \mathcal{W}\}$. According to equation 12 in Bubeck et al. (2013), Y_l has the parameter

$$p \leq \frac{3B}{N^{\epsilon}\mathcal{W}^{1+\epsilon}}$$

Choosing $\mathcal{W}(u,t) = f(t)\left(\frac{1}{N}\right)^{g(t)}$, where $f(t)$ is a slow-growing function, and $g(t)$ is a slow-decaying function, yields,

$$p \leq \frac{3B}{N^{\epsilon} f(t)^{1+\epsilon}\left(\frac{1}{N}\right)^{g(t)(1+\epsilon)}}$$

Since $f(t)$ is slow-growing and $g(t)$ is slow-decaying, we are guaranteed to have a t_0 such that, for all $t > t_0$, we have $g(t) < \frac{\epsilon}{1+\epsilon}$ and $f(t)^{1+\epsilon} > 12B$. For such $t > t_0$, we get,

$$p \leq \left(\frac{1}{4}\right)\left(\frac{3B}{12f(t)^{1+\epsilon}}\right)\left(\frac{1}{N^{\epsilon - g(t)(1+\epsilon)}}\right) \leq \frac{1}{4}$$

Finally, using Hoeffding inequality for binomial random variable,

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}(i,u,t) - \mu > \mathcal{W}(u,t)\right) = \mathbb{P}\left(\sum_{j=1}^{q} X_j\right) &\leq \exp\left(-2q(\frac{1}{2} - p)^2\right) \\ &\leq \exp\left(\frac{-q}{8}\right) = \exp\left(\frac{-32\log(t)}{8}\right) = t^{-4}\end{aligned}$$

Note that this inequality is valid for all time instances $t > t_0$ and $u > u_0$ where t_0 is a distribution dependent constant parameter and $u_0 = \lceil 32\log(t) \rceil$, an increasing function.

This inequality is useful in establishing an upper bound on the probability of events E_1 and E_2, similar to case 1. We have,

$$\mathbb{P}(E_1) \leq \mathbb{P}(\exists u \in [t] : \mathcal{U}(i^*, u, t) \leq \mu^*) \leq t.t^{-4} = t^{-3} \quad \text{by union bound over } u$$

Similarly, $\mathbb{P}(E_2) \leq t^{-3}$.

We define u_i' as done in the the proof of theorem 2. However, for the above distribution oblivious concentration inequality to hold, we have an additional constraint of $u > u_0$. Hence, in this case we choose $u_i = \max(u_i', u_0, t_0)$.

Similar to the previous two cases, we get,

$$\mathbb{E}[T_i(t)] \leq u_i + 4 \quad \text{but here} \quad u_i = \max\left\{\left(\frac{2f(t)}{\Delta_i}\right)^{\frac{1}{g(t)}} 32\log(t), 32\log(t), t_0\right\}$$

However, we observe that t_0 is a constant and thus the first two terms (u_i', u_0) will be more than t_0 after a time instance, say t_1'. Moreover, the first function is faster growing than the second function, since $\left(\frac{2f(t)}{\Delta_i}\right)^{\frac{1}{g(t)}}$ is increasing with time instance t. Denote t_1'' as the threshold time. Define $t_1 = \max(t_1'.t_1'')$. Hence,

$$\mathbb{E}[T_i(t)] \leq \left(\frac{2f(t)}{\Delta_i}\right)^{\frac{1}{g(t)}} 32\log(t) + 4 \quad \forall t > t_{min}(\nu)$$

where the instance dependent threshold $t_{min}(\nu) = \max(t_0, t_1)$.

Thus, we get the regret upper bound as

$$\boxed{R_t(\nu) \leq \sum_{i:\Delta_i>0} \left(\Delta_i \left(\frac{2f(t)}{\Delta_i}\right)^{\frac{1}{g(t)}} 32\log(t) + 4\Delta_i\right) \quad \forall t > t_{min}(\nu)}$$

It is left to show that the above regret bound is indeed *consistent*. We show that there exists appropriate choices of $f(t)$ and $g(t)$ so that the overall regret expression can be made as close to logarithmic as we want. □

D.1 Proof of corollary 1

Proof. We see that $e^{0.5(\log\Phi(t))^{1-c}}$ is an increasing function for $c \in (0,1)$. Hence, we choose $f(t) = 0.5e^{0.5(\log\Phi(t))^{1-c}}$ and $g(t) = \frac{1}{\log^c(\Phi(t))}$. Also there exists t_0 such that for all $t > t_0$, $\frac{1}{\Delta_i} \leq e^{0.5(\log\Phi(t))^{1-c}}$ since LHS is a constant while RHS is an increasing function of t. Thus, we have,

$$\frac{f(t)}{\Delta_i} \leq e^{(\log\Phi(t))^{1-c}} \quad \forall t > t_0$$

Again, there exists t_1 such that LHS (and hence RHS) is greater than 1.

Finally for all $t > t_{min}$, where $t_{min} = \max(t_0, t_1)$, we have,

$$\left(\frac{f(t)}{\Delta_i}\right)^{\frac{1}{g(t)}} \leq \left(e^{(\log\Phi(t))^{1-c}}\right)^{\log^c(\Phi(t))} = \Phi(t) \quad \forall t > t_{min}$$

□