Towards Fully Adaptive Regret Minimization in Heavy-Tailed Bandits

Gianmarco Genalti
Politecnico di Milano
gianmarco.genalti@polimi.it

Lupo Marsigli
Politecnico di Milano
lupo.marsigli@mail.polimi.it

Nicola Gatti
Politecnico di Milano
nicola.gatti@polimi.it

Alberto Maria Metelli
Politecnico di Milano
albertomaria.metelli@polimi.it

Abstract

Heavy-tailed distributions naturally arise in many settings, from finance to telecommunications. While regret minimization under sub-Gaussian or bounded support rewards has been widely studied, learning on heavy-tailed distributions only gained popularity over the last decade. In the stochastic heavy-tailed bandit problem, an agent learns under the assumption that the distributions have finite moments of maximum order $1 + \epsilon$ which are uniformly bounded by a constant u, for some $\epsilon \in (0, 1]$. To the best of our knowledge, literature only provides algorithms requiring these two quantities as an input. In this paper we study the stochastic *adaptive* heavy-tailed bandit, a variation of the standard setting where both ϵ and u are unknown to the agent. We show that adaptivity comes at cost, introducing two lower bounds on the regret of any adaptive algorithm implying an higher regret w.r.t. the standard setting. Finally, we introduce a specific distributional assumption and provide `Adaptive Robust UCB`, a regret minimization strategy matching the known lower bound for the heavy-tailed MAB problem.

1 Introduction

In this paper, we investigate the stochastic *multi-armed bandit problem* (MAB, Auer et al., 2002; Lattimore and Szepesvári, 2020) under the assumption of *heavy-tailed* (HT) reward distributions. In the classic stochastic multi-armed bandit setting Robbins (1952), an agent has access to a set of K possible actions (*i.e.*, *arms*). Each arm $i \in [K] \coloneqq \{1, \dots, K\}$ is associated with a reward probability distribution ν_i, having finite mean μ_i. At every round $t \in [T]$, being T a learning horizon, after an action I_t is selected, a reward X_t is sampled from ν_{I_t} and revealed to the agent. The goal of the agent is to minimize its *expected regret* after T rounds, defined as

$$R_T = T \max_{i \in [K]} \mu_i - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{I_t}\right], \tag{1}$$

where $\Delta_i \coloneqq \max_{j \in [K]} \mu_j - \mu_i$ is the sub-optimality gap for all $i \in [K]$ and the expectation is taken w.r.t. the randomness of the reward and the possible randomness of the algorithm.

Most of the existing works assume that the reward probability distributions ν_i are *sub-Gaussian*. Under this assumption, the tails of the distribution present a strong decay (at least as fast as that of the Gaussian distribution). An important implication is that every moment of finite order is finite. While this assumption enables the application of powerful theoretical tools and, consequently, strong regret guarantees, it is often limiting in many practical scenarios such as, for example, financial

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

environments (Gagliolo and Schmidhuber, 2011) or network routing problems (Liebeherr et al., 2012). In settings, where uncertainty has a significant impact, *heavy-tailed distributions* naturally arise. In these cases, the tails decay slower than a Gaussian, and the moment-generating function is no longer assumed to be finite. As a consequence, the moments of any finite order might not exist.

In this work, we investigate the regret minimization problem for MAB with heavy tails, according to the setting introduced in the seminal work (Bubeck et al., 2013a). We assume moments of order up to $1+\epsilon$, with $\epsilon \in (0, 1]$ to be finite and uniformly bounded by a constant u, namely

$$\mathbb{E}_{X\sim\nu_i}[|X|^{1+\epsilon}] \leq u < +\infty, \qquad \forall i \in [K]. \tag{2}$$

In (Bubeck et al., 2013a), the authors show that even if the variance is finite (*i.e.*, $\epsilon = 1$) but the higher order moments are not, the same guarantees of order $\sum_{i\in[K]:\Delta_i>0} \frac{\log T}{\Delta_i}$ attained in the classic sub-Gaussian stochastic bandit setting (Auer et al., 2002) can be achieved in the heavy-tailed bandit problem. However, if $\epsilon < 1$, then the dependency on the suboptimality gaps Δ_i deteriorates and the following upper bound on regret is of order $\sum_{i\in[K]:\Delta_i>0} \frac{\log T}{\Delta_i^{1/\epsilon}}$. Finally, the authors show that the dependency on $\Delta_i^{1/\epsilon}$ is unavoidable, by deriving the corresponding lower bound. From a worst-case perspective, this translates into a regret bound of order (apart from logarithmic terms) $K^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$, which degenerates to linear when $\epsilon \to 0$.

In the heavy-tailed MAB problem, it is canonical to assume the knowledge of both ϵ and u and, to the best of the authors' knowledge, every regret minimization strategy in the literature requires them as an algorithm's input.[1] The goal of this work is to answer the following question:

> *Is it possible to provide a novel regret minimization strategy in the heavy-tailed bandit problem that does not require any prior knowledge on ϵ nor u, but still achieves comparable performances to other approaches knowing them?*

In this work, we will show that in general it is not possible to achieve the same order of performance while being adaptive to the aforementioned unknown quantities. Fortunately, we will show that under a specific distributional assumption, the answer is, instead, affirmative. In particular, we will discuss the role of the *truncated non-positivity* assumption (Huang et al., 2022), and show that when this assumption is violated it is not possible anymore to guarantee the existence of an adaptive algorithm w.r.t. ϵ nor u achieving state-of-the-art performances. We introduce `Adaptive Robust UCB`, an algorithm based on the *optimism in the face of uncertainty* principle that is capable of being *fully adaptive* w.r.t. the two parameters ϵ and u that characterize the reward distributions. In particular, we propose a modification of the well-known `Robust UCB` algorithm from (Bubeck et al., 2013a), showing that under our assumption we are able to attain the same theoretical guarantees.

2 Setting

We now introduce the stochastic heavy-tailed multi-armed bandit problem. Formally speaking, there are $K \geq 2$ available actions and a sequence of T rounds. To each arm $i_t \in [K]$, we associate a probability distribution ν_i satisfying (2). At each round $t \in [T]$, the agent can choose an index i and subsequently collects a reward of X_t, which is an independent sample from ν_i. The agent is allowed to make a decision at round t by considering all history up to time $t-1$. We remark that distributions $\{\nu_i\}_{i\in[K]}$ only admit finite moments up to order $1+\epsilon$, with $\epsilon \in (0, 1]$, thus distributions with infinite variance are allowed in this problem formulation. The goal of the agent is to minimize the regret as defined in (1). In the heavy-tailed bandits literature, is customary to assume the knowledge on both ϵ and the upper bound on the $(1+\epsilon)$-th order moment u, which is assumed to be common for all $\{\nu_i\}_{i\in[K]}$ without loss of generality. The main contribution of this paper is twofold: first, we show that in general it is not possible to achieve the same order of performance of the state-of-the-art approaches while being unaware of these two quantities; then we propose an algorithm, called `Adaptive Robust UCB`, that under a specific distributional assumption called *truncated non-positivty* is able to match the order of the regret lower bound for the classic scenario even when both ϵ and u are unknown. Thus, in our setting, we will consider both quantities to be unknown to the agent. From now on, we will refer to any algorithm operating without this knowledge as *adaptive w.r.t. ϵ or u*, depending on which one is unknown (possibly both).

[1] Even if ϵ and u are not requested for some estimators (*e.g.*, median of means), they are necessary for the construction of the upper confidence bound.

3 Lower Bound on the Regret for Adaptive Heavy-Tailed Bandits

In this section, we state a lower bound on the expected regret that any adaptive algorithm (w.r.t. to either u or ϵ) can achieve in the heavy-tailed bandit problem. We start by stating the lower bound on the regret when ϵ and u are known.

Theorem 1 (Lower Bound on Regret for Stochastic Heavy-Tailed Bandit, adapted from Bubeck et al. (2013a))**.** *For any fixed T, there exists a set of K distributions satisfying* (2) *and such that for any algorithm, one has*

$$R_T \geq 0.01 K^{\frac{\epsilon}{1+\epsilon}} u^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}. \tag{3}$$

This result is independent of the problem instance and show how the dependency on T deteriorates as $\epsilon \to 0$. In the particular scenario in which variance is finite, *i.e.* $\epsilon = 1$, the lower bound achieves the same order as the one for classic stochastic multi-armed bandit problems (Lattimore and Szepesvári, 2020). The following results show that any algorithm unaware of ϵ or u, respectively, cannot achieve the same regret order as the one stated in 1. We start by stating the regret lower bound for any algorithm adaptive w.r.t. u.

Theorem 2 (Lower Bound on Regret for Stochastic Adaptive Heavy-Tailed Bandit, unknown u)**.** *For any fixed T, there exist two sets of distributions satisfying* (2) *with u and u' (assume $u' > u$ without loss of generality), respectively, and such that for any algorithm adaptive w.r.t. to the $1+\epsilon$-th order moment, one has*

$$\max\left\{\frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq C_1 \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(1+\epsilon)^2}}, \tag{4}$$

where $R(T)$ and $R'(T)$ are the regrets achieved by this algorithm in the two instances, respectively, and C_1 is a constant independent of u or u'.

Some remarks are in order. This result states that there exist two particular heavy-tailed bandit problem instances s.t. no algorithm can match the lower bound on regret presented in (3) on both, and instead some regret is accrued in a way that is proportional to the ratio of the two $1+\epsilon$-th order moments of those instances. In our construction (more details can be found in Appendix A.1), the gap between u and u' can be taken arbitrarily large, and thus the regret gap with the non-adaptive lower bound presented in (4) can be arbitrarily large. In particular, this result shows that is not possible to be adaptive in u without the risk of incurring in an arbitrarily large regret bound.
Next, we present a similar result concerning adaptivity w.r.t. to the maximum finite order moment ϵ.

Theorem 3 (Lower Bound on Regret for Stochastic Adaptive Heavy-Tailed Bandit, unknown ϵ)**.** *For any fixed T, there exist two sets of distributions satisfying* (2) *with ϵ and ϵ' (assume $\epsilon' < \epsilon$ without loss of generality), respectively, and such that for any algorithm adaptive w.r.t. to the order ϵ of the maximum order finite moment $u = 1$, one has*

$$\max\left\{\frac{R(T)}{T^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq C_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon)(1+\epsilon')^2}}, \tag{5}$$

where $R(T)$ and $R'(T)$ are the regrets achieved by this algorithm in the two instances, respectively, and C_2 is a constant independent of ϵ or ϵ'.

Differently from Theorem 2, where u and u' could take arbitrarily high values on the positive semi-axis of real numbers, the values of ϵ and ϵ' are known to belong to the set $(0, 1]$ and thus, for any fixed T, the term on the right-hand side of (5) cannot grow arbitrarily. Modern statistical literature present methods to adapt to any unknown quantity for which lower and upper bounds are known while controlling finite-time convergence (Lepskii, 1992), thus, to be adaptive w.r.t. unknown ϵ is an easier task than adapting to an unknown u. We can observe that by searching for the maximum value of the right-hand side of (5), we get that for $\epsilon = 1$ and $\epsilon' = \frac{1}{3}$ the gap's order is $\approx T^{\frac{1}{16}}$.

Any algorithm adaptive w.r.t. either u or ϵ has a higher regret lower bound than the one of the non-adaptive heavy-tailed bandit problem. We remark that the two bounds introduced in this section refers to adaptivity w.r.t. only one of the unknown quantities. It's unknown if simultaneous adaptivity on both implies an even higher lower bound, so we leave this as an open question for future investigations.

4 A Fully Adaptive Approach for Heavy-Tailed Bandits

In this section, we finally give an answer to our original research question, *i.e.* whether there's an algorithm adaptive w.r.t. both ϵ and u matching the standard setting's lower bound stated in (3). We already shown how adaptivity has a cost, and thus the lower bound presented in (3) is not achievable by any algorithm unaware of at least one of these quantities. Luckily, it is possible to restrict the set of adaptive heavy-tailed bandit problem instances under analysis to a special set, that will be defined in a short, on which our algorithm, namely `Adaptive Robust UCB` (shortly `AdaR-UCB`), is able to achieve a regret order matching the lower bound for the standard heavy-tailed bandit problem.

4.1 The Truncated Non-Positivity Assumption

We start by stating a key assumption, namely the *truncated non-positivity assumption*.

Assumption 1 (Truncated Non-Positivity)**.** *Given a set of K distributions satisfying* (2)*, let ν_1 be the distribution of the optimal arm, namely $\mu_1 \geq \mu_i \ \ \forall i \geq 1$, then*

$$\mathbb{E}_{\nu_1}[X \mathbb{1}_{\{|X|>M\}}] \leq 0 \quad \forall M \geq 0. \tag{6}$$

Some remarks are in order. This assumption requires the optimal arm of a heavy-tailed bandit instance to have more mass on the negative semi-axis, but still allows the distribution to have an arbitrary support covering, potentially, all $\mathbb{R}$. A similar version of this assumption, called *truncated non-negativity*, appeared in Huang et al. (2022) in the context of heavy-tailed bandits: the authors discuss the weak nature of this assumption comparing it to more stronger assumptions that are common in the literature. The two lower bounds (4) and (5) have been obtained by introducing two instances that violate this assumption (see Appendix [A.1,A.2]), and thus the lower bound on regret for the adaptive heavy-tailed bandit problem under the truncated non-positivity assumption can be smaller than the ones presented in the previous section. However, it is possible to show (see Appendix A.3) that forcing the truncated non-positivity assumption doesn't result in an improvement of the lower bound (3). In the next section, we show that under this assumption it is possible to be adaptive w.r.t. both ϵ and u while attaining the best regret order achievable in the heavy-tailed bandit problem.

4.2 A Fully Adaptive Algorithm: `Adaptive Robust UCB`

We are now ready to introduce Algorithm 1, namely `Adaptive Robust UCB` (shortly, `AdaR-UCB`), an algorithm able to operate in the heavy-tailed bandit problem *without any prior knowledge on ϵ nor u*. `AdaR-UCB` is an optimism in the face of uncertainty based algorithm, built upon the `Robust UCB` strategy from Bubeck et al. (2013a) using a modified version of the *trimmed mean estimator*. Trimmed mean is a common estimator in the heavy-tailed statistics literature, where observations are averaged while cutting-off values outside of a limited and bounded set of the form $[-M, M]$, thus being more robust w.r.t. empirical mean to extreme values. More formally, we define the trimmed mean estimator for the mean of a set of observations $\mathbf{X} = \{X_1, \dots X_s\}$ as

$$\hat{\mu}_s(\mathbf{X}) = \frac{1}{s} \sum_{j \in [s]} X_j \mathbb{1}_{\{|X_j| \leq M\}} \tag{7}$$

where $M > 0$ is a given threshold. In the `Robust UCB` algorithm, the trimmed mean estimator replaces sample average in a standard optimism in the face of uncertainty strategy, by selecting at each round t the action i maximising the sum of the estimator with a proper upper confidence bound. `AdaR-UCB` operates in the same way, but while in `Robust UCB` the threshold choice is driven by the values of ϵ and u, `AdaR-UCB` computes a proxy threshold $\widehat{M}$ for M without resorting to either ϵ or u (or any estimation of them).

`AdaR-UCB` operates over T rounds, however in Algorithm 1 we presented an interaction with the environment lasting only $\lfloor \frac{T}{2} \rfloor$ rounds. Indeed, for each round t, `AdaR-UCB` choose a single arm, but collects two rewards from it instead of one, that's the reason two different sets of collected rewards have been introduced: $\mathbf{X}_i$ and $\mathbf{X}'_i$ for each arm $i \in [K]$. The reason behind this choice lies in the fact that threshold $\widehat{M}_{i,s_i,t}$ and trimmed mean estimator $\widehat{\mu}_{i,s_i,t}$ need to be computed from independent samples of data. This design choice will ensure that the concentration inequalities built on both $\widehat{M}_{i,s_i,t}$ and $\widehat{\mu}_{i,s_i,t}$ to hold properly (see Appendix A.4 and A.5 for more details) at the cost of a 2 factor in the final regret of the algorithm.

Algorithm 1: `Adaptive Robust UCB (AdaR-UCB)`

Initialize $s_i \leftarrow 0$, $\mathbf{X}_i \leftarrow \emptyset$, $\mathbf{X}'_i \leftarrow \emptyset$, $\hat{\mu}_{i,0,1} \leftarrow +\infty \ \ \forall i \in [K]$.
for $t \in [\lfloor \frac{T}{2} \rfloor]$ **do**
  for $i \in [K]$ **do**
    Compute threshold $\widehat{M}_{i,s_i,t}$ solving

$$\frac{1}{s_i} \sum_{j \in [s]} \frac{\min\{(X'_{i,j})^2, \widehat{M}^2_{i,s_i,t}\}}{\widehat{M}^2_{i,s_i,t}} - \frac{9 \log(t^4)}{s_i} = 0$$

    Compute trimmed observations $\mathbf{Y_i} \leftarrow \{X_{i,1} \mathbb{1}_{\{|X_{i,1}| \leq \widehat{M}_{i,s_i,t}\}}, \dots, X_{i,s_i} \mathbb{1}_{\{|X_{i,s_i}| \leq \widehat{M}_{i,s_i,t}\}}\}$
    Compute trimmed mean estimator $\widehat{\mu}_{i,s_i,t}(\mathbf{X}_i) \leftarrow \frac{1}{s_i} \sum_{j \in [s_i]} Y_{i,j}$
  end
  Select an action

$$i_t \in \arg\max_{i \in [K]} \widehat{\mu}_{i,s_i,t} + 2\sqrt{\frac{V_{s_i}(\mathbf{Y_i}) \log(t^4)}{s_i}} + 19 \frac{\widehat{M}_{i,s_i,t} \log(t^4)}{s_i}$$

  Play action i_t and receive an observation X_t
  Update samples $\mathbf{X}_{i_t} \leftarrow \mathbf{X}_{i_t} \cup \{X_t\}$
  Play action i_t and receive an observation X'_t
  Update samples $\mathbf{X}'_{i_t} \leftarrow \mathbf{X}'_{i_t} \cup \{X'_t\}$
  Update number of pulls $s_{i_t} \leftarrow s_{i_t} + 1$
end

We start by stating the main theoretical result about `AdaR-UCB`, *i.e.* its upper bound on regret.

Theorem 4 (Upper Bound on Regret for `AdaR-UCB`)**.** *Given a heavy-tailed bandit problem instance satisfying Assumption 1, the regret of `AdaR-UCB` then satisfies:*

$$R_T \leq \sum_{i:\Delta_i > 0} \left(216 \left(\frac{36u}{\Delta_i} \right)^{\frac{1}{\epsilon}} \log T + 10\Delta_i \right). \tag{8}$$

First, we point out that *this result provides a positive answer to our initial research question*. The proof of Theorem 4 follows similar steps to the result provided by Bubeck et al. (2013a) concerning the upper bound on regret for `Robust UCB`. To be adaptive w.r.t. both ϵ and u brings additional difficulties in constructing the algorithm and proving its theoretical guarantees (see Appendix A.2 and A.1 for more details). In particular, in the next paragraphs, we introduce a new form for the concentration inequality of the trimmed mean estimator, which is explicitly dependent on the threshold choice but not on ϵ nor u. For this result, we discuss the role of Assumption 1, and finally provide an adaptive way to compute $\widehat{M}$.

Finally, as customary in the bandit literature, we also provide an instance-independent version of the upper bound on regret of `AdaR-UCB`.

Theorem 5 (Instance-Independent Upper Bound on Regret for `AdaR-UCB`)**.** *For any heavy-tailed bandit problem instance satisfying Assumption 1, the regret of `AdaR-UCB` then satisfies:*

$$R_T \leq 2T^{\frac{1}{1+\epsilon}} (216K \log T)^{\frac{\epsilon}{1+\epsilon}} (36u)^{\frac{1}{1+\epsilon}}. \tag{9}$$

4.3 Concentration Inequality for Trimmed Mean Estimator

In this section we state a concentration inequality for the trimmed mean estimator (7), which is explicitly dependent on the threshold value M. This concentration result is a key ingredient to prove the theoretical performances of our approach.

Theorem 6 (Concentration Inequality for Trimmed-Mean Estimator)**.** *Given a set of i.i.d. observations $\mathbf{X} = \{X_1, \dots, X_s\}$, and given a threshold $M > 0$, under Assumption 1 we get that for any given $\delta > 0$:*

$$\mathbb{P}\left(\mu - \widehat{\mu}_s(\mathbf{X}) \leq 2\sqrt{\frac{V_s(\mathbf{Y}) \log(\delta^{-1})}{s}} + 19 \frac{M \log(\delta^{-1})}{s} \right) \geq 1 - \delta \tag{10}$$

where $\mathbf{Y} = \{X_1 \mathbb{1}_{\{|X| \leq M\}}, \dots, X_s \mathbb{1}_{\{|X| \leq M\}}\}$ *is the trimmed version of* $\mathbf{X}$ *and* $V_s(\mathbf{Y})$ *is its sample variance.*

For the algorithm's execution we will set $\delta^{-1} = t^3$. The result above can be obtained by decomposing the gap between the true mean and the estimator in a bias-variance fashion by the means of the trimmed variable $\mathbf{Y}$ then, noting that under Assumption 1 bias can be neglected, by bounding the variance of $\mathbf{Y}$ (which is bounded by construction) using the well-known Empirical Bernstein bound (Maurer and Pontil, 2009). More details can be found in Appendix A.4. Theorem 6 show that trimmed mean estimator achieves a sub-Gaussian type concentration rate in function of M. This bound is used at line 8 of Algorithm 1 to compute the upper confidence bounds over the trimmed mean estimators, this allows `AdaR-UCB` to choose an action following the optimism in the face of uncertainty paradigm. Our goal is now to seek for a proper value of $\widehat{M}$ s.t. this concentration result is powerful enough for `AdaR-UCB` to achieve an upper bound on regret matching the lower bound (3), without requiring the knowledge of u nor ϵ for the threshold's construction.

4.4 Computing the Threshold

We now discuss the choice of the threshold M, showing that the chosen threshold $M_{s,t}$ depends on both the number of observations and the round t and thus being dynamic in time. In particular, following a procedure which is similar to the one presented in Wang et al. (2021), for each round $t \in [T]$ we compute the value of $\widehat{M}_{s,t}$ as the solution (in M) of

$$f_{s,t}(\mathbf{X}; M) = \frac{1}{s} \sum_{j \in [s]} \frac{\min\{X_j^2, M^2\}}{M^2} - \frac{9 \log(t^3)}{s} = 0. \tag{11}$$

In line 4 of Algorithm 1 we make use of this procedure to choose the thresholds for all available actions. Exploiting the fact that the first addendum is monotonically decreasing after some point, and under the assumption that $\sum_{i=1}^{s} \mathbb{1}_{|X_i|>0} > 9 \log\left(t^3\right)$, (11) admits a unique positive solution. A reader might notice that we are solving the sample version of this equation instead of the population version $\mathbb{E}_X[f_{s,t}(\mathbf{X}; M)] = 0$ (as in Wang et al. (2021)). The solution $\widetilde{M}_{s,t}$ to the population version of (11) satisfies the following inequality:

$$\widetilde{M}_{s,t} \leq \mathbb{E}_X[\min\{X^2, \widetilde{M}_{s,t}^2\}] \frac{s}{9 \log(t^3)} \leq \mathbb{E}_X[|X|^{1+\epsilon}] \widetilde{M}_{s,t}^{1-\epsilon} \frac{s}{9 \log(t^3)} = u \widetilde{M}_{s,t}^{1-\epsilon} \frac{s}{9 \log(t^3)} \tag{12}$$

and thus, by making explicit $\widetilde{M}_{s,t}$, we obtain the following bound:

$$\widetilde{M}_{s,t} \leq \left(\frac{us}{9 \log(t^3)} \right)^{\frac{1}{1+\epsilon}}. \tag{13}$$

We point out that the upper bound presented above is the value that authors choose as threshold for the trimmed mean `Robust UCB` algorithm in Bubeck et al. (2013a), however this choice of $\widehat{M}_{s,t}$ requires the knowledge of both u and ϵ, making their estimator unfeasible in the adaptive heavy-tailed bandit problem. It can be shown (see Appendix A.5) that the solution of the sample version can be bounded (in high probability) the same way as in (13), and thus allowing us to perform a proper analysis while dealing with this source of uncertainty.

5 Conclusions

In this paper we studied the *adaptive heavy-tailed bandit* problem, a variation on the classical heavy-tailed bandit problem where the no information is provided to the agent regarding the moments of the distribution, not even which of them are finite. The first results concern the intrinsic difficulty of the setting, for which two novel lower bounds have been provided. In particular, we proved that without any additional assumption no algorithm can match the performances of the non-adaptive setting. Finally, we provided a novel algorithm, namely `Adaptive Robust UCB`, that under a specific distributional assumption over the optimal arm is able to achieve the state-of-the-art performances of the standard heavy-tailed bandit problem. Future directions of investigation regard the role of the *truncated non-positivity* assumption: in particular, we wonder if is it possible to find a weaker assumption ensuring this kind of performances for an algorithm.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

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

Herbert Robbins. Some aspects of the sequential design of experiments. 1952.

Matteo Gagliolo and Jürgen Schmidhuber. Algorithm portfolio selection as a bandit problem with unbounded losses. *Annals of Mathematics and Artificial Intelligence*, 61:49–86, 2011.

Jörg Liebeherr, Almut Burchard, and Florin Ciucu. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2):1010–1024, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013a.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning*, pages 9173–9200. PMLR, 2022.

OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR, 2013b.

Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

A Proofs of Theoretical Results

In this section, we prove the main theoretical results outlined in the paper.

We start by proving the two novel lower bounds provided for the adaptive heavy-tailed bandit problem, namely Theorem 2 and Theorem 3.

A.1 Proof of Theorem 2

We start by constructing two heavy-tailed bandit instances with a common (and known to the algorithm) maximum order of moment ϵ, but where $u' > u$.

Base Instance

$$\begin{cases} \nu_1 = {\delta_0}^2, \\ \nu_2 = \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \delta_0 + \left(1 - \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}\right) \delta_{u^{\frac{1}{\epsilon}} \Delta^{-\frac{1}{\epsilon}}}, \end{cases} \tag{14}$$

where $\Delta \in (0, u^{\frac{1}{1+\epsilon}})$. Thus, we have $\mu_1 = 0$ and $\mu_2 = \Delta$. Furthermore, $\mathbb{E}_{\nu_1}[|X|^{1+\epsilon}] = 0$ and $\mathbb{E}_{\nu_2}[|X|^{1+\epsilon}] = u$. Therefore, optimal arm is arm 2.

Alternative instance

$$\begin{cases} \nu_1' = \left(1 - (2\Delta)^{1+\frac{1}{\epsilon}} (u')^{-\frac{1}{\epsilon}}\right) \delta_0 + (2\Delta)^{1+\frac{1}{\epsilon}} (u')^{-\frac{1}{\epsilon}} \delta_{(u')^{\frac{1}{\epsilon}} (2\Delta)^{-\frac{1}{\epsilon}}}, \\ \nu_2' = \nu_2, \end{cases} \tag{15}$$

where $\Delta \in (0, \frac{1}{2}(u')^{\frac{1}{1+\epsilon}})$. Thus we have $\mu_1' = 2\Delta$ and $\mu_2' = \Delta$. Furthermore, $\mathbb{E}_{\nu_1'}[|X|^{1+\epsilon}] = u'$ and $\mathbb{E}_{\nu_2'}[|X|^{1+\epsilon}] = u$. Therefore, optimal arm is arm 1.

Suppose that a matching, adaptive algorithms in u and u' exist. In such a case (see (3)), we will have that the expected regret of the base instance $R(T)$ is of order $(uT)^{\frac{1}{1+\epsilon}}$, while the regret of the alternative instance $R'(T)$ is of order $(u'T)^{\frac{1}{1+\epsilon}}$. Thus, the following is satisfied for a constant c that does not depend on T:

$$\max\left\{\frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \leq c,$$

We will prove that this is not the case and, specifically, that for any algorithm:

$$\max\left\{\frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq f(T, \epsilon, u, u'),$$

being f a function increasing in T. This suffices to show the inexistence of an algorithm adaptive in u matching the minimax lower bound (3).

The proof is quite technical and merges the approach of (Bubeck et al., 2013b, Theorem 5) with that of (Lattimore and Szepesvári, 2020, Chapters 14.2, 14.3).

First, we observe that:

$$\max\left\{\frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq \frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}} = \frac{\Delta \mathbb{E}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}}, \tag{16}$$

where $\mathbb{E}[N_1(T)]$ is the expected number of times arm 1 is pulled over the horizon T.

Second, recalling which are the optimal arms in the two instances and that $u' > u$, we have:

$$\begin{aligned} &\max\left\{\frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq \\ &\qquad \geq (u'T)^{-\frac{1}{\epsilon+1}} \max\left\{\frac{\Delta T}{2}\mathbb{P}\left(N_1(T) \geq T/2\right), \frac{\Delta T}{2}\mathbb{P}'\left(N_1(T) < T/2\right)\right\} \\ &\qquad \geq \frac{\Delta}{4}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \left(\mathbb{P}\left(N_1(T) \geq T/2\right) + \mathbb{P}'\left(N_1(T) < T/2\right)\right) \\ &\qquad \geq \frac{\Delta}{8}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \exp\left(-\mathbb{E}[N_1(T)] D_{KL}(\nu_1 \| \nu_1')\right). \end{aligned} \tag{17}$$

where we used Bretagnolle-Huber inequality and divergence decomposition, together with $\max\{a,b\} \geq \frac{1}{2}(a+b)$ for $a,b \geq 0$. Let us now compute the KL-divergence, noting that $\nu_1 \ll \nu_1'$:

$$\begin{aligned} D_{KL}(\nu_1 \| \nu_1') &= \nu_1(0) \log \frac{\nu_1(0)}{\nu_1'(0)} \\ &= \log \frac{1}{1-(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}}} \leq c(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}}, \end{aligned} \tag{18}$$

for $\Delta \in (0, \left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}} (u')^{\frac{1}{1+\epsilon}})$ and some c to be calculated (we can take $c=2$).

Putting together Equations (16), (17) and (18), we have:

$$\begin{aligned} &\max\left\{ \frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{(u'T)^{\frac{1}{1+\epsilon}}} \right\} \geq \\ &\quad \geq \max\left\{ \frac{\Delta \mathbb{E}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{\Delta}{8}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \exp\left(-c\mathbb{E}[N_1(T)](2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right) \right\} \\ &\quad \geq \frac{\Delta}{2}\left(\frac{\mathbb{E}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}} + \frac{1}{8}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \exp\left(-c\mathbb{E}[N_1(T)](2\Delta)^{\frac{1+\epsilon}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right) \right) \\ &\quad \geq \frac{\Delta}{2} \min_{x \in [0,T]} \left\{ \frac{x}{(uT)^{\frac{1}{1+\epsilon}}} + \frac{1}{8}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \exp\left(-cx(2\Delta)^{\frac{1+\epsilon}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right) \right\} =: g(x) \end{aligned}$$

The latter is a convex function of x and the minimization can be carried out in closed form vanishing the derivative choosing x^* s.t.

$$\frac{1}{(uT)^{\frac{1}{1+\epsilon}}} - \frac{c}{8}(u')^{-\left(\frac{1}{\epsilon+1}+\frac{1}{\epsilon}\right)} T^{\frac{\epsilon}{\epsilon+1}} (2\Delta)^{\frac{1+\epsilon}{\epsilon}} \exp\left(-cx^*(2\Delta)^{\frac{1+\epsilon}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right) = 0,$$

with simple calculations, we obtain

$$x^* = c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}} \log\left(\frac{T u^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}} c(2\Delta)^{\frac{1+\epsilon}{\epsilon}}} \right),$$

which leads to:

$$\begin{aligned} g(x^*) &= \frac{\Delta}{2}(uT)^{-\frac{1}{\epsilon+1}} c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}} \log\left(\frac{T u^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}} c(2\Delta)^{\frac{1+\epsilon}{\epsilon}}} \right) + \\ &\quad + \frac{\Delta}{2}\frac{1}{8}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \frac{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}{T u^{\frac{1}{\epsilon+1}}} c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}} \\ &= \frac{\Delta}{2}(uT)^{-\frac{1}{\epsilon+1}} c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}} \left[\log\left(\frac{T u^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}} c(2\Delta)^{\frac{1+\epsilon}{\epsilon}}} \right) + 1 \right] \\ &= \frac{\Delta}{2}(uT)^{-\frac{1}{\epsilon+1}} c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}} \log\left(\frac{T u^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}} ec(2\Delta)^{\frac{1+\epsilon}{\epsilon}}} \right). \end{aligned}$$

We take Δ (which is $< \left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}} (u')^{\frac{1}{1+\epsilon}}$ for sufficiently large T) as:

$$\frac{T u^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}} c(2\Delta)^{\frac{1+\epsilon}{\epsilon}}} = e^{\epsilon},$$

resulting in:

$$\Delta = 2^{\frac{2\epsilon-1}{1+\epsilon}} e^{\frac{\epsilon^2}{1+\epsilon}} (cT)^{-\frac{\epsilon}{\epsilon+1}} u^{-\frac{\epsilon}{(\epsilon+1)^2}} (u')^{\frac{1+2\epsilon}{(\epsilon+1)^2}}.$$

This imply, after some calculations, that:

$$g(x^*) = c^{-\frac{\epsilon}{\epsilon+1}} 2^{-\frac{2\epsilon+5}{\epsilon+1}} (1+\epsilon) e^{-\frac{\epsilon}{\epsilon+1}} u^{-\frac{\epsilon}{(\epsilon+1)^2}} (u')^{\frac{\epsilon}{(\epsilon+1)^2}} \geq C_1 \cdot \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}},$$

where C_1 is a value independent of T and both u and u'. Finally, we have that

$$\max\left\{\frac{R(T)}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq C_1 \cdot \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}}.$$

Since $u' > u$ can be taken arbitrarily large, we have that the right-hand side of this inequality can be arbitrarily large. We conclude the proof by noting that, when $\epsilon = 0$, the gap vanishes: however, this is irrelevant since for this scenario no sublinear regret can be achieved by any bandit algorithm.

□

A.2 Proof of Theorem 3

We now prove the lower bound for adaptivity w.r.t. ϵ.

We start by constructing two heavy-tailed bandit instances with different maximum orders of moment ϵ and ϵ', where $0 < \epsilon' < \epsilon < 1$. For the sake of simplicity, but without loss of generality, we will assume a common (and known to the algorithm) maximum moment of $u = 1$.

Base Instance

$$\begin{cases} \nu_1 = \delta_0, \\ \nu_2 = (1 + \Delta\gamma - \gamma^{1+\epsilon})\delta_0 + (\gamma^{1+\epsilon} - \Delta\gamma)\delta_{1/\gamma}, \end{cases} \tag{19}$$

where $\Delta \in [0, \frac{1}{2}]$ and $\gamma = (2\Delta)^{\frac{1}{\epsilon}}$. Thus, we have $\mu_1 = 0$ and $\mu_2 = \Delta$. Furthermore, $\mathbb{E}_{\nu_1}[|x|^\alpha] = 0$ and $\mathbb{E}_{\nu_2}[|x|^\alpha] = 2^{\frac{1-\alpha}{\epsilon}}\Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}$, which are guaranteed to be bounded by constant only if $\alpha \leq \epsilon + 1$. Thus, this instance admits moments finite only up to order $\epsilon + 1$. Therefore, optimal arm is arm 2.

Alternative instance

$$\begin{cases} \nu_1' = (1 - (\gamma')^{1+\epsilon'})\delta_0 + (\gamma')^{1+\epsilon'}\delta_{1/\gamma'}, \\ \nu_2' = \nu_2, \end{cases} \tag{20}$$

where $\Delta \in [0, \frac{1}{2}]$ and $\gamma = (2\Delta)^{\frac{1}{\epsilon'}}$. Thus, we have $\mu_1 = 2\Delta$ and $\mu_2 = \Delta$. Furthermore, $\mathbb{E}_{\nu_1}[|x|^\alpha] = (2\Delta)^{\frac{1+\epsilon'-\alpha}{\epsilon'}}$ and $\mathbb{E}_{\nu_2}[|x|^\alpha] = 2^{\frac{1-\alpha}{\epsilon}}\Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}$, which are guaranteed to be bounded by constant only if $\alpha \leq \epsilon' + 1$. Thus, this instance admits moments finite only up to order $\epsilon' + 1$. Therefore, optimal arm is arm 1.

Suppose that a matching, adaptive algorithms in ϵ and ϵ' exist. In such a case (see (3)), we will have that the expected regret of the base instance $R(T)$ is of order $T^{\frac{1}{1+\epsilon}}$, while the regret of the alternative instance $R'(T)$ is of order $T^{\frac{1}{1+\epsilon'}}$. Thus, the following is satisfied for a constant c that does not depend on T:

$$\max\left\{\frac{R(T)}{T^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{T^{\frac{1}{1+\epsilon'}}}\right\} \leq c,$$

We will prove that this is not the case and, specifically, that for any algorithm:

$$\max\left\{\frac{R(T)}{T^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq f(T, \epsilon, \epsilon'),$$

being f a function increasing in T. This suffices to show the non-existence of an algorithm adaptive in ϵ matching the minimax lower bound (3).

The proof is quite technical and emulates the analyses and steps performed to prove Theorem 2.

First, we observe that:

$$\max\left\{\frac{R(T)}{T^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq \frac{R(T)}{T^{\frac{1}{1+\epsilon}}} = \frac{\Delta\mathbb{E}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}}, \tag{21}$$

where $\mathbb{E}[N_1(T)]$ is the expected number of times arm 1 is pulled over the horizon T.

Second, recalling which are the optimal arms in the two instances and that $\epsilon' < \epsilon$, we have:

$$\begin{aligned}
\max\left\{\frac{R(T)}{T^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq \\
&\geq T^{-\frac{1}{\epsilon'+1}} \max\left\{\frac{\Delta T}{2}\mathbb{P}\left(N_1(T) \geq \frac{T}{2}\right), \frac{\Delta T}{2}\mathbb{P}'\left(N_1(T) < \frac{T}{2}\right)\right\} \\
&\geq \frac{\Delta}{4} T^{\frac{\epsilon'}{\epsilon'+1}} \left(\mathbb{P}\left(N_1(T) \geq \frac{T}{2}\right) + \mathbb{P}'\left(N_1(T) < \frac{T}{2}\right)\right) \\
&\geq \frac{\Delta}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \exp\left(-\mathbb{E}[N_1(T)] D_{KL}(\nu_1 \| \nu_1')\right).
\end{aligned} \tag{22}$$

where we used Bretagnolle-Huber inequality and divergence decomposition, together with $\max\{a, b\} \geq \frac{1}{2}(a+b)$ for $a, b \geq 0$. Let us now compute the KL-divergence, noting that $\nu_1 \ll \nu_1'$:

$$\begin{aligned}
D_{KL}(\nu_1 \| \nu_1') &= \nu_1(0) \log \frac{\nu_1(0)}{\nu_1'(0)} \\
&= \log \frac{1}{1 - (2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}} \leq c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}},
\end{aligned} \tag{23}$$

for $\Delta \in [0, 1/4]$ and some c to be calculated (we can take $c = 2$).

Putting together Equations (21), (22) and (23), we have:

$$\begin{aligned}
\max\left\{\frac{R(T)}{T^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{T^{\frac{1}{1+\epsilon'}}}\right\} &\geq \max\left\{\frac{\Delta \mathbb{E}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}}, \frac{\Delta}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \exp\left(-c\mathbb{E}[N_1(T)](2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right\} \\
&\geq \frac{\Delta}{2}\left(\frac{\mathbb{E}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}} + \frac{1}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \exp\left(-c\mathbb{E}[N_1(T)](2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right) \\
&\geq \frac{\Delta}{2} \min_{x \in [0,T]}\left\{\frac{x}{T^{\frac{1}{1+\epsilon}}} + \frac{1}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \exp\left(-cx(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right\} =: g(x).
\end{aligned}$$

The latter is a convex function of x and the minimization can be carried out in closed form vanishing the derivative choosing x^* s.t.

$$\frac{1}{T^{\frac{1}{1+\epsilon}}} - \frac{c}{8} T^{\frac{\epsilon'}{\epsilon'+1}} (2\Delta)^{\frac{1+\epsilon'}{\epsilon'}} \exp\left(-cx^*(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right) = 0,$$

with simple calculations, we obtain

$$x^* = c^{-1}(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \log\left(\frac{T^{\frac{1}{\epsilon+1} + \frac{\epsilon'}{1+\epsilon'}}}{8} c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right),$$

which leads to:

$$\begin{aligned}
g(x^*) &= \frac{\Delta}{2} T^{-\frac{1}{\epsilon+1}} c^{-1}(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \log\left(\frac{T^{\frac{1}{\epsilon+1} + \frac{\epsilon'}{1+\epsilon'}}}{8} c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right) + \\
&\quad + \frac{\Delta}{2}\frac{1}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \left[\frac{T^{\frac{1}{\epsilon+1} + \frac{\epsilon'}{1+\epsilon'}}}{8} c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right]^{-1} \\
&= \frac{\Delta}{2} T^{-\frac{1}{\epsilon+1}} c^{-1}(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \left[\log\left(\frac{T^{\frac{1}{\epsilon+1} + \frac{\epsilon'}{1+\epsilon'}}}{8} c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right) + 1\right] \\
&= \frac{\Delta}{2} T^{-\frac{1}{\epsilon+1}} c^{-1}(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \log\left(\frac{T^{\frac{1}{\epsilon+1} + \frac{\epsilon'}{1+\epsilon'}}}{8} ec(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)
\end{aligned}$$

We take Δ (which is $< 1/4$ for sufficiently large T) as

$$\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{8} c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}} = 1,$$

resulting in:

$$\Delta = 2^{\frac{2\epsilon'-1}{1+\epsilon'}} c^{-\frac{\epsilon'}{1+\epsilon'}} T^{-\frac{\epsilon'}{1+\epsilon'}\left(\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}\right)}.$$

This imply, after some calculations, that:

$$\begin{aligned} g(x^*) &= 2^{\left(\frac{2\epsilon'-1}{1+\epsilon'}-1-3\right)} c^{\left(-\frac{\epsilon'}{1+\epsilon'}-1+1\right)} T^{\left[-\frac{\epsilon'}{1+\epsilon'}\left(\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}\right)+\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}-\frac{1}{\epsilon+1}\right]} \\ &= 2^{\frac{-2\epsilon'-5}{1+\epsilon'}} c^{-\frac{\epsilon'}{1+\epsilon'}} T^{-\frac{\epsilon'}{1+\epsilon'}\left(\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}-1\right)} \\ &= 2^{\frac{-2\epsilon'-5}{1+\epsilon'}} c^{-\frac{\epsilon'}{1+\epsilon'}} T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}} \geq C_2 \cdot T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}}. \end{aligned}$$

where C_2 is a value independent on T. Finally, we have that:

$$\max\left\{\frac{R(T)}{T^{\frac{1}{1+\epsilon}}}, \frac{R'(T)}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq C_2 \cdot T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}}.$$

We conclude the proof by noting that the dependence on T vanishes when $\epsilon' = 0$. This is correct, since even when knowing that $\epsilon' = 0$ the regret lower bound is necessarily linear. Thus, we can just focus on ϵ.

□

A.3 Regret Lower Bound under Assumption 1

One may wonder whether enforcing the truncated non-positive assumption (Assumption 1) leads to a strictly smaller lower bound on the regret in the heavy-tailed bandit problem. In this section, we show that this is not the case.

We employ the same construction as in Lattimore and Szepesvári (2020), but replace the Gaussian distributions with mixtures of Dirac deltas. We define:

$$\nu_\Delta = \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \delta_0 + \left(1 - \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}\right) \delta_{-u^{\frac{1}{\epsilon}} \Delta^{-\frac{1}{\epsilon}}}, \tag{24}$$

The two instances are constructed by the means of (24).

Base Instance μ

$$\begin{cases} \nu_1 = \nu_{2\Delta}, \\ \nu_j = \nu_{3\Delta}, \qquad j \neq 1. \end{cases}$$

Alternative Instance μ'

$$\begin{cases} \nu'_1 = \nu_{2\Delta}, \\ \nu'_i = \nu_{\Delta}, \\ \nu'_j = \nu_{3\Delta}, \qquad j \neq 1, i, \end{cases}$$

where $i \in \text{argmin}_{j \neq 1} \mathbb{E}_\mu[N_j(T)]$.

Both instances satisfy Assumption 1.

We have:

$$R(T) + R'(T) \geq \frac{\Delta T}{2}\left(\mathbb{P}_\mu\left(N_1 \leq \frac{T}{2}\right) + \mathbb{P}_{\mu'}\left(N_1 > \frac{T}{2}\right)\right)$$

Using the Bretagnolle-Huber inequality, we get

$$R(T) + R'(T) \geq \frac{\Delta T}{2} \exp\left(-\frac{1}{2} D_{KL}(\mu||\mu')\right).$$

We develop the Kullback-Leibler divergence between the two instances:

$$\begin{aligned}
D_{KL}(\mu||\mu') &= \sum_{j=1}^{k} \mathbb{E}[N_j(T)] D_{KL}(\nu_j||\nu'_j) \\
&= \mathbb{E}[N_i(T)] D_{KL}(\nu_i||\nu'_i) \\
&\overset{(*)}{\leq} \frac{T}{k-1} D_{KL}(\nu_i||\nu'_i) \\
&= \frac{T}{k-1}\left((3\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \log\left(\frac{(3\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}}{(\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}}\right) + \right. \\
&\qquad \left. + (1-(3\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}) \log\left(\frac{1-(3\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}}{1-(\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}}\right)\right) \\
&\leq \frac{T}{k-1}(3\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \log(3^{1+\frac{1}{\epsilon}}),
\end{aligned}$$

where the step marked by $(*)$ follows from the fact that i is the least pulled arm in instance μ.

Plugging this result, we finally get:

$$\begin{aligned}
R(T) + R'(T) &\geq \frac{\Delta T}{2} \exp\left(-\frac{1}{2} D_{KL}(\mu||\mu')\right) \\
&\geq \frac{\Delta T}{2} \exp\left(-\frac{1}{2}\frac{T}{k-1}(3\Delta)^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \log(3^{1+\frac{1}{\epsilon}})\right).
\end{aligned}$$

We conclude the proof by noting that $\max\{x, y\} > \frac{1}{2}(x+y)$ and setting $\Delta = \frac{1}{2}\left(\frac{k-1}{T} u^{\frac{1}{\epsilon}} \frac{1}{\log(3^{1+\frac{1}{\epsilon}})}\right)^{\frac{\epsilon}{1+\epsilon}}$.

Finally, we have:

$$\max\{R(T), R'(T)\} \geq c u^{\frac{1}{\epsilon}} T^{\frac{1}{1+\epsilon}}$$

for some constant c independent of T.

This proves that even under Assumption 1 is not possible to further improve the regret lower bound from Bubeck et al. (2013a).

A.4 Proof of Theorem 6

With probability at least $1-\delta$:

$$\begin{aligned}
\mu - \widehat{\mu}_s(\mathbf{X}) &= \mathbb{E}[X_1] - \frac{1}{s}\sum_{t=1}^{s} X_t \mathbb{1}_{|X_t| \leq M_{s,t}} \\
&= \frac{1}{n}\sum_{t=1}^{n} \left(\mathbb{E}[X_1] - \mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M_{s,t}}\right]\right) + \frac{1}{n}\sum_{t=1}^{n} \left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M_{s,t}}\right] - X_t \mathbb{1}_{|X_t| \leq M_{s,t}}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[X_t \mathbb{1}_{|X_t| > M_{s,t}}] + \frac{1}{s}\sum_{t=1}^{s} \left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M_{s,t}}\right] - X_t \mathbb{1}_{|X_t| \leq M_{s,t}}\right) \\
&\overset{(*)}{\leq} \frac{1}{s}\sum_{t=1}^{s} \left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M_{s,t}}\right] - X_t \mathbb{1}_{|X_t| \leq M_{s,t}}\right) \\
&\overset{(**)}{\leq} \sqrt{\frac{2V_s(\mathbf{Y})\log(2\delta^{-1})}{s}} + \frac{14 M_{s,t}\log(2\delta^{-1})}{3(s-1)} \\
&\leq \sqrt{\frac{4V_s(\mathbf{Y})\log(\delta^{-1})}{s}} + \frac{56 M_{s,t}\log(\delta^{-1})}{3s}
\end{aligned}$$

Note that in step $(*)$ we used Assumption 1 to make the first term vanish. In step $(**)$, instead, we used Empirical Bernstein Inequality (Maurer and Pontil, 2009). We also use the facts that $\log(2\delta^{-1}) \leq 2\log(\delta^{-1})$ for $\delta \leq \frac{1}{2}$ and $\frac{1}{s-1} \leq \frac{2}{s}$ in the last step.

□

A.5 Properties of Empirical Adaptive Threshold $\widehat{M}_{s,t}$

In Wang et al. (2021), the authors provide an upper bound on the solution $\widetilde{M}_{s,t}$ to the population version of (11)[3]:

$$\frac{\mathbb{E}[\min\{X_1^2, M^2\}]}{M^2} - \frac{\log(\delta^{-1})}{s} = 0. \tag{25}$$

The result is reported in (13): unfortunately this bound is unknown in the adaptive heavy-tailed bandit problem, since it includes both ϵ and u. However, neither computing the solution of (25) in feasible, since it would require the knowledge of an expected value, which is a theoretical quantity. Furthermore, approximating the expected value with a sample mean would not have a straightforward effect on the equation's solution. The goal of this section is to study the behavior of the solution to the sample version of (25), namely (11).

In particular, we will provide an high-probability upper bound on its value. This result is particularly important to prove Theorem 4, since it lets us provide another concentration inequality (which is dependent on both ϵ and u, but independent on $\widehat{M}_{s,t}$) on the trimmed mean estimator.

We start by recalling an important result from Wang et al. (2021).

Proposition 7 (Uniqueness of Solution for (11))**.** *Let $\mathbf{X} = (X_1, \dots, X_s)$ be independent and identically distributed random variables, and $\delta > 0$, if*

$$0 < 9\log\left(\delta^{-1}\right) < \sum_{i=1}^{s} \mathbb{1}_{|X_i|>0}, \tag{26}$$

then Equation (11) *admits a unique positive solution.*

This result is crucial to establish a well-posed computation of the empirical threshold $\widehat{M}_{s,t}$, since it excludes ambiguities.

However, the most important goal, is to bound $\widehat{M}_{s,t}$ the most similarly to $\widetilde{M}_{s,t}$. Luckily, the following result states that, in fact, the solution to (11) can be bounded the same way as the solution to (25).

Theorem 8 (Bounds for $\widehat{M}_{s,t}$)**.** *Let $\mathbf{X} = (X_1, \dots, X_s)$ be independent and identically distributed random variables satisfying* (2)*, and let $\widehat{M}_{s,t}$ be the (random) solution of* (11)*, then we have that, with probability at least $1 - \delta$:*

$$\widehat{M}_{s,t} \leq \left(\frac{us}{\log(\delta^{-1})}\right)^{\frac{1}{1+\epsilon}}. \tag{27}$$

and:

$$\mathbb{P}\left(|X_1| > \widehat{M}_{s,t}\right) \leq \frac{49\log(\delta^{-1})}{s}. \tag{28}$$

Proof. This proof makes use of the concentration inequality for self-normalizing random variable (Maurer, 2006; Maurer and Pontil, 2009).

Let $\mathbf{X} = (X_1, \dots, X_s)$ be independent and identically distributed random variables satisfying (2), and $M > 0$, we then introduce

$$U_{i,M} \coloneqq \min\left\{\left(\frac{X_i}{M}\right)^2, 1\right\}$$

[3]In (11) we set $\delta^{-1} = t^3$, from now on we will refer to it considering the version with a generic $\delta^{-1} > 2$.

random variable in $[0, 1]$. Furthermore, let

$$Z_M(\mathbf{X}) = \sum_{i=1}^{s} U_{i,M},$$

random variable in $[0, s]$ and such that $U_{i,M} = \frac{1}{s} Z_M(\mathbf{X})$ for all $i \in [s]$.

We start by showing that $Z_M(\mathbf{X})$ satisfies the assumptions of Theorem 13 from Maurer (2006), in particular, let $a \geq 1$:

$$Z_M(\mathbf{X}) - \inf_{y \in \mathbb{R}} Z_M(\mathbf{X}_{y,k}) \leq 1 \quad \forall k \in [s], \tag{29}$$

$$\sum_{k=1}^{s} \left(Z_M(\mathbf{X}) - \inf_{y \in \mathbb{R}} Z_M(\mathbf{X}_{y,k}) \right)^2 \leq a Z_M(\mathbf{X}), \tag{30}$$

where $\mathbf{X}_{y,k}$ is obtained by substituting y to the k-th element of $\mathbf{X}$.

Equation (29) follows from

$$Z_M(\mathbf{X}) - \inf_{y \in \mathbb{R}} Z_M(\mathbf{X}_{y,k}) = U_{k,M} - \inf_{y \in \mathbb{R}} \min \left\{ \left(\frac{y}{M} \right)^2, 1 \right\} = U_{k,M} \leq 1 \quad \forall k \in [s].$$

Similarly, we set $a = 1$ and obtain (30) as follows:

$$\begin{aligned}
\sum_{k=1}^{s} \left(Z_M(\mathbf{X}) - \inf_{y \in \mathbb{R}} Z_M(\mathbf{X}_{y,k}) \right)^2 &= \sum_{k=1}^{s} \left(U_{k,M} - \inf_{y \in \mathbb{R}} \min \left\{ \left(\frac{y}{M} \right)^2, 1 \right\} \right)^2 \\
&\leq \sum_{k=1}^{s} U_{k,M}^2 \\
&\leq \sum_{k=1}^{s} U_{k,M} \\
&= Z_M(\mathbf{X}).
\end{aligned}$$

Using Theorem 13 from Maurer (2006) with $a = 1$, for the right tail of the distribution, the following holds:

$$\mathbb{P}\left(\mathbb{E}[Z_M(\mathbf{X})] - Z_M(\mathbf{X}) > s\epsilon \right) \leq \exp\left(\frac{-\epsilon^2 s^2}{2\mathbb{E}[Z_M(\mathbf{X})]} \right)$$

which implies

$$\mathbb{P}\left(\mathbb{E}[U_{i,M}] - U_{i,M} > \epsilon \right) \leq \exp\left(\frac{-\epsilon^2 s^2}{2s\mathbb{E}[U_{i,M}]} \right) = \exp\left(\frac{-\epsilon^2 s}{2\mathbb{E}[U_{i,M}]} \right) \quad \forall i \in [s].$$

By letting $\epsilon \coloneqq 2\sqrt{\frac{2\mathbb{E}[U_{i,M}] \log(2\delta^{-1})}{s}}$ for all $i \in [s]$, we get that

$$\mathbb{P}\left(\mathbb{E}[U_{i,M}] - U_{i,M} > \sqrt{\frac{2\mathbb{E}[U_{i,M}] \log(2\delta^{-1})}{s}} \right) \leq \frac{\delta}{2} \quad \forall i \in [s],$$

which implies

$$\mathbb{P}\left(\sqrt{\mathbb{E}[U_{i,M}]} - \sqrt{U_{i,M}} > 2\sqrt{\frac{\log(2\delta^{-1})}{s}} \right) \leq \frac{\delta}{2} \quad \forall i \in [s].$$

A similar inequality holds for the left tail:

$$\mathbb{P}\left(Z_M(\mathbf{X}) - \mathbb{E}[Z_M(\mathbf{X})] > s\epsilon \right) \leq \exp\left(\frac{-\epsilon^2 s^2}{2\mathbb{E}[Z_M(\mathbf{X})] + \epsilon s} \right),$$

with similar steps, we get

$$\mathbb{P}\left(\sqrt{U_{i,M}} - \sqrt{\mathbb{E}[U_{i,M}]} > 2\sqrt{\frac{\log(\delta^{-1})}{s}}\right) \leq \frac{\delta}{2} \quad \forall i \in [s].$$

Using a union bound over the two inequalities on the left and the right tail, we finally get

$$\mathbb{P}\left(\left|\sqrt{\mathbb{E}[U_{i,M}]} - \sqrt{U_{i,M}}\right| > 2\sqrt{\frac{\log(\delta^{-1})}{s}}\right) \leq \delta \quad \forall i \in [s]. \tag{31}$$

Let us now define $\widehat{M}$ random variable such that

$$U_{i,\widehat{M}} = \frac{c\log(\delta^{-1})}{s} \quad \forall i \in [s],$$

where $c > 0$.

We now plug $U_{i,\widehat{M}}$ in (31) and get that, with probability at least $1-\delta$, and for all $i \in [K]$,

$$\begin{aligned}
\sqrt{\frac{c\log(\delta^{-1})}{s}} &= \sqrt{U_{i,\widehat{M}}} \\
&\geq \sqrt{\mathbb{E}\left[U_{i,\widehat{M}}\right]} - 4\sqrt{\frac{\log(\delta^{-1})}{s}} \\
&\geq \sqrt{\mathbb{P}\left(|X_i| > \widehat{M}\right)} - 4\sqrt{\frac{\log(\delta^{-1})}{s}},
\end{aligned}$$

which implies that, with probability at least $1-\delta$:

$$\sqrt{\mathbb{P}\left(|X_i| > \widehat{M}\right)} \leq (\sqrt{c}+4)^2\sqrt{\frac{\log(\delta^{-1})}{s}} \quad \forall i \in [s].$$

On the other side, plugging $U_{i,\widehat{M}}$ in (31) also yields that for every $i \in [K]$:

$$\begin{aligned}
\sqrt{\frac{c\log(\delta^{-1})}{s}} &= \sqrt{U_{i,\widehat{M}}} \\
&\leq \sqrt{\mathbb{E}\left[U_{i,\widehat{M}}\right]} + 4\sqrt{\frac{\log(\delta^{-1})}{s}} \\
&\leq \sqrt{\frac{u}{\widehat{M}^{1+\epsilon}}} + 4\sqrt{\frac{\log(\delta^{-1})}{s}},
\end{aligned}$$

which implies

$$\widehat{M} \leq \left(\frac{us}{(\sqrt{c}-4)^2\log(\delta^{-1})}\right)^{\frac{1}{1+\epsilon}}.$$

Choosing $c = 9$ concludes the proof. □

A.6 Threshold-Independent Concentration Inequality for Trimmed Mean Estimator

In this section we provide another concentration inequality for the trimmed mean estimator defined in (7). Differently from the concentration inequality (10), this result is independent on the chosen threshold, but instead have an explicit dependency on both ϵ and u. This result will play a key role in the regret analysis of `Adaptive Robust UCB`.

Theorem 9 (Threshold-Independent Concentration Inequality for Trimmed-Mean Estimator)**.** *Given a set of i.i.d. observations* $\mathbf{X} = \{X_1, \dots, X_s\}$ *satisfying* (2)*, the following holds:*

$$\mathbb{P}\left(\widehat{\mu}_s(\mathbf{X}) - \mu \leq 18u^{\frac{1}{1+\epsilon}}\left[\frac{\log(\delta^{-1})}{s}\right]^{\frac{\epsilon}{1+\epsilon}}\right) \geq 1-\delta. \tag{32}$$

Proof. With probability at least $1 - \frac{\delta}{2}$:

$$\begin{aligned}
\widehat{\mu}_s(\mathbf{X}) - \mu &= \frac{1}{s}\sum_{i=1}^{s} X_i \mathbb{1}_{|X_i| \le M_{s,t}} - \mathbb{E}[X_1] \\
&= \frac{1}{s}\sum_{i=1}^{s} \left(X_i \mathbb{1}_{|X_i| \le M_{s,t}} - \mathbb{E}\left[X_i \mathbb{1}_{|X_i| \le M_{s,t}}\right]\right) - \frac{1}{s}\sum_{i=1}^{s} \left(\mathbb{E}[X_1] - \mathbb{E}\left[X_t \mathbb{1}_{|X_i| \le M_{s,t}}\right]\right) \\
&= \frac{1}{s}\sum_{i=1}^{s} \left(X_i \mathbb{1}_{|X_i| \le M_{s,t}} - \mathbb{E}\left[X_i \mathbb{1}_{|X_i| \le M_{s,t}}\right]\right) - \frac{1}{s}\sum_{i=1}^{s} \mathbb{E}[X_i \mathbb{1}_{|X_i| > M_{s,t}}] \\
&\le \frac{1}{s}\sum_{i=1}^{s} \left(X_i \mathbb{1}_{|X_i| \le M_{s,t}} - \mathbb{E}\left[X_i \mathbb{1}_{|X_i| \le M_{s,t}}\right]\right) + \frac{1}{s}\sum_{i=1}^{s} \mathbb{E}[|X_i| \mathbb{1}_{|X_i| > M_{s,t}}] \\
&\overset{(*)}{\le} \frac{1}{s}\sum_{i=1}^{s} \left(X_i \mathbb{1}_{|X_i| \le M_{s,t}} - \mathbb{E}\left[X_i \mathbb{1}_{|X_i| \le M_{s,t}}\right]\right) + \\
&\qquad + \frac{1}{s}\sum_{i=1}^{s} \left(\mathbb{E}\left[|X_i|^{1+\epsilon}\right]^{\frac{1}{1+\epsilon}}\right) \left(\mathbb{E}\left[\left(\mathbb{1}_{|X_i| > M_{s,t}}\right)^{\frac{1+\epsilon}{\epsilon}}\right]^{\frac{\epsilon}{1+\epsilon}}\right) \\
&\overset{(**)}{\le} \sqrt{\frac{2M_{s,t}^{1-\epsilon} u \log\left(2\delta^{-1}\right)}{s}} + \frac{M_{s,t} \log\left(2\delta^{-1}\right)}{3s} + \frac{1}{s}\sum_{i=1}^{s} \left(u^{\frac{1}{1+\varepsilon}}\right) \left(\mathbb{E}\left[\mathbb{1}_{|X_i| > M_{s,t}}\right]^{\frac{\epsilon}{1+\varepsilon}}\right) \\
&\le \sqrt{\frac{2M_{s,t}^{1-\epsilon} u \log\left(2\delta^{-1}\right)}{s}} + \frac{M_{s,t} \log\left(2\delta^{-1}\right)}{3s} + u^{\frac{1}{1+\varepsilon}} \left(\frac{1}{s}\sum_{i=1}^{s} \mathbb{P}\left(|X_i| > M_{s,t}\right)^{\frac{\epsilon}{1+\varepsilon}}\right).
\end{aligned}$$

Step $(*)$ follows from Hölder inequality, while step $(**)$ is a consequence of Bernstein Inequality for bounded random variables.

To proceed further, we make use of Theorem 8 with a confidence level of $\frac{\delta}{2}$.

In particular, with probability at least $1 - \delta$ (by union bound):

$$\begin{aligned}
\widehat{\mu}_s(\mathbf{X}) - \mu &\le \sqrt{\frac{2M_s^{1-\epsilon} u \log\left(2\delta^{-1}\right)}{s}} + \frac{M_{s,t} \log\left(2\delta^{-1}\right)}{3s} + \\
&\qquad + u^{\frac{1}{1+\varepsilon}} \left(\frac{1}{s}\sum_{i=1}^{s} \mathbb{P}\left(|X_i| > M_{s,t}\right)^{\frac{\epsilon}{1+\varepsilon}}\right) \\
&\overset{(27)}{\le} \sqrt{\frac{2\left(\frac{us}{\log(2\delta^{-1})}\right)^{\frac{1-\epsilon}{1+\epsilon}} u \log\left(2\delta^{-1}\right)}{s}} + \frac{\left(\frac{us}{\log(2\delta^{-1})}\right)^{\frac{1}{1+\epsilon}} \log\left(2\delta^{-1}\right)}{3s} + \\
&\qquad + u^{\frac{1}{1+\varepsilon}} \left(\frac{1}{s}\sum_{i=1}^{s} \mathbb{P}\left(|X_i| > M_{s,t}\right)^{\frac{\epsilon}{1+\varepsilon}}\right) \\
&\overset{(28)}{\le} \sqrt{\frac{2\left(\frac{us}{\log(2\delta^{-1})}\right)^{\frac{1-\epsilon}{1+\epsilon}} u \log\left(2\delta^{-1}\right)}{s}} + \frac{\left(\frac{us}{\log(2\delta^{-1})}\right)^{\frac{1}{1+\epsilon}} \log\left(2\delta^{-1}\right)}{3s} + \\
&\qquad + \frac{u^{\frac{1}{1+\varepsilon}}}{s}\sum_{i=1}^{s} \left(\frac{49 \log\left(2\delta^{-1}\right)}{s}\right)^{\frac{\epsilon}{1+\varepsilon}} \\
&\le \left(2 + \frac{\sqrt{2}}{3} + 7\sqrt{2}\right) u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{n}\right]^{\frac{\epsilon}{1+\epsilon}} \\
&\le 15 u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{n}\right]^{\frac{\epsilon}{1+\epsilon}}.
\end{aligned}$$

□

A.7 Proof of Theorem 4

We now prove the upper bound on regret for `AdaR-UCB` under Assumption 1.

We define for every arm $i \in [K]$ and every $t \in [T]$, the upper confidence bound as

$$B_{i,N_i(t-1),t} = \widehat{\mu}_{i,N_i(t-1),t} + 2\sqrt{\frac{V_{N_i(t-1)}(\mathbf{Y}_i)\log(t^3)}{N_i(t-1)}} + 19\frac{\widehat{M}_{i,N_i(t-1),t}\log(t^3)}{N_i(t-1)},$$

where $N_i(t-1)$ is the number of times arm i has been pulled up to time $t-1$. Note that $N_i(t-1)$ is a random variable for every i and every t, where the randomness is inherited from the stochastic nature of the environment.

Each time an arm i is chosen, `AdaR-UCB` pulls it two times in a row in order to collect independent samples for estimator and threshold. Since the two collections of samples are independent and never used jointly, we consider $N_i(t-1)$ as the number of times arm i *has been chosen*, and not the number of times it has been effectively pulled (which would be the double).

We now show that if $i_t = i$, for any i such that $\Delta_i > 0$, then one of the following four inequalities is true:

$$\text{either } B_{i^*,N_{i^*}(t-1),t} \leq \mu^*, \tag{33}$$

$$\text{or} \quad \widehat{\mu}_{i,N_i(t-1),t}(\mathbf{X}_i) > \mu_i + 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}}, \tag{34}$$

$$\text{or } N_i(t-1) < 36\left(\frac{36u}{(\Delta_i)^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}\log(t^3), \tag{35}$$

$$\text{or } \sqrt{V_{N_i(t-1)}(\mathbf{Y}_i)} > \sqrt{\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y}_i)]} + 2\widehat{M}_{i,N_i(t-1),t}\sqrt{\frac{2\log(t^3)}{N_i(t-1)}}. \tag{36}$$

Indeed, assume that all four inequalities are false. Then we have

$$\begin{aligned}
B_{i^*,N_{i^*}(t-1),t} &\overset{(33)}{>} \mu^* = \mu_i + \Delta_i \\
&\overset{(34)}{\geq} \widehat{\mu}_{i,N_i(t-1),t}(\mathbf{X}_i) - 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + \Delta_i \\
&\overset{(*)}{\geq} \widehat{\mu}_{i,N_i(t-1),t}(\mathbf{X}_i) + 2\sqrt{\frac{V_{N_i(t-1)}(\mathbf{Y}_i)\log(t^3)}{N_i(t-1)}} + 19\frac{\widehat{M}_{i,N_i(t-1),t}\log(t^3)}{N_i(t-1)} \\
&= B_{i,N_i(t-1),t}.
\end{aligned}$$

The step marked with $(*)$ is a consequence of the fact that both (35) and (36) are false. In particular, we need to show that

$$\Delta_i \geq 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{V_{N_i(t-1)}(\mathbf{Y})\log(t^3)}{N_i(t-1)}} + 19\frac{\widehat{M}_{N_i(t-1)}\log(t^3)}{N_i(t-1)}. \tag{$*$}$$

To do so, we make use of Theorem 8 and the following fact:

$$\begin{aligned}
\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y}_i)] &= \mathbb{E}\left[\mathbf{Y}_i^2\right] - \mathbb{E}\left[\mathbf{Y}_i\right]^2 \leq \mathbb{E}\left(X_t^2\mathbb{1}_{|X_t|\leq\widehat{M}_t}\right) \\
&= \mathbb{E}\left(|X_t|^{1+\epsilon}|X_t|^{1-\epsilon}\mathbb{1}_{|X_t|\leq\widehat{M}_t}\right) \\
&\leq \mathbb{E}\left(|X_t|^{1+\epsilon}\widehat{M}_t^{1-\varepsilon}\right) \\
\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y}_i)] &\leq u\widehat{M}_t^{1-\epsilon}.
\end{aligned} \tag{37}$$

Now, we make use of the fact that (35) and (36) are false together with inequalities (27) and (37):

$$\Delta_i \overset{(35)}{\geq} 36^{\frac{\epsilon}{1+\epsilon}}(36u)^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}}$$

$$
\begin{aligned}
&= 36u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} \\
&= (15+2+19)u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} \\
&= 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{\log(t^3)u\left(\frac{uN_i(t-1)}{\log(t^3)}\right)^{\frac{1-\epsilon}{1+\epsilon}}}{N_i(t-1)}} + 19\frac{\left(\frac{uN_i(t-1)}{\log(t^3)}\right)^{\frac{1}{1+\epsilon}}\log(t^3)}{N_i(t-1)} \\
&\overset{(27)}{\geq} 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{\log(t^3)u\widehat{M}^{1-\epsilon}_{i,N_i(t-1),t}}{N_i(t-1)}} + 19\frac{\widehat{M}_{i,N_i(t-1),t}\log(t^3)}{N_i(t-1)} \\
&\overset{(37)}{\geq} 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y_i})]\log(t^3)}{N_i(t-1)}} + 19\frac{\widehat{M}_{i,N_i(t-1),t}\log(t^3)}{N_i(t-1)} \\
&= 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{\log(t^3)}{T_i(t-1)}}\left[\sqrt{\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y})]} + 2\widehat{M}_{N_i(t-1)}\sqrt{\frac{2\log(t^3)}{N_i(t-1)}}\right] \\
&\qquad + 19\frac{\widehat{M}_{N_i(t-1)}\log(t^3)}{N_i(t-1)} \\
&\overset{(36)}{\geq} 15u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^3)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{V_{N_i(t-1)}(\mathbf{Y}_i)\log(t^3)}{N_i(t-1)}} + 19\frac{\widehat{M}_{i,N_i(t-1),t}\log(t^3)}{N_i(t-1)}. \qquad (*)
\end{aligned}
$$

Finally, as a consequence of $(*)$, we have

$$B_{i^*,N_{i^*}(t-1),t} \overset{(*)}{>} B_{i,N_i(t-1),t}$$

but this is a contradiction to the fact that $i_t = i$. Thus, statements (33) to (36) cannot be false simultaneously.

We make use of a union over all the possible values of $N_i(t-1)$ and of the previously introduced concentration inequalities to bound with $\frac{1}{t^3}$ the probabilities of events (33), (34) and (36) to be true:

$$\mathbb{P}\left(\{\text{(33) is true}\} \text{ or } \{\text{(34) is true}\} \text{ or } \{\text{(36) is true}\} \ \forall N_i(t-1) \in [t]\ \right) \leq 3\sum_{s=1}^{t}\frac{1}{t^3} = \frac{3}{t^2}.$$

To proceed, we introduce the quantity

$$v \coloneqq \left\lceil 108\frac{(36u)^{\frac{1}{\varepsilon}}}{\Delta_i^{\frac{1+\varepsilon}{\epsilon}}}\log T\right\rceil.$$

It's now time to bound the expected number of times each arm is pulled:

$$
\begin{aligned}
\mathbb{E}[N_i(T)] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{i_t=i}\right] &\leq v + \mathbb{E}\left[\sum_{t=v+1}^{T}\mathbb{1}_{i_t=i \text{ and } \{\text{(35) is false }\}}\right] \\
&\leq v + \mathbb{E}\left[\sum_{t=v+1}^{T}\left(\mathbb{1}_{i_t=i \text{ and } \{\text{(33) or (34) or (36) is true}\}}\right)\right] \qquad (38) \\
&\leq v + \sum_{t=v+1}^{T}\frac{3}{t^2} \\
&\leq v + 5
\end{aligned}
$$

We can now conclude the proof using the regret decomposition (and considering that the effective number of pulls is doubled):

$$R_T \leq \sum_{i:\Delta_i>0}\left(216\left(\frac{36u}{\Delta_i^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}\log T + 10\Delta_i\right).$$

□

A.8 Proof of Theorem 5

The result can be proven by assuming $\log T \geq \max_{i\in[K]} \frac{10}{216}\left(\frac{\Delta_i^{1+\epsilon}}{36u}\right)^{\frac{1}{\epsilon}}$.

We have:

$$
\begin{aligned}
R_T &= \sum_{i:\Delta_i>0} 2\Delta_i \mathbb{E}[N_i(T)]^{\frac{\varepsilon}{1+\varepsilon}} \mathbb{E}[N_i(T)]^{\frac{1}{1+\varepsilon}} \\
&\overset{(38)}{\leq} \sum_{i:\Delta_i>0} 2\Delta_i \mathbb{E}[N_i(T)]^{\frac{1}{1+\varepsilon}} \left(108\left(\frac{36u}{\Delta_i^{1+\epsilon}}\right)^{\frac{1}{\epsilon}} \log T + 5\right)^{\frac{\varepsilon}{1+\varepsilon}} \\
&\overset{(*)}{\leq} \sum_{i:\Delta_i>0} 2\Delta_i \mathbb{E}[N_i(T)]^{\frac{1}{1+\varepsilon}} \left(216\left(\frac{36u}{\Delta_i^{1+\epsilon}}\right)^{\frac{1}{\epsilon}} \log T\right)^{\frac{\varepsilon}{1+\varepsilon}} \\
&= \left(216(36u)^{\frac{1}{\epsilon}} \log T\right)^{\frac{\varepsilon}{1+\varepsilon}} \sum_{i:\Delta_i>0} 2\mathbb{E}[N_i(T)]^{\frac{1}{1+\varepsilon}} \\
&\overset{(**)}{\leq} 216^{\frac{1}{1+\epsilon}} (36u)^{\frac{1}{1+\epsilon}} (\log T)^{\frac{\varepsilon}{1+\varepsilon}} K^{\frac{\varepsilon}{1+\varepsilon}} 2\left(\sum_{i:\Delta_i>0} \mathbb{E}[N_i(T)]\right)^{\frac{1}{1+\varepsilon}} \\
&\leq 2T^{\frac{1}{1+\varepsilon}} (216K \log T)^{\frac{\varepsilon}{1+\varepsilon}} (36u)^{\frac{1}{1+\epsilon}}.
\end{aligned}
$$

The step marked with $(*)$ follows from the starting assumption on $\log T$. The step marked with $(**)$ follows from Hölder inequality.

□