POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Towards Fully-Adaptive Regret Minimazion in Heavy-Tailed Bandits

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Lupo Marsigli

Advisor: Prof. Alberto Maria Metelli

Co-advisor: Dott. Gianmarco Genalti

Academic year: 2022-2023

1. Introduction

In this thesis, we investigate the stochastic *multi-armed bandit problem* (MAB) [3] under the assumption of *heavy-tailed* (HT) reward distributions. In the classic stochastic multi-armed bandit setting, an agent has access to a set of K possible actions (*i.e.*, *arms*). Each arm $i \in [K] \coloneqq \{1, \dots, K\}$ is associated with a reward probability distribution ν_i, having finite mean μ_i. At every round $t \in [T]$, being T a learning horizon, after an action I_t is selected, a reward X_t is sampled from ν_{I_t} and revealed to the agent. The goal of the agent is to minimize its *expected regret* after T rounds, defined as:

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

where $\Delta_i \coloneqq \max_{j \in [K]} \mu_j - \mu_i$ is the sub-optimality gap for all $i \in [K]$.

Most of the existing works assume that the reward probability distributions ν_i are *sub-Gaussian*. Under this assumption, the tails of the distribution present a strong decay (at least as fast as that of the Gaussian distribution). An important implication is that every moment of finite order is finite. While this assumption enables the application of powerful theoretical tools and, consequently, strong regret guarantees, it is often limiting in many practical scenarios such as, for example, financial environments or network routing problems. In settings where uncertainty has a significant impact, *heavy-tailed distributions* naturally arise. In these cases, the tails decay slower than a Gaussian, and the moment-generating function is no longer assumed to be finite. As a consequence, the moments of any finite order might not exist.

In this work, we investigate the regret minimization problem for heavy-tailed MAB, according to the setting introduced in the seminal work [1].

Assumption 1. *Given a bandit instance $(\nu_i)_{i \in [K]}$, we assume moments of order up to $1 + \epsilon$, with $\epsilon \in (0, 1]$, to be finite and uniformly bounded by a constant u, namely:*

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

In the heavy-tailed MAB problem, it is canonical to assume the knowledge of both ϵ and u and, to the best of our knowledge, every regret minimization strategy in the stochastic HT bandit literature which is optimal in the instance-dependent case, requires at least one of them as

an algorithm's input.
The goal of this work is to answer the following question:
Is it possible to provide a novel regret minimization strategy in the heavy-tailed bandit problem that does not require any prior knowledge on ϵ nor u, but still achieves comparable performances to other approaches knowing them?
In this work, we will prove that in general it is not possible to achieve the same order of performance while being adaptive to the aforementioned unknown quantities. Fortunately, we will show that under a specific distributional assumption, the answer is, instead, affirmative. In particular, we will discuss the role of the *truncated non-positivity* assumption, and show that, when this assumption is violated, it is not possible anymore to guarantee the existence of an adaptive algorithm with respect to ϵ nor u achieving state-of-the-art performances. We introduce `Adaptive Robust UCB` (shortly `AdaR-UCB`), an algorithm based on the *optimism in the face of uncertainty* principle that is capable of being *fully adaptive* w.r.t. the two parameters ϵ and u that characterize the reward distributions. In particular, we propose a modification of the well-known `Robust UCB` algorithm from [1], showing that under our assumption we are able to attain the same theoretical guarantees.

2. Setting

We now introduce the stochastic heavy-tailed multi-armed bandit problem. Formally speaking, there are $K \geq 2$ available actions and a sequence of T rounds. To each arm $i \in [K]$, we associate a probability distribution ν_i satisfying Assumption 1. At each round $t \in [T]$, the agent can choose an index I_t and subsequently collects a reward of X_t, which is an independent sample from ν_{I_t}. The agent is allowed to make a decision at round t by considering all history up to time $t-1$. We remark that distributions $(\nu_i)_{i\in[K]}$ only admit finite moments up to order $1+\epsilon$, with $\epsilon \in (0, 1]$, thus, distributions with infinite variance are allowed in this problem formulation. The goal of the agent is to minimize the regret as defined in Equation (1).

In the heavy-tailed bandits literature, is customary to assume the knowledge on both ϵ and the upper bound on the $(1+\epsilon)$-th order moment u, which is assumed to be common for all $(\nu_i)_{i\in[K]}$ without loss of generality. In our setting, we will remove this constraint which is unfeasible in real-world, considering both quantities to be unknown to the agent. From now on, we will refer to any algorithm operating without this knowledge as *adaptive w.r.t. ϵ or u*, depending on which one is unknown (possibly both). In a specific bandit setting, we will also say that the upper bound on regret suffered by an algorithm *matches* the lower bound, if the two have the same order in T up to constants. In this case, we can refer to the algorithm as *tight* or *optimal*.

3. Related Works

The stochastic heavy-tailed bandit model was first introduced by Bubeck et al. [1], who designed `Robust UCB`, an algorithm that assumes ϵ and u are both known to the agent. Lately, more contributions started to propose approaches towards adaptive settings, at first only w.r.t u, still assuming ϵ as known. Other recent research tried to tackle the fully-adaptive setting, *i.e.*, considering both the parameters unknown. Despite that, these works usually assessed different performance targets than regret minimization, so that the only work that currently attempted to get closer to our goal is [2]. Huang et al. [2022] presented a fully adaptive algorithm which is optimal in the instance-independent case, only under a weak assumption on the losses. For the instance-dependent one, on the contrary, they proposed another approach that still requires no prior knowledge, but gives a sub-optimal regret compared to [1]. And here is where the thesis enters the game, since, to the best of our knowledge, nobody yet presented any approach that is feasible in real-world, *i.e.* any fully adaptive algorithm for the stochastic HT bandit problem, which achieves a tight instance-dependent regret and, simultaneously, an instance-independent one optimal at most up to logarithmic terms in T.

4. Lower Bound on the Regret for Adaptive Heavy-Tailed Bandits

In this section, we state a lower bound on the expected regret that any adaptive algorithm (w.r.t.

either u or ϵ) can achieve in the heavy-tailed bandit problem. We start by stating the lower bound on the regret when ϵ and u are known.

Theorem 4.1 (Lower Bounds on Regret for Stochastic Heavy-Tailed Bandit, adapted from [1]). *For any algorithm and for any fixed T, there exists a set of K distributions satisfying Assumption 1, such that:*

$$R_T \geq \Omega\left(\sum_{i:\Delta_i>0}\left(\frac{u}{\Delta_i}\right)^{\frac{1}{\epsilon}}\log T\right), \tag{3}$$

$$R_T \geq \Omega\left((uT)^{\frac{1}{1+\epsilon}}K^{\frac{\epsilon}{1+\epsilon}}\right). \tag{4}$$

The following two results show that any algorithm unaware of ϵ or u, respectively, cannot achieve the same regret order as the one depicted in Theorem 4.1.

We first state the regret lower bound for any algorithm adaptive with respect to u.

Theorem 4.2 (Lower Bound on Regret for Stochastic Adaptive Heavy-Tailed Bandit, unknown u). *For any algorithm adaptive w.r.t. to the $(1+\epsilon)$-th order moment of reward distributions, and for any fixed T, there exist two stochastic heavy-tailed bandit instances satisfying (2) with u and u' respectively (assume $u' > u$ without loss of generality), such that:*

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

where R_T and R'_T are the regrets suffered by this algorithm in the two instances, respectively, and C_1 is a constant independent of u, u' and T.

This result states that there exist two particular heavy-tailed bandit problem instances s.t. no algorithm can match the lower bound on regret presented in (4) on both, and instead some regret is accrued in a way that is proportional to the ratio between u' and u. Since it can be taken arbitrarily large, it is not possible to be adaptive in u without the risk of incurring in an arbitrarily big regret bound.

Next, we present a similar result concerning adaptivity with respect to the maximum order $1+\epsilon$ of finite moment.

Theorem 4.3 (Lower Bound on Regret for Stochastic Adaptive Heavy-Tailed Bandit, unknown ϵ). *For any algorithm adaptive with respect to ϵ, with the maximum order finite moment u known, and for any fixed T, there exist two stochastic heavy-tailed bandit instances satisfying (2) with ϵ and ϵ' respectively (assume $\epsilon' < \epsilon$ without loss of generality), such that:*

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

where R_T and R'_T are the regrets suffered by this algorithm in the two instances, respectively, and C_2 is a constant independent of ϵ, ϵ' and T.

Differently from Theorem 4.2, since the values of ϵ and ϵ' are known to belong to the set $(0, 1]$, then, for any fixed T, the term on the right-hand side of (5) cannot grow arbitrarily. For instance, when $\epsilon = 1$ and $\epsilon' = \frac{1}{3}$, the gap's order is $\approx T^{\frac{1}{16}}$, which gives an intuition on how being adaptive with respect to unknown ϵ is an easier task rather than adapting to unknown u.

To wrap up, we have shown how any algorithm adaptive w.r.t. either u or ϵ has a higher regret lower bound than the one of the non-adaptive heavy-tailed bandit problem. We remark that the two bounds introduced in this section refer to adaptivity with respect to only one of the unknown quantities. As a future research direction, it could be interesting to investigate if simultaneous adaptivity to both quantities implies an even higher lower bound.

5. A Fully Adaptive Approach for Bandits with Heavy Tails

In this section, we finally give an answer to our original research question, *i.e.*, whether there exists an algorithm adaptive w.r.t. both ϵ and u matching the standard setting's lower bound stated in Theorem 4.1. In the previous section, we already shown how adaptivity has a cost, and thus the lower bound presented in Theorem 4.1 is not achievable by any algorithm unaware of at least one of these quantities. Luckily, it is possible to restrict the set of adaptive heavy-tailed bandit problem instances under analysis to a special set, that will be defined in a short, on which our algorithm, `Adaptive Robust UCB`, is able to achieve a regret order matching the instance dependent lower bound for the standard heavy-tailed bandit problem.

5.1. The Truncated Non-Positivity Assumption

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

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

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

This assumption, intuitively, requires the optimal arm of a heavy-tailed bandit instance to have more mass on the negative semi-axis, but still allows the distribution to have an arbitrary support covering, potentially, all $\mathbb{R}$.
The two lower bounds in Theorems 4.2 and 4.3 have been obtained by introducing four instances that violate this assumption, and thus the lower bound on regret for the adaptive heavy-tailed bandit problem under the truncated non-positivity assumption can be smaller than the ones presented in Section 4. However, it is possible to show that forcing the truncated non-positivity assumption does not result in an improvement of the lower bounds in Theorem 4.1.

5.2. A Fully Adaptive Algorithm: `Adaptive Robust UCB`

We are now ready to introduce Algorithm 1, namely `Adaptive Robust UCB`, an algorithm able to operate in the heavy-tailed bandit problem *without any prior knowledge on ϵ nor u.* `AdaR-UCB` is an optimism in the face of uncertainty based algorithm, built upon the `Robust UCB` strategy from [1] using a modified version of the *trimmed mean estimator.* This estimator is a common one in the heavy-tailed statistics literature, since it allows robustness against extreme values.
More formally, we define the trimmed mean estimator for the mean of a set of independent observations $\mathbf{X} = \{X_1, \dots, X_s\}$ as:

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

where $M > 0$ is a given threshold.

Algorithm 1 `Adaptive Robust UCB`

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

$$\frac{1}{s_i} \sum_{j \in [s_i]} \min \left\{ \frac{(X'_{i,j})^2}{\widehat{M}^2_{i,s_i,t}}, 1 \right\} - 25 \frac{\log(t^4)}{s_i} = 0$$

    Compute trimmed observations $\mathbf{Y_{i,t}}$, with its j-th component $Y_{i,j,t}$:

$$Y_{i,j,t} \leftarrow X_{i,j} \mathbb{1}_{|X_{i,j}| \leq \widehat{M}_{i,s_i,t}} \quad \forall j \in [s_i]$$

    Compute trimmed mean estimator:

$$\hat{\mu}_{i,s_i,t} \leftarrow \frac{1}{s_i} \sum_{j \in [s_i]} Y_{i,j,t}$$

    Compute sample variance $V_{i,s_i,t}(\mathbf{Y_{i,t}})$ as:

$$\frac{1}{s_i(s_i-1)} \sum_{l,j \in [s_i]} \frac{(Y_{i,l,t} - Y_{i,j,t})^2}{2}$$

  end for
  Select an action i_t as:

$$\operatorname*{argmax}_{i \in [K]} \left\{ \hat{\mu}_{i,s_i,t} + 2\sqrt{\frac{V_{i,s_i,t}(\mathbf{Y_{i,t}}) \log(t^4)}{s_i}} + 19 \frac{\widehat{M}_{i,s_i,t} \log(t^4)}{s_i} \right\}$$

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

In the `Robust UCB` algorithm, the trimmed mean estimator replaces sample average in a standard optimism in the face of uncertainty strategy, by selecting at each round t the action i maximising the sum of the estimator with a proper upper confidence bound. `AdaR-UCB` operates in the same way, but while in `Robust UCB` the thresh-

old choice is driven by the values of ϵ and u, `AdaR-UCB` computes a proxy threshold $\widehat{M}$ for M without resorting to either ϵ or u (or any estimation of them). In particular, our chosen threshold $\widehat{M}_{s,t}$ is dynamic in time and can be computed as the solution (in M) of $f_{s,t}(\mathbf{X}; M) = 0$, *i.e.*,

$$\frac{1}{s}\sum_{j\in[s]}\frac{\min\left\{X_j^2, M^2\right\}}{M^2} - \frac{25\log\left(t^4\right)}{s} = 0. \quad (8)$$

We remark that, since Equation (8) has some randomness introduced by $(X_j)_{j\in[s]}$, the threshold is a positive random variable, and not simply a parameter.

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

Theorem 5.1 (Upper Bound on Regret for `AdaR-UCB`). *Given a heavy-tailed bandit problem instance satisfying Assumption 2, the regret of* ***`AdaR-UCB`*** *then satisfies:*

$$R_T \leq \sum_{i:\Delta_i>0}\left(160\left(\frac{40u}{\Delta_i}\right)^{\frac{1}{\varepsilon}}\log T + 7\Delta_i\right). \quad (9)$$

First, we point out that *this result provides a positive answer to our initial research question*, since the upper bound matches the order of the regret lower bound for the classic scenario, even when both ϵ and u are unknown.
Finally, as customary in the bandit literature, we also provide an instance-independent version of the upper bound on regret of `AdaR-UCB`.

Theorem 5.2 (Instance-Independent Upper Bound on Regret for `AdaR-UCB`). *Given any heavy-tailed bandit problem instance with K arms that satisfies Assumption 2, if horizon T is such that:*

$$\log T \geq \max_{i\in[K]}\left\{\frac{7\Delta_i^{\frac{1+\varepsilon}{\varepsilon}}}{160(40u)^{\frac{1}{\varepsilon}}}\right\},$$

then the regret of ***`AdaR-UCB`*** *satisfies:*

$$R_T \leq T^{\frac{1}{1+\epsilon}}(320K\log T)^{\frac{\epsilon}{1+\epsilon}}(40u)^{\frac{1}{1+\epsilon}}. \quad (10)$$

Thus, we have showed that, under Assumption 2, it is possible to be adaptive w.r.t. both ϵ and u while attaining the best regret order achievable in the heavy-tailed bandit problem.

6. Numerical Simulations

Given the theoretical novelties presented, we need to validate empirically the performance of `AdaR-UCB` algorithm, by comparing it with some state of the art regret minimization algorithms, e.g., `UCB1` [3] and `Robust UCB` [1]. The instances considered have rewards distributed as generalized Pareto with infinite variance, which is the custom heavy-tailed distribution in literature. The probability density function can present only one tail, either on the positive or negative axis, or can be double-tailed.
We start considering heavy-tailed bandit instances having an optimal arm that is truncated non-positive, satisfying Assumption 2.
For *simulation 1*, we evaluate the performances of our three reference algorithms on an instance with all the arms distributed as Pareto with negative tail. The expected cumulative regrets are reported in Figure 1, over a time horizon $T = 25000$. We note that `Robust UCB` is always run with the right values of parameters u and ϵ, computed analytically.

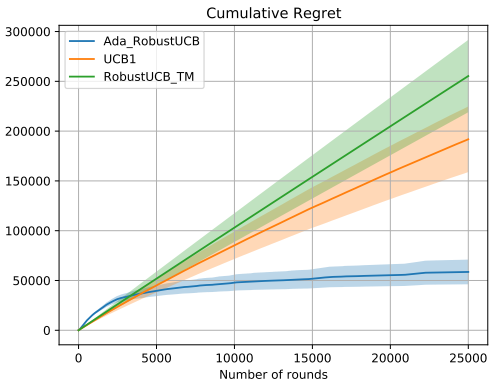

Figure 1: Simulation 1 - Numerical results - Baseline regret comparison, 20 runs (shaded areas are standard deviations).

In this case, `AdaR-UCB` performs way better than `UCB1` and `Robust UCB`, with a clearly sub-linear regret that tends to flatten fast. On the other side, `UCB1` and `Robust UCB` algorithms show a regret behaviour only slightly sub-linear, almost linear, with a slower convergence with respect to our algorithm.
This trend of behaviour replicates also for a more difficult instance that still satisfies Assumption 2, but has Pareto rewards both with negative tail only, and double-tailed. The regret suffered by the algorithms in *simulation 2* is reported in

Figure 2, with `Robust UCB` consistently showing poor performances, even if we have theoretical results that guarantee a logarithmic regret.

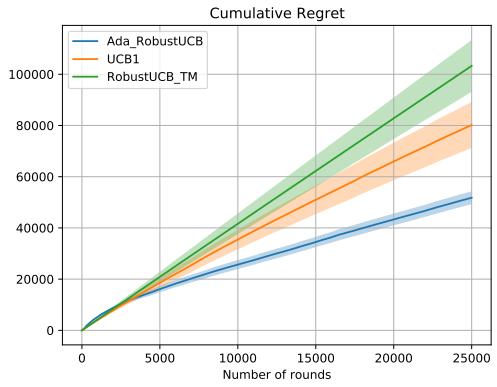

Figure 2: Simulation 2 - Numerical results of regret comparison, 20 runs.

Eventually, we can consider some bandit instances such that the optimal arm does not satisfy Assumption 2. In this case, the theoretical results presented in Section 5 for `AdaR-UCB` do not hold, thus we do not have any guarantee on the behaviour of its cumulative regret. The performance of `AdaR-UCB` is not predictable in this setting, with its regret curve that can grow fast and flatten with time, can be logarithmic, slightly sub-linear or even linear. Considering an instance with mixed Pareto distributions, the execution of the algorithms suffers regret results as in Figure 3.

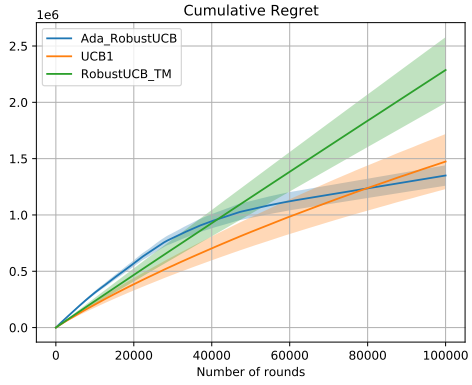

Figure 3: Simulation 3 - Numerical results of regret comparison, 20 runs, T=100000.

Even in more general heavy-tailed instances, `AdaR-UCB` algorithm still shows a slope of cumulative regret which decreases faster than `UCB1` and `Robust UCB`.

7. Conclusions

In this thesis, we studied the *adaptive heavy-tailed bandit* problem, a variation on the classical heavy-tailed bandit problem where no information is provided to the agent regarding the moments of the distribution, not even which of them are finite.
The first results concern the intrinsic difficulty of the setting, for which two novel lower bounds have been provided. In particular, we proved that without any additional assumption no algorithm can match the performances of the non-adaptive setting.
Finally, we provided a novel algorithm, namely `Adaptive Robust UCB` (`AdaR-UCB`), that, under a specific distributional assumption over the optimal arm, is able to achieve the state-of-the-art performances of the standard heavy-tailed bandit problem.
We validated numerically the design choices of our solution in a synthetic environment. In general, for heavy-tailed bandit instance, `AdaR-UCB` outperforms the other two well-known baselines algorithms, namely `Robust UCB` and `UCB1`.

Future directions of investigation regard the role of the *truncated non-positivity* assumption. In particular, we wonder if is it possible to find a weaker assumption ensuring this kind of performances for an algorithm. Moreover, future work should also provide theoretical guarantees on how `AdaR-UCB` performs on bandit instances not satisfying Assumption 2.

References

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

[2] 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.

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