Proceedings of Machine Learning Research vol 291:1–5, 2025

Open Problem: Regret Minimization in Heavy-Tailed Bandits with Unknown Distributional Parameters

Gianmarco Genalti GIANMARCO.GENALTI@POLIMI.IT
Politecnico di Milano

Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Politecnico di Milano

Abstract

The heavy-tailed bandit problem (Bubeck et al., 2013), is a variant of the stochastic multi-armed bandit problem where 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, most of the proposed approaches crucially rely on the knowledge of both ϵ and u. Recent works have highlighted that adapting to such parameters when they are unknown is harder than adapting to the subgaussian constant or the rewards range in non-heavy-tailed bandits. It is known that it is not possible to adapt to either ϵ or u without either (i) incurring extra regret or (ii) enforcing additional assumptions. However, it remains an *open question* what the best attainable performance is when no additional assumptions are provided. Moreover, the assumptions proposed in the literature are not comparable, as none of them is strictly weaker than the others. Thus, another *open question* is about the nature of the assumptions needed to compensate for this cost.

Keywords: multi-armed bandit, heavy-tailed noise, regret minimization, adaptation.

1. Introduction and Setting

In the *stochastic multi-armed bandit* (MAB) problem (Lattimore and Szepesvári, 2020), a learner is repeatedly faced with a finite set of *actions*. At each round, the learner chooses one action and receives a *reward* sampled from an unknown associated probability distribution. The learner's goal is to maximize its expected cumulative reward or, equivalently, minimize its expected *regret* w.r.t. the clairvoyant decision-maker that knows the reward expectation. In the most common settings, the distributions are assumed *subgaussian* or with bounded support, *i.e.,* an exponential decrease in the density function dominates the tails, and, consequently, every moment of the distribution is finite. In *heavy-tailed* MABs (HTMAB, Bubeck et al., 2013), instead, the following scenario arises.

Definition 1 (Regret Minimization in Heavy-Tailed Bandits) *In the heavy-tailed multi-armed bandit problem (HTMAB), a learner is faced with $K \in \mathbb{N}$ arms repeatedly for $T \in \mathbb{N}$ rounds. Every time an action $i \in [K]$ is selected, a reward X is sampled from the distribution ν_i satisfying*

$$\mathbb{E}_{X\sim\nu_i}[|X|^{1+\epsilon}] \leq u, \tag{1}$$

for some $\epsilon \in (0, 1]$ and $u \in \mathbb{R}^+$. If ν_i satisfies (1) for every $i \in [K]$, we call $\boldsymbol{\nu} := \{\nu_i\}_{i\in[K]}$ an HTMAB. We denote with $\mathcal{E}(\epsilon, u)$ the set of HTMABs with all reward distributions $\boldsymbol{\nu}$ satisfying

Equation (1) with certain values of ϵ and u. Let $\mu_i \coloneqq \mathbb{E}_{X\sim\nu_i}[X]$ and I_t be the action selected at round $t \in [T]$. Then, the learner's goal is to minimize the expected cumulative regret, defined as:

$$\mathbb{E}[R_T(\boldsymbol{\nu})] \coloneqq \mathbb{E}\Bigg[\sum_{t\in[T]} (\mu^* - \mu_{I_t})\Bigg], \qquad \textit{where} \qquad \mu^* \coloneqq \max_{i\in[K]} \mu_i.$$

We denote with $\Delta_i \coloneqq \mu_* - \mu_i$ the suboptimality gap and with $\mathcal{E}(\epsilon, u, \boldsymbol{\Delta})$ the subset set of HTMABs in $\mathcal{E}(\epsilon, u)$ having $\boldsymbol{\Delta} = (\Delta_i)_{i\in[K]}$ as suboptimality gaps. When the learner *knows* both ϵ and u, there exist several algorithms that achieve an upper bound on the expected regret in the order of $\sup_{\boldsymbol{\nu}\in\mathcal{E}(\epsilon,u,\boldsymbol{\Delta})} \mathbb{E}[R_T(\boldsymbol{\nu})] \le \mathcal{O}(u^{\frac{1}{\epsilon}} \Delta_i^{-\frac{1}{\epsilon}} \ln T)$ (*instance-dependent* bound) and $\sup_{\boldsymbol{\nu}\in\mathcal{E}(\epsilon,u)} \mathbb{E}[R_T(\boldsymbol{\nu})] \le \mathcal{O}((uT)^{\frac{1}{1+\epsilon}} (K \ln T)^{\frac{\epsilon}{1+\epsilon}})$ (*worst-case* bound). The $\mathcal{O}$ notation only hides universal constants. These upper bounds have been proved to be *tight*, *i.e.*, they can only be improved up to constants or logarithmic terms (Bubeck et al., 2013).

The knowledge over ϵ and u is crucial, and no algorithm can achieve such tight bounds without knowing them or relying on additional assumptions. We are now ready to state our first two open questions:

Open Question 1 *What is the best regret rate that can be incurred in the HTMAB problem without any knowledge on ϵ and u and without any additional assumption?*

Open Question 2 *There exists an algorithm that can achieve such a bound without knowing ϵ and u and without any additional assumption?*

Answering these requires both a minimax lower bound (over unknown (ϵ, u)) and an algorithm with a matching upper bound. Equivalently, find suitable f, f', g, g' such that

$$\sup_{\epsilon\in(0,1],\, u\in\mathbb{R}^+} \sup_{\boldsymbol{\nu}\in\mathcal{E}(\epsilon,u)} \frac{\mathbb{E}[R_T(\boldsymbol{\nu})]}{f'(\epsilon, u)} \substack{\geq \\ <} g'(T, K), \qquad \sup_{\epsilon\in(0,1],\, u\in\mathbb{R}^+} \sup_{\boldsymbol{\nu}\in\mathcal{E}(\epsilon,u,\boldsymbol{\Delta})} \frac{\mathbb{E}[R_T(\boldsymbol{\nu})]}{f(\epsilon, u)} \substack{\geq \\ <} g(T, K, \boldsymbol{\Delta}).$$

2. Existing Results

In this section, we discuss the existing literature on the problem of adaptation in MABs. In Table 1, we provide a comprehensive comparison of the approaches that tackled the problem of (partial) adaptation to ϵ and u in HTMABs. We point out some relevant facts.

The Cost of Adaptation. Without any additional assumption, it is not possible to achieve a regret upper bound matching the lower bound from Bubeck et al. (2013). This has been proven in Theorems 2 and 3 of Genalti et al. (2024). However, these results are qualitative, and account for either adaptation in ϵ or u, not simultaneously both. Still it is not known whether these lower bounds are tight and no algorithm is currently able to match them.

Different Assumptions. The approaches of Table 1 rely on different assumptions. The algorithms presented in Ashutosh et al. (2021) rely on a technical assumption that requires the time horizon T to be larger than a quantity depending on both u and ϵ. Both `R-UCB-TEA` and `R-UCB-MoM` only have been analyzed from an instance-dependent perspective. Moreover, their assumption is the only one depending on ϵ and u (making it harder to validate, in practice, without the knowledge of ϵ and

Algorithm	Regret Bounds				ϵ-adaptive		u-adaptive		Assumption
	Instance-dependent	Matching?[§]	Worst-case	Matching?[¶]	Estimator	Algorithm	Estimator	Algorithm	
`OptHTINF` (Huang et al., 2022)	$\sum_{i:\Delta_i>0} \left(\frac{u^2}{\Delta_i^{2-\epsilon}}\right)^{1/\epsilon} \log T$	✗	$K^{\frac{\epsilon}{2}} u^{\frac{1}{1+\epsilon}} T^{\frac{2-\epsilon}{2}}$	✗	✓	✓	✓	✓	Truncated Non-Negativity
`AdaTINF` (Huang et al., 2022)	—	—	$K^{\frac{\epsilon}{1+\epsilon}} u^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$	✓	✓	✓	✓	✓	Truncated Non-Negativity
`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. $\begin{matrix} g(T)<\frac{\epsilon}{1+\epsilon} \\ f(T)>(12u)^{\frac{1}{1+\epsilon}} \end{matrix}$
`RMM-UCB` (Tamás et al., 2024)	$\sum_{i:\Delta_i>0} \left(\frac{u}{\Delta_i}\right)^{1/\epsilon} \log^2 T$	✗	$K^{\frac{\epsilon}{1+\epsilon}} u^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} (\log T)^{\frac{3\epsilon}{1+\epsilon}}$	✓	✓	✓	✓	✓	Simmetric Distribution
`uniINF` [⋆] (Chen et al., 2024)	$K\left(\frac{u^{1+\epsilon}}{\Delta_{min}}\right)^{1/\epsilon} \log T$	✗	$K^{\frac{\epsilon}{1+\epsilon}} u^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} (\log T)^{\frac{\epsilon}{1+\epsilon}}$	✓	✓	✓	✓	✓	Truncated Non-Negativity
`AdaR-UCB` (Genalti et al., 2024)	$\sum_{i:\Delta_i>0} \left(\left(\frac{u}{\Delta_i}\right)^{1/\epsilon} + \frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)}\right) \log T$	✓	$K^{\frac{\epsilon}{1+\epsilon}} u^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} (\log T)^{\frac{\epsilon}{1+\epsilon}}$ $+ \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathbb{P}_{\nu_i}(X\neq 0)} \log T$	✓	✓	✓	✓	✓	Truncated Non-Positivity

[‡] f and g are to be given in input. Choosing an optimal value of those would require knowing ϵ and u.

[⋆] $\Delta_{min} := \min_{i\in[K]} \Delta_i$.

[§] 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.

u). In Tamás et al. (2024), the authors require the *symmetry* of the reward distribution w.r.t. the mean. Their algorithm, `RMM-UCB`, achieves nearly optimal worst-case guarantees (up to logarithmic terms). The remaining works use the *truncated non-negativity/positivity* (negativity for losses, positivity for rewards), introduced in Huang et al. (2022). The truncated non-positivity assumption (TNP, Genalti et al. (2024)) requires that the optimal action $i^* \in \arg\max_{i\in[K]} \mu_i$ satisfy:

$$\mathbb{E}_{\nu_{i^*}}[X\mathbb{1}_{\{|X|>M\}}] \leq 0, \quad \text{for every } M \in \mathbb{R}^+.$$

When dealing with losses instead of reward, we flip the inequality and obtain the truncated non-negativity assumption. In Theorem 4 of Genalti et al. (2024), the authors prove that this assumption does not reduce the known lower bound rate, and thus it does not make the problem easier. Intuitively, this assumption enforces that an estimation procedure based on truncations does not negatively bias the best action's estimated mean reward. Condition (1) allows for many pathological distributions, such as the ones used in the lower bound proof of Bubeck et al. (2013). It implies that the left tail of the distribution is heavier than the right one. Also, it imposes that the *belly* of the distribution is well-behaved. Remarkably, the Pareto distribution does not satisfy this assumption (the left tail does not exist), but the trimmed mean estimator's bias can be bounded in such a way that algorithms like `AdaR-UCB` obtain tight regret bounds even if the assumption is violated. To see this, we can directly plug the CDF of a Pareto distribution into the proof of Lemma 2 of Genalti et al. (2024). In the step marked with $(*)$, the first addendum is bounded with 0 using the assumption. Instead, one can express it explicitly, and it has the same order of the variance term. Thus, the TNP assumption fails somehow to encompass all the distributions for which adaptation is possible without additional costs. On the other hand, the other assumptions are not comparable with TNP, and are neither weaker nor stronger. For instance, the *symmetry assumption* also excludes the Pareto distribution. Similarly, the assumptions used by Ashutosh et al. (2021) are dependent on u and ϵ, which may not be desirable in applications. We conclude this paragraph with another interesting open question:

Open Question 3 *Is there an assumption that is "better" than the others, or that encompasses all of the distributions for which adaptation comes at no additional cost?*

Adaptation in Non-Heavy-Tailed MABs. In Cowan et al. (2018), the authors propose an algorithm with optimal guarantees when the distribution is Gaussian with unknown variance. In Lattimore (2017), an upper bound on the kurtosis is required to adapt to the unknown range and variance with tight guarantees. Finally, in Hadiji and Stoltz (2023), the authors propose a fully adaptive approach where no knowledge of the range or the variance is provided to the algorithm, and no additional assumptions are introduced. Their algorithm, named `AHB`, modifies the well-known `AdaHedge` strategy to make it adaptive to the range. Moreover, they provide an important impossibility result. Let $\Phi_{free}(T)$ and $\Phi_{dep}(T)$ be the worst-case and the asymptotic instance-dependent regret bound rates, respectively. Then, we have:

$$\Phi_{dep}(T)\Phi_{free}(T) \geq T. \tag{2}$$

This result has two major implications. First, logarithmic regret cannot be achieved. Second, to achieve a $\sqrt{T}$ worst-case rate, it is necessary to pay a $\sqrt{T}$ instance-dependent rate. This trade-off partially characterizes the price of adaptation. Their approach, `AHB`, controls this trade-off using a hyperparameter that weights the distribution-free rate against the distribution-dependent one. A natural idea would be to translate the bound in (2) to the heavy-tailed scenario together with the `AHB` approach. If we consider the problem of adapting to u only (which is the analogy of adapting to the range or the subgaussianity constant), we can get the following:

$$\Phi_{dep}(T)\Phi_{free}(T)^{\frac{1+\epsilon}{\epsilon}} \geq T^{\frac{1+\epsilon}{\epsilon}}. \tag{3}$$

A formal statement (together with a proof) of this result can be found in Genalti and Metelli (2025). If we impose the two rates to be equal, *i.e.* $\Phi_{dep} = \Phi_{free}$, we get that $\Phi_{free}(T) \geq \Omega\left(T^{\frac{1+\epsilon}{1+2\epsilon}}\right)$. When $\epsilon = 1$, we have $\Omega\left(T^{\frac{2}{3}}\right)$, which is higher than the $\Omega\left(\sqrt{T}\right)$ lower bound obtained in bounded range bandits.

3. Conclusions

Heavy-tailed distributions are suitable models for representing real-world phenomena characterized by high variability and challenging estimation problems. However, knowledge (and estimation) of their parameters, such as ϵ and u, is often unrealistic, and their estimation is not possible. Therefore, regret minimization in this adaptive setting is an important problem, and understanding the inherent complexity is certainly of interest to the COLT community, just as in the non-heavy-tailed case, which is currently well understood.

References

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.

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

Yu Chen, Jiatai Huang, Yan Dai, and Longbo Huang. uniinf: Best-of-both-worlds algorithm for parameter-free heavy-tailed mabs. *arXiv preprint arXiv:2410.03284*, 2024.

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.

Gianmarco Genalti and Alberto Maria Metelli. A regret lower bound for u-adaptive heavy-tailed bandits. Technical Note, available at https://gianmarcogenalti.github.io/papers/lowerboundheavytail/technicalnoteHT.pdf, 2025.

Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, and Alberto Maria Metelli. (ε, u)-adaptive regret minimization in heavy-tailed bandits. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1882–1915. PMLR, 2024.

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.

Ambrus Tamás, Szabolcs Szentpéteri, and Balázs Csanád Csáji. Data-driven upper confidence bounds with near-optimal regret for heavy-tailed bandits. *arXiv preprint arXiv:2406.05710*, 2024.