(ϵ, u)-Adaptive Regret Minimization in Heavy-Tailed Bandits

Gianmarco Genalti GIANMARCO.GENALTI@POLIMI.IT
Lupo Marsigli LUPO.MARSIGLI@MAIL.POLIMI.IT
Nicola Gatti NICOLA.GATTI@POLIMI.IT
Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Department of Computer Science, Politecnico di Milano

Editors: Shipra Agrawal and Aaron Roth

Abstract

Heavy-tailed distributions naturally arise in several settings, from finance to telecommunications. While regret minimization under subgaussian or bounded rewards has been widely studied, learning with heavy-tailed distributions only gained popularity over the last decade. In this paper, we consider the setting in which the reward distributions have finite absolute raw moments of maximum order $1 + \epsilon$, uniformly bounded by a constant $u < +\infty$, for some $\epsilon \in (0, 1]$. In this setting, we study the regret minimization problem when ϵ and u are unknown to the learner and it has to adapt. First, we show that adaptation comes at a cost and derive two negative results proving that the same regret guarantees of the non-adaptive case cannot be achieved with no further assumptions. Then, we devise and analyze a fully data-driven trimmed mean estimator and propose a novel adaptive regret minimization algorithm, `AdaR-UCB`, that leverages such an estimator. Finally, we show that `AdaR-UCB` is the first algorithm that, under a known distributional assumption, enjoys regret guarantees nearly matching those of the non-adaptive heavy-tailed case.

Keywords: bandits, heavy-tailed distributions, adaptivity

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 stochastic MAB setting (Robbins, 1952), a learner has access to a set of $K \in \mathbb{N}_{\geq 2}$ actions (*i.e.*, *arms*). Each arm $i \in [K] \coloneqq \{1, \dots, K\}$ is associated with a reward probability distribution $\nu_i \in \Delta(\mathbb{R})$,[1] having finite expected value $\mu_i \coloneqq \mathbb{E}_{X \sim \nu_i}[X]$ (*i.e.*, *expected reward*). We denote with $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K) \in \Delta(\mathbb{R})^K$ a specific bandit instance. Let $T \in \mathbb{N}$ be the learning horizon, at every round $t \in [T]$, the learner selects an arm $I_t \in [K]$ and, in response, the environment reveals the $X_t \sim \nu_{I_t}$ (*i.e.*, *reward*) sampled from the distribution ν_{I_t}. The performance of a learner running an algorithm `Alg` is quantified by the *(expected cumulative) regret* over T rounds, defined as:

$$R_T(\texttt{Alg}, \boldsymbol{\nu}) \coloneqq T\mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{I_t}\right], \tag{1}$$

[1] Given a measurable space $(\mathcal{X}, \mathcal{F})$, we denote with $\Delta(\mathcal{X})$ the set of probability measures over $(\mathcal{X}, \mathcal{F})$.

where we assume, without loss of generality, that 1 is the optimal arm and $\Delta_i := \mu_1 - \mu_i$ is the suboptimality gap of arm $i \in [K]$, and the expectation is taken w.r.t. the randomness of the reward and the possible randomness of the algorithm.

Although most of the literature in stochastic MABs usually assumes a convenient tail property of the reward distributions, like *subgaussian* (Lattimore and Szepesvári, 2020) or *bounded* (Auer et al., 2002) rewards, in many practical scenarios, such as financial environments (Gagliolo and Schmidhuber, 2011) or network routing problems (Liebeherr et al., 2012), where uncertainty has a significant impact, heavy-tailed distributions naturally arise. In these cases, the tails decay slower than a Gaussian, the moment-generating function is no longer finite, and the moments of all finite orders might not exist. This prevents the application of standard concentration tools, such as Hoeffding's inequality (Boucheron et al., 2013), calling for more complex technical tools.

This work investigates the regret minimization problem for MAB with HT reward distributions, according to the setting introduced in the seminal work (Bubeck et al., 2013a). We assume that the *absolute raw moments* of the reward distributions of order up to $1+\epsilon$, with $\epsilon \in (0, 1]$ (*i.e.,* moment order) is finite and uniformly bounded by a constant $u \in \mathbb{R}_{\geq 0}$ (*i.e.,* moment bound), namely:

$$\boldsymbol{\nu} \in \mathcal{P}_{\text{HT}}(\epsilon, u)^K := \left\{ \boldsymbol{\nu} \in \Delta(\mathbb{R})^K \; : \; \mathop{\mathbb{E}}_{X \sim \nu_i} \left[|X|^{1+\epsilon} \right] \leq u, \;\; \forall i \in [K] \right\}, \tag{2}$$

In Bubeck et al. (2013a), the authors assume that ϵ *and* u *are known to the learner*. They show that if the variance is finite (*i.e.*, $\epsilon = 1$), but the higher order moments are not, the same (apart from constants) instance-dependent regret guarantees of order $O\big(\sum_{i:\Delta_i>0} \frac{\sigma^2}{\Delta_i} \log T\big)$ attained in the subgaussian setting (Lattimore and Szepesvári, 2020) can be achieved. However, for general $\epsilon \in (0, 1)$, the instance-dependent regret becomes of order $O\big(\sum_{i:\Delta_i>0} \big(\frac{u}{\Delta_i}\big)^{1/\epsilon} \log T\big)$, showing the detrimental effect of ϵ on the dependence of the suboptimality gaps. Moreover, they show that these regret guarantees are tight (up to constant terms) by deriving the corresponding asymptotic lower bound. From a worst-case regret perspective, the presented results translate into a regret bound of order $\widetilde{O}\big(K^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}}\big)$, for sufficiently large T, matching the lower bound, up to logarithmic terms. This regret bound degenerates to linear when $\epsilon \to 0$, *i.e.,* when only absolute moment of order 1 exists. However, these matching results are obtained thanks to the knowledge of both ϵ and u, *i.e.,* by *non-adaptive* algorithms. Indeed, ϵ and u are needed by the *algorithm* to drive exploration via the optimistic index, and, in some cases, to construct the expected reward *estimator* too. Nevertheless, recent works have shed light on the possibility of removing this knowledge at the cost of additional assumptions (e.g., Lee et al., 2020; Ashutosh et al., 2021; Huang et al., 2022). In particular, Huang et al. (2022) introduce the *truncated non-negativity* assumption designed for losses, that, by converting it for rewards, leads to the following *truncated non-positivity* assumption.

Assumption 1 (Truncated Non-Positivity) *Let $\boldsymbol{\nu}$ be a bandit. For the optimal arm 1, we have:*

$$\mathbb{E}_{X \sim \nu_1} \left[X \mathbb{1}_{\{|X| > M\}} \right] \leq 0, \quad \forall M \geq 0. \tag{3}$$

This assumption requires that the optimal arm 1 has a larger probability mass on the negative semi-axis but still allows the distribution to have an arbitrary support covering, potentially, the whole

$\mathbb{R}$. To the best of the authors' knowledge, this is the only assumption in literature truly independent of the values of ϵ and u. Additionally, as discussed by Huang et al. (2022), it is relatively weak if compared to other standard assumptions in the bandit literature. Under Assumption 1, *without the knowledge of* ϵ *and* u, Huang et al. (2022) provide an (ϵ, u)-*adaptive*[2] regret minimization algorithm, `AdaTINF`, that succeeds in matching the worst-case regret lower bound of Bubeck et al. (2013a) derived for the non-adaptive case. However, no instance-dependent analysis is provided of `AdaTINF`[3] and the following research questions remain open:

Research Question 1 *Is Assumption 1 needed to devise* (ϵ, u)*-adaptive algorithms (with unknown* (ϵ, u)*) matching the worst-case lower bound of order* $\Omega\big(K^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}}\big)$ *(*i.e., *as if we knew* (ϵ, u)*)?*

Research Question 2 *Is it possible, under Assumption 1, to devise* (ϵ, u)*-adaptive algorithms (with unknown* (ϵ, u)*) matching the instance-dependent regret lower bound of order* $\Omega\big(\sum_{i:\Delta_i>0}\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}\log T\big)$ *(*i.e., *as if we knew* (ϵ, u)*)?*

Original Contributions. In this paper, we investigate the regret minimization problem in heavy-tailed bandits giving up the knowledge of ϵ and u. Specifically, we address Research Question 1 and Research Question 2. The original contributions of the paper are summarized as follows:

- In Section 3, we address Research Question 1, by characterizing the challenges of the regret-minimization problem in HT bandits without knowing ϵ and u. In particular, we provide two *negative results* (Theorems 2 and 3), showing that, without any additional assumption, there exists no (ϵ, u)-adaptive algorithm that achieves the same worst-case regret guarantees as if ϵ or u were known (Bubeck et al., 2013a). These results provide a first justification of Assumption 1. Furthermore, we show how Assumption 1 does not reduce the complexity of the regret minimization problem even in the non-adaptive case (Theorem 4). These results rely on accurately defined HT bandit instances and information theory tools for deriving the lower bounds.
- In Section 4, we enhance the *trimmed mean* estimator, commonly used in HT bandits, to make it fully data-driven. Indeed, in the seminal paper (Bubeck et al., 2013a), both (i) the trimming threshold and (ii) the upper confidence bound were computed thanks to the knowledge of ϵ and u. Taking inspiration from Huber regression (Wang et al., 2021), we overcome (i) the need for ϵ and u in the estimator by developing a novel approach to recover an estimated threshold via *root-finding*. Leveraging an analysis based on the *self-bounding functions* (Maurer, 2006; Maurer and Pontil, 2009), we control the accuracy of the estimated threshold (Lemma 5). In particular, we show that our threshold underestimates (in high probability) the one proposed by Bubeck et al. (2013a). Furthermore, we overcome (ii) by resorting to *empirical Bernstein inequality* (Maurer and Pontil, 2009). This way, differently from Bubeck et al. (2013a), we use the empirical variance to eliminate the dependence on ϵ and u in the upper confidence bound (Lemma 1), preserving the desirable concentration properties of the delivered estimate (Theorem 6).

[2] We use the word *adaptive* to qualify algorithms that do not know the values of ϵ and/or u.

[3] Huang et al. (2022) actually provide algorithm `OptHTINF` with an instance-dependent analysis that, however, does not match the asymptotic lower bound of Bubeck et al. (2013a).

- In Section 5, we address Research Question 2, by devising and analyzing a novel (ϵ, u)-adaptive regret minimization algorithm, Adaptive Robust UCB (AdaR-UCB, Algorithm 1), that operates without the knowledge of ϵ and u. AdaR-UCB is an *optimistic anytime* algorithm that builds upon Robust UCB of Bubeck et al. (2013a), leveraging our trimmed mean estimator with estimated threshold. First, we show that, under Assumption 1, AdaR-UCB attains an instance-dependent regret bound of order $O\big(\sum_{i:\Delta_i>0}\big(\big(\frac{u}{\Delta_i}\big)^{1/\epsilon}+\frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\big)\log T\big)$ (Theorem 7). This result shows that AdaR-UCB nearly matches the instance-dependent lower bound of Bubeck et al. (2013a) for the non-adaptive case, apart from the second logarithmic term, which, however, does not depend on the reciprocal of the suboptimality gaps, and originates from an additional forced exploration needed for computing the empirical threshold.[4] Moreover, we show that AdaR-UCB suffers a worst-case regret bound of order $\widetilde{O}\big(K^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}}\big)$ (Theorem 8), matching, up to logarithmic terms, the minimax lower bound of the non-adaptive case (Bubeck et al., 2013a). To the best of authors' knowledge, AdaR-UCB is the first (ϵ, u)-adaptive algorithm for HT bandits that nearly matches both the instance-dependent and worst-case lower bounds of the non-adaptive case, under conditions (Assumption 1) not explicitly formulated in terms of ϵ and u.

In Section 2 we provide an up-to-date literature review on *adaptivity* in heavy-tailed bandits. The proofs of the results presented in the main paper are reported in Appendix B.

2. Related Works

During the last ten years, the stochastic heavy-tailed bandit problem has been steadily increasing in popularity. In this section, we summarize the main contributions, with a particular focus on partially adaptive approaches. Table 1 provides a comprehensive comparison.

Bubeck et al. (2013a) represents the most influential work in this area, formally introducing the setting, deriving both instance-dependent and worst-case lower bounds, and proposing the first non-adaptive algorithm, namely Robust UCB. Such an algorithm can be instanced with three robust estimators: *trimmed mean* (TM), *median of means* (MoM), and *Catoni estimator* (Catoni) achieving near-optimal regret guarantees from both instance-dependent and worst-case cases. The first minimax optimal algorithm was proposed in Wei and Srivastava (2020), namely Robust MOSS, removing the $(\log T)^{\frac{\epsilon}{1+\epsilon}}$. Instead, in Agrawal et al. (2021), the authors propose $\mathtt{KL}_{\mathtt{inf}}$-UCB attaining an asymptotically-optimal instance-dependent upper bound, highlighting the dependence on the instance with the KL-divergence, similarly as Garivier and Cappé (2011) for non-heavy-tailed bandits. These algorithms, however, require the knowledge of ϵ and u, *i.e.,* they are non-adaptive.[5] In Ashutosh et al. (2021), the authors show that *adaptivity* comes at a cost in both subgaussian and heavy-tailed bandits. In particular, logarithmic instance-dependent regret is unachievable when no further information on the environment is available. They introduce two algorithms, namely R-UCB-TEA and R-UCB-MoM, exploiting the TM and the MoM estimators, respectively. Although, in principle, they do not require the knowledge of ϵ or u for execution,

[4] A similar additional term appears in the instantiation of Robust UCB with Catoni estimator (Bubeck et al., 2013a).

[5] The *truncated mean* requires the knowledge of ϵ and u in the construction of the expected reward estimator too.

Algorithm	**Regret Bounds**				**ϵ-adaptive**		**u-adaptive**		**Assumption**
	Instance-dependent	Matching?[§]	Worst-case	Matching?[¶]	Estimator	Algorithm	Estimator	Algorithm	
`Robust UCB` - TM (Bubeck et al., 2013a)	$\sum_{i:\Delta_i>0}\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}(\log T)^{\frac{\epsilon}{1+\epsilon}}$	✓	✗	✗	✗	✗	—
`Robust UCB` - MoM * (Bubeck et al., 2013a)	$\sum_{i:\Delta_i>0}\left(\frac{v}{\Delta_i}\right)^{1/\epsilon}\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}v^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}(\log T)^{\frac{\epsilon}{1+\epsilon}}$	✓	✓	✗	✓	✗	—
`Robust UCB` - Catoni * (Bubeck et al., 2013a)	$\sum_{i:\Delta_i>0}\left(\frac{v}{\Delta_i}+\Delta_i\right)\log T$	✓	$\sqrt{vKT\log T}+\sum_{i:\Delta_i>0}\Delta_i\log T$	✓	✗	✗	✓	✗	$\epsilon=1$ only
`Robust MOSS` (Wei and Srivastava, 2020)	$\sum_{i:\Delta_i>0}\log\left(\frac{T\Delta_i^{\frac{1+\epsilon}{\epsilon}}}{K}\right)\frac{1}{\Delta_i^{1/\epsilon}}$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$	✓	✗	✗	✗	✗	—
$\texttt{KL}_{\texttt{inf}}\texttt{-UCB}$ † (Agrawal et al., 2021)	$\sum_{i:\Delta_i>0}\frac{\log T}{D_{\mathrm{KL}}^{\inf}(\nu_i,\mu_1)}$	✓	—	—	✓	✗	✓	✗	—
`APE`^{2} (Lee et al., 2020)	$\sum_{i:\Delta_i>0}\left(e^u+(T\Delta_i^{\frac{1+\epsilon}{\epsilon}}\log K)^{\frac{1+\epsilon}{\epsilon\log K}}\right)\frac{1}{\Delta_i^{1/\epsilon}}$	✗	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}\log Te^u$	✗	✗	✗	✓	✓	—
`MR-APE`^{2} (Lee and Lim, 2022)	$\sum_{i:\Delta_i>0}\left(Ke^u+\log\left(\frac{T\Delta_i^{\frac{1+\epsilon}{\epsilon}}}{K}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\frac{1}{\Delta_i^{1/\epsilon}}$	✗	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}e^u$	✗	✗	✗	✓	✓	—
`HTINF` (Huang et al., 2022)	$\sum_{i:\Delta_i>0}\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$	✓	✗	✗	✗	✗	Assumption 1
`OptHTINF` (Huang et al., 2022)	$\sum_{i:\Delta_i>0}\left(\frac{u^2}{\Delta_i^{2-\epsilon}}\right)^{1/\epsilon}\log T$	✗	$K^{\frac{\epsilon}{2}}u^{\frac{1}{1+\epsilon}}T^{\frac{2-\epsilon}{2}}$	✗	✓	✓	✓	✓	Assumption 1
`AdaTINF` (Huang et al., 2022)	—	—	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$	✓	✓	✓	✓	✓	Assumption 1
`R-UCB-TEA` ‡ (Ashutosh et al., 2021)	$\sum_{i:\Delta_i>0}\frac{f(T)}{1-\frac{2}{\Delta_i\log f(t)}}\log T$	✗	—	—	✓	✓	✓	✓	T s.t. $3u\log f(T)<2f(T)^\epsilon$
`R-UCB-MoM` ‡ (Ashutosh et al., 2021)	$\sum_{i:\Delta_i>0}\Delta_i\left(\frac{2f(T)}{\Delta_i}\right)^{\frac{1}{g(T)}}\log T$	✗	—	—	✓	✓	✓	✓	T s.t. $g(T)<\frac{\epsilon}{1+\epsilon}$ $f(T)>(12u)^{\frac{1}{1+\epsilon}}$
`AdaR-UCB` **(ours)**	$\sum_{i:\Delta_i>0}\left(\left(\frac{u}{\Delta_i}\right)^{1/\epsilon}+\frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq0)}\right)\log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}(\log T)^{\frac{\epsilon}{1+\epsilon}}$ $+\sum_{i:\Delta_i>0}\frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq0)}\log T$	✓	✓	✓	✓	✓	Assumption 1

* The bound depends on the centered absolute moment $v := \max_{i\in[K]}\mathbb{E}_{X\sim\nu_i}[|X-\mu_i|^{1+\epsilon}]$ of order $1+\epsilon$.
† $D_{\mathrm{KL}}^{\inf}(\eta, x) := \inf\{D_{\mathrm{KL}}(\eta,\kappa) : \kappa \in \mathcal{P}_{\mathrm{HT}}(\epsilon, u) \text{ and } \mathbb{E}_{X\sim\kappa}[X] \geq x\}$.
‡ f and g are to be given in input. Choosing an optimal value of those would require knowing ϵ and u.
[§] Matching the instance-dependent lower bound for the non-adaptive case w.r.t. T, $1/\Delta_i$, u (or v), and ϵ, up to constants.
[¶] Matching worst-case lower bound for the non-adaptive case w.r.t. T, K, u (or v), and ϵ, up to logarithmic terms.

Table 1: Comparison with the state-of-the-art. The regret bounds are deprived by constants.

logarithmic regret cannot be achieved but only approached with arbitrary precision. Moreover, the bounds hold only for a learning horizon T larger than a threshold depending on ϵ and u. No worst-case analysis is presented.

The closest work to ours is Huang et al. (2022) where the authors introduce the *adversarial heavy-tailed bandits*, in which an adversary chooses HT distributions for the losses. They first introduce the *truncated non-negativity* (analogous to our Assumption 1 for rewards) representing, to the best of authors' knowledge, the only assumption not explicitly related to ϵ and u. Three algorithms are provided: `HTINF`, `OptHTINF`, and `AdaTINF`, all analyzed under this assumption. `HTINF` requires knowledge of both ϵ and u and it is nearly optimal. Differently, both `OptHTINF` and `AdaTINF` are (ϵ, u)-adaptive. However, the instance-dependent bound of `OptHTINF` exposes an inconvenient dependence on $\left(\frac{u^2}{\Delta_i^{2-\epsilon}}\right)^{1/\epsilon}$ and the worst-case bound scales with $T^{\frac{2-\epsilon}{2}}$, both failing to match the lower bounds of the non-adaptive setting. Finally, the worst-case bound of AdaTINF matches the non-adaptive lower bound. However, the authors show that no logarithmic instance-dependent regret can be obtained by AdaTINF. Finally, Lee et al. (2020) introduce the APE^2 algorithm, adaptive in u, that, unfortunately, does not achieve logarithmic instance-dependent regret that, instead, scales with $T^{\frac{1}{1+\epsilon}} \log T$ and displays an inconvenient exponential dependence e^u. A modified version of this algorithm, namely MR-APE^2, introduced in Lee and Lim (2022), succeeds in removing the polynomial dependence on T, now poly-logarithmic $(\log T)^{\frac{1+\epsilon}{\epsilon}}$, but maintains the dependence on e^u.

3. Minimax Lower Bounds for Adaptive Heavy-Tailed Bandits

In this section, we address Research Question 1, by analyzing the challenges of the (ϵ, u)-adaptive regret minimization problem, *i.e.,* without the knowledge of ϵ and u. We start by revising the minimax regret lower bound derived in Bubeck et al. (2013a) for the *non-adaptive* case, *i.e.,* when ϵ and u are known (Theorem 1). Then, in Section 3.1, we provide two novel *negative results* showing that achieving the same worst-case regret guarantees when either u (Theorem 2) or ϵ (Theorem 3) are unknown is not possible. [6] Finally, in Section 3.2, we derive a new minimax regret lower bound under the *truncated non-positivity* (Assumption 1), illustrating how, even in the non-adaptive case, this assumption does not lead to smaller regret lower bounds.

Let us start by recalling the minimax regret lower bound for the non-adaptive case.

Theorem 1 (Minimax lower bound – non-adaptive, Bubeck et al. (2013a)) *Fix $\epsilon \in (0, 1]$ and $u \geq 0$. For every algorithm $\texttt{Alg}$, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $K \in \mathbb{N}_{\geq 2}$, it holds that:*

$$\sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^K} R_T(\texttt{Alg}, \boldsymbol{\nu}) \geq c_0 K^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}}, \tag{4}$$

where $c_0 > 0$ is a constant independent of u, ϵ, K, and T.

This result shows how the dependency on T deteriorates as ϵ approaches 0 and, instead, when the variance is finite, *i.e.,* $\epsilon = 1$, the lower bound displays the same order as the one for stochastic MABs with subgaussian rewards (Lattimore and Szepesvári, 2020).

3.1. Negative Results about Adaptivity

We now move to our negative results about the possibility of matching the minimax regret lower bound of the non-adaptive setting using (ϵ, u)-adaptive algorithms. The following result shows that any u-adaptive algorithm cannot achieve the same regret as in the non-adaptive case in Theorem 1.

[6] We remark that from the instance-dependent regret perspective, a negative answer to the possibility of achieving logarithmic regret with adaptive algorithms for heavy-tailed bandits has been already provided in (Ashutosh et al., 2021, Theorem 1). Thus, our focus is on the minimax regret perspective.

Theorem 2 (Minimax lower bound – u-adaptive) *Fix $\epsilon \in (0,1]$. For every algorithm $\texttt{Alg}$, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $K \in \mathbb{N}_{\geq 2}$, it holds that:*

$$\sup_{u \geq 0} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^K} \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{u^{\frac{1}{1+\epsilon}}} = +\infty. \tag{5}$$

More precisely, for every $u' \geq u \geq 0$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon, u')$ such that:

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{u^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{(u')^{\frac{1}{1+\epsilon}}} \right\} \geq c_1 \left(\frac{u'}{u} \right)^{\frac{\epsilon}{(1+\epsilon)^2}} T^{\frac{1}{1+\epsilon}}, \tag{6}$$

where $c_1 > 0$ is a constant independent of u, u', and T.

Some remarks are in order. First, let us observe that for proving the negative result, we have studied the ratio $R_T(\texttt{Alg}, \boldsymbol{\nu})/u^{\frac{1}{1+\epsilon}}$. Indeed, if there exists a u-adaptive regret minimization algorithm matching the lower bound for the non-adaptive case (Theorem 1), this ratio would not depend on u anymore. It is convenient to start commenting on Theorem 2 from the lower bound in Equation (6). Here, we show the existence of two heavy-tailed bandit instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$, characterized by the same moment order ϵ but possibly different moment bounds $u' \geq u$, for which any algorithm suffers (apart from constants) a regret that preserves the dependence on T but introduces a dependence on the ratio u'/u. Since we can make the ratio arbitrarily large by varying $u, u' \in \mathbb{R}_{\geq 0}$, we conclude the statement in Equation (5) showing that the minimax lower bound degenerates to infinity. This shows that, with no additional assumptions, there exists no u-adaptive algorithm matching the lower bound for the non-adaptive case (Theorem 1).

We now present the counterpart negative result concerning adaptivity to the moment order ϵ.

Theorem 3 (Minimax lower bound – ϵ-adaptive) *Fix $u = 1$. For every algorithm $\texttt{Alg}$, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $K \in \mathbb{N}_{\geq 0}$, it holds that:*

$$\sup_{\epsilon \in (0,1]} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^K} \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}} \geq c_2 T^{\frac{1}{16}}. \tag{7}$$

More precisely, for every $\epsilon, \epsilon' \in (0,1]$ with $\epsilon' \leq \epsilon$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon', u)$ such that:

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}} \right\} \geq c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon)(1+\epsilon')^2}}, \tag{8}$$

where $c_2 > 0$ is a constant independent of ϵ, ϵ', and T.

Differently from Theorem 2, here we target the ratio $R_T(\texttt{Alg}, \boldsymbol{\nu})/T^{\frac{1}{1+\epsilon}}$ for deriving the negative result. Indeed, having fixed $u = 1$, if an ϵ-adaptive algorithm exists matching the lower bound of Theorem 1, then, the considered ratio would not depend on T anymore. Starting from the lower bound of Equation (8), we observe that there exist two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$, with ϵ and ϵ' as moment orders, for which the ratio is lower bounded by a function dependent on T. Since $\epsilon \geq \epsilon'$, the exponent to which T is raised is non-negative and, consequently, the lower bound is a non-decreasing function of T. By letting ϵ and ϵ' range in $[0, 1)$, we obtain the minimax bound of Equation (7), displaying a gap of order $T^{\frac{1}{16}}$, which is attained by taking $\epsilon = 1$ and $\epsilon' = 1/3$. This result shows that there exists no ϵ-adaptive algorithm able to suffer the same regret as in the non-adaptive case of Theorem 1.

Combining Theorem 2 with Theorem 3, we conclude the non-existence of an (ϵ, u)-adaptive algorithm suffering the same regret guarantees as in the non-adaptive case. It is worth noting that the constructions employed for deriving the lower bounds presented in this section violate Assumption 1.

3.2. Minimax Lower Bound under Assumption 1

The results presented above show that, if our goal is to match the worst-case bound of the non-adaptive case of Theorem 1, we surely need to enforce additional assumptions. Huang et al. (2022) succeeds in this task by using the *truncated non-positivity* assumption (or more precisely, its dual version for losses). We may wonder whether enforcing Assumption 1 radically simplifies the problem. In the following, we show that this is not the case, by deriving a novel minimax lower bound for the non-adaptive case under this assumption of the same order as that of Theorem 1.

Theorem 4 (Minimax lower bound under Assumption 1 - non-adaptive) *Fix $\epsilon \in (0, 1]$ and $u \geq 0$. For every algorithm $\texttt{Alg}$, sufficiently large learning horizon $T \in \mathbb{N}$, and every number of arms $K \in \mathbb{N}_{\geq 2}$, it holds that:*

$$\sup_{\substack{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^K \\ \boldsymbol{\nu} \text{ fulfills Assumption } 1}} R_T(\texttt{Alg}, \boldsymbol{\nu}) \geq c_3 K^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}}, \tag{9}$$

where $c_3 > 0$ is a constant independent of u, ϵ, K and T.

Since (i) achieving the regret of Theorem 1 without further assumptions is not possible (Theorems 2 3) and (ii) Assumption 1 does not change the complexity of the non-adaptive case (Theorem 4), it makes sense to search for adaptive algorithms matching Theorem 1 under Assumption 1.

4. Trimmed Mean Estimator with Empirical Threshold

In this section, we present our novel *trimmed mean with empirical threshold* estimator, in which the threshold is computed from data. The trimmed mean estimator (Bickel, 1965), common in heavy-tailed statistics, cuts off the observations outside a predefined interval $[-M, M]$ with $M \geq 0$, named *trimming threshold*. Given a set of $s \in \mathbb{N}_{\geq 1}$ i.i.d. random variables $\mathbf{X} = \{X_1, \dots, X_s\}$, with expected value $\mu := \mathbb{E}[X_1]$, the trimmed mean estimator with threshold M is defined as:

$$\widehat{\mu}_s(\mathbf{X}; M) := \frac{1}{s} \sum_{j \in [s]} X_j \mathbb{1}_{\{|X_j| \leq M\}}. \tag{10}$$

The following result shows that, under truncated non-positivity (Assumption 1), it is possible to design an upper confidence bound on μ based on the trimmed mean estimator $\widehat{\mu}_s(\mathbf{X}; M)$ that can be computed with no knowledge of ϵ and u, depending only on the trimming threshold M.

Lemma 1 ((ϵ, u)-free Upper Confidence Bound) *Let $\delta \in (0, 1/2)$ and $\mathbf{X} = \{X_1, \dots, X_s\}$ be a set of $s \in \mathbb{N}_{\geq 2}$ i.i.d. random variables satisfying $X_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, $\mu := \mathbb{E}[X_1]$, and $M > 0$ be a (possibly random) trimming threshold independent of $\mathbf{X}$. Then, under Assumption 1, it holds that:*

$$\mathbb{P}\left(\mu - \widehat{\mu}_s(\mathbf{X}; M) \leq \sqrt{\frac{2V_s(\mathbf{X}; M) \log \delta^{-1}}{s}} + \frac{10M \log \delta^{-1}}{s}\right) \geq 1 - 2\delta, \tag{11}$$

where $V_s(\mathbf{X}; M)$ is the sample variance of the trimmed random variables, defined as:

$$V_s(\mathbf{X}; M) := \frac{1}{s-1} \sum_{j \in [s]} (X_j \mathbb{1}_{\{|X_j| \leq M\}} - \widehat{\mu}_s(\mathbf{X}; M))^2. \tag{12}$$

The result is obtained by applying the *empirical Bernstein's inequality* (Maurer and Pontil, 2009) and it is a *one-sided inequality* because of the nature of Assumption 1. From an algorithmic perspective, this enables us to build an optimistic index that does not require knowing the values of ϵ and u and represents the essential role of Assumption 1 in our `AdaR-UCB` algorithm.

The next step consists of computing the trimming threshold M in a fully data-driven way. Notice that the trimming threshold in `Robust UCB` (Bubeck et al., 2013a) is selected thanks to the knowledge of ϵ and u as $\widetilde{M}_s(\delta) = \left(\frac{us}{\log \delta^{-1}}\right)^{\frac{1}{1+\epsilon}}$. Instead, we follow a procedure similar to that of Wang et al. (2021) for Huber regression, and we estimate an *empirical trimming threshold* via a root-finding problem. Specifically, given a set of $s \in \mathbb{N}_{\geq 1}$ i.i.d. random variables $\mathbf{X}' = \{X_1', \dots, X_s'\}$ (independent of $\mathbf{X}$), the empirical threshold $\widehat{M}_s(\delta)$ is the solution of the equation:[7]

$$f_s(\mathbf{X}'; M, \delta) := \frac{1}{s} \sum_{j \in [s]} \frac{\min\{(X_j')^2, M^2\}}{M^2} - \frac{c \log \delta^{-1}}{s} = 0, \tag{13}$$

where $c > 0$ is a hyperparameter that will be set later. If the number of non-zero samples X_j' is sufficiently large, *i.e.*, $\sum_{j \in [s]} \mathbb{1}_{\{X_j' \neq 0\}} > c \log \delta^{-1}$ (see Proposition 9 for details), Equation (13) admits a unique positive solution, that we denote as $\widehat{M}_s(\delta)$.[8] A reader might notice that we are solving the "sample version" of the "population version" equation $\mathbb{E}[f_s(\mathbf{X}'; M, \delta)] = 0$. Denoting with $M_s(\delta)$ the solution (when it exists) of this latter equation, we can establish a meaningful relation between $M_s(\delta)$ and the threshold $\widetilde{M}_s(\delta)$ used by `Robust UCB` (Bubeck et al., 2013a):

$$c \log \delta^{-1} = \mathbb{E}[\min\{(X_1')^2 / M_s(\delta)^2, 1\}] \leq \mathbb{E}[|X_1'|^{1+\epsilon}] M_s(\delta)^{-1-\epsilon} \tag{14}$$

$$\implies M_s(\delta) \leq \left(\frac{us}{c \log \delta^{-1}}\right)^{\frac{1}{1+\epsilon}} = c^{-\frac{1}{1+\epsilon}} \widetilde{M}_s(\delta). \tag{15}$$

In practice, however, we cannot solve the "population" equation $\mathbb{E}[f_s(\mathbf{X}'; M, \delta)] = 0$ and we need to resort to the "sample version", delivering $\widehat{M}_s(\delta)$. The following result shows that $\widehat{M}_s(\delta)$ behaves (in high probability) analogously to $M_s(\delta)$, for a suitable choice of c.

Theorem 5 (Bounds on $\widehat{M}_s(\delta)$) *Let $\delta \in (0, 1/2)$ and $\mathbf{X}' = \{X_1', \dots, X_s'\}$ be a set of $s \in \mathbb{N}_{\geq 1}$ i.i.d. random variables satisfying $X_1' \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, and let $\widehat{M}_s(\delta)$ be the (random) positive root*

[7] The sets $\mathbf{X}$ and $\mathbf{X}'$ are chosen to have the same cardinality s to provide more readable results.

[8] An efficient algorithm for solving Equation (13) and its computational complexity analysis are reported in Appendix C.

of Equation (13) with $c > 2$. Then, if $\widehat{M}_s(\delta)$ exists, with probability at least $1 - 2\delta$, it holds that:

$$\widehat{M}_s(\delta) \leq \left(\frac{us}{(\sqrt{c} - \sqrt{2})^2 \log \delta^{-1}} \right)^{\frac{1}{1+\epsilon}} \quad \textit{and} \quad \mathbb{P}\left(|X_1| > \widehat{M}_s(\delta) \right) \leq (\sqrt{c} + \sqrt{2})^2 \frac{\log \delta^{-1}}{s}. \tag{16}$$

The proof relies on the concentration inequalities for *self-bounding functions* (Maurer, 2006; Maurer and Pontil, 2009). By selecting $c > (1 + \sqrt{2})^2$, we have that, with probability $1 - 2\delta$, the empirical threshold $\widehat{M}_s(\delta)$ is smaller than $\widetilde{M}_s(\delta)$, used in `Robust UCB`. Furthermore, in Bubeck et al. (2013a), the particular form of the *deterministic* threshold $\widetilde{M}_s(\delta)$ allows the authors to apply *Bernstein's inequality* and obtain a concentration bound explicitly depending on ϵ and u (Lemma 1):

$$\mathbb{P}\left(\left| \widehat{\mu}_s(\mathbf{X}; \widetilde{M}_s(\delta)) - \mu \right| \leq 4 u^{\frac{1}{1+\epsilon}} \left(\frac{\log \delta^{-1}}{s} \right)^{\frac{\epsilon}{1+\epsilon}} \right) \geq 1 - 2\delta, \tag{17}$$

We now show that using the *random* threshold $\widehat{M}_s(\delta)$, instead, still allows achieving analogous guarantees with just a slightly larger constant.

Theorem 6 ((ϵ, u)-dependent Concentration Bound) *Let $\delta \in (0, 1/4)$, $\mathbf{X} = \{X_1, \dots, X_{s/2}\}$, and $\mathbf{X}' = \{X'_1, \dots, X'_{s/2}\}$ be two independent sets of $s/2 \in \mathbb{N}_{\geq 2}$ i.i.d. random variables satisfying $X_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, $\mu := \mathbb{E}[X_1]$, and let $\widehat{M}_s(\delta)$ be the (random) positive root of Equation (13) with $c = (1 + \sqrt{2})^2$. Then, if $\widehat{M}_s(\delta)$ exists, it holds that:*

$$\mathbb{P}\left(\left| \widehat{\mu}_s(\mathbf{X}; \widehat{M}_s(\delta)) - \mu \right| \leq 8 u^{\frac{1}{1+\epsilon}} \left(\frac{\log \delta^{-1}}{s} \right)^{\frac{\epsilon}{1+\epsilon}} \right) \geq 1 - 4\delta. \tag{18}$$

First, we notice that, at the price of a slightly larger constant, this concentration bound displays the same behavior as that of Equation (17). However, remarkably, our estimator is fully data-driven as the trimming threshold $\widehat{M}_s(\delta)$ is computed from dataset $\mathbf{X}'$ (*i.e.,* half of the available samples) with no knowledge of ϵ and u. Second, differently from Lemma 1, this is a *double-sided inequality* and holds even without Assumption 1. From a technical perspective, this result is proved by resorting to *Bernstein's inequality* and Theorem 5 to control the values of the estimated threshold $\widehat{M}_s(\delta)$.

Summarizing, we conclude that the trimmed mean estimator $\widehat{\mu}_s(\mathbf{X}; \widehat{M}_s(\delta))$ with the empirical threshold $\widehat{M}_s(\delta)$ fulfills two important properties: (i) under Assumption 1, it enjoys an upper confidence bound that is fully empirical and (ϵ, u)-free (Lemma 1). This bound will be used in the implementation of the `AdaR-UCB` algorithm; (ii) it enjoys (up to constants) the same concentration properties as the truncated mean with the (ϵ, u)-dependent threshold $\widetilde{M}_s(\delta)$ (Theorem 6). This bound, instead, will be used in the analysis of the `AdaR-UCB` algorithm.

5. An (ϵ, u)-Adaptive Approach for Heavy-Tailed Bandits

In this section, we address Research Question 2 by presenting `Adaptive Robust UCB` (`AdaR-UCB`, Algorithm 1), an (ϵ, u)-adaptive *anytime* regret minimization algorithm able to operate in the heavy-tailed bandit problem *with no prior knowledge on ϵ or u*, and providing its regret analysis.

5.1. The AdaR-UCB Algorithm

AdaR-UCB (Algorithm 1) is based on the optimism in-the-face-of-uncertainty principle, and built upon the Robust UCB strategy from Bubeck et al. (2013a) leveraging the estimator presented in Section 4. AdaR-UCB keeps track of the number of times every arm $i \in [K]$ has been selected $N_i(\tau)$ and maintains two disjoint sets of rewards $\mathbf{X}_i(\tau)$ and $\mathbf{X}'_i(\tau)$ (line 1). Specifically, $\mathbf{X}'_i(\tau)$ will be employed to compute the empirical threshold, while $\mathbf{X}_i(\tau)$ for the trimmed mean estimator. The algorithm operates over $\lfloor T/2 \rfloor$ rounds, indexed by τ, and, in every round $\tau \in \lfloor T/2 \rfloor$, it collects *two* samples from the selected arm I_τ (the time index is $t = 2\tau$). Specifically, AdaR-UCB first computes the *upper confidence bound* index $B_i(\tau)$ for every arm $i \in [K]$. If the condition for the existence of the positive root of Equation (13) is not verified (line 4), the index $B_i(\tau)$ is set to $+\infty$ (line 5), forcing the algorithm to pull arm i. Instead, if the condition is verified, the empirical threshold is computed $\widehat{M}_i(\tau) \leftarrow \widehat{M}_{N_i(\tau-1)}(\tau^{-3})$ (line 7) according to Equation (13) with $c = (1+\sqrt{2})^2$ using the dataset $\mathbf{X}'_i(\tau-1)$ and selecting $\delta = \tau^{-3}$. Then, the algorithm employs it (line 8) to compute the trimmed mean estimator $\widehat{\mu}_i(\tau) \leftarrow \widehat{\mu}_{N_i(\tau-1)}(\mathbf{X}_i(\tau-1); \widehat{M}_i(\tau))$ (as in Equation 10) and variance estimator $V_i(\tau) \leftarrow V_{N_i(\tau-1)}(\mathbf{X}_i(\tau-1); \widehat{M}_i(\tau))$ (as in Equation 12) using the samples from the other dataset $\mathbf{X}_i(\tau-1)$. These quantities are then employed for the optimistic index computation $B_i(\tau)$ (line 9) according to the empirical bound of Lemma 1. The optimistic arm I_τ is then played *twice* (line 12) and the two collected samples are used to augment the reward sets $\mathbf{X}_i(\tau)$ and $\mathbf{X}'_i(\tau)$, respectively (lines 13-14), and the arm pull counters $N_i(\tau)$ (line 15).

5.2. Regret Analysis

In this section, we provide the regret analysis of AdaR-UCB under the truncated non-positivity assumption (Assumption 1). We start with the instance-dependent regret bound.

Theorem 7 (Instance-Dependent Regret bound of AdaR-UCB) *Let $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^K$ and $T \in \mathbb{N}_{\geq 2}$ be the learning horizon. Under Assumption 1, AdaR-UCB suffers a regret bounded as:*

$$R_T(\textit{AdaR-UCB}, \boldsymbol{\nu}) \leq \sum_{i:\Delta_i>0} \left[\left(120 \left(\frac{u}{\Delta_i} \right)^{\frac{1}{\epsilon}} + \frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)} \right) \log \frac{T}{2} + 20\Delta_i \right]. \tag{19}$$

Some observations are in order. We notice that the dependence on ϵ and u match the instance-dependent lower bound for the non-adaptive case (Bubeck et al., 2013a). Note that the trimming threshold estimation requires in AdaR-UCB the *forced exploration* (line 5) and leads to the additional logarithmic term $\sum_{i:\Delta_i>0} \frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)} \log \frac{T}{2}$ that grows proportionally to the suboptimality gap Δ_i and inversely with the probability $\mathbb{P}_{\nu_i}(X \neq 0)$ of sampling a non-zero reward. This is explained by the condition (line 4) for the existence of a positive trimming threshold that requires a sufficiently large number of non-zero rewards. It is worth noting that for *absolutely continuous* reward distributions, *i.e.,* the ones we are interested in the heavy-tail setting, we have $\mathbb{P}_{\nu_i}(X \neq 0) = 1$. Moreover, if there is an arm i s.t. $\mathbb{P}_{\nu_i}(X \neq 0) = 0$, then based on Assumption 1, this arm is considered optimal. Consequently, AdaR-UCB achieves low regret by repeatedly selecting this arm. In such a case, this

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

```
Initialize counters $N_i(0) = 0$, reward sets $\mathbf{X}_i(0) = \{\}$, $\mathbf{X}'_i(0) = \{\}$ for every $i \in [K]$, $\tau \leftarrow 1$, $t \leftarrow 2\tau$
while $\tau \le \lfloor T/2 \rfloor$ do
    for $i \in [K]$ do
        if $N_i(\tau-1) = 0$ or $\sum_{X' \in \mathbf{X}'_i(\tau-1)} \mathbb{1}_{\{X' \neq 0\}} \le 4 \log \tau^{-3}$ then
            Compute the optimistic index: $B_i(\tau) = +\infty$
        else
            Compute the trimming threshold: $\widehat{M_i}(\tau) \leftarrow \widehat{M}_{N_i(\tau-1)}(\tau^{-3})$ solving the equation
             $f(\mathbf{X}'_i(\tau-1); M, \tau^{-3}) = 0$ (Eq. 13 with $c = (1+\sqrt{2})^2$)
            Compute the trimmed mean estimator: $\widehat{\mu}_i(\tau) \leftarrow \widehat{\mu}_{N_i(\tau-1)}(\mathbf{X}_i(\tau-1); \widehat{M_i}(\tau))$ (Eq. 10) and
             the variance estimator $V_i(\tau) \leftarrow V_{N_i(\tau-1)}(\mathbf{X}_i(\tau-1); \widehat{M_i}(\tau))$ (Eq. 12)
            Compute the optimistic index:
                $$B_i(\tau) = \widehat{\mu}_i(\tau) + \sqrt{\frac{2V_i(\tau) \log \tau^3}{N_i(\tau-1)}} + \frac{10\widehat{M_i}(\tau) \log \tau^3}{N_i(\tau-1)}$$
        end
    end
    Select arm $I_\tau \in \arg\max_{i \in [K]} B_i(\tau)$, play it twice, and receive rewards $X$ and $X'$
    Update reward sets $\mathbf{X}_{I_\tau}(\tau) = \mathbf{X}_{I_\tau}(\tau-1) \cup \{X\}$, $\mathbf{X}_i(\tau) = \mathbf{X}_i(\tau-1)$ for every $i \neq I_\tau$
    Update reward sets $\mathbf{X}'_{I_\tau}(\tau) = \mathbf{X}'_{I_\tau}(\tau-1) \cup \{X'\}$, $\mathbf{X}'_i(\tau) = \mathbf{X}'_i(\tau-1)$ for every $i \neq I_\tau$
    Update counters $N_i(\tau) = |\mathbf{X}_i(\tau)|$ for every $i \in [K]$, $\tau \leftarrow \tau + 1$, $t \leftarrow 2\tau$
end
```

additional regret term reduces to $24 \sum_{i:\Delta_i > 0} \Delta_i \log \frac{T}{2}$, a term that was present in the regret bound of `Robust UCB` with the Catoni estimator too.[9] In general, we are unsure whether this term is unavoidable or an artifact of our algorithm and/or analysis. From a technical perspective, the proof of Theorem 7 follows similar steps to the result provided by Bubeck et al. (2013a) concerning the upper bound on regret for `Robust UCB`, although additional care is needed to control simultaneously the concentration of the empirical threshold and of the trimmed mean estimator. In conclusion, this result positively answers our Research Question 2, showing how `AdaR-UCB` nearly matches the instance-dependent lower bound for the non-adaptive case.

Finally, to complement the analysis, we provide the worst-case regret bound for `AdaR-UCB`.

Theorem 8 (Worst-Case Regret bound of `AdaR-UCB`) *Let $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^K$ and $T \in \mathbb{N}_{\ge 2}$ be the learning horizon. Under Assumption 1, `AdaR-UCB` suffers a regret bounded as:*

$$R_T(\texttt{AdaR-UCB}, \boldsymbol{\nu}) \le 46 \left(K \log \frac{T}{2} \right)^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i > 0} \left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)} \log \frac{T}{2} + 20\Delta_i \right).$$

This result matches the lower bound from Bubeck et al. (2013a), up to logarithmic terms.

[9] We remark that this second term, although logarithmic, does not depend on the reciprocal of the suboptimality gaps and it is negligible when $1/\Delta_i \gg 1$ compared to the first one.

6. Conclusions

In this paper, we studied the (ϵ, u)-*adaptive* heavy-tailed bandit problem, where no information on moments of the reward distribution, not even which of them are finite, is provided to the learner. Focusing on two appealing research questions, we have: (i) shown that, with no further assumptions, no adaptive algorithm can achieve the same worst-case regret guarantees as in the non-adaptive case; (ii) devised a novel algorithm (`AdaR-UCB`), based on a fully data-driven estimator, enjoying nearly optimal instance-dependent and worst-case regret, under the truncated non-positivity assumption.

Future directions include: (i) investigating the role of the truncated non-positivity assumption, especially, whether weaker assumptions can be formulated; (ii) characterizing the *limits of* ϵ*-adaptivity*, *i.e.*, the best performance attainable by an ϵ-adaptive algorithm *without additional assumption*; (iii) understanding whether the forced exploration for the empirical threshold computation in `AdaR-UCB` (and the corresponding regret term) is unavoidable.

Acknowledgments

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.3 DD. 341 - 15 March 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006, and by the EU Horizon project ELIAS (European Lighthouse of AI for Sustainability, No. 101120237).

References

Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR, 2021.

Kumar Ashutosh, Jayakrishnan Nair, Anmol Kagrecha, and Krishna Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR, 2021.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

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

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Nearly optimal catoni's m-estimator for infinite variance. In *International Conference on Machine Learning*, pages 1925–1944. PMLR, 2022.

Peter J Bickel. On some robust estimates of location. *The Annals of Mathematical Statistics*, pages 847–858, 1965.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

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

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

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

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.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in k–armed bandits. *Journal of Machine Learning Research*, 24(13):1–33, 2023.

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.

Tor Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. *Advances in Neural Information Processing Systems*, 30, 2017.

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*, 2022.

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*, 33:8452–8462, 2020.

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

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.

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

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

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

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.

Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits. *IEEE Control Systems Letters*, 5(4):1423–1428, 2020.

Appendix A. Additional Related Works

In this section, we provide additional related works concerning adaptivity in statistics via Lepskii method and adaptivity in the case of subgaussian bandits.

A.1. Adaptivity via Lepskii Method

In Bhatt et al. (2022), authors provide a novel technique to extend Catoni's M-estimator (Catoni, 2012) to the infinite variance setting. In principle, their procedure relies on the knowledge of both ϵ and the centered moment v, however, they propose a strategy based on the Lepskii method (Lepskii, 1992) to adapt to unknown v. While the Lepskii method is a popular choice in the adaptive statistics literature, we point out how it requires an upper bound on the quantity to estimate. Indeed, this method can be safely applied when adapting to unknown ϵ (since it can be at most 1), but when it comes to u (or the centered moment v), requiring an upper bound makes the approach *not* fully adaptive.

A.2. Adaptivity in Subgaussian Bandits

In the literature of subgaussian stochastic bandits, σ (the subgaussian proxy) is usually assumed to be known by the agent. However, many works consider settings in which this quantity is unknown. In this section, we discuss standard approaches to adapt to σ (or estimate it) in subgaussian bandits, and show the additional difficulties implied by the heavy-tailed setting.

The main difference between σ and u is that the former can be estimated from data while guaranteeing strong convergence properties. In Audibert et al. (2009), authors introduce `UCB-V`, a variation of the well-known `UCB-1` algorithm capable of using a data-driven estimation of the variance while keeping optimal performance. As customary in most of the literature, rewards are assumed to be bounded in a known range. However, in heavy-tailed bandits, it is not possible to make such an assumption, and the estimation of u cannot be carried on. Other works try to relax the assumption of bounded rewards by the means of other assumptions, *e.g.,* a known upper bound on kurtosis (Lattimore, 2017), or Gaussian rewards (Cowan et al., 2018).

Without additional assumptions, dealing with both the unknown range of the rewards and unknown σ comes at a cost. As shown in Hadiji and Stoltz (2023), when the range of the rewards is unknown and no additional knowledge on the distributions is available, it is impossible to be simultaneously optimal in both the instance-dependent sense and the worst-case one. The existence of such a trade-off shows how difficult is, even in subgaussian bandits, to attain optimal performances when no knowledge is given on the environment. As a consequence, also in fully adaptive heavy-tailed bandits, such an impossibility result holds. However, as we have discussed, thanks to a specific assumption not involving ϵ nor u we can provide optimal regret guarantees in both cases.

Appendix B. Proofs and Derivations

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

B.1. Lower Bounds

Theorem 2 (Minimax lower bound – u-adaptive) *Fix $\epsilon \in (0, 1]$. For every algorithm $\texttt{Alg}$, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $K \in \mathbb{N}_{\geq 2}$, it holds that:*

$$\sup_{u \geq 0} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^K} \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{u^{\frac{1}{1+\epsilon}}} = +\infty. \tag{5}$$

More precisely, for every $u' \geq u \geq 0$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon, u')$ such that:

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{u^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{(u')^{\frac{1}{1+\epsilon}}} \right\} \geq c_1 \left(\frac{u'}{u} \right)^{\frac{\epsilon}{(1+\epsilon)^2}} T^{\frac{1}{1+\epsilon}}, \tag{6}$$

where $c_1 > 0$ is a constant independent of u, u', and T.

Proof We start by constructing two heavy-tailed bandit instances with a common maximum order of moment ϵ, but where $u' \geq u$. We use δ_x to denote the Dirac delta distribution centered on x

Base instance

$$\boldsymbol{\nu} = \begin{cases} \nu_1 = \delta_0, \\ \nu_2 = \left(1 - \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}\right) \delta_0 + \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \delta_{u^{\frac{1}{\epsilon}} \Delta^{-\frac{1}{\epsilon}}}, \end{cases} \tag{20}$$

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

Alternative instance

$$\boldsymbol{\nu}' = \begin{cases} \nu_1' = \left(1 - (2\Delta)^{1+\frac{1}{\epsilon}} (u')^{-\frac{1}{\epsilon}}\right) \delta_0 + (2\Delta)^{1+\frac{1}{\epsilon}} (u')^{-\frac{1}{\epsilon}} \delta_{(u')^{\frac{1}{\epsilon}} (2\Delta)^{-\frac{1}{\epsilon}}}, \\ \nu_2' = \nu_2, \end{cases} \tag{21}$$

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

We seek to prove that for any algorithm $\texttt{Alg}$, it holds that:

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}} \right\} \geq f(T, \epsilon, u, u'),$$

being f a function increasing in T. The proof 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:

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}} \right\} \geq \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}} = \frac{\Delta \mathbb{E}_{\texttt{Alg}, \boldsymbol{\nu}}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}}, \tag{22}$$

where $\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_1(T)]$ 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 $u' \geq u$, we have:

$$\begin{aligned}
&\max\left\{\frac{R_T(\texttt{Alg},\boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg},\boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq \\
&\qquad \geq (u'T)^{-\frac{1}{\epsilon+1}}\frac{\Delta T}{2}\max\left\{\mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}}\left(N_1(T) \geq T/2\right), \mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}'}\left(N_1(T) < T/2\right)\right\} \\
&\qquad \geq \frac{\Delta}{4}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\left(\mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}}\left(N_1(T) \geq T/2\right) + \mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}'}\left(N_1(T) < T/2\right)\right) \\
&\qquad \geq \frac{\Delta}{8}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\exp\left(-\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_1(T)]D_{\text{KL}}(\nu_1\|\nu_1')\right).
\end{aligned} \tag{23}$$

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 $\nu_1 \ll \nu_1'$:

$$\begin{aligned}
D_{\text{KL}}(\nu_1\|\nu_1') &= \nu_1(0)\log\frac{\nu_1(0)}{\nu_1'(0)} \\
&= \log\frac{1}{1-(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}}} \leq c(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}},
\end{aligned} \tag{24}$$

for $\Delta \in (0, \left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}}(u')^{\frac{1}{1+\epsilon}})$ and some constant $c \in (1,2)$. Putting together Equations (22), (23) and (24), we have:

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

The latter is a convex function of x and the minimization can be carried out in closed form, vanishing the derivative and finding:

$$x^* = c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\log\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}c(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right),$$

which leads to:

$$g(x^*) = \frac{\Delta}{2}(uT)^{-\frac{1}{\epsilon+1}}c^{-1}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\log\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}ec(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right).$$

We choose Δ such that:

$$\frac{Tu^{\frac{1}{\epsilon+1}}}{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}} c(2\Delta)^{\frac{1+\epsilon}{\epsilon}} = e^{\epsilon},$$

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 implies, after some calculations, that:

$$g(x^*) = c^{-\frac{\epsilon}{\epsilon+1}} 2^{-\frac{2\epsilon+5}{\epsilon+1}} (1+\epsilon) e^{-\frac{\epsilon}{\epsilon+1}} u^{-\frac{\epsilon}{(\epsilon+1)^2}} (u')^{\frac{\epsilon}{(\epsilon+1)^2}} \geq c_1 \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}},$$

where $c_1 > 0$ is a value independent of T and both u and u'. Finally, we have that

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{(u'T)^{\frac{1}{1+\epsilon}}} \right\} \geq c_1 \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}}.$$

We observe that $\Delta < \left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}} (u')^{\frac{1}{1+\epsilon}}$ for sufficiently large T. This concludes the proof of the second statement. For the first statement, we observe that, since $u' \geq u$ can be taken arbitrarily large, the right-hand side of this inequality can be arbitrarily large. ■

Theorem 3 (Minimax lower bound – ϵ-adaptive) *Fix $u = 1$. For every algorithm $\texttt{Alg}$, sufficiently large learning horizon $T \in \mathbb{N}$, and number of arms $K \in \mathbb{N}_{\geq 0}$, it holds that:*

$$\sup_{\epsilon \in (0,1]} \sup_{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^K} \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}} \geq c_2 T^{\frac{1}{16}}. \tag{7}$$

More precisely, for every $\epsilon, \epsilon' \in (0,1]$ with $\epsilon' \leq \epsilon$, under the same conditions above, there exist two instances $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)$ and $\boldsymbol{\nu}' \in \mathcal{P}_{HT}(\epsilon', u)$ such that:

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}} \right\} \geq c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon)(1+\epsilon')^2}}, \tag{8}$$

where $c_2 > 0$ is a constant independent of ϵ, ϵ', and T.

Proof 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

$$\boldsymbol{\nu} = \begin{cases} \nu_1 = \delta_0, \\ \nu_2 = (1 + \Delta\gamma - \gamma^{1+\epsilon})\delta_0 + (\gamma^{1+\epsilon} - \Delta\gamma)\delta_{1/\gamma} \end{cases}, \tag{25}$$

where $\Delta \in [0, 1/2]$ and $\gamma = (2\Delta)^{\frac{1}{\epsilon}}$. Thus, we have $\mu_1 = 0$ and $\mu_2 = \Delta$. Furthermore, $\mathbb{E}_{X\sim\nu_1}[|X|^{\alpha}] = 0$ and $\mathbb{E}_{X\sim\nu_2}[|X|^{\alpha}] = 2^{\frac{1-\alpha}{\epsilon}} \Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}$, which are guaranteed to be bounded by a constant smaller than 1 only if $\alpha \leq \epsilon + 1$. Thus, this instance admits moments finite only up to order $\epsilon + 1$, *i.e.,* $\boldsymbol{\nu} \in \mathcal{P}(\epsilon, 1)^2$. Moreover, the optimal arm is arm 2.

Alternative instance

$$\boldsymbol{\nu}' = \begin{cases} \nu_1' = (1-(\gamma')^{1+\epsilon'})\delta_0 + (\gamma')^{1+\epsilon'}\delta_{1/\gamma'}, \\ \nu_2' = \nu_2 \end{cases}, \tag{26}$$

where $\Delta \in [0,1/2]$ and $\gamma' = (2\Delta)^{\frac{1}{\epsilon'}}$. Thus, we have $\mu_1' = 2\Delta$ and $\mu_2' = \Delta$. Furthermore, $\mathbb{E}_{X\sim\nu_1'}[|x|^\alpha] = (2\Delta)^{\frac{1+\epsilon'-\alpha}{\epsilon'}}$ and $\mathbb{E}_{X\sim\nu_2'}[|x|^\alpha] = 2^{\frac{1-\alpha}{\epsilon}}\Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}$, which are guaranteed to be bounded by a constant smaller than 1 only if $\alpha \le \epsilon' + 1$. Thus, this instance admits moments finite only up to order $\epsilon' + 1$, *i.e.*, $\boldsymbol{\nu} \in \mathcal{P}(\epsilon', 1)^2$. Moreover, the optimal arm is arm 1.

We will prove, that for any algorithm $\texttt{Alg}$ it holds that:

$$\max\left\{\frac{R_T(\texttt{Alg},\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg},\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\} \ge f(T,\epsilon,\epsilon'),$$

being f a function increasing in T. The proof emulates the analyses and steps performed to prove Theorem 2. First, we observe that:

$$\max\left\{\frac{R_T(\texttt{Alg},\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg},\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\} \ge \frac{R_T(\texttt{Alg},\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}} = \frac{\Delta\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}}, \tag{27}$$

where $\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_1(T)]$ 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:

$$\begin{aligned} &\max\left\{\frac{R_T(\texttt{Alg},\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg},\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\} \ge \\ &\qquad \ge T^{-\frac{1}{\epsilon'+1}} \max\left\{\frac{\Delta T}{2}\mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}}\left(N_1(T) \ge \frac{T}{2}\right), \frac{\Delta T}{2}\mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}'}\left(N_1(T) < \frac{T}{2}\right)\right\} \\ &\qquad \ge \frac{\Delta}{4} T^{\frac{\epsilon'}{\epsilon'+1}} \left(\mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}}\left(N_1(T) \ge \frac{T}{2}\right) + \mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}'}\left(N_1(T) < \frac{T}{2}\right)\right) \\ &\qquad \ge \frac{\Delta}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \exp\left(-\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_1(T)] D_{\mathrm{KL}}(\nu_1 \| \nu_1')\right). \end{aligned} \tag{28}$$

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

$$\begin{aligned} D_{\mathrm{KL}}(\nu_1\|\nu_1') &= \nu_1(0)\log\frac{\nu_1(0)}{\nu_1'(0)} \\ &= \log\frac{1}{1-(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}} \le c(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}, \end{aligned} \tag{29}$$

for $\Delta \in [0,1/4]$ and some constant $c \in (1,2)$. Putting together Equations (27), (28) and (29), we have:

$$\begin{aligned} &\max\left\{\frac{R_T(\texttt{Alg},\boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg},\boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}}\right\} \\ &\qquad \ge \max\left\{\frac{\Delta\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}}, \frac{\Delta}{8}T^{\frac{\epsilon'}{\epsilon'+1}}\exp\left(-c\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_1(T)](2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right\} \end{aligned}$$

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

The latter is a convex function of x and the minimization can be carried out in closed form vanishing the derivative and obtaining:

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

which leads to:

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

We take Δ such that:

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

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:

$$g(x^*) = 2^{\frac{-2\epsilon'-5}{1+\epsilon'}} c^{-\frac{\epsilon'}{1+\epsilon'}} T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}} \geq c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}}.$$

where $c_2 > 0$ is a value independent of T and can be always selected to be ϵ and ϵ'. Finally, we have that:

$$\max \left\{ \frac{R_T(\texttt{Alg}, \boldsymbol{\nu})}{T^{\frac{1}{1+\epsilon}}}, \frac{R_T(\texttt{Alg}, \boldsymbol{\nu}')}{T^{\frac{1}{1+\epsilon'}}} \right\} \geq c_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}}.$$

We observe that $\Delta < 1/4$ for sufficiently large T. We conclude by observing that the exponent of T is maximized by taking $\epsilon = 1$ and $\epsilon' = 1/3$. ■

Theorem 4 (Minimax lower bound under Assumption 1 - non-adaptive) *Fix $\epsilon \in (0, 1]$ and $u \geq 0$. For every algorithm $\texttt{Alg}$, sufficiently large learning horizon $T \in \mathbb{N}$, and every number of arms $K \in \mathbb{N}_{\geq 2}$, it holds that:*

$$\sup_{\substack{\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon,u)^K \\ \boldsymbol{\nu} \text{ fulfills Assumption 1}}} R_T(\texttt{Alg}, \boldsymbol{\nu}) \geq c_3 K^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}}, \tag{9}$$

where $c_3 > 0$ is a constant independent of u, ϵ, K and T.

Proof We will construct instances using the following prototype of reward distribution, defined for $y \in (0, u^{\frac{1}{1+\epsilon}})$ and $\Delta \in (0, u^{\frac{1}{1+\epsilon}})$:

$$\rho_y = \left(1 - y^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \right) \delta_0 + \left(y^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \right) \delta_{-u^{\frac{1}{\epsilon}} \Delta^{-\frac{1}{\epsilon}}}. \tag{30}$$

The two instances are constructed by the means of Equation (30). Note that we have:

$$\mathop{\mathbb{E}}_{X\sim\rho_y}[X] = -y^{1+\frac{1}{\epsilon}}\Delta^{-\frac{1}{\epsilon}}, \tag{31}$$

$$\mathop{\mathbb{E}}_{X\sim\rho_y}[|X|^{1+\epsilon}] = y^{1+\frac{1}{\epsilon}}\Delta^{-1-\frac{1}{\epsilon}}u \le u, \tag{32}$$

for every $0 \le y \le \Delta$.

Base instance

$$\boldsymbol{\nu} = \begin{cases} \nu_1 = \rho_{\left(\frac{2}{3}\right)^{\frac{\epsilon}{1+\epsilon}}\Delta}, & \\ \nu_j = \rho_\Delta, & j \in [K] \setminus \{1\}. \end{cases}$$

Alternative instance

$$\boldsymbol{\nu}' = \begin{cases} \nu'_1 = \rho_{\left(\frac{2}{3}\right)^{\frac{\epsilon}{1+\epsilon}}\Delta}, & \\ \nu'_i = \rho_{\left(\frac{1}{3}\right)^{\frac{\epsilon}{1+\epsilon}}\Delta}, & \\ \nu'_j = \rho_\Delta, & j \in [K] \setminus \{1, i\}, \end{cases}$$

where $i \in \mathrm{argmin}_{j\neq 1}\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}'}[N_j(T)]$. For the base instance, we have $\mu_1 = -2\Delta/3$ and $\mu_j = -\Delta$ for all $j \neq 1$; whereas for the alternative instance $\mu'_j = \mu_j$ for all $j \neq i$ and $\mu'_i = -\Delta/3$. Both instances satisfy Assumption 1, being the support a subset made of non-positive numbers. Moreover, for the base instance, the optimal arm is 1 and for the alternative instance, the optimal arm is i. Using the Bretagnolle-Huber inequality, we obtain:

$$\begin{aligned} R_T(\texttt{Alg}, \boldsymbol{\nu}) + R_T(\texttt{Alg}, \boldsymbol{\nu}') &\ge \frac{\Delta T}{6}\left(\mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}}\left(N_1 \le \frac{T}{2}\right) + \mathbb{P}_{\texttt{Alg},\boldsymbol{\nu}'}\left(N_1 > \frac{T}{2}\right)\right) \\ &\ge \frac{\Delta T}{6}\mathrm{exp}\left(-\mathop{\mathbb{E}}_{\texttt{Alg},\boldsymbol{\nu}}[N_i(T)]D_{\mathrm{KL}}(\nu_i||\nu'_i)\right) \end{aligned}$$

We recall that by the definition of i, we have that $\mathbb{E}_{\texttt{Alg},\boldsymbol{\nu}}[N_i(T)] \le \frac{T}{K-1}$. We now compute the Kullback-Leibler divergence between the two instances:

$$\begin{aligned} D_{\mathrm{KL}}(\nu_i||\nu'_i) &= \Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}\log\left(\frac{\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}{\frac{1}{3}\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}\right) + \underbrace{(1-\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}})\log\left(\frac{1-\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}{1-\frac{1}{3}\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}\right)}_{\le 0} \\ &\le \Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}\log 3. \end{aligned}$$

Plugging this result, we finally get:

$$R_T(\texttt{Alg}, \boldsymbol{\nu}) + R_T(\texttt{Alg}, \boldsymbol{\nu}') \ge \frac{\Delta T}{6}\exp\left(-\frac{T}{K-1}\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}\log 3\right).$$

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

$$\max\{R_T(\texttt{Alg}, \boldsymbol{\nu}), R_T(\texttt{Alg}, \boldsymbol{\nu}')\} \geq c_3 K^{\frac{\epsilon}{1+\epsilon}} (uT)^{\frac{1}{1+\epsilon}},$$

for some constant $c_3 > 0$ independent of T, u, ϵ and K. ■

B.2. Estimator

Lemma 2 ((ϵ, u)-free Upper Confidence Bound) *Let $\delta \in (0, 1/2)$ and $\mathbf{X} = \{X_1, \dots, X_s\}$ be a set of $s \in \mathbb{N}_{\geq 2}$ i.i.d. random variables satisfying $X_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, $\mu := \mathbb{E}[X_1]$, and $M > 0$ be a (possibly random) trimming threshold independent of $\mathbf{X}$. Then, under Assumption 1, it holds that:*

$$\mathbb{P}\Big(\mu - \widehat{\mu}_s(\mathbf{X}; M) \leq \sqrt{\frac{2V_s(\mathbf{X}; M)\log \delta^{-1}}{s}} + \frac{10M \log \delta^{-1}}{s}\Big) \geq 1 - 2\delta, \tag{11}$$

where $V_s(\mathbf{X}; M)$ is the sample variance of the trimmed random variables, defined as:

$$V_s(\mathbf{X}; M) := \frac{1}{s-1} \sum_{j \in [s]} (X_j \mathbb{1}_{\{|X_j| \leq M\}} - \widehat{\mu}_s(\mathbf{X}; M))^2. \tag{12}$$

Proof Since M is computed independently of $\mathbf{X}$, the trimmed samples $X_i \mathbb{1}_{\{|X_i| \leq M\}}$ remain independent. Thus, with probability at least $1 - \delta$, we have:

$$\begin{aligned}
\mu - \widehat{\mu}_s(\mathbf{X}; M) &= \mathbb{E}[X_1] - \frac{1}{s}\sum_{t=1}^{s} X_t \mathbb{1}_{|X_t| \leq M} \\
&= \frac{1}{s}\sum_{t=1}^{s} \left(\mathbb{E}[X_1] - \mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M}\right]\right) + \frac{1}{s}\sum_{t=1}^{s} \left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M}\right] - X_t \mathbb{1}_{|X_t| \leq M}\right) \\
&= \frac{1}{s}\sum_{t=1}^{s} \mathbb{E}[X_t \mathbb{1}_{|X_t| > M}] + \frac{1}{s}\sum_{t=1}^{s} \left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M}\right] - X_t \mathbb{1}_{|X_t| \leq M}\right) \\
&\overset{(*)}{\leq} \frac{1}{s}\sum_{t=1}^{s} \left(\mathbb{E}\left[X_t \mathbb{1}_{|X_t| \leq M}\right] - X_t \mathbb{1}_{|X_t| \leq M}\right) \\
&\overset{(**)}{\leq} \sqrt{\frac{2V_s(\mathbf{Y})\log 2\delta^{-1}}{s}} + \frac{14M \log 2\delta^{-1}}{3(s-1)} \\
&\leq \sqrt{\frac{2V_s(\mathbf{Y})\log 2\delta^{-1}}{s}} + \frac{10M \log 2\delta^{-1}}{s}
\end{aligned}$$

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) recalling that the trimmed random variables range in $[-M, M]$. We also use $\frac{1}{s-1} \leq \frac{2}{s}$ in the last step for $s \geq 2$. ■

Proposition 9 (Uniqueness of Solution of Equation (13), Wang et al. (2021)) *Let $\mathbf{X} = \{X_1, \dots, X_s\}$ be a set of real numbers. If:*

$$0 < c \log \delta^{-1} < \sum_{j \in [s]} \mathbb{1}_{\{X_i \neq 0\}}, \tag{33}$$

then Equation (13) admits a unique positive solution.

Theorem 5 (Bounds on $\widehat{M}_s(\delta)$) *Let $\delta \in (0, 1/2)$ and $\mathbf{X}' = \{X'_1, \dots, X'_s\}$ be a set of $s \in \mathbb{N}_{\geq 1}$ i.i.d. random variables satisfying $X'_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, and let $\widehat{M}_s(\delta)$ be the (random) positive root of Equation (13) with $c > 2$. Then, if $\widehat{M}_s(\delta)$ exists, with probability at least $1 - 2\delta$, it holds that:*

$$\widehat{M}_s(\delta) \leq \left(\frac{us}{(\sqrt{c} - \sqrt{2})^2 \log \delta^{-1}} \right)^{\frac{1}{1+\epsilon}} \quad \textit{and} \quad \mathbb{P}\left(|X_1| > \widehat{M}_s(\delta) \right) \leq (\sqrt{c} + \sqrt{2})^2 \frac{\log \delta^{-1}}{s}. \tag{16}$$

Proof The proof makes use of the concentration inequality for self-bounding random variables (Maurer, 2006; Maurer and Pontil, 2009). Let $M > 0$, for every $i \in [\![s]\!]$, we define the random variable:

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

that ranges in $[0, 1]$. Furthermore, let: $Z_M(\mathbf{X}) \coloneqq \sum_{i=1}^{s} U_{i,M}$, ranging in $[0, s]$. Let us denote $\overline{U}_M(\mathbf{X}) \coloneqq Z_M(\mathbf{X})/s$, we observe that, given these definitions, the equation we want to solve for non-zero roots becomes:

$$\overline{U}_M(\mathbf{X}) - \frac{c \log \delta^{-1}}{s} = 0. \tag{34}$$

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

$$Z_M(\mathbf{X}) - \inf_{y \in \mathbb{R}} Z_M(\mathbf{X}_{y,k}) \leq 1, \quad \forall k \in [s], \tag{35}$$

$$\sum_{k=1}^{s} \left(Z_M(\mathbf{X}) - \inf_{y \in \mathbb{R}} Z_M(\mathbf{X}_{y,k}) \right)^2 \leq a Z_M(\mathbf{X}), \tag{36}$$

where $\mathbf{X}_{y,k}$ is obtained by replacing with y the k-th element X_k of the set $\mathbf{X}$. Indeed, Equation (35) follows as:

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

Similarly, we set $a = 1$ and obtain Equation (36) as follows:

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

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

since $U_{k,M} \leq 1$. Using Theorem 13 from Maurer (2006) with $a = 1$, for the right tail of the distribution, we have for every $\epsilon > 0$:

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

By letting $\epsilon = \sqrt{\frac{2\mathbb{E}[\overline{U}_M(\mathbf{X})] \log 2\delta^{-1}}{s}}$ and recalling the definition of $\overline{U}_M(\mathbf{X})$, we obtain:

$$\mathbb{P}\left(\mathbb{E}[\overline{U}_M(\mathbf{X})] - \overline{U}_M(\mathbf{X}) > \sqrt{\frac{2\mathbb{E}[\overline{U}_M(\mathbf{X})] \log \delta^{-1}}{s}}\right) \leq \delta,$$

which implies, after some algebraic manipulations (see Theorem 10 of (Maurer and Pontil, 2009)), the following:

$$\mathbb{P}\left(\sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} - \sqrt{\overline{U}_M(\mathbf{X})} > \sqrt{\frac{2\log\delta^{-1}}{s}}\right) \leq \delta.$$

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 obtain:

$$\mathbb{P}\left(\sqrt{\overline{U}_M(\mathbf{X})} - \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} > \sqrt{\frac{2\log\delta^{-1}}{s}}\right) \leq \delta.$$

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

$$\mathbb{P}\left(\left|\sqrt{\overline{U}_M(\mathbf{X})} - \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]}\right| > \sqrt{\frac{2\log\delta^{-1}}{s}}\right) \leq 2\delta. \tag{37}$$

Let us now define $\widehat{M}_s(\delta)$ random variable corresponding to the solution of the equation:

$$\overline{U}_{\widehat{M}_s(\delta)}(\mathbf{X}) = \frac{c\log\delta^{-1}}{s},$$

where $c > 0$. To control the bounds on $\widehat{M}$, we define the following auxiliary (non-random) quantities:

$$\sqrt{\overline{U}_M^+} := \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} + \sqrt{\frac{2\log\delta^{-1}}{s}} \quad \text{and} \quad \sqrt{\overline{U}_M^-} := \sqrt{\mathbb{E}[\overline{U}_M(\mathbf{X})]} - \sqrt{\frac{2\log\delta^{-1}}{s}}. \tag{38}$$

Thanks to Equation 37, we have, for every $M \geq 0$, that $\mathbb{P}(\overline{U}_M^- \leq \overline{U}_M(\mathbf{X}) \leq \overline{U}_M^+) \geq 1 - 2\delta$. Furthermore, let $M^+(\delta), M^-(\delta) > 0$, the solutions of the following (non-random) equations:

$$U^+_{M^+(\delta)} = \frac{c \log \delta^{-1}}{s} \qquad \text{and} \qquad U^-_{M^-(\delta)} = \frac{c \log \delta^{-1}}{s}. \tag{39}$$

Since $\mathbb{P}(\overline{U}_M^- \leq \overline{U}_M(\mathbf{X}) \leq \overline{U}_M^+) \geq 1-2\delta$, it follows that $\mathbb{P}(M^-(\delta) \leq \widehat{M}_s(\delta) \leq M^+(\delta)) \geq 1-2\delta$. We now proceed at lower bounding $M^-(\delta)$ and upper bounding $M^+(\delta)$:

$$\begin{aligned}
\sqrt{\frac{c \log \delta^{-1}}{s}} &= \sqrt{U^-_{M^-(\delta)}} && (40)\\
&= \sqrt{\mathbb{E}[\overline{U}_{M^-(\delta)}(\mathbf{X})]} - \sqrt{\frac{2 \log \delta^{-1}}{s}} && (41)\\
&\geq \sqrt{\mathbb{P}\left(|X_1| \geq M^-(\delta)\right)} - \sqrt{\frac{2 \log \delta^{-1}}{s}} && (42)\\
&\geq \sqrt{\mathbb{P}\left(|X_1| \geq \widehat{M}_s(\delta)\right)} - \sqrt{\frac{2 \log \delta^{-1}}{s}}, && (43)
\end{aligned}$$

where the last but one inequality follows from:

$$\mathbb{E}[\overline{U}_M(\mathbf{X})] = \mathbb{E}\left[\min\left\{\left(\frac{X_1}{M}\right)^2, 1\right\}\right] \geq \mathbb{P}\left(\left(\frac{X_1}{M}\right)^2 \geq 1\right) = \mathbb{P}\left(|X_1| \geq M\right), \tag{44}$$

and the last inequality holds with probability $1 - \delta$ and follows from the fact that $\widehat{M}_s(\delta) \geq M^-(\delta)$. Similarly, we have:

$$\begin{aligned}
\sqrt{\frac{c \log \delta^{-1}}{s}} &= \sqrt{U^+_{M^+(\delta)}} && (45)\\
&= \sqrt{\mathbb{E}[\overline{U}_{M^-(\delta)}(\mathbf{X})]} + \sqrt{\frac{2 \log \delta^{-1}}{s}} && (46)\\
&\leq \sqrt{\frac{u}{(M^+(\delta))^{1+\epsilon}}} + \sqrt{\frac{2 \log \delta^{-1}}{s}} && (47)\\
&\leq \sqrt{\frac{u}{(\widehat{M}_s(\delta))^{1+\epsilon}}} + \sqrt{\frac{2 \log \delta^{-1}}{s}}, && (48)
\end{aligned}$$

where the last but one inequality follows from:

$$\mathbb{E}[\overline{U}_M(\mathbf{X})] = \mathbb{E}\left[\min\left\{\left(\frac{X_1}{M}\right)^2, 1\right\}\right] \leq M^{-1-\epsilon}\mathbb{E}\left[|X_1|^{1+\epsilon}\right] \leq M^{-1-\epsilon}u, \tag{49}$$

and the last inequality holds with probability $1 - \delta$ and follows from the fact that $\widehat{M}_s(\delta) \leq M^+(\delta)$. Thus, with probability $1 - 2\delta$, we have for $c > 2$:

$$\mathbb{P}\left(|X_1| > \widehat{M}_s(\delta)\right) \leq (\sqrt{c} + \sqrt{2})^2 \frac{\log \delta^{-1}}{s} \quad \text{and} \quad \widehat{M}_s(\delta) \leq \left(\frac{us}{(\sqrt{c} - \sqrt{2})^2 \log \delta^{-1}}\right)^{\frac{1}{1+\epsilon}}. \tag{50}$$

■

Theorem 6 ((ϵ, u)-dependent Concentration Bound) *Let $\delta \in (0, 1/4)$, $\mathbf{X} = \{X_1, \dots, X_{s/2}\}$, and $\mathbf{X}' = \{X'_1, \dots, X'_{s/2}\}$ be two independent sets of $s/2 \in \mathbb{N}_{\geq 2}$ i.i.d. random variables satisfying $X_1 \sim \nu \in \mathcal{P}_{HT}(\epsilon, u)$, $\mu \coloneqq \mathbb{E}[X_1]$, and let $\widehat{M}_s(\delta)$ be the (random) positive root of Equation (13) with $c = (1+\sqrt{2})^2$. Then, if $\widehat{M}_s(\delta)$ exists, it holds that:*

$$\mathbb{P}\left(\left| \widehat{\mu}_s(\mathbf{X}; \widehat{M}_s(\delta)) - \mu \right| \leq 8u^{\frac{1}{1+\epsilon}} \left(\frac{\log \delta^{-1}}{s} \right)^{\frac{\epsilon}{1+\epsilon}} \right) \geq 1 - 4\delta. \tag{18}$$

Proof The result is obtained by combining an application of Bernstein's inequality and the bounds on the threshold $\widehat{M}_s(\delta)$ of Lemma 5. Furthermore since $\widehat{M}_s(\delta)$ is independent of $\mathbf{X}$, we can condition on the value of $\widehat{M}_s(\delta)$. With probability $1 - \delta$, we have:

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

where step $(*)$ follows from Hölder inequality, while step $(**)$ is a consequence of Bernstein's inequality for bounded random variables. To proceed further, we use Lemma 5 in union bound with the previously applied inequality. Thus, with probability at least $1 - 3\delta$, we have:

$$\widehat{\mu}_s(\mathbf{X}; \widehat{M}_s(\delta)) - \mu \leq$$

$$
\begin{aligned}
&\leq \sqrt{\frac{2\left(\frac{us}{(\sqrt{c}-\sqrt{2})^2 \log\delta^{-1}}\right)^{\frac{1-\epsilon}{1+\epsilon}} u \log\left(\delta^{-1}\right)}{s}} + \frac{\left(\frac{us}{(\sqrt{c}-\sqrt{2})^2 \log\delta^{-1}}\right)^{\frac{1}{1+\epsilon}} \log\left(\delta^{-1}\right)}{3s} \\
&\qquad\qquad\qquad\qquad + u^{\frac{1}{1+\epsilon}} \left((\sqrt{c}+\sqrt{2})^2 \frac{\log\delta^{-1}}{s} \right)^{\frac{\epsilon}{1+\epsilon}} \\
&\leq \left(\frac{\sqrt{2}}{(\sqrt{c}-\sqrt{2})^{\frac{1-\epsilon}{1+\epsilon}}} + \frac{1}{3(\sqrt{c}-\sqrt{2})^{\frac{2}{1+\epsilon}}} + (\sqrt{c}+\sqrt{2})^{\frac{2\epsilon}{1+\epsilon}} \right) u^{\frac{1}{1+\epsilon}} \left(\frac{\log\delta^{-1}}{n} \right)^{\frac{\epsilon}{1+\epsilon}} \\
&\leq 5.6 u^{\frac{1}{1+\epsilon}} \left(\frac{\log\delta^{-1}}{s} \right)^{\frac{\epsilon}{1+\epsilon}},
\end{aligned}
$$

where in the last passage we set $c = (1+\sqrt{2})^2$ and bounded the resulting expression for $\epsilon \in (0,1]$. A symmetric derivation leads to the second inequality. A union bound combined with renaming $s \leftarrow s/2$ and using $5.6\sqrt{2} \leq 8$, concludes the proof. ■

Theorem 7 (Instance-Dependent Regret bound of `AdaR-UCB`) *Let $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^K$ and $T \in \mathbb{N}_{\geq 2}$ be the learning horizon. Under Assumption 1, `AdaR-UCB` suffers a regret bounded as:*

$$
R_T(\texttt{AdaR-UCB}, \boldsymbol{\nu}) \leq \sum_{i:\Delta_i>0} \left[\left(120 \left(\frac{u}{\Delta_i} \right)^{\frac{1}{\epsilon}} + \frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)} \right) \log \frac{T}{2} + 20\Delta_i \right]. \tag{19}
$$

Proof For notational convenience, in this derivation, we will perform the substitution $T \leftarrow \lfloor T/2 \rfloor$ and $t \leftarrow \tau$. For every arm $i \in [K]$ and round $t \in [T]$, let us define the event:

$$
\mathcal{E}_{i,t} \coloneqq \left\{ \sum_{X \in \mathbf{X}'_i(t-1)} \mathbb{1}_{\{X \neq 0\}} - 4 \log t^3 > 0 \right\}. \tag{51}
$$

Under event $\mathcal{E}_{i,t}$ we do not incur in the forced exploration (FE) in line 4 ensuring that every arm has collected at least $4 \log t^3$ nonzero samples in $\mathbf{X}'_i$. Thus, we can decompose the expected number of pulls as follows:

$$
\begin{aligned}
\mathbb{E}[N_i^{\text{ALL}}(T)] &= \mathbb{E}\left[\sum_{t \in [T]} \mathbb{1}_{\{I_t = i \text{ and } \mathcal{E}_{i,t}\}} \right] + \mathbb{E}\left[\sum_{t \in [T]} \mathbb{1}_{\{I_t = i \text{ and } \mathcal{E}_{i,t}^{\complement}\}} \right] && (52) \\
&= \mathbb{E}[N_i(T)] + \mathbb{E}[N_i^{\text{FE}}(T)]. && (53)
\end{aligned}
$$

Part I: Bounding the expected number of pulls for forced exploration. We first bound the expected number of pulls $\mathbb{E}[N_i^{\text{FE}}(T)]$ due to the forced exploration. Considering only the samples collected due to forced exploration, thanks to independence among these samples, we can see the required number of pulls as a sum of geometric random variables. Thus, we can compute an upper bound on the expectation as:

$$\mathbb{E}_{\nu_i}[N_i^{\mathrm{FE}}(T)] \leq \frac{4\log T^3}{\mathbb{P}_{\nu_i}(|X|>0)}. \tag{54}$$

Part II: Bounding the expected number of pulls for optimistic exploration. We define for every arm $i \in [K]$ and every round $t \in [T]$, the upper confidence bound as:

$$B_i(t) = \widehat{\mu}_i(t) + \sqrt{\frac{2V_i(t)\log t^3}{N_i(t-1)}} + \frac{10\widehat{M}_i(t)\log t^3}{N_i(t-1)},$$

where $N_i(t-1)$ is the number of times arm i has been pulled up to time $t-1$, i.e., $N_i(t-1) = |\mathbf{X}_i(t-1)|$. We now show that if $I_t = i$, for an arm i such that $\Delta_i > 0$, then, one of the following four inequalities is true:

$$\text{either } B_1(t) \leq \mu_1, \tag{55}$$

$$\text{or}\quad \widehat{\mu}_i(t) > \mu_i + 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}}, \tag{56}$$

$$\text{or}\quad N_i(t-1) < 20\left(\frac{u}{\Delta_i^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}\log t^3, \tag{57}$$

$$\text{or } \sqrt{V_i(t)} > \sqrt{\mathbb{E}[V_i(t)]} + 2\widehat{M}_i(t)\sqrt{\frac{\log t^3}{N_i(t-1)}}, \tag{58}$$

$$\text{or } \widehat{M}_i(t) \geq \left(\frac{uN_i(t-1)}{\log t^3}\right)^{\frac{1}{1+\epsilon}}. \tag{59}$$

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

$$\begin{aligned}
B_1(t) &\overset{(55)}{>} \mu_1 = \mu_i + \Delta_i \\
&\overset{(56)}{\geq} \widehat{\mu}_i(t) - 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \Delta_i \\
&\overset{(*)}{\geq} \widehat{\mu}_i(t) + \sqrt{\frac{2V_i(t)\log t^3}{N_i(t-1)}} + \frac{10\widehat{M}_i(t)\log t^3}{N_i(t-1)} \\
&= B_i(t).
\end{aligned}$$

The step marked with $(*)$ is a consequence of the fact that both (57), (58) and (59) are false. In particular, we need to show that

$$\Delta_i \geq 5.6u^{\frac{1}{1+\epsilon}}\left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2V_i(t)\log t^3}{N_i(t-1)}} + \frac{10\widehat{M}_i(t)\log t^3}{N_i(t-1)}. \tag{$*$}$$

To do so, we make use of the following inequality derived by exploiting the independence between $\mathbf{X}_i(t-1)$ and $\mathbf{X}'_i(t-1)$:

$$\mathbb{E}[V_i(t)] \leq \mathbb{E}\left[X^2 \mathbb{1}_{|X| \leq \widehat{M}_i(t)}\right] \leq \mathbb{E}\left[|X|^{1+\epsilon}\right] \widehat{M}_i(t)^{1-\varepsilon} \leq u \widehat{M}_i(t)^{1-\epsilon}. \tag{60}$$

Now, we make use of the fact that (57), (58), and (59) are false together with (60):

$$\begin{aligned}
\Delta_i &\overset{(57)}{\geq} 20 u^{\frac{1}{1+\epsilon}} \left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} \\
&\geq (5.6 + \sqrt{2} + 10 + 2\sqrt{2}) u^{\frac{1}{1+\epsilon}} \left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} \\
&= 5.6 u^{\frac{1}{1+\epsilon}} \left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2 \log t^3 u \left(\frac{u N_i(t-1)}{\log t^3}\right)^{\frac{1-\epsilon}{1+\epsilon}}}{N_i(t-1)}} + \frac{(10 + 2\sqrt{2}) \left(\frac{u N_i(t-1)}{\log t^3}\right)^{\frac{1}{1+\epsilon}} \log t^3}{N_i(t-1)} \\
&\overset{(59)}{\geq} 5.6 u^{\frac{1}{1+\epsilon}} \left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2 \log t^3 u \widehat{M}_i(t)^{1-\epsilon}}{N_i(t-1)}} + \frac{(10 + 2\sqrt{2}) \widehat{M}_i(t) \log t^3}{N_i(t-1)} \\
&\overset{(60)}{\geq} 5.6 u^{\frac{1}{1+\epsilon}} \left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2\mathbb{E}[V_i(t)] \log t^3}{N_i(t-1)}} + \frac{(10 + 2\sqrt{2}) \widehat{M}_i(t) \log t^3}{N_i(t-1)} \\
&\overset{(58)}{\geq} 5.6 u^{\frac{1}{1+\epsilon}} \left(\frac{\log t^3}{N_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} + \sqrt{\frac{2 \log t^3}{N_i(t-1)}} \left[\sqrt{V_i(t)} - 2\widehat{M}_i(t) \sqrt{\frac{\log t^3}{N_i(t-1)}}\right] \\
&\qquad + \frac{(10 + 2\sqrt{2}) \widehat{M}_i(t) \log t^3}{N_i(t-1)} \\
&\geq 5.6 u^{\frac{1}{1+\epsilon}} \left[\frac{\log t^3}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{V_i(t) \log t^3}{N_i(t-1)}} + \frac{10 \widehat{M}_i(t) \log t^3}{N_i(t-1)}. \qquad (*)
\end{aligned}$$

Finally, as a consequence of $(*)$, we have $B_1(t) > B_i(t)$ but this is a contradiction since $T_t = i$. Thus, statements (55) to (59) cannot be false simultaneously. We now proceed with a union bound 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 (55), (56), (58), and (59) to be true:

$$\begin{aligned}
&\mathbb{P}\left(\exists N_i(t-1) \in [t] \; : \; \{(55) \text{ is true}\} \text{ or } \{(56) \text{ is true}\} \text{ or } \{(58) \text{ is true}\} \text{ or } \{(59) \text{ is true}\}\right) \leq \\
&\qquad \leq 6 \sum_{s=1}^{t} \frac{1}{t^3} = \frac{6}{t^2},
\end{aligned}$$

where for (58), we used the second inequality of Theorem 10 of (Maurer and Pontil, 2009) (bounding $1/(n-1) \leq 2/n$) and for (59), we used Theorem 5. To proceed, we introduce the quantity:

$$v \coloneqq \left\lceil 60 \left(\frac{u}{\Delta_i^{1+\varepsilon}}\right)^{\frac{1}{\varepsilon}} \log T \right\rceil.$$

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

$$\begin{aligned}
\mathbb{E}[N_i(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t=i\text{and}\mathcal{E}_{i,t}\}}\right] &\leq v + \mathbb{E}\left[\sum_{t=v+1}^{T} \mathbb{1}_{\{I_t=i \text{ and } \{(57) \text{ is false }\}\}}\right] \\
&\leq v + \mathbb{E}\left[\sum_{t=v+1}^{T} \mathbb{1}_{\{I_t=i \text{ and } \{(55) \text{ or } (56) \text{ or } (58) \text{ or } (59) \text{ is true}\}\}}\right] \qquad (61)\\
&\leq v + \sum_{t=v+1}^{T} \frac{6}{t^2} \\
&\leq v + 10.
\end{aligned}$$

We now conclude the proof using the regret decomposition, considering the forced exploration through Equation (54) and that the effective number of pulls is doubled:

$$R_T(\texttt{AdaR-UCB}, \boldsymbol{\nu}) \leq \sum_{i:\Delta_i>0} \left[\left(120\left(\frac{u}{\Delta_i}\right)^{\frac{1}{\epsilon}} + \frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\right)\log\frac{T}{2} + 20\Delta_i\right].$$

■

Theorem 8 (Worst-Case Regret bound of AdaR-UCB) *Let $\boldsymbol{\nu} \in \mathcal{P}_{HT}(\epsilon, u)^K$ and $T \in \mathbb{N}_{\geq 2}$ be the learning horizon. Under Assumption 1, AdaR-UCB suffers a regret bounded as:*

$$R_T(\texttt{AdaR-UCB}, \boldsymbol{\nu}) \leq 46\left(K\log\frac{T}{2}\right)^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\log\frac{T}{2} + 20\Delta_i\right).$$

Proof Let us fix $\Delta > 0$, to be chosen later. We have:

$$\begin{aligned}
R_T(\texttt{AdaR-UCB}, \boldsymbol{\nu}) &= \sum_{i\in[K]} \Delta_i\left(2\mathbb{E}[N_i(T/2)] + \mathbb{E}_{\nu_i}[N_i^{\text{FE}}(T/2)]\right) \\
&= \sum_{i:\Delta_i\leq\Delta} 2\Delta_i\mathbb{E}[N_i(T/2)] + \sum_{i:\Delta_i>\Delta} 2\Delta_i\mathbb{E}[N_i(T/2)] + \sum_{i:\Delta_i>0}\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\log\frac{T}{2} \\
&\leq \Delta T + \sum_{i:\Delta_i>\Delta} 2\Delta_i\left(60\left(\frac{u}{\Delta_i^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}\log\frac{T}{2} + 10\right) + \sum_{i:\Delta_i>0}\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\log\frac{T}{2} \\
&\leq \Delta T + 2K\left(60\left(\frac{u}{\Delta}\right)^{\frac{1}{\epsilon}}\log\frac{T}{2}\right) + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\log\frac{T}{2} + 20\Delta_i\right) \\
&\overset{(*)}{\leq} 120^{\frac{\epsilon}{1+\epsilon}}(1+\epsilon)\epsilon^{-\frac{\epsilon}{1+\epsilon}}\left(K\log\frac{T}{2}\right)^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\log\frac{T}{2} + 20\Delta_i\right) \\
&\overset{(**)}{\leq} 46\left(K\log\frac{T}{2}\right)^{\frac{\epsilon}{1+\epsilon}}(uT)^{\frac{1}{1+\epsilon}} + \sum_{i:\Delta_i>0}\left(\frac{24\Delta_i}{\mathbb{P}_{\nu_i}(X \neq 0)}\log\frac{T}{2} + 20\Delta_i\right),
\end{aligned}$$

where the step marked with $(*)$ follows by a proper choice of Δ minimizing the bound:

$$T - 120Ku^{\frac{1}{\epsilon}}\epsilon^{-1}\Delta^{-\frac{1+\epsilon}{\epsilon}} \log\frac{T}{2} = 0 \implies \Delta = \left(\frac{120Ku^{\frac{1}{\epsilon}} \log\frac{T}{2}}{\epsilon T}\right)^{\frac{\epsilon}{1+\epsilon}},$$

and step marked with $(**)$ follows by bounding simple numerical bounds. ■

Appendix C. Efficient Numerical Resolution of Equation (13)

In this appendix, we present a computationally efficient strategy that can be implemented in Algorithm 1 to execute line 7, *i.e.*, the solution of the root-finding problem. In particular, to solve the equation:

$$f_s(\mathbf{X}'; M, \delta) \coloneqq \frac{1}{s} \sum_{j \in [s]} \frac{\min\{(X'_j)^2, M^2\}}{M^2} - \frac{c \log \delta^{-1}}{s} = 0. \tag{13}$$

Algorithm 2: Computationally Efficient Threshold Estimation.

```
1 Reward set $\mathbf{X}' = \{X'_1, \dots, X'_s\}$, time counter $\tau$, machine tolerance $\eta > 0$.
2 Initialize counter $h \leftarrow 0$, initial guess $x_0 \leftarrow \eta$, initial value $y_0 \leftarrow f_s(\mathbf{X}'; x_0, \tau^{-3})$.
3 while $y_h > 0$ do
4     $x_{h+1} \leftarrow 2x_h$
5     $y_{h+1} \leftarrow f_s(\mathbf{X}'; x_{h+1}, \tau^{-3})$
6     $h \leftarrow h + 1$
7 end
8 Return $x_h$ .
```

We propose Algorithm 2 to find an upper bound $\bar{M}_s(\tau^{-3})$ on the true solution $\widehat{M}_s(\tau^{-3})$ which is based on *bisection*. The strategy works as follows. We provide the minimum numerical tolerance of our machine $\eta > 0$, start from an initial guess $x_0 = \eta$, then, if this guess is an underestimation (*i.e.*, $f_s(\cdot, x_0)$ yields a positive value y_0) we proceed to iteratively double our guess until the real threshold has been passed (lines 4-6). In line 8, we return the final guess x_h. If the initial guess is already an overestimation of the threshold (*i.e.*, $f_s(\cdot, x_0)$ yields a negative value y_0), we simply have $x_0 = x_h = \eta$.

We point out that, by construction, the output of Algorithm 2 can be *at most* two times the true solution to Equation (13), *i.e.*, $\bar{M}_s(\tau^{-3}) \leq 2\widehat{M}_s(\tau^{-3})$. Thus, regret guarantees for Algorithm 1 remain the same (up to numerical constants) even when performing this approximation of the threshold. In particular, in the proof of Theorem 7, we can modify (59) as follows:

$$\bar{M}_i(t) \geq 2 \left(\frac{u N_i(t-1)}{\log t^3} \right)^{\frac{1}{1+\epsilon}},$$

and the final result remains the same up to multiplicative constants.

We now characterize the computational complexity of Algorithm 2, *i.e.*, the maximum number of steps to be performed before returning a solution.

Proposition 10 (Upper Bound on the Number of Steps of Algorithm 2) *Let η be the minimum numerical tolerance, and assume $\eta \leq \widehat{M}_s(\tau^{-3})$. Then, in at most $\bar{h}_{\eta,\tau}(\epsilon, u)$ steps such that:*

$$\bar{h}_{\eta,\tau}(\epsilon, u) = \log_2 \left(\frac{1}{\eta} \left(\frac{us}{\log(\tau^3)} \right)^{\frac{1}{1+\epsilon}} \right),$$

Algorithm 2, returns a solution $x_{\bar{h}_{\eta,\tau}(\epsilon,u)}$ s.t.

$$\mathbb{P}\left(\frac{x_{\bar{h}_{\eta,\tau}(\epsilon,u)}}{\widehat{M}_s(\tau^{-3})} \in [1,2] \right) \geq 1 - \frac{2}{\tau^3}.$$

Proposition 10 states an upper bound for the number of steps of Algorithm 2 as a function of both ϵ and u. However, we remark that these two are not required as input to the numerical solver. Moreover, it emerges a dependence on the inverse of the numerical tolerance of the machine on which the algorithm is run. Thanks to the logarithm, this dependence hardly becomes an issue. If we consider a very small tolerance of 10^{-16} (which is the standard tolerance of many programming languages) the number of steps becomes:

$$\bar{h}_{\eta,\tau}(\epsilon, u) = \log_2 \left(\left(\frac{us}{\log(\tau^3)} \right)^{\frac{1}{1+\epsilon}} \right) + 16 \log_2(10),$$

which is totally reasonable.