Data-Driven Upper Confidence Bounds with Near-Optimal Regret for Heavy-Tailed Bandits

Ambrus Tamás[1,2] **Szabolcs Szentpéteri**[1] **Balázs Csanád Csáji**[1,2]

[1] Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN)
[2] Department of Probability Theory and Statistics, Eötvös Loránd Universtiy (ELTE), Budapest, Hungary

Abstract

Stochastic multi-armed bandits (MABs) provide a fundamental reinforcement learning model to study sequential decision making in uncertain environments. The upper confidence bounds (UCB) algorithm gave birth to the renaissance of bandit algorithms, as it achieves near-optimal regret rates under various moment assumptions. Up until recently most UCB methods relied on concentration inequalities leading to confidence bounds which depend on moment parameters, such as the variance proxy, that are usually unknown in practice. In this paper, we propose a new distribution-free, data-driven UCB algorithm for symmetric reward distributions, which needs no moment information. The key idea is to combine a refined, one-sided version of the recently developed resampled median-of-means (RMM) method with UCB. We prove a near-optimal regret bound for the proposed anytime, parameter-free RMM-UCB method, even for heavy-tailed distributions.

1 Introduction

In this paper we study *stochastic multi-armed bandits* (MABs) which are special online learning models (Lattimore and Szepesvári, 2020). They are fundamental to study the notorious *exploration-exploitation dilemma* of reinforcement learning, and also have a wide range of direct applications. They were initially introduced to study sequential clinical trials, but recent applications include recommender systems, portfolio optimization, adaptive routing, clustering, anomaly detection and Monte Carlo tree search (Bouneffouf et al., 2020).

The theory of MABs has a rich history. They were introduced by Robbins (1952), nevertheless, the modern renaissance of bandit algorithms started with the publication of the *upper confidence bounds* (UCB) method by Auer et al. (2002). Since then, there has been continued interest in bandit algorithms (Bubeck et al., 2012; Lattimore and Szepesvári, 2020).

MABs are studied under diverse conditions and have an ever growing number of variants. We focus on stochastic MABs with finitely many arms, which were extensively studied in the past (Lattimore and Szepesvári, 2020). We deal with UCB algorithms which aim to find the right balance between exploration and exploitation based on the optimism principle, by choosing the arm with the highest UCB at each round. A vast amount of research has been done under various assumptions on the moment generating function of the arms' reward distributions (Bubeck et al., 2012). Famously, for subgaussian bandits Auer et al. (2002) showed that the optimal regret rate is $O(\sqrt{n})$, which can be achieved by UCB type algorithms.

In the heavy-tailed regime, for rewards with bounded central moment of order $1 + a$, Bubeck et al. (2013) showed that the regret is at least $O(n^{1/(1+a)})$. They also presented a near-optimal robust UCB algorithm, which however uses a as well as the moment bound M. On the other hand, these (hyper) parameters are typically *unknown* in practice, similarly to the variance proxies of subgaussian rewards. These call for data-driven, *parameter-free* methods. Cesa-Bianchi et al. (2017) removed the dependence on M for $a = 1$ by applying a robust estimator with random exploration. In (Lee et al., 2020) an adaptively perturbed exploration (APE^2) scheme was proposed which uses only the value of a and not the value of M. Wei and Srivastava (2021) developed the robust version of minimax optimal strategy for stochastic bandits (MOSS) and proved optimal regret bounds if M is given for the

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

agent. In (Lee and Lim, 2024) the MOSS is combined with a $1+a$-robust estimator to achieve minimax optimality without using the knowledge of M, however, their method essentially relies on the knowledge of a. Fully data-driven methods are presented in (Huang et al., 2022) and (Genalti et al., 2024). Huang et al. (2022) guarantee an $O(\log(T))$ gap-dependent bound for stochastically constrained environments without the knowledge of a and M for the Optimistically Tsallis Implicitly Normalized Forecaster (OptTINF) algorithm and provides minimax adversarial regret for the Adaptive Tsallis Implicitly Normalized Forecaster algorithm (AdaTINF). The method of Genalti et al. (2024) is based on an adaptive trimmed mean estimator and under the truncated non-positivity assumption near-optimal regret bounds are proved. For a risk-aware best arm identification problem Kagrecha et al. (2019) developed a completely parameter-free (distribution oblivious) method.

1.1 Stochastic Multi-Armed Bandits

A finite-armed stochastic bandit problem consists of an environment and an agent. The environment is described by the set of arms $[K] \doteq \{1, \dots, K\}$ and the corresponding reward distributions $\{\nu_1, \dots, \nu_K\}$. At each round $t \in \mathbb{N}$ the agent picks an arm I_t from action set $[K]$ and receives a random reward X_t drawn independently from distribution ν_{I_t}. Each distribution ν_i has a finite expected value μ_i, and for notational simplicity let $\mu_1 > \mu_i$ for $i \geq 2$ and let $\Delta_i \doteq \mu_1 - \mu_i$ for $i \in [K]$, which is called the suboptimality gap of arm i. The goal is to minimize the (expected) regret

$$R_n \doteq n \max_{i \in [K]} \mu_i - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right], \tag{1}$$

or equivalently to maximize the (expected) gained rewards. The main challenge of the problem is that the reward distributions are unknown for the agent, henceforth the actions need to be chosen based on previously observed (random) rewards. Thus, the exploitation of the gathered information and the exploration of the (stochastic) environment have to be balanced.

1.2 Data-Driven Bandits

The standard approach to stochastic MABs is to apply an UCB-style algorithm. Most of these methods strongly rely on some posed moment assumptions. The method of Auer et al. (2002) uses the form of the moment generating function and also the moments of the reward distributions. However, in practice these moment values are typically *unknown*. A recent paper of Khorasani and Weyer (2023) proposed a *data-driven* approach, which is able to overcome this challenge. They introduced the so-called *maximum average randomly sampled* (MARS) algorithm, which does not rely on the moment generating function or on the moments, hence it is parameter-free. They prove theoretical guarantees and achieve competitive performance results for *symmetric* bandits. MARS achieves a regret bound of

$$\begin{aligned} R_n &\leq \sum_{i:\Delta_i>0} \left(3 + \frac{3\log(n)}{\tau_i}\right)\Delta_i, \qquad \text{with} \\ \tau_i &\doteq -\log\left(\frac{1}{2} + \frac{1}{2}\exp(-\psi_i^*(\Delta_i))\right), \end{aligned} \tag{2}$$

on a *fixed horizon*, where $\psi_i^*(z) \doteq \sup_{\lambda \in \mathbb{R}}(\lambda z - \psi_i(\lambda))$ is the Legendre-Fenchel transform of the moment generating function ψ_i of distribution ν_i.

1.3 Heavy-Tailed Bandits

If the reward distributions are heavy-tailed, then providing UCBs for the means and proving regret bounds are challenging. Bubeck et al. (2013) do not assume that the moment generating function exists for every ν_i. They pose the much milder moment assumption of

$$\mathbb{E}_{\zeta_i \sim \nu_i}[|\zeta_i - \mathbb{E}[\zeta_i]|^{1+a}] \leq M < \infty, \tag{3}$$

with some known $a > 0$ and M. Then, for a heavy-tailed UCB algorithm they prove that

$$R_n \leq \sum_{i:\Delta_i>0} \left(2c\left(\frac{M}{\Delta_i}\right)^{1/a} \log(n) + 5\Delta_i\right), \tag{4}$$

for some constant c. This implies a rate of $n^{1/(1+a)}$ for R_n. Bubeck et al. (2013) also prove that there is an environment of K distributions such that for every strategy the regret has the lower bound

$$R_n \geq \text{const} \cdot K^{a/(1+a)} n^{1/(1+a)}. \tag{5}$$

A limitation of their methods is that the agent needs to know a and M to construct the UCBs at each round.

1.4 Main Contributions

In this paper, we build upon the recently developed *resampled median-of-means* (RMM) estimator (Tamás et al., 2024), which can build *non-asymptotically exact* confidence intervals for the mean of a symmetric distribution, without any moment information, such that it ensures *optimal subgaussian rates* for the sizes of the intervals under heavy-tailed moment conditions. Using RMM, we propose an *anytime*, *parameter-free* UCB algorithm which is efficient even for the case of *heavy-tailed rewards*. We introduce a *one-sided* version of the RMM method to build UCBs with *exact* confidence levels, assuming only that the rewards are

distributed symmetrically about their means. The data-driven MARS algorithm of Khorasani and Weyer (2023) can be seen as a special case of our construction, i.e., it corresponds to the case of taking the (resampled) median of only one mean ($k = 1$). We prove an *exponential concentration bound* for the one-sided RMM.

We also prove a *regret bound* for the proposed RMM-UCB algorithm whose rate is *optimal up to a logarithmic factor*, even under mild, *heavy-tailed* moment conditions. Unlike most previous constructions, RMM-UCB achieves this near-optimal regret rate *without any a priori information on the moments*, particularly, without the knowledge of parameters a and M in (3).

RMM-UCB is also compared numerically with several baseline, concentration inequality based UCB methods, as well as with recent, state-of-the-art and data-driven UCB algorithms on heavy-tailed bandit problems.

2 Median-of-Means Type Upper Confidence Bounds

The fundamental ingredient of an UCB algorithm is the confidence bound construction. In this section we present a new, *one-sided* version of the recently developed resampled median-of-means (RMM) method (Tamás et al., 2024) to construct exact UCBs for the mean of a symmetric, heavy-tailed distribution. We assume that an i.i.d. sample $\mathcal{D}_0 \doteq \{X_i\}_{i=1}^n$ is given from an unknown distribution Q_X which is *symmetric* about an unknown parameter μ, that is

A1. $X_1, \dots, X_n$ are i.i.d.

A2. Q_X is symmetric about μ.

We aim at constructing non-asymptotically valid upper confidence bounds for μ. We use no further assumptions other than symmetry and manage to reach any user-chosen (rational) confidence level. We also prove non-asymptotic, probably approximately correct (PAC) bounds for the distances between the UCB and μ under the extra assumption that

A3. $\mathbb{E}[|X - \mathbb{E}[X]|^{1+a}] = M < \infty$, for $a \in (0, 1]$.

Let $\operatorname{med}(X)$ denote a median of random variable X. Let the *ordered sample* be denoted by

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}. \tag{6}$$

Then, the *empirical median* of sample $X_1, \dots, X_n$ is

$$\operatorname{med}(\mathcal{D}_0) \doteq \begin{cases} X_{(n/2)} & \text{if } n \text{ is even,} \\ X_{(\lfloor n/2 \rfloor + 1)} & \text{if } n \text{ is odd.} \end{cases} \tag{7}$$

If Q_X is symmetric, $\mu = \operatorname{med}(X)$. In this paper we assume that $\mathbb{E}[X_1] = \mu$, however, most of the presented results regarding the confidence level holds even without this reasonably mild moment assumption.

2.1 Remarks on the Empirical Mean

Let us denote the well-known empirical mean by $\bar{\mu}_n$. By the celebrated Gauss-Markov theorem (Plackett, 1949) if $\sigma^2 \doteq D^2(X_1) < \infty$, then $\bar{\mu}_n$ is BLUE: it is the "best" linear unbiased estimator, i.e., $\bar{\mu}_n$ has the lowest variance among linear unbiased estimators. By the central limit theorem (CLT) if $\sigma^2 < \infty$, then $\sqrt{n}(\bar{\mu}_n - \mu) \to Z$ in distribution, where Z is a zero mean normal variable with variance σ^2. This convergence provides gaussian-type asymptotic bounds, but for bandits we typically seek finite sample guarantees. If X is σ-subgaussian, i.e., $\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$, then one has a non-asymptotic subgaussian concentration inequality for the mean. If the distribution is not subgaussian, Chebyshev's inequality provides an exponentially weaker bound w.r.t. the confidence parameter. Chebyshev's bound is tight in some sense, i.e., for every sample size n and confidence parameter there exists a distribution with variance σ^2 on which the empirical mean performs poorly, therefore the empirical mean is not subgaussian for μ in the general case.

2.2 Median-of-Means

It is a somewhat surprising result that one can construct subgaussian estimators for the mean of heavy-tailed distribution without assuming subgaussianity for the distribution of the sample. One of the most well-known estimator of this kind is the so-called median-of-means (MoM) estimator introduced by Nemirovsky and Yudin (1983). Let $k \leq n$ be an integer smaller than n. For the MoM estimator one needs to partition the dataset into k groups of almost the same size, i.e., let $\tilde{n} = \lfloor n/k \rfloor$ and for $\ell = 1, \dots, k$ if $\tilde{n}k + \ell \leq n$ let $B_\ell \doteq \{X_\ell, \dots, X_{\ell+\tilde{n}k}\}$, and $B_\ell \doteq \{X_\ell, \dots, X_{\ell+(\tilde{n}-1)k}\}$ otherwise. Partition $B_1, \dots, B_k$ can be defined in many ways as long as $|B_\ell| \geq \tilde{n}$ holds. The MoM estimator is defined by

$$\widehat{\mu}(\mathcal{D}_0) \doteq \operatorname{med}\left(\frac{1}{|B_1|} \sum_{i \in B_1} X_i, \dots, \frac{1}{|B_k|} \sum_{i \in B_k} X_i \right). \tag{8}$$

The MoM estimator reaches optimal finite sample performance in the following sense (Bubeck et al., 2013; Devroye et al., 2016) and (Lugosi and Mendelson, 2019):

Theorem 2.1. *Assume A1 and A3 and let* $\mathbb{E}[X_1] = \mu$. *The MoM estimator* $\widehat{\mu}$ *with* $k = \lceil 8 \log(2/\delta) \rceil$ *satisfies*

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \leq 8 \left(\frac{12 M^{1/a} \log(1/\delta)}{n} \right)^{a/(1+a)} \right) > 1 - \delta.$$

Moreover, for any mean estimator μ_n, *sample size* $n \in \mathbb{N}$ *and* $\delta > 0$ *there exists a distribution with mean* μ *and* $(1+a)$*th central moment* M *such that*

$$\mathbb{P}\left(|\mu_n - \mu| > \left(\frac{M^{1/a} \log(2/\delta)}{n} \right)^{a/(1+a)} \right) \geq \delta.$$

The second part provides a lower bound for the rate w.r.t. δ and n, hence the MoM estimator achieves an optimal convergence rate up to a constant factor.

2.3 One-Sided Resampled Median-of-Means

In this section, we introduce the *one-sided* version of the recently developed resampled MoM (RMM) method (Tamás et al., 2024), which then can be used to construct *exact* upper confidence bounds for the means of symmetric variables. In order to help understanding the intuitions behind the construction, we start by introducing the method as a statistical hypothesis test.

For a given $\theta \in \mathbb{R}$, let us consider the hypotheses

$$H_0 : \mu \leq \theta \qquad \text{vs} \qquad H_1 : \mu > \theta. \tag{9}$$

We assume conditions A1 and A2 and construct non-asymptotically exact UCBs based on a rank test. Then, we prove exponential PAC-bounds w.r.t. k for the distances between the UCBs and the true parameter under A3. Note that from A3 it follows that $\mathbb{E}X_1 = \mu < \infty$. In particular for $a = 1$ assumption A3 requires $\sigma^2 < \infty$. It is one of the main aims of this paper to deal with cases where $a < 1$, i.e., when the observed distribution is heavy-tailed. We emphasize that the presented hypothesis test can decide about H_0 without using the knowledge of a and M. These values are only included in the theoretical analysis of the proposed UCB.

Let p be the desired (rational) significance level. Let us find integers r and m such that $p = r/m$. We present a resampling algorithm to test H_0 for any $\theta \in \mathbb{R}$. Let $X(\theta) \doteq \alpha(X-\theta)+\theta$ be a random variable where α is a random sign, i.e., a Rademacher variable independent of X. One can immediately see that $\mathbb{E}[X(\theta)] = \theta$ and $X(\theta)$ is symmetrical about θ. Let $W = X-\mu$ be the centered version of X and $\mathcal{W}_0 \doteq \{W_i\}_{i=1}^n$, where $W_i = X_i - \mu$ for $i \in [n]$. Note that $\{W_i\}_{i=1}^n$ are not observed. One of our main observations is that $X(\mu) = \mu + \alpha W$ and $X = \mu + W$ have the same distribution, because W is symmetric about zero. Furthermore, one can prove that X and $X(\mu)$ are conditionally i.i.d. w.r.t. $\{|W_i|\}_{i=1}^n$, hence they are also exchangeable (Csáji et al., 2015). On the other hand, if $\theta \neq \mu$, then the distribution of X differs from the distribution of $X(\theta)$, e.g., $X(\theta)$ is symmetrical about θ whereas X is symmetrical about μ. We aim at constructing UCBs by utilizing this difference. Our procedure generates alternative samples and compares them to the original one with the help of a so-called *ranking function* (Csáji et al., 2015):

Definition 2.1 (ranking function)**.** *Let $\mathbb{A}$ be a measurable space, a (measurable) function $\psi : \mathbb{A}^m \to [m]$ is called a ranking function if for all $(a_1, \ldots, a_m) \in \mathbb{A}^m$, it satisfies the following two properties:*

P1 For all permutation μ of set $\{2, \ldots, m\}$, we have

$$\psi(a_1, a_2, \ldots, a_m) = \psi(a_1, a_{\mu(2)}, \ldots, a_{\mu(m)}), \tag{10}$$

that is function ψ is invariant with respect to reordering the last $m-1$ terms of its arguments.

P2 For all $i, j \in [m]$, if $a_i \neq a_j$, then we have

$$\psi(a_i, \{a_k\}_{k\neq i}) \neq \psi(a_j, \{a_k\}_{k\neq j}), \tag{11}$$

where the simplified notation is justified by P1.

The value of a ranking function is called the *rank*. An important observation about the ranks of exchangeable random elements is that all ranks have the same probability (Csáji and Tamás, 2019, Lemma 1):

Lemma 2.2. *Let $\xi_1, \ldots, \xi_m$ be a.s. pairwise different exchangeable random elements and let ψ be a ranking function. Then $\psi(\xi_1, \ldots, \xi_m)$ has a discrete uniform distribution on $[m]$, $\forall\, i : \mathbb{P}(\psi(\xi_1, \ldots, \xi_m) = i) = 1/m$.*

In our case, let $\{\alpha_{i,j}\}$ be i.i.d. Rademacher random variables for $i \in [n]$, $j \in [m-1]$ and

$$\mathcal{D}_j(\theta) \doteq \{\, \alpha_{1,j}(X_1-\theta)+\theta, \ldots, \alpha_{n,j}(X_n-\theta)+\theta \,\} \tag{12}$$

be parameter dependent *alternative samples* for $j = 1, \ldots, m-1$ and $\mathcal{D}_0(\theta) \doteq \mathcal{D}_0$ for $\theta \in \mathbb{R}$. We can observe that each $\mathcal{D}_j(\theta)$ is an i.i.d. sample from the distribution of $X(\theta) \doteq \alpha(X-\theta)+\theta$, for $j \neq 0$.

Specifically, the original sample $\mathcal{D}_0$ and the alternative samples $\mathcal{D}_j(\theta)$ are random vectors in $\mathbb{R}^n$. Observe that these datasets can be identical. This poses a technical challenge in ranking. We use a tie-breaking permutation π on set $[m-1]_0 \doteq \{0, \ldots, m-1\}$ generated independently from $\mathcal{D}_0$ and $\{\alpha_{i,j}\}$ to resolve this issue.

Let $\mathcal{D}_j^\pi(\theta) \doteq (\mathcal{D}_j(\theta), \pi(j))$ denote the extended datasets for $j = 0, \ldots, m-1$. It is easy to prove that $\mathcal{D}_0^\pi(\mu), \ldots, \mathcal{D}_{m-1}^\pi(\mu)$ are exchangeable, hence we obtain an exact hypothesis test if we reject H_0 if and only if $\psi(\mathcal{D}_0^\pi(\theta), \ldots, \mathcal{D}_{m-1}^\pi(\theta)) > m-r$. We refer to (Tamás et al., 2024, Theorem 3.1):

Theorem 2.3. *For any ranking function ψ and integers $r \leq m$, under assumptions A1 and A2, we have*

$$\mathbb{P}\big(\, \psi\big(\, \mathcal{D}_0^\pi(\mu), \ldots, \mathcal{D}_{m-1}^\pi(\mu) \,\big) > m-r \,\big) = \frac{r}{m}. \tag{13}$$

Let us also define *reference variables* using the MoM estimator $\widehat{\mu}(\cdot)$, for $j = 0, \ldots, m-1$, as

$$S_j(\theta) \doteq \widehat{\mu}(\mathcal{D}_j(\theta)) - \theta. \tag{14}$$

Then, we decide about H_0 by comparing $S_0(\theta)$ to $S_j(\theta)$ for $j \in [m-1]$ with a *ranking function* R. Notice that

Algorithm 1: One-Sided RMM for the Mean

Inputs: i.i.d. sample $\mathcal{D}_0$, rational significance level p, integer MoM parameter k (number of partitions)

1: Choose integers $1 \leq r \leq m$ such that $p = r/m$.
2: Generate $n(m-1)$ independent Rademacher signs $\{\alpha_{i,j}\}$ for $j \in [m-1]$ and $i \in [n]$.
3: Generate a random permutation π on $[m-1]_0$ independently from $\mathcal{D}_0$ and $\{\alpha_{i,j}\}$ uniformly from the symmetrical group.
4: Construct alternative samples for $j \in [m-1]$ by

$$\mathcal{D}_j(\theta) \doteq \{\alpha_{1,j}(X_1 - \theta) + \theta, \ldots, \alpha_{n,j}(X_n - \theta) + \theta\},$$

and let $\mathcal{D}_j^\pi \doteq (\mathcal{D}_j(\theta), \pi(j))$ for $j \in [m-1]_0$.
5: Compute the reference variables for $j \in [m-1]_0$:

$$S_j(\theta) \doteq \widehat{\mu}(\mathcal{D}_j(\theta)) - \theta.$$

6: Compute the rank

$$R(\theta) = 1 + \sum_{j=1}^{m-1} \mathbb{I}\big(S_0(\theta) \prec_\pi S_j(\theta)\big).$$

7: Reject H_0 if and only if $R(\theta) > m - r$.

if $\theta = \mu$, then each reference variable has the same distribution, while if $\theta \neq \mu$, then $S_0(\theta)$ is farther from θ than $S_j(\theta)$ with high probability for $j = 1, \ldots, m-1$.

In conclusion, let us construct a ranking function as

$$R(\theta) \doteq 1 + \sum_{j=1}^{m-1} \mathbb{I}\big(S_0(\theta) \prec_\pi S_j(\theta)\big), \tag{15}$$

where $\prec_\pi$ is defined as the standard $<$ ordering with tie-breaking (Csáji et al., 2015), formally:

$$\begin{gathered} S_j(\theta) \prec_\pi S_k(\theta) \\ \Updownarrow \\ S_j(\theta) < S_k(\theta) \text{ or } \big(S_j(\theta) = S_k(\theta) \text{ and } \pi(j) < \pi(k)\big), \end{gathered} \tag{16}$$

where π is the random permutation. The one-sided RMM test rejects θ if and only if $R(\theta) > m - r$. The step by step procedure is presented in Algorithm 1. One of our main results is that our rank test admits an exact confidence level (type I error probability):

Theorem 2.4. *Assume A1 and A2, then we have*

$$\mathbb{P}\,(R(\mu) > m - r) = \frac{r}{m}. \tag{17}$$

Proof. Theorem 2.4 is a direct consequence of Theorem 2.3, because $\mathcal{D}_0^\pi(\mu), \mathcal{D}_1(\mu)^\pi, \ldots, \mathcal{D}_{m-1}^\pi(\mu)$ are exchangeable and R is a ranking function. □

2.4 Resampled Median-of-Means Upper Confidence Bounds

Let us use the proposed hypothesis test to construct upper confidence bounds for μ. We build a confidence set out of those parameters that are accepted by Algorithm 1. For the RMM test we need to generate random signs, however, this procedure does not need to be repeated for every parameter. We use the same set of random signs for every θ. Hence, the confidence region for expectation μ is defined by $\Theta_n \doteq \{\,\theta \in \mathbb{R} : R(\theta) \leq m - r\,\}$. This leads to the upper confidence bound:

$$U \doteq \sup\{\,\theta \in \mathbb{R} : R(\theta) \leq m - r\,\}. \tag{18}$$

We will show that confidence set Θ_n is an interval, and it is either $(-\infty, U)$ or $(-\infty, U]$ depending on the tie-breaking permutation π. Note that U can be infinite.

An important consequence of Theorem 2.4 is as follows:

Corollary 2.4.1. *Set Θ_n is an exact confidence region for μ with significance level r/m, that is,*

$$\mathbb{P}(\,\mu \in \Theta_n\,) = 1 - \frac{r}{m}. \tag{19}$$

It can also be shown that the inclusion probability for $\theta > \mu$ goes to zero with an exponential rate as the sample size tends to infinity (Tamás et al., 2024). The whole procedure of the UCB construction is described by Algorithm 2. First, we provide a formula for U for $m = 2$ and $r = 1$ by Lemma 2.5, which can be efficiently computed from the data, then we prove Lemma 2.6, which presents a finite representation for Θ_n for any rational confidence probability.

Lemma 2.5. *Assume A1 and A2. For $m = 2$*

$$U = \mathrm{med}_{\ell \in [k]} \frac{\widehat{\mu}(\mathcal{D}_0) - \frac{1}{|B_\ell|}\sum_{i \in B_\ell} \alpha_{i,1} X_i}{1 - \frac{1}{|B_\ell|}\sum_{i \in B_\ell} \alpha_{i,1}}, \tag{20}$$

where we define $\pm c/0 = \pm\infty$, for all $c > 0$, and for notational simplicity, $0/0 = \mathrm{sign}(\pi(1) - \pi(0)) \cdot \infty$.

Lemma 2.6. *Under assumptions A1 and A2, for any $p = r/m$ with $1 \leq r < m$, we have $U = U_{(m-r)}$, with*

$$U_j \doteq \mathrm{med}_{\ell \in [k]} \frac{\widehat{\mu}(\mathcal{D}_0) - \frac{1}{|B_\ell|}\sum_{i \in B_\ell} \alpha_{i,j} X_i}{1 - \frac{1}{|B_\ell|}\sum_{i \in B_\ell} \alpha_{i,j}}, \tag{21}$$

for $j = 1, \ldots, m-1$, where $U_{(1)} \prec_\pi \cdots \prec_\pi U_{(m-1)}$ is ordered, and for notational simplicity we use $\frac{\pm c}{0} = \pm\infty$ for all $c > 0$, and $0/0 = \mathrm{sign}(\pi(1) - \pi(0)) \cdot \infty$.

Proof. Because of Lemma 2.5 we have

$$\{\theta : S_0(\theta) \succ_\pi S_j(\theta)\} = (-\infty, U_j]_\pi \tag{22}$$

Algorithm 2: RMM Upper Confidence Bound

Inputs: i.i.d. sample $\mathcal{D}_0$, rational significance level p, integer MoM parameter k

1: Choose integers $1 \le r < m$ such that $p = r/m$.
2: Generate $n(m-1)$ independent Rademacher signs $\{\alpha_{i,j}\}$ for $(i,j) \in [n] \times [m-1]$.
3: Generate a random permutation π on $[m-1]_0$ independently from $\mathcal{D}_0$ and $\{\alpha_{i,j}\}$ uniformly from the symmetrical group.
4: For all $(\ell, j) \in [k] \times [m-1]$, if $S_0(\theta) = S_j^{(\ell)}(\theta)$, then let $U_{\ell,j} \doteq \operatorname{sign}(\pi(j) - \pi(0)) \cdot \infty$; else

$$U_{\ell,j} = \frac{\widehat{\mu}(\mathcal{D}_0) - \frac{1}{|B_\ell|}\sum_{i \in B_\ell} \alpha_{i,j} X_i}{1 - \frac{1}{|B_\ell|}\sum_{i \in B_\ell} \alpha_{i,j}},$$

where $\frac{\pm c}{0} = \pm\infty$ for all $c > 0$.
5: For all $(\ell, j) \in [k] \times [m-1]$, let $U_j \doteq \operatorname{med}_{\ell \in [k]} U_{\ell,j}$.
6: Return $U \doteq U_{(m-r)}$, where $U_{(1)}, \dots, U_{(m-1)}$ are ordered w.r.t. $\prec_\pi$.

for $j = 1, \dots, m-1$, where U_j is included if $\pi(0) > \pi(j)$. Henceforth, it follows that

$$\begin{aligned}\Theta_n &\doteq \left\{\theta : 1 + \sum_{j=1}^{m-1} \mathbb{I}(S_0(\theta) \prec_\pi S_j(\theta)) \le m - r\right\} \\ &= (-\infty, U_{(m-r)}]_\pi,\end{aligned} \tag{23}$$

therefore $U = U_{(m-r)}$. □

The next theorem (its proof can be found in the supplementary material) presents a PAC bound on $U - \mu$.

Theorem 2.7. *Assume A1, A2 and A3. For $1 \le r \le m$ user-chosen integers and for*

$$\Theta_n \doteq \{\, \theta : R(\theta) \le m - r \,\}, \tag{24}$$

for every $n \in \mathbb{N}$, integer $k \le n$ and $\tilde{n} = \lfloor n/k \rfloor$ we have

$$\begin{aligned}&\mathbb{P}\left(U - \mu > 4\left(\frac{(12M)^{1/a}}{\tilde{n}}\right)^{\frac{a}{1+a}}\right) \\ &\le (m-r)\left(2k\exp(-\tilde{n}/8) + 2\exp(-k/8)\right).\end{aligned} \tag{25}$$

3 Resampled Median-of-Means Based UCB Algorithm

We present an important application of our method by deriving a new UCB algorithm for the problem of stochastic MABs. The upper confidence algorithms and their near-optimality were first published in (Auer et al., 2002) under subgaussian assumptions on the reward distributions. We elaborate on the work of (Bubeck et al., 2013), where heavy-tailed distribution were considered and near-optimal bounds were given. Our method is proved to have the same regret rate up to a logarithmic factor, however, unlike their approach, RMM-UCB does not need the knowledge of the moments nor its order to apply our algorithm under symmetricity.

Algorithm 3: RMM-UCB policy

Inputs: number of arms K

1: Pull each arm once.
2: At round t let $p_t = 1/\lceil 1 + t\log^2(t) \rceil$ and $k_t^{(i)} \doteq \lfloor 17\log(t) \wedge \sqrt{T_i(t-1)} \rfloor$.
3: At round $t \in \mathbb{N}$ for every $i \in [K]$ compute $U_i(T_i(t-1), p_t, k_t^{(i)})$ as in (27)
4: Choose arm $I_t \doteq \arg\max_{i \in [K]} U_i(T_i(t-1), p_t, k_t^{(i)})$.

Let us consider a stochastic MAB with K arms. The unknown reward distributions, $\nu_1, \dots, \nu_K$, have finite expected values, denoted by $\mu_1, \dots, \mu_K$ as above. We pose only mild assumptions on the environment:

A4. *For $i \in [K]$ distribution ν_i is symmetric about μ_i.*

A5. *For every $i \in [K]$ there exists $a_i \in (0, 1]$ such that*

$$\mathbb{E}_{X_i \sim \nu_i}[|X_i - \mu_i|^{1+a_i}] = M_i < \infty.$$

We consider heavy-tailed reward distributions with different moment parameters which are not known for the agent. At every round t from 1 to a finite horizon n we must choose an arm I_t and then obtain a reward from ν_{I_t}. Our main goal is to minimize the *regret* (1).

Let us present the *robust resampling median-of-means based* upper confidence bound (RMM-UCB) policy, see Algorithm 3. Observe that for constructing UCBs with Algorithm 2 one needs a sample $\mathcal{D}$, a (rational) confidence level p and a MoM parameter k. Let

$$A_i(t) = \{s : s \le t, I_s = i\}, \tag{26}$$

$T_i(t) = \sum_{s=1}^t \mathbb{I}(I_s = i)$ and $\mathcal{D}_i(t-1) = \{X_i\}_{i \in A_i(t-1)}$. and let us use the simplified notation

$$U_i(T_i(t-1), p, k) \doteq U(\mathcal{D}_i(t-1), p, k), \tag{27}$$

for $i \in [K]$, $t \in \mathbb{N}$, $p \in (0,1)$ and $k \le T_i(t-1)$. Then, the RMM-UCB algorithm pulls each arm once in the beginning and at round t we let $p_t = 1/\lceil 1 + t\log^2(t) \rceil$,

$$k_t^{(i)} \doteq \lfloor 17\log(t) \wedge \sqrt{T_i(t-1)} \rfloor, \tag{28}$$

compute $U_i(T_i(t-1), p_t, k_t^{(i)})$ for $i \in [K]$ and choose the arm with the highest UCB. The method guarantees that $k_t^{(i)} \le T_i(t-1)$ for all $t \in \mathbb{N}$ and $i \in [K]$, therefore the UCBs are always well-defined. We emphasize that the algorithm is anytime and parameter-free, i.e., we

do not need to a priori know the horizon nor any hyperparameters to apply the method.

In Theorem 3.2 we present our main result, a near-optimal regret bound for the RMM-UCB algorithm. The main ingredients of the proof of Theorem 3.2 are the regret decomposition lemma (Lattimore and Szepesvári, 2020) and the lemma that follows.

Lemma 3.1. *If we apply the RMM-UCB policy, then for $i \in [K]$, $i \neq 1$, for $c_i \doteq 4^{\frac{1+a_i}{a_i}} \cdot 12^{1/a_i}$ and some constant $C > 0$ we have*

$$\mathbb{E}[T_i(n)] \leq \max\left(c_i\left(\frac{M_i}{\Delta_i^{1+a_i}}\right)^{1/a_i}, 17^2\right)\log^2(n) + C.$$

Its proof can be found in the supplementary material.

Theorem 3.2. *Assume A4 and A5 and let $c_i = 4^{\frac{1+a_i}{a_i}} \cdot 12^{1/(1+a_i)}$ for $i \in [K]$. Then, for the (expected) regret of the RMM-UCB policy, we have*

$$R_n \leq \tag{29}$$
$$\sum_{i:\Delta_i>0}\left(\max\left\{c_i\left(\frac{M_i}{\Delta_i^{1+a_i}}\right)^{1/a_i}, 17^2\right\}\log^2(n) + C\right)\cdot\Delta_i.$$

Additionally, if $a = a_i$ and $M = M_i$ for $i \in [K]$, then for n large enough we have

$$R_n \leq n^{\frac{1}{1+a}}(K2c\log^3(n))^{\frac{a}{1+a}}M^{\frac{1}{1+a}}. \tag{30}$$

Proof. By Lemma 3.1 and Lemma 5.5 Equation (29) follows. For Equation (30) one can follow the proof of (Bubeck et al., 2013, Proposition 1), which relies mostly on Hölder's inequality. If n is large enough, i.e., if for all $i \in [K]$, we have

$$\begin{aligned} c\frac{M^{1/a}}{\Delta_i^{(1+a)/a}}\log(n) &\geq 17^2, \\ c\frac{M^{1/a}}{\Delta_i^{(1+a)/a}}\log^3(n) &\geq C, \end{aligned} \tag{31}$$

then, we get

$$\begin{aligned} R_n &\leq \sum_{i:\Delta_i>0}\Delta_i(\mathbb{E}[T_i(n)])^{\frac{1}{1+a}}(\mathbb{E}[T_i(n)])^{\frac{a}{1+a}} \\ &\leq \sum_{i:\Delta_i>0}\Delta_i(\mathbb{E}[T_i(n)])^{\frac{1}{1+a}}\cdot \\ &\quad\left(\max\left(c\frac{M^{1/a}}{\Delta_i^{(1+a)/a}}, 17^2\right)\log^2(n) + C\right)^{\frac{a}{1+a}} \\ &\leq \sum_{i:\Delta_i>0}\Delta_i(\mathbb{E}[T_i(n)])^{\frac{1}{1+a}}\left(2c\frac{M^{1/a}}{\Delta_i^{(1+a)/a}}\log^3(n)\right)^{\frac{a}{1+a}} \\ &\leq \left(\sum_{i:\Delta_i>0}\mathbb{E}[T_i(n)]\right)^{\frac{1}{1+a}}K^{\frac{a}{1+a}}(2c\log^3(n))^{\frac{a}{1+a}}M^{\frac{1}{1+a}} \\ &\leq n^{\frac{1}{1+a}}(K2c\log^3(n))^{\frac{a}{1+a}}M^{\frac{1}{1+a}}. \end{aligned} \tag{32}$$

In the last step we used the fact that $\sum_{i=1}^K T_i(n) = n$ holds for all $n \in \mathbb{N}$. □

4 Numerical Experiments

In this section, we empirically evaluate the performance of the proposed RMM-UCB algorithm and compare it with baseline concentration inequality based approaches, such as Vanilla UCB (Lattimore and Szepesvári, 2020), Median-of-Means UCB (Bubeck et al., 2013), and Truncated Mean UCB (Bubeck et al., 2013). Furthermore, RMM-UCB is also compared with state-of-the-art data-driven approaches, such as the Perturbed-History Exploration (PHE) method (Kveton et al., 2019) and the Maximum Average Randomly Sampled (MARS) algorithm (Khorasani and Weyer, 2023), as well as with the Minimax Optimal Robust Adaptively Perturbed Exploration (MR-APE) approach (Lee and Lim, 2024), a heavy-tailed bandit solution.

We consider a stochastic bandit setting with $K = 2$ arms, where $\mu^* = \mu_1 = 1$, $\mu_2 = \mu_1 - \Delta$ and $\Delta \in (0, 1]$ determines the suboptimality gap. As we study the MAB problem in case of heavy-tailed distributions and we assume symmetric reward distributions, in the first set of experiments, the rewards of both arms are sampled from a symmetrized Pareto distribution, more specifically $\forall\, i \in [K] : \nu_i \sim S(X - 1)$, where S is sampled from the Rademacher distribution and X is from a Pareto distribution with scale parameter 1 and shape parameter $\alpha_p = 1.05 + \varepsilon_p$.

In all algorithms we set the round dependent confidence parameter as $\delta_t = \lceil 1 + t\log^2(t)\rceil$ ($p_t = 1/\delta_t$) and in case of the median-of-means estimators, the rewards of arm i are divided into $k_t = \lfloor 17\log(t) \wedge \sqrt{T_i(t-1)}\rfloor$ partitions as in Algorithm 3. For the moment upper bound parameter M of the Truncated Mean and Median-of-Means UCB algorithms, we used the best possible one, i.e., $M = \mathbb{E}[|\nu_i|^{1+\varepsilon}]$. The PHE hyperparameter was set to $a = 5.1$, while we chose uniform perturbations, $p = 1 + \varepsilon$ for the (non-centered) absolute moment and hyperparameters $c = 0.5$, $\epsilon = 0$ for the MR-APE.

The average cumulative regret of each algorithm from 100 independently generated trajectories in case of $\varepsilon_p = 0.1, \varepsilon_p = 0.5$ and suboptimality gaps $\Delta = 0.1$, $\Delta = 0.5$, are shown in Figure 1. As we investigate bandit problems with heavy-tailed distributions, showing the standard deviations of the cumulative regrets in the same figure would reduce comprehensibility. Neverthe-

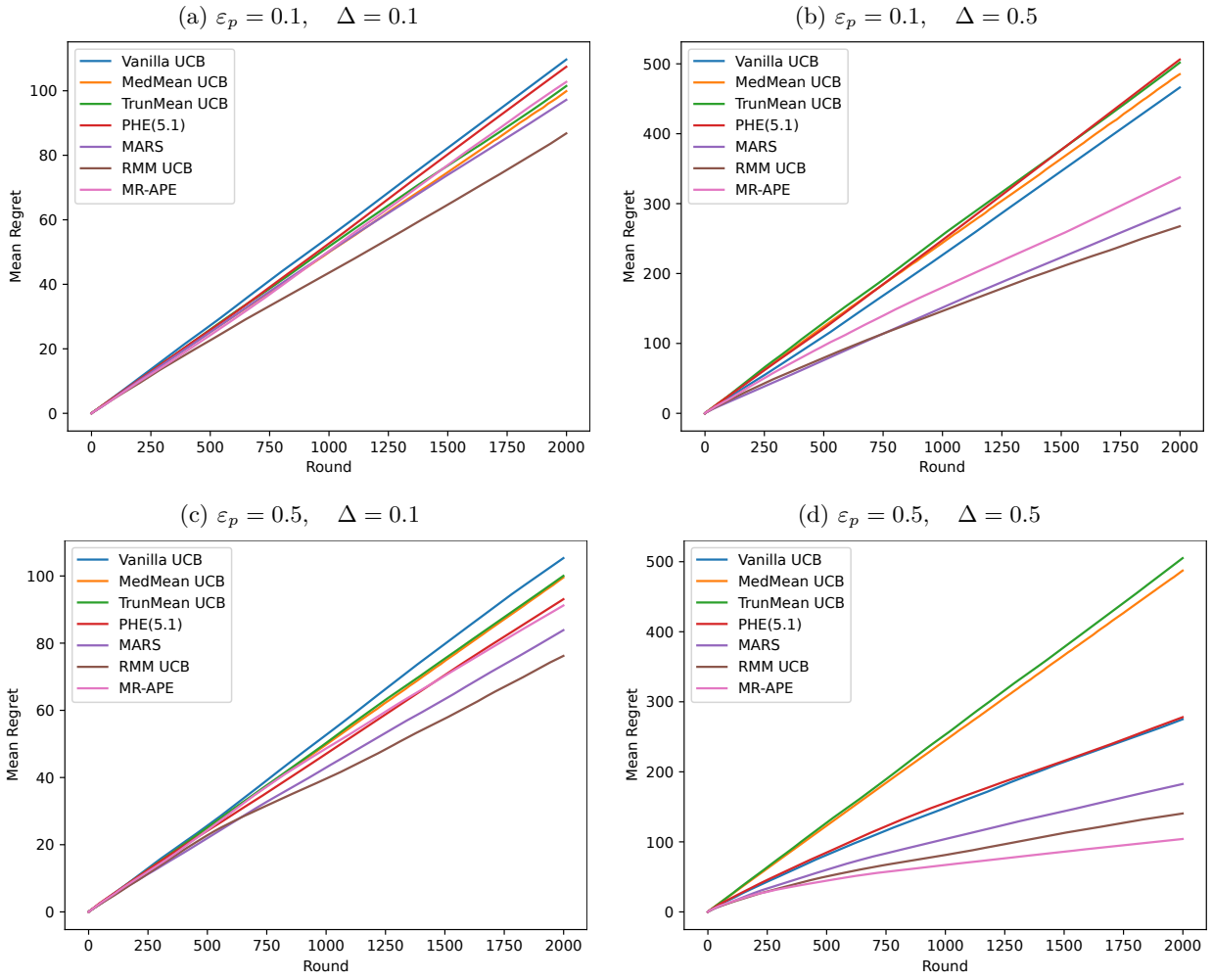

Figure 1: Average cumulative regrets of symmetric Pareto bandits with $\varepsilon_p = \{0.1, 0.5\}$ and $\Delta = \{0.1, 0.5\}$.

less, the standard deviations are illustrated on separate figures which are presented in the supplements.

It can be observed that for a small suboptimality gap $\Delta = 0.1$, the proposed RMM-UCB method outperforms all of the other algorithms, which have similar average regrets, for both ε_p values. In case $\Delta = 0.5$, the RMM-UCB, MARS and MR-APE algorithms perform significantly better than the other solutions, however, for $\varepsilon_p = 0.1$, RMM-UCB still demonstrates the best performance. For the final case, when $\varepsilon_p = 0.5$ and $\Delta = 0.5$, MR-APE outperforms RMM-UCB.

We also investigated the performance of the aforementioned algorithms for *non-symmetric* heavy-tailed arm distributions. In this case, $\forall i \in [K] : \nu_i \sim (2B-1)(X-1)$, where B is sampled from the Bernoulli distribution with $p = 0.8$ and X is from a Pareto distribution with parameter ε_p. All the other experimental settings and parameters were the same. The average cumulative regret of each algorithm for $\varepsilon_p = \{0.1, 0.5\}$ and $\Delta = 0.5$ are illustrated in Figure 2. It can be seen that in both cases, RMM-UCB has a smaller average regret than most algorithms, the two exceptions are MR-APE and MARS, the latter one is a special case of RMM-UCB, with the choice $\forall t : k_t = 1$. Also, for $\varepsilon_p = 0.1$, RMM-UCB shows a similar performance as MR-APE. The standard deviations of the achieved regrets can be found in the supplementary material.

An additional numerical experiment with $K = 5$ arms, presented in the supplementary material, shows similar results to the $K = 2$ case. More specifically, for $\varepsilon_p = 0.1, \Delta = 0.1$, the proposed RMM-UCB method has the lowest average cumulative regret, while for $\varepsilon_p = 0.5, \Delta = 0.5$, only MR-APE outperforms RMM-UCB.

We should emphasize that MR-APE is not a fully data-driven solution, as it requires a moment parameter of the arm distributions. Moreover, MR-APE is not an anytime algorithm, either, since it needs to know the horizon. On the other hand, the proposed RMM-UCM

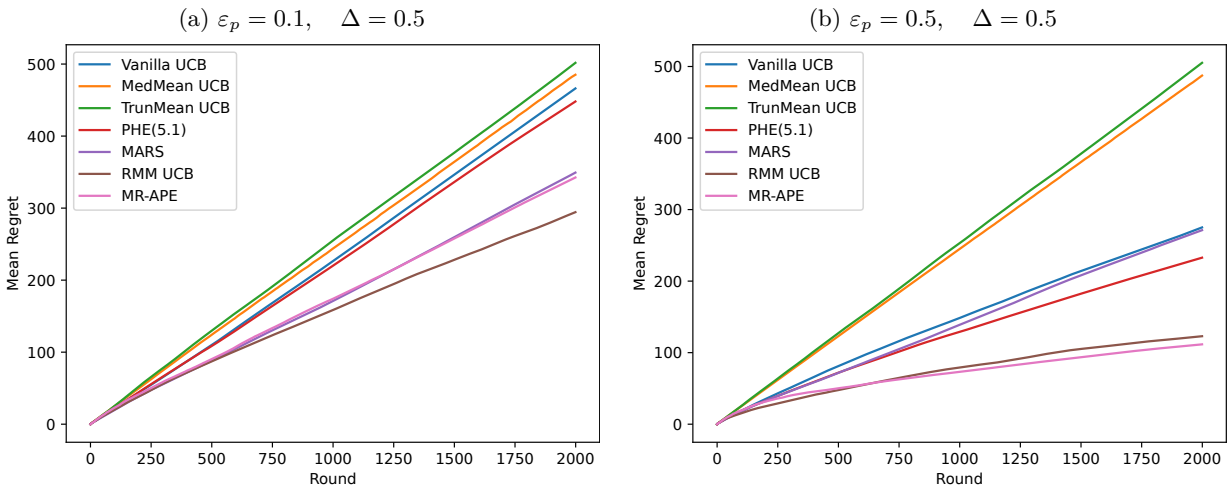

Figure 2: Average cumulative regrets of non-symmetric Pareto bandits with $\Delta = 0.5$ and $\varepsilon_p = \{0.1, 0.5\}$.

is both an anytime and data-driven method, which should be taken into account for a fair comparison.

Overall, these experimental results indicate that the proposed anytime data-driven RMM-UCB method can achieve lower cumulative regrets than existing solutions in case of difficult bandit problems, i.e., with small suboptimality gaps and heavy-tailed reward distributions, while also having the advantage that it does not require any knowledge of moment or distribution parameters.

5 Discussion

Stochastic multi-armed bandits (MABs) are fundamental and extensively studied online learning problems, which are especially important for reinforcement learning, and also have a wide range of direct applications.

UCB style algorithms serve as one of the standard ways to solve them, as they can achieve near-optimal regret for a large number of MAB variants. The core and most important part of UCB methods is the construction of the upper confidence bounds. The standard way to deduce such confidence bounds is to rely on concentration inequalities based on some moment assumptions. However, in many cases we do not have any a priori information on the reward distributions, we may not know any of their moment parameters either. This motivates completely data-driven methods which are free from such hyper-parameters that need to be tuned.

In this paper, we proposed a *parameter-free*, anytime, heavy-tailed UCB algorithm which achieves the same optimal regret rate, up to a logarithmic term, as previous parameter-dependent solutions. We combined the advantages of the recently developed *resampled median-of-means* (RMM) estimator (Tamás et al., 2024) with UCB to ensure similar rates as in (Bubeck et al., 2012) on a data-driven manner, assuming symmetric reward distributions, as in (Khorasani and Weyer, 2023). In order to achieve this, we also introduced and studied a one-sided version of RMM. Finally, we presented several numerical experiments, which indicate that RMM-UCB can outperform most previous algorithms on difficult MAB problems, i.e., when the suboptimality gap is small and the reward distributions are heavy-tailed.

Acknowledgements

This research was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory, and by the TKP2021-NKTA-01 grant of the National Research, Development and Innovation Office, Hungary.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47:235–256, 2002.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on Applications of Multi-Armed and Contextual Bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2020.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits With Heavy Tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann Exploration Done Right. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pages 6284–6293, 2017.

Balázs Csanád Csáji, Marco Claudio Campi, and Erik Weyer. Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non–Asymptotic Confidence Regions in Linear Regression models. *IEEE Transactions on Signal Processing*, 63 (1):169–181, 2015.

Balázs Csanád Csáji and Ambrus Tamás. Semi-Parametric Uncertainty Bounds for Binary Classification. In *58th IEEE Conference on Decision and Control (CDC), Nice, France*, pages 4427–4432, 2019.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian Mean Estimators. *Annals of Statistics*, 44(6):2695–2725, 2016.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ε, u)-Adaptive Regret Minimization in Heavy-Tailed Bandits. In *37h Annual Conference on Learning Theory (COLT)*, pages 1882–1915. PMLR, 2024.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive Best-of-Both-Worlds Algorithm for Heavy-Tailed Multi-Armed Bandits. In *39th International Conference on Machine Learning (ICML)*, pages 9173–9200. PMLR, 2022.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna P Jagannathan. Distribution Oblivious, Risk-Aware Algorithms for Multi-Armed Bandits with Unbounded Rewards. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 11269–11278. PMLR, 2019.

Masoud M Khorasani and Erik Weyer. Maximum Average Randomly Sampled: A Scale Free and Non-parametric Algorithm for Stochastic Bandits. In *Advances in Neural Information Processing Systems (NeurIPS) 37*, 2023.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-History Exploration in Stochastic Multi-Armed Bandits. In *28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2786–2793, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Kyungjae Lee and Sungbin Lim. Minimax Optimal Bandits for Heavy Tail Rewards. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4): 5280–5294, 2024.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy Tailed Rewards. *Advances in Neural Information Processing Systems (NeurIPS) 33*, pages 8452–8462, 2020.

Gábor Lugosi and Shahar Mendelson. Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey. *Foundations of Computational Mathematics*, 19:1145–1190, 2019.

Arkadij S Nemirovsky and David B Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Robin L Plackett. A Historical Note on the Method of Least Squares. *Biometrika*, 36(3/4):458–460, 1949.

Herbert Robbins. Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematics Society*, 58(527-535), 1952.

Szabolcs Szentpéteri and Balázs Csanád Csáji. Sample Complexity of the Sign-Perturbed Sums Identification Method: Scalar Case. In *22nd IFAC World Congress*, pages 10363–10370, 2023.

Ambrus Tamás, Szabolcs Szentpéteri, and Balázs Csanád Csáji. Data-Driven Confidence Intervals with Optimal Rates for the Mean of Heavy-Tailed distributions. In *27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3439–3447. PMLR, 2024.

Lai Wei and Vaibhav Srivastava. Minimax Policy for Heavy-Tailed Bandits. *IEEE Control Systems Letters*, 5(4):1423–1428, 2021.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] The main theoretical result of the paper quantifies the regret rate of the method, with respect to the sample size.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable] We included the pseudocode of the new algorithms.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material for Data-Driven Upper Confidence Bounds with Near-Optimal Regret for Heavy-Tailed Bandits

This appendix contains the proofs and the experiments which were left out from the paper due to lack of space.

Proof of Lemma 2.5

Proof. The graph of reference function

$$S_0(\theta) = \widehat{\mu}(\mathcal{D}_0) - \theta \tag{33}$$

is a line with slope -1, whereas

$$\begin{aligned} S_1(\theta) &= \widehat{\mu}(\mathcal{D}_1(\theta)) - \theta \\ &= \operatorname{med}\left(\frac{1}{|B_1|}\sum_{i\in B_1}(\alpha_{i,1}(X_i-\theta)+\theta), \dots, \frac{1}{|B_k|}\sum_{i\in B_k}(\alpha_{i,1}(X_i-\theta)+\theta)\right) - \theta \\ &= \operatorname{med}\left(\frac{1}{|B_1|}\sum_{i\in B_1}\alpha_{i,1}(X_i-\theta), \dots, \frac{1}{|B_k|}\sum_{i\in B_k}\alpha_{i,1}(X_i-\theta)\right), \end{aligned} \tag{34}$$

which is the median of linear functions with slopes between -1 and $+1$. Let

$$S_1^\ell(\theta) \doteq \frac{1}{|B_\ell|}\sum_{i\in B_\ell}\alpha_{i,1}(X_i-\theta) \text{ for } \ell = 1,\dots,k \tag{35}$$

be sub-sample linear functions. It is easy to see that

$$\{\theta : S_1^\ell(\theta) < S_0(\theta)\} = (-\infty, \nu_\ell), \tag{36}$$

where if S_0 and S_1^ℓ are not parallel, then ν_ℓ is the intersection of S_0 with S_1^ℓ, in which case

$$\begin{aligned} \nu_\ell &= \frac{\widehat{\mu}(\mathcal{D}_0) - \frac{1}{|B_\ell|}\sum_{i\in B_\ell}\alpha_{i,1}X_i}{1 - \frac{1}{|B_\ell|}\sum_{i\in B_\ell}\alpha_{i,1}} \\ &= \mu + \frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{|B_\ell|}\sum_{i\in B_\ell}\alpha_{i,1}W_i}{\frac{1}{|B_\ell|}\sum_{i\in B_\ell}(1-\alpha_{i,1})} \\ &= \mu + \frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{|B_\ell|}\sum_{i\in B_\ell}\alpha_{i,1}W_i}{\frac{2}{|B_\ell|}Z_\ell}. \end{aligned} \tag{37}$$

If S_0 and S_1^ℓ are parallel, then the extended "intersection" points are

$$\nu_\ell \doteq \begin{cases} +\infty & \text{if } \forall\theta : S_1^\ell(\theta) \prec_\pi S_0(\theta) \\ -\infty & \text{if } \forall\theta : S_0(\theta) \prec_\pi S_1^\ell(\theta). \end{cases} \tag{38}$$

These values are equivalent to the formula of (37) with $\pm c/0 = \pm\infty$ and slightly abuse of notation $0/0 = \operatorname{sign}(\pi(1)-\pi(0))\cdot\infty$. By

$$\begin{aligned} \{\theta : S_0(\theta) \succ_\pi S_1(\theta)\} &= \bigcap_{I\subseteq[n],|I|=\lfloor k/2\rfloor}\ \bigcap_{\ell\in I}\{\theta : S_0(\theta) \succ_\pi S_1^\ell(\theta)\} \\ &= \bigcap_{I\subseteq[n],|I|=\lfloor k/2\rfloor+1}\ \bigcap_{\ell\in I}(-\infty,\nu_\ell]_\pi = (-\infty,\nu]_\pi, \end{aligned} \tag{39}$$

where $(-\infty, \nu]_\pi$ is closed from the right depending on π, i.e., $(-\infty, \nu]_\pi \doteq (-\infty, \nu]$ if $\pi(0) > \pi(1)$ and closed otherwise. Henceforth,

$$U = \operatorname{med}_{1\leq \ell \leq k} \nu_\ell = \mu + \operatorname{med}_{1\leq \ell \leq k} \left(\frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{|B_\ell|} \sum_{i\in B_\ell} \alpha_{i,1} W_i}{\frac{2}{\tilde{n}} Z_\ell} \right) = \mu + V,$$

where V is defined by

$$V \doteq \operatorname{med}_{1\leq \ell \leq k} \left(\frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{|B_\ell|} \sum_{i\in B_\ell} \alpha_{i,1} W_i}{\frac{2}{|B_\ell|} Z_\ell} \right) \tag{40}$$

is a valid 50% UCB. □

Proof of Theorem 2.7

The proof is based on Theorem 5.1 and the union bound. It is essentially the same as the proof of (Szentpéteri and Csáji, 2023, Theorem 3). □

Theorem 5.1. *Assume A1-A3. For $m = 2$ and*

$$U \doteq \sup \{\, \theta : R(\theta) = 1 \}, \tag{41}$$

for every $k \leq n$, $n \in \mathbb{N}$ and $\tilde{n} = \lfloor n/k \rfloor$, we have

$$\mathbb{P}\left(U - \mu > 4(12M)^{1/(1+a)} \left(\frac{1}{\tilde{n}} \right)^{\frac{a}{1+a}} \right) \leq 2k \exp(-\tilde{n}/8) + 2\exp(-k/8) \tag{42}$$

Proof. The proof consists of two parts as we bound

$$\mathbb{P}(U - \mu > \varepsilon) \tag{43}$$

with two different terms. Let

$$Z_\ell \doteq \frac{1}{2} \sum_{i\in B_\ell} (1 - \alpha_i) \quad \text{for} \quad \ell = 1, \ldots, k. \tag{44}$$

We observe that $\{Z_\ell\}_{\ell=1}^k$ are independent binomial variables with parameters $|B_\ell|$, $1/2$ and expected value $|B_\ell|/2$ for $\ell \in [k]$. Let us denote the events that follow by

$$A_\ell \doteq \left\{ |Z_\ell - \mathbb{E}Z_\ell| \leq |B_\ell|/4 \right\} \quad \text{for} \quad \ell = 1, \ldots, k \quad \text{and} \quad A \doteq \bigcap_{\ell=1}^k A_\ell. \tag{45}$$

By the law of total probability

$$\begin{aligned} \mathbb{P}(U - \mu > \varepsilon) &= \mathbb{P}(U - \mu > \varepsilon \,|\, A)\mathbb{P}(A) + \mathbb{P}(U - \mu > \varepsilon \,|\, \bar{A})\mathbb{P}(\bar{A}) && (46) \\ &\leq \mathbb{P}(\{U - \mu > \varepsilon\} \cap A) + \mathbb{P}(\bar{A}). && (47) \end{aligned}$$

The second term is bounded from above by the union bound and Hoeffding's inequality as

$$\begin{aligned} \mathbb{P}(\,\bar{A}\,) &\leq \mathbb{P}\left(\bigcup_{\ell=1}^k \bar{A}_\ell \right) \leq \sum_{\ell=1}^k \mathbb{P}(|Z_\ell - \mathbb{E}Z_\ell| > |B_\ell|/4) \\ &\leq \sum_{\ell=1}^k 2\exp\left(-\frac{|B_\ell|}{8} \right) \leq 2k \exp\left(-\frac{\tilde{n}}{8} \right) \end{aligned} \tag{48}$$

For the first term recall that

$$\mathbb{P}(U - \mu > \varepsilon \,|\, A) = \mathbb{P}(V > \varepsilon \,|\, A).$$

We observe that Z_ℓ is independent of the nominator, because $\alpha_{i,1}W_i$ and $\alpha_{i,1}$ are independent for each $i \in [n]$. Let us consider the following events:

$$B = \{V > \varepsilon\}, \quad \tilde{B} = \left\{ \operatorname{med}_{\ell\in[k]} \left(\frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{|B_\ell|}\sum_{i\in B_\ell} \alpha_{i,1}W_i}{\frac{1}{2}} \right) > \varepsilon \right\}. \tag{49}$$

Our key observation is that

$$B \cap A \subseteq \tilde{B} \cap A, \tag{50}$$

because if V is positive, then decreasing Z_ℓ in the denominator for every $\ell = 1, \dots, k$ down to $\tilde{n}/4$, increases the median. Consequently

$$\begin{aligned}
&\mathbb{P}(B \cap A) \le \mathbb{P}(\tilde{B} \cap A) \\
&\le \mathbb{P}\left(\operatorname{med}_{\ell\in[k]} \left(\widehat{\mu}(\mathcal{W}_0) - \frac{1}{|B_\ell|}\sum_{i\in B_\ell} \alpha_{i,1}W_i \right) > \varepsilon/2 \right) \\
&= \mathbb{P}\left(\widehat{\mu}(\mathcal{W}_0) - \operatorname{med}_{\ell\in[k]} \left(\frac{1}{|B_\ell|}\sum_{i\in B_\ell} \alpha_{i,1}W_i \right) > \varepsilon/2 \right) \\
&\le \mathbb{P}\left(\{\widehat{\mu}(\mathcal{W}_0) > \varepsilon/4\} \cup \left\{ - \operatorname{med}_{\ell\in[k]} \frac{1}{|B_\ell|}\sum_{i\in B_\ell} \alpha_{i,1}W_i > \varepsilon/4 \right\} \right) \\
&\le \mathbb{P}\big(\widehat{\mu}(\mathcal{W}_0) > \varepsilon/4\big) + \mathbb{P}\left(\operatorname{med}_{\ell\in[k]} \frac{1}{|B_\ell|}\sum_{i\in B_\ell} -\alpha_{i,1}W_i > \varepsilon/4 \right) \\
&= 2 \cdot \mathbb{P}\big(\widehat{\mu}(\mathcal{W}_0) > \varepsilon/4\big).
\end{aligned} \tag{51}$$

hence for $\varepsilon = 4\left(\frac{(12M)^{1/a}}{\tilde{n}}\right)^{\frac{a}{1+a}}$ by Theorem 5.4 we have

$$\mathbb{P}\left(U - \mu > 4\left(\frac{(12M)^{1/a}}{\tilde{n}}\right)^{\frac{a}{1+a}} \right) \le 2k \exp\left(-\frac{\tilde{n}}{8} \right) + 2\exp\left(-\frac{k}{8} \right) \tag{52}$$

□

Proof of Lemma 3.1

Proof. Our proof strongly relies on (Bubeck et al., 2013) and (Khorasani and Weyer, 2023). First observe that if $I_t = i$ for $i \neq 1$, then at least one of the following inequalities holds

$$U_1(T_1(t-1), p_t, k_t^{(1)}) \le \mu_1, \tag{53}$$

$$U_i(T_i(t-1), p_t, k_t^{(i)}) > \mu_i + M_i^{\frac{1}{1+a_i}} \left(\frac{c_i k_t^{(i)}}{T_i(t-1)} \right)^{\frac{a_i}{1+a_i}}, \tag{54}$$

$$T_i(t-1) < u_i \tag{55}$$

with $c_i \doteq 4^{\frac{1+a_i}{a_i}} \cdot 12^{1/a_i}$ and

$$u_i \doteq \left\lceil \max\left(c_i \left(\frac{M_i}{\Delta_i^{(1+a_i)}} \right)^{1/a_i}, 17^2 \right) \log^2(n) \right\rceil.$$

If all of them were false, then

$$\begin{aligned}
&U_1(T_1(t-1), p_t, k_t^{(1)}) > \mu_1 = \mu_i + \Delta_i \ge \mu_i + M_i^{1/(1+a_i)} \left(\frac{c_i \log^2(n)}{T_i(t-1)} \right)^{a_i/(1+a_i)} \\
&\ge \mu_i + M_i^{1/(1+a_i)} \left(\frac{c k_t^{(i)}}{T_i(t-1)} \right)^{a_i/(1+a_i)} \ge U_i(T_i(t-1), p_t, k_t^{(i)}),
\end{aligned} \tag{56}$$

which contradicts $I_t = i$. We prove that (53) or (54) occur with small probability. Recall that $U_i = \mu_1 + V_i$. Let us denote the bad events above as

$$B_t^{(1)} \doteq \{V_1(T_1(t-1), p_t, k_t^{(1)}) \leq 0\},$$
$$B_t^{(2)} \doteq \left\{V_i(T_i(t-1), p_t, k_t^{(i)}) > M_i^{1/(1+a_i)} \left(\frac{c_i k_t^{(i)}}{T_i(t-1)}\right)^{a_i/(1+a_i)}\right\}.$$

and the good events as

$$G_t \doteq \{\, T_i(t-1) < u_i \,\}$$

for $t = 1, \ldots, n$. By the first observation we have

$$\begin{aligned}\mathbb{E}[T_i(n)] &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}(I_t = i)\right] = \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}(I_t = i)\mathbb{I}(G_t)\right] + \mathbb{E}\left[\sum_{t=u_i}^{n} \mathbb{I}(I_t = i)\mathbb{I}(\bar{G}_t)\right] \\ &\leq u_i + \sum_{t=u_i}^{n} \mathbb{P}((B_t^{(1)} \cup B_t^{(2)}) \cap \bar{G}_t) \leq u_i + \sum_{t=u_i}^{n} \mathbb{P}(B_t^{(1)}) + \sum_{t=u_i}^{n} \mathbb{P}(B_t^{(2)} \cap \bar{G}_t).\end{aligned} \tag{57}$$

For each $i \in [K]$ by Theorem 2.4

$$\mathbb{P}(B_t^{(1)}) = p_t = \frac{1}{\lceil 1 + t\log^2(t)\rceil}. \tag{58}$$

Additionally, we have $\sqrt{u_i} \geq 17\log(t)$ for $t \leq n$, henceforth by Theorem 2.7 we have

$$\begin{aligned}&\mathbb{P}(B_t^{(2)} \cap \bar{G}_t) \\ &\leq \mathbb{P}\left(V_i(T_i(t-1), p_t, k_t^{(i)}) > M_i^{\frac{1}{1+a_i}} \left(\frac{c_i k_t^{(i)}}{T_i(t-1)}\right)^{\frac{a_i}{1+a_i}}, T_i(t-1) \geq u_i\right) \\ &= \sum_{s=u_i}^{t} \mathbb{P}\left(V_i(s, p_t, k_t^{(i)}) > M_i^{\frac{1}{1+a_i}} \left(\frac{c_i k_t^{(i)}}{s}\right)^{\frac{a_i}{1+a_i}} \mid T_i(t-1) = s\right) \cdot \mathbb{P}(T_i(t-1) = s) \\ &\leq \sum_{s=u_i}^{t} 2\lceil t\log^2(t)\rceil \left(k_t^{(i)} \exp\left(-\frac{s}{k_t^{(i)} 8}\right) + \exp\left(-\frac{k_t^{(i)}}{8}\right)\right) \cdot \mathbb{P}(T_i(t-1) = s) \\ &\leq \sum_{s=u_i}^{t} 2\lceil t\log^2(t)\rceil \cdot \\ &\left(\sqrt{s} \wedge 17\log(t) \exp\left(-\frac{s}{\lfloor \sqrt{s} \wedge 17\log(t)\rfloor \cdot 8}\right) + \exp\left(-\frac{\sqrt{s} \wedge 17\log(t)}{8}\right)\right) \cdot \mathbb{P}(T_i(t-1) = s) \\ &\leq \sum_{s=u_i}^{t} 2\lceil t\log^2(t)\rceil \left(17\log(t)\exp\left(-\frac{\sqrt{s}}{8}\right) + \exp\left(-\frac{\sqrt{s} \wedge 17\log(t)}{8}\right)\right) \cdot \mathbb{P}(T_i(t-1) = s) \\ &\leq 2\lceil t\log^2(t)\rceil \left(17\log(t)\exp\left(-\frac{\sqrt{u_i}}{8}\right) + \exp\left(-\frac{\sqrt{u_i} \wedge 17\log(t)}{8}\right)\right) \\ &\leq 2\lceil t\log^2(t)\rceil \left(17\log(t)\exp\left(-\frac{17\log(t)}{8}\right) + \exp\left(-\frac{17\log(t)}{8}\right)\right) \\ &\leq 2\lceil t\log^2(t)\rceil \left(17\log(t)\frac{1}{t^{2+\gamma}} + \frac{1}{t^{2+\gamma}}\right),\end{aligned} \tag{59}$$

where $\gamma = 1/8$. We can observe that the right hand side is summable in t. Hence, the key quantity can be bounded from above as

$$\mathbb{E}[T_i(n)] \leq u_i + \sum_{t=u_i}^{n} \left(\frac{1}{1 + \lceil t\log^2(t)\rceil} + \frac{\tilde{c}\log^3(t)}{t^{1+\gamma}}\right) \leq u_i + C. \tag{60}$$

□

Auxiliary Results

We refer to fundamental results from (Bubeck et al., 2013):

Lemma 5.2. *Assume A4 and A5. Let $\bar{\mu}$ be the empirical mean. Then for any $\delta \in (0,1)$*

$$\mathbb{P}\left(\bar{\mu} \leq \mu + \left(\frac{3M}{\delta n^a}\right)^{\frac{1}{1+a}}\right) \geq 1 - \delta. \tag{61}$$

Theorem 5.3 from Bubeck et al. (2013) is our main tool to prove the concentration inequality for MoM estimates.

Theorem 5.3. *Let $\delta \in (0,1)$ and $a \in (0,1]$. Assume A1, A3 and $n = k\tilde{n}$. Let $k = \lfloor \min(8\log(e^{1/8}/\delta), n/2) \rfloor$, then for the MoM estimator $\widehat{\mu}$ we have*

$$\mathbb{P}\left(\widehat{\mu} \leq \mu + (12M)^{\frac{1}{1+a}}\left(\frac{16\log(e^{1/8}/\delta)}{n}\right)^{\frac{a}{1+a}}\right) \geq 1 - \delta. \tag{62}$$

We use a reparameterized version of Theorem 5.3. The proof for almost equal group sizes is included for the sake of completeness.

Theorem 5.4. *Assume A1 and A3. Let $k \leq n$ and $\tilde{n} = \lfloor n/k \rfloor$, then for the MoM estimator $\widehat{\mu}$ we have*

$$\mathbb{P}\left(\widehat{\mu} > \mu + (12M)^{\frac{1}{1+a}}\left(\frac{1}{\tilde{n}}\right)^{\frac{a}{1+a}}\right) \leq \exp\left(-\frac{k}{8}\right). \tag{63}$$

Proof. Let $\eta > 0$ and $Y_\ell \doteq \mathbb{I}(\bar{\mu}_\ell > \mu + \eta)$ for $\bar{\mu}_\ell \doteq \frac{1}{|B_\ell|}\sum_{i \in B_\ell} X_i$ and $\ell = 1, \ldots, k$. By Lemma 5.2 Y_ℓ is a Bernoulli variable with

$$p_\ell \leq \frac{3M}{|B_\ell|^a \eta^{1+a}}. \tag{64}$$

For

$$\eta = (12M)^{1/(1+a)}\left(\frac{1}{\tilde{n}}\right)^{\frac{a}{1+a}} \tag{65}$$

we have $p_\ell \leq 1/4$ for all $\ell \in [k]$. Finally, by Hoeffding's inequality for a binomial variable Z with parameters k and $1/4$ we have

$$\mathbb{P}(\widehat{\mu} > \mu + \eta) \leq \mathbb{P}(Z \geq k/2) \leq \exp\left(-\frac{k}{8}\right). \tag{66}$$

□

The regret decomposition lemma (Lattimore and Szepesvári, 2020, Lemma 4.5) is one of the main tools to prove Theorem 3.2.

Lemma 5.5. *For any policy and stochastic bandit environment $\{\nu_1, \ldots, \nu_K\}$ and horizon $n \in \mathbb{N}$ the regret R_n of the policy in the environment satisfies*

$$R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]. \tag{67}$$

Numerical Experiments

Experiment: Five armed bandit

Here, we consider a stochastic bandit problem with $K = 5$ arms, where $\mu^* = \mu_1 = 1$, $\forall i \in \{2, \ldots, 5\} : \mu_i = \mu_1 - i\Delta$ and $\Delta \in (0,1]$ determines the suboptimality gap. The reward distribution of the arms are sampled from a symmetric Pareto distribution given in Section 4. All the other experimental settings and parameters are also the same as in Section 4. The average cumulative regret of each algorithm for $\varepsilon_p = 0.1, \Delta = 0.1$ and $\varepsilon_p = 0.5$, $\Delta = 0.5$, are shown in Figure 3, while their corresponding standard deviations are illustrated in Figure 4. Similarly to the case of $K = 2$, for $\varepsilon_p = 0.1, \Delta = 0.1$ the RMM-UCB has the best average cumulative regret, while for $\varepsilon_p = 0.5, \Delta = 0.5$ only the MR-APE algorithms outperforms the proposed RMM-UCB. We would emphasize again that MR-APE is not a data-driven nor an anytime solution, while RMM-UCB is.

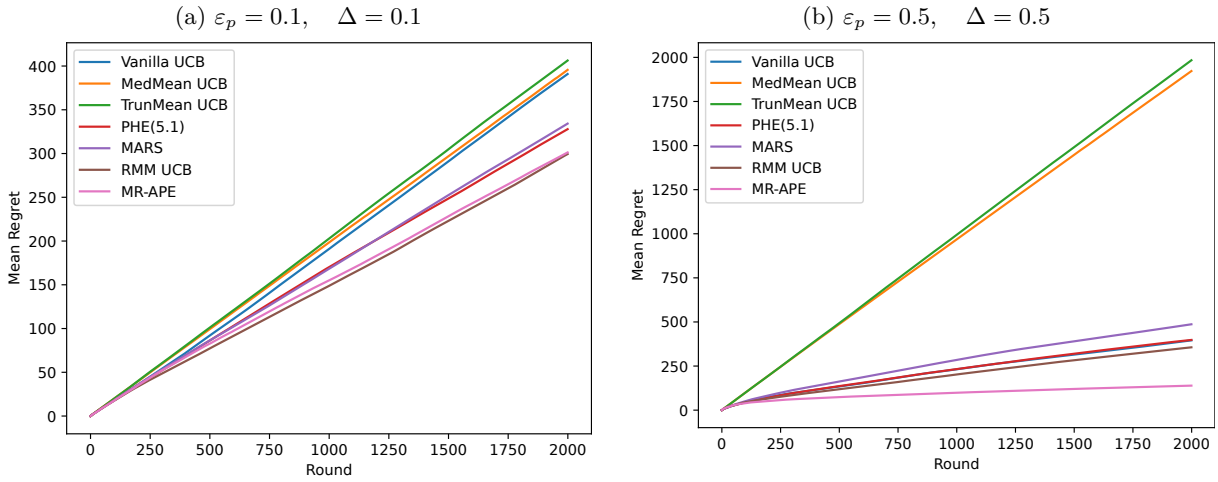

Figure 3: Average cumulative regrets of symmetric Pareto bandits with $K = 5$, $\varepsilon_p = 0.1, \Delta = 0.1$ and $\varepsilon_p = 0.5$, $\Delta = 0.5$.

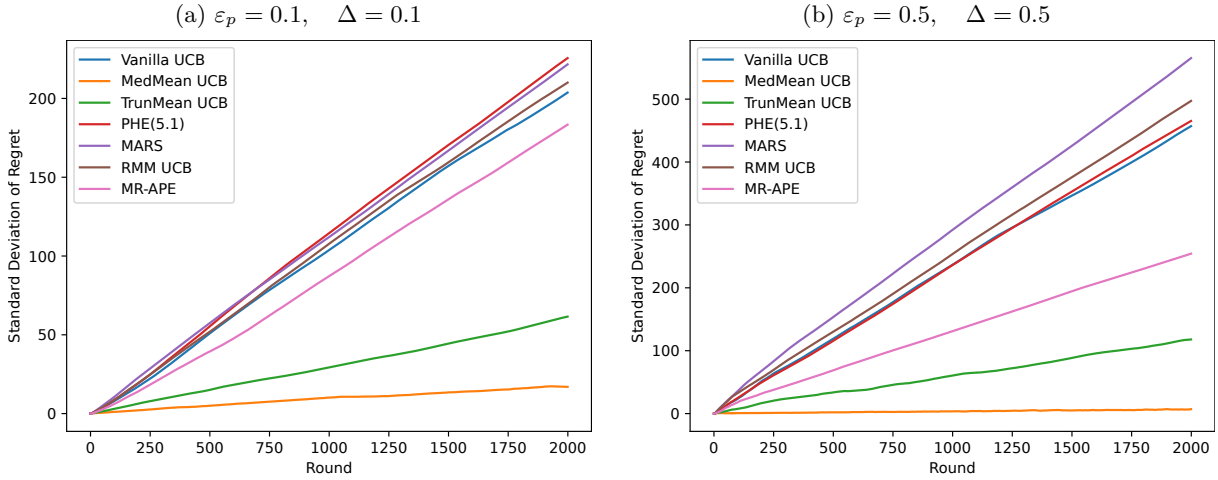

Figure 4: Standard deviations of cumulative regrets of symmetric Pareto bandits with $K = 5$, $\varepsilon_p = 0.1, \Delta = 0.1$ and $\varepsilon_p = 0.5$, $\Delta = 0.5$.

Experiment: Standard deviation of cumulative regrets

In this experiment we investigate the standard deviations of the cumulative regrets for the same algorithms and simulation setting as in Section 4. The results for different ε_p and Δ parameters in case of $K = 2$ are shown in Figure 5. It can be seen that the heavy-tailed UCB algorithms based on concentration inequalities, the Median-of-Means UCB and the Truncated Mean UCB, have the lowest standard deviation of cumulative regret. However, as our previous results showed, their performance considering the average cumulative regret is poor, and the upper confidence bounds generated by these algorithms are very conservative. These results also illustrate that for the other algorithms, in case of a small gap, $\Delta = 0.1$, the standard deviations are roughly the same, while for a larger gap, $\Delta = 0.5$, RMM-UCB and MR-APE have the least standard deviations for the cumulative regret.

The standard deviations of the cumulative regrets for non-symmetric Pareto rewards with $\Delta = 0.5$ and $\varepsilon_p = \{0.1, 0.5\}$ are illustrated in Figure 6. In this case, the Median-of-Means UCB and the Truncated Mean UCB algorithms have the lowest standard deviation of cumulative regrets, as well. This result also shows that MR-APE and MARS (a special case of RMM UCB) have a smaller standard deviation for the cumulative regret than RMM-UCB.

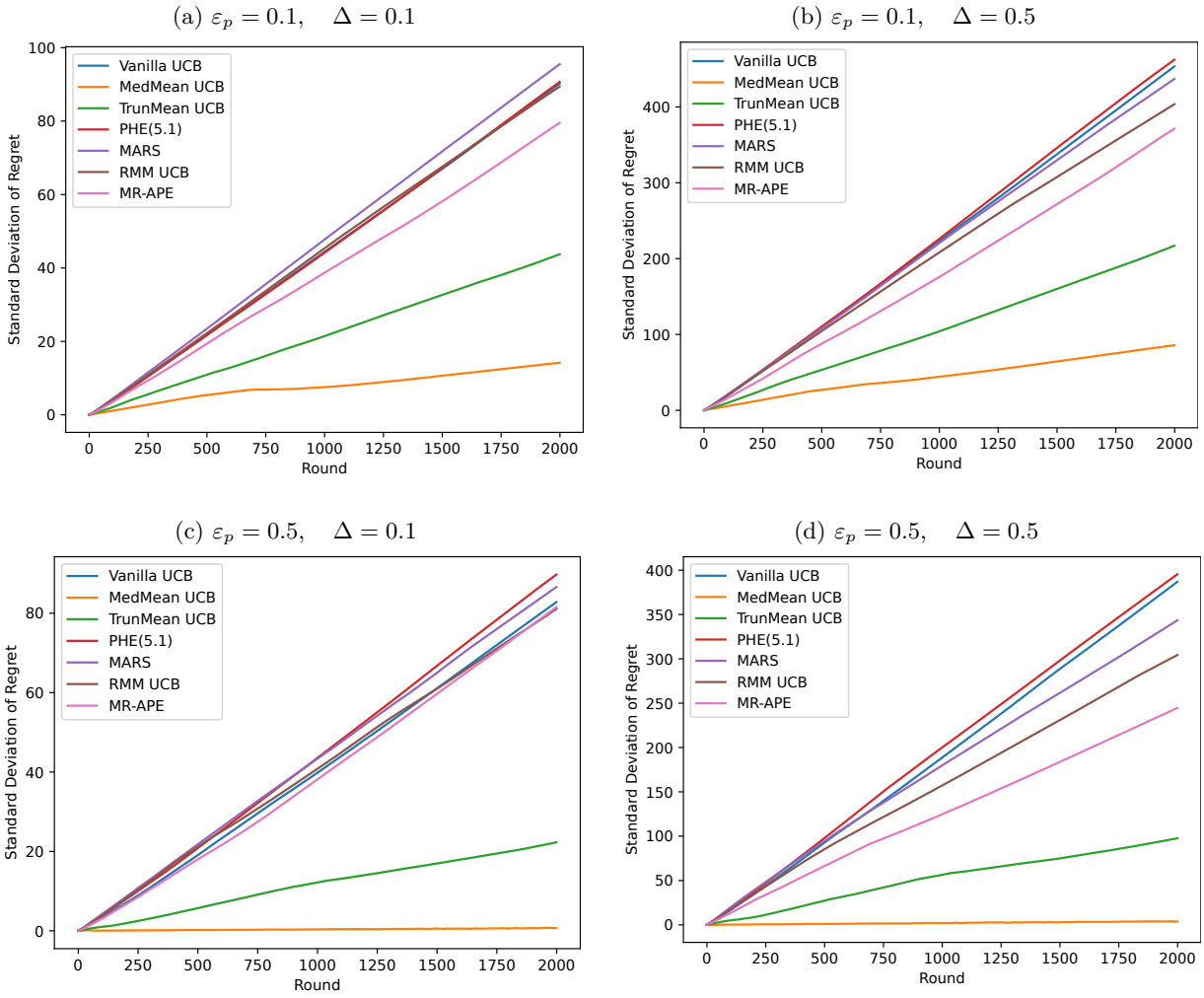

Figure 5: Standard deviation of cumulative regrets for symmetric Pareto bandits; $\varepsilon_p = \{0.1, 0.5\}$ and $\Delta = \{0.1, 0.5\}$.

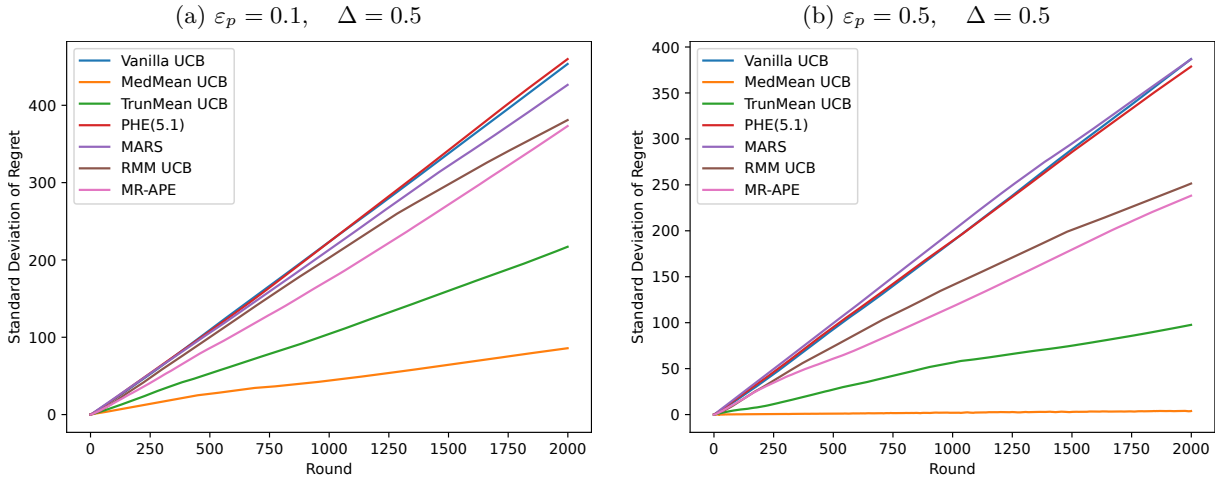

Figure 6: Standard deviations of cumulative regrets of non-symmetric Pareto bandits; $\Delta = 0.5$ and $\varepsilon_p = \{0.1, 0.5\}$.