POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Towards Fully-Adaptive Regret Minimization in Heavy-Tailed Bandits

Tesi di Laurea Magistrale in
Mathematical Engineering

Author: **Lupo Marsigli**

Student ID: 968978
Advisor: Prof. Alberto Maria Metelli
Co-advisors: Dott. Gianmarco Genalti
Academic Year: 2022-2023

A mia mamma,

Abstract

The stochastic multi-armed bandit problem is well understood in settings where the reward distributions are subgaussian or have bounded support. In this thesis, we revisit the classic regret minimization problem in the stochastic bandit setting where the arm distributions are allowed to be heavy-tailed. We work under the weaker assumption that the distributions have finite moments of maximum order $1 + \epsilon$, with $\epsilon \in (0, 1]$, which are uniformly bounded by a constant u. These heavy-tailed distributions naturally arise in common real-world contexts where uncertainty has a significant impact, including finance, telecommunications and network traffic. Thus, it is important to understand and develop a general theory and efficient algorithms, both computationally and statistically, that have wider applicability and are tailored to handle such practical scenarios effectively. Any approach that requires prior knowledge on the distribution parameters is not applicable in these cases, since ϵ and u are not known in real-world, and not even easily quantifiable. In this work, we move towards the *adaptive heavy-tailed bandit setting*, where no information is provided to the agent regarding the moments of the distribution, not even which of them are finite. Besides proving that no algorithm, without any additional assumption, can match the performances of the non-adaptive setting, the main novelty we introduce in the thesis is `AdaR-UCB`. It is an algorithm based on the *optimism in the face of uncertainty* principle that is capable of being *fully adaptive* with respect to the two parameters ϵ and u. Under a specific but not restrictive distributional assumption, namely the *truncated non-positivity*, it is able to achieve performances comparable to the optimal approaches knowing the aforementioned quantities. To the best of our knowledge, this is the first regret minimization strategy in the stochastic adaptive heavy-tailed bandit setting that does not require any prior knowledge on ϵ nor u, but still achieves optimality, with a regret upper bound matching the well known lower bound for the standard non-adaptive setting. Finally, we evaluate numerically and compare our algorithm on a range of heavy-tailed bandit instances, noting that it outperforms several state-of-the-art baselines.

Keywords: Multi-Armed-Bandit, Distributions with Heavy Tails, Regret Minimization, Full Adaptivity, Optimism in the Face of Uncertainty, Truncated Non-Positivity.

Abstract in lingua italiana

Il modello Multi-Armed Bandit stocastico è stato ampiamente studiato in ambienti dove le ricompense sono variabili aleatorie subgaussiane o con supporto finito. In questa tesi, rivisitiamo il classico problema di minimizzazione del *regret* nel contesto stocastico dove le distribuzioni delle azioni sono *a coda pesante*. Lavoriamo sotto l'assunzione che abbiano momento finito di massimo ordine $1 + \epsilon$, con $\epsilon \in (0, 1]$, maggiorato uniformemente dalla costante u. Queste distribuzioni con code pesanti emergono naturalmente in contesti del mondo reale, quali la finanza, le telecomunicazioni e le reti di traffico. È quindi importante sviluppare una teoria generale e algoritmi efficienti che abbiano un'applicabilità più ampia e gestiscano efficacemente questi scenari pratici. Qualsiasi approccio che richieda una conoscenza a priori sui parametri delle distribuzioni non è applicabile in questi casi, poiché ϵ e u non sono noti né quantificabili nel mondo reale. In questo lavoro, ci orientiamo verso lo studio dei modelli stocastici *bandit* in un contesto adattivo, in cui non viene fornita all'agente alcuna informazione sui momenti delle distribuzioni, nemmeno su quali siano finiti. Oltre a dimostrare che nessun algoritmo, senza alcuna assunzione aggiuntiva, è in grado di eguagliare le prestazioni degli approcci non adattivi, la principale novità introdotta nella tesi è `AdaR-UCB`. Si tratta di un algoritmo basato sul principio dell'*ottimismo di fronte all'incertezza*, che è in grado di essere *completamente adattivo* rispetto ai due parametri ϵ e u. Sotto una specifica ma non restrittiva assunzione di *non-positività troncata*, è in grado di raggiungere risultati paragonabili agli approcci ottimali che conoscono le suddette quantità. Al meglio delle nostre conoscenze, questa è la prima strategia di minimizzazione del *regret*, nell'ambito del modello *bandit* stocastico adattativo a code pesanti, che non richiede alcuna conoscenza a priori né di ϵ né di u, ma raggiunge comunque l'ottimalità, con un limite superiore sul *regret* che combacia con il ben noto limite inferiore per il problema standard non adattativo. Infine, valutiamo numericamente e confrontiamo il nostro algoritmo su una serie di istanze a coda pesante, notando che empiricamente ha performance migliori dello stato dell'arte in letteratura.

Parole chiave: Modello Multi-Armed-Bandit, Distribuzioni con Code Pesanti, Minimizzazione del *Regret*, Completa Adattività, Non-Positività Troncata.

Contents

1 | Introduction

In numerous practical scenarios, decision makers perform a series of choices over time, where each one can impact future outcomes and decisions. These challenges, termed as *sequential decision-making tasks* [54], are prevalent in areas like automated control, robotics, gaming, and financial sectors.

In solving these kinds of problems, Reinforcement Learning (RL) algorithms [90] are employed, as a specific differentiation of Machine Learning (ML) techniques [78]. In a typical RL task, an agent has to take actions in an environment. The execution of an action has multiple consequences. It generates a reward for the agent and changes the state of the environment. The goal of the agent is to choose actions to maximize the cumulative reward. Of course, the difficulty for the agent is that the consequences of the actions are stochastic and, sometimes, it may be better to sacrifice immediate reward to gain a larger long-term reward. Such tasks, indeed, necessitate the decision-making entity to consider the long-term implications of its choices and the unpredictable nature of its surrounding. The dynamics of the environment are not known, and the agent's knowledge is incomplete, implying that the decision strategy must be learnt through interaction.

The concept of Multi-Armed Bandit (MAB) [14] has emerged as a cornerstone in the realm of these decision-making algorithms under uncertainty, since it provides simple mathematical models for these type of challenges and dilemmas we all face. In particular, bandit problems are less general than RL problems because of their nature in which the state of both the environment and the agent is fixed, i.e., it is not influenced by the chosen action.

The name "Multi-Armed Bandit" is derived from the analogy of a gambler at a casino, who faces multiple slot machines (or "one-armed bandits"), each one promising a potential reward with unknown payout probabilities. The gambler wants to maximize her earnings by playing the arm with highest winning probability, but, since no prior information is given, she initially may want to spend her money trying different slot machines and record all the rewards she received. As soon as there is enough evidence that an option is better than another, the gambler will play consistently the best identified arm only, with the

purpose of maximizing the total payoff she gains during this process.

The fundamental dilemma a learner faces when choosing between uncertain options is captured here, since the decision maker operates under uncertainty and performs choices that will maximize long-term gains. Each pull of a slot machine arm represents a decision, and the main challenge lies in deciding whether to *explore* a new machine or *exploit* a machine that has provided good returns in the past. The gambler must weigh the immediate reward of a known machine against the potential of a higher payout from an untried machine.
Should one explore an option that looks inferior or exploit by going with the option that looks best currently? Finding the right balance between exploration and exploitation is at the heart of all bandit problems. This dilemma is known in the MAB literature as the *exploration/exploitation dilemma*. A MAB strategy should find the optimal trade-off between exploration and exploitation, to maximize the total reward the learner obtains over a horizon of time.

Bandit algorithms fall under the umbrella of Online Learning (OL) [44], a subfield of Machine Learning characterized by a fundamentally different approach compared to traditional ML solutions. Classical ML paradigms often work in a batch (or offline) learning fashion. For instance, in a supervised learning task, a model is trained following a specific algorithm and using a predefined set of data (learning phase), and it is then deployed in production with the target of predicting on new data. Such a paradigm mandates the availability of the entire dataset before training, and the training typically occurs offline due to its intensive computational demands. These offline methods face drawbacks like inefficiency in time and space, and scalability for large-scale applications, because the model usually has to be retrained to incorporate new training data.

In contrast to batch learning algorithms, OL is suited for data arriving in sequential order, where a learner aims to learn and update the best predictor for future data at every step. Since the predictive model can be updated instantly for any new data instances, OL algorithms are far more efficient and scalable for large-scale real-world data analysis tasks, where big data are arriving at a high rate. Therefore, an online problem setting is characterized by getting a single data point at a time. MAB problems adhere to this paradigm since we start with no data, we select an action, observe the corresponding outcome, update our model and iterate. A MAB policy is an algorithm which chooses the next arm to play based on the current model, which incorporates all the information available and obtained from any data we have observed so far. OL algorithms are usually evaluated computing the *regret*, which corresponds to the loss in cumulative reward the decision-maker suffers with respect to the maximum possible reward.

In the realm of MABs, the *stochastic* bandit problem is a significant variant where rewards are generated independently of past rewards, and based on fixed but unknown probability distributions. Unlike other settings where rewards might be deterministic, stochastic bandits introduce an element of randomness, making the exploration-exploitation trade-off even more challenging.

Traditional MAB problems assume reward distributions which are easy to treat theoretically, but not frequent in practical cases. In many real-world scenarios and domains, including finance, telecommunications, and internet traffic, these distributions exhibit *heavy-tailed* (HT) characteristics [88], where extreme values (or "tail" values) are more probable than what is predicted by classic distributions, i.e., subgaussian ones (see Section 2.3). These large values occur at an anomalously high frequency, which means, in the context of MAB, that there can be rare but extremely high rewards. Traditional stochastic MAB algorithms might not be well-suited for such scenarios because they exploit classical estimators which are badly affected from "tail" values. For this reason, they might lead to not enough exploration, with the risk of not encountering the mentioned high rewards. Studying heavy-tailed stochastic MAB settings is necessary to develop algorithms that can effectively handle such reward distributions and ensure that the potential of receiving high rewards is not overlooked.

It is also relevant to note that the mathematical formulation of bandit problems, which will be presented in the next chapter, leads to a rich structure that is strictly connected to other branches of mathematics. Over the decades, this topic has evolved into a multidisciplinary research area with contributions from computer science, operations research, economics, and statistics, and this is what have fascinated us the most to start an in-depth analysis of it.

1.1. Applications

The significance of the general MAB framework extends beyond the casino. Multi-armed bandits serve as foundational models for various real-world scenarios where decisions must be made sequentially under uncertainty.

Thompson's initial exploration of bandit problems in 1933 [91] was driven by the challenges in *clinical trials* design, where multiple treatments for a specific ailment existed. The dilemma was to select the best treatment in terms of effect on an upcoming patient.

However, with the advent of modern technology, the applications of bandit problems have expanded significantly, especially in the industrial sector. In the following, we survey

some of them.

- *Online platforms* are prime candidates for bandit algorithms since they can be tailored based on a user's specific sequence of interactions. In online recommendation systems [57], the goal is to suggest different items in order to understand the tastes of the clients (explorative action) while maximizing their interests for the items proposed (exploitative action), thus being suitable to be formulated with the MAB model.
- *Advert placement*, where the challenge is about determining the most suitable advertisement to showcase to a website's incoming visitor. Firms may decide among different multiple versions of their ads in order to learn which one is the most effective [85]. Similarly, the task of website optimization revolves around selecting the best design elements for a webpage, such as typography, visuals, and overall layout. The success of these choices is often gauged by user actions, like clicks or other target behaviors. However, it is worth noting that these scenarios are way more complex than the foundational bandit problem. The collection of ads might evolve, the feedbacks from users might be delayed, and the additional information on context may be necessary [66].
- *Source routing*, where the learner tries to direct internet traffic sending packets through the shortest path on a communication network [29]. The chosen route for each packet can vary, and the cost is typically the time required for packet delivery, influenced by the congestion on the chosen route's components. There are several possible routes and the learner must choose one for each packet to minimize the transmission cost.
- *Game-playing problems*, in which decisions are made by simulating potential game progressions with search trees. Bandit algorithms, especially those tailored for tree-structured bandit problems, can enhance the exploration of the vast game progression tree by concentrating on the most promising paths from the root to the leaves. This approach was effectively employed in the MoGo gaming [40] enabling it to play the game of Go at an elite level, with a strategy [55] inspired from a well-known bandit algorithm.

These are just some examples of practical applications, but the list might be way broader, including waiting problems and resource allocation, such that bandit problems are way more relevant that what we could expect at a first sight.

1.2. Motivations

The study of heavy-tailed multi-armed bandits is crucial for several reasons. Most of the existing works assume that the probability distributions of the rewards are parametric, e.g. subgaussian (see Section 2.3), or bounded. While this assumption enables the use of strong theoretical tools, it is often limiting in many practical scenarios such as, for example, financial environments [37] or network routing problems [67], where it is rarely the case that the arm distributions are bounded or have tails with a strong decay. In particular, it is well known that the stock returns in developed economies follow heavy-tailed distributions that typically have finite moments of order at most 4 [36]. Higher moments are not guaranteed to exist. Furthermore, daily exchange rates and income or wealth distributions may have heavier tails with finite moments of order less than 2 [80]. Thus, it is important to understand and develop a general theory and efficient algorithms, both computationally and statistically, that have wider applicability and are tailored to handle such real-world scenarios effectively.

Moreover, with the rise of Big Data and the increasing complexity of modern systems, the occurrence of outliers is becoming more common. Applications like risk assessment, anomaly detection, and recommendation systems can benefit from algorithms that consider heavy-tailed behavior.

In traditional stochastic MAB, the exploration-exploitation trade-off is relatively straightforward due to the well-behaved distributions of rewards. However, in heavy-tailed scenarios, the potential for rare but extremely high rewards makes the decision-making process more complex. Algorithms need to be more explorative to encounter these rare high rewards, and, in this case, they are also likely to be more robust and adaptable. For instance, [93] shows that for heavy-tailed reward distributions, the regret of traditional algorithms behaves differently compared to light-tailed distributions, emphasizing the need for specialized approaches in scenarios with heavy tails.

In a nutshell, while the traditional MAB literature provides a foundational understanding of the exploration-exploitation dilemma, the study of heavy-tailed multi-armed bandits is essential for a more comprehensive and realistic approach to sequential decision-making in environments with uncertain and extreme outcomes.

To conclude, current available approaches tackling the stochastic HT bandit setting assume the knowledge of the real parameters that characterize the reward distributions. This is a strong requirement since in practical cases these parameters are usually not known. On one side, it is not possible to be completely sure if real-world samples are

generated or not from a heavy-tailed distribution, so even knowing its parameters seems like an over-complicated unrealistic task. This fact gives a proper driving force to the study of new algorithms which are unaware of these quantities, that, from now on, we will refer to as *fully adaptive* with respect to the parameters. Thus, in our setting, we will consider them to be unknown to the agent.

1.3. Goal of the research

As mentioned above, commonly adopted algorithms in the setting of stochastic heavy-tailed multi-armed bandits have currently practical limitations. In particular, it is canonical to assume the knowledge of two parameters, mentioned in the following as ϵ and u and, to the best of our knowledge, every optimal regret minimization strategy in the literature requires at least one of them as an algorithm's input.

The main open problem we want to target with this work is to extend the current results trying to analyze whether it is feasible to develop an approach which is fully adaptive with respect to ϵ and u, not requiring their prior knowledge but still achieving comparable performances to other approaches knowing them. More in detail, it is relevant:

- to understand if there is any additional underlying assumption needed;
- to stress any theoretical guarantees underlying the possible novelty in algorithmic approaches;
- to validate how a potential new algorithm performs empirically.

1.4. Contributions

The main contribution of this research is twofold. Firstly, we show that in general it is not possible to achieve the same order of performance of the state-of-the-art approaches while being unaware of the aforementioned two quantities. Secondly, we will show that, under a specific but not restrictive distributional assumption, this is indeed possible.

More precisely, we will discuss the role of the *truncated non-positivity* assumption [45], and show that, when this assumption is violated, is not possible anymore to guarantee the existence of an adaptive algorithm w.r.t. ϵ nor u achieving comparable performance to optimal approaches in literature.

We will also propose `Adaptive Robust UCB` (shortly `AdaR-UCB`), an algorithm based on the *optimism in the face of uncertainty* principle that is capable to be *fully adaptive* w.r.t.

the two parameters characterizing the reward distributions. We show that, under the distributional assumption mentioned, it is able to attain the same theoretical guarantees of the well known `RobustUCB` algorithm from [22], matching the order of the regret lower bound for the classic heavy-tailed scenario.

To wrap up, we provide a theoretical analysis to bound the regret measure that a learner might incur while using `AdaR-UCB` algorithm, and perform a wide experimental campaign to compare what we proposed with state-of-the-art policies for heavy-tailed and traditional MAB problems. Therefore, our research is a blend of learning theory novelties, algorithm design and numerical evidencies.

1.5. Thesis Structure

The contents of this thesis are organized in the following six chapters.

We start, in Chapter 2, with an overview of the different aspects of stochastic MAB problems. We introduce here the mathematical notation and formulation for the heavy-tailed setting, together with the objective and metric to measure performances. Moreover, we outline the main concentration inequalities that will be relevant and used in successive chapters. Eventually, we introduce the concept of lower bounds and some well-established theoretical results and approaches to describe the properties and difficulties of a bandit setting.

In Chapter 3, we depict the landscape of the state-of-the-art adaptive heavy-tailed bandit methods. We start presenting the most known algorithms for the standard stochastic setting and the heavy-tailed one, and then we move straightforward to a detailed review of the most remarkable extensions to approaches requiring less prior knowledge on the rewards distributions. A comparative discussion of the properties of these algorithms is provided here. The goal of this chapter is to guide the reader in a conscious understanding of the fundamental motivations of this work.

In Chapter 4, we study how the lower bounds on regret change moving from a standard non-adaptive setting to an adaptive one. It is relevant to stress this point to understand the theoretical guarantees we could expect from the approach we propose in the new real-world setting, where the parameters characterizing the rewards are not known.

Chapter 5 is finally devoted to answer our main research question, with an extensive description of `AdaR-UCB` algorithm. We first describe in details a probabilistic reasoning to retrieve an adaptive robust estimator, followed by an illustration of the unique weak distributional assumption that is required to drive our analysis. The focus of this chapter

is on the theoretical guarantees supporting `AdaR-UCB` performance, with its pros and cons underlined. Many concentration inequalities on the new estimator are proved, to then conclude with a detailed proof of the main optimality result.

Chapter 6 is the one where our proposed solution is finally validated empirically. We use a simulated setting to understand the possible numerical results on regret performance in all the circumstances, both when our key assumption is satisfied and when not. We make here different experiments to compare `AdaR-UCB` performances against baseline methods.

In Chapter 7, we summarize the most relevant achievements of this thesis and we highlight the points of strength and weakness of the proposed approach. Furthermore, we suggest possible extensions of this work.

Appendix A reports more details on the notation used throughout the thesis, with other basic functional analysis inequalities. Appendix B, instead, briefly outlines few more side derivations and possible concerns that were not raised in the main text.

2 | Preliminaries

In the research, we investigate the stochastic multi-armed bandit problem under the assumption of heavy-tailed reward distributions. Thus, we present here the mathematical setting and formulations to understand properly the next chapters.

2.1. Stochastic Multi-Armed Bandits Model

A bandit problem is a sequential game between a learner and an environment. The game is played over an horizon of $T \in \mathbb{N}$ rounds, and in each round $t \in [T]$[1], the learner first chooses an action I_t (called "arm") from a given set $\mathcal{A}$, and the environment then reveals a reward $X_t \in \mathbb{R}$. The reward is sampled from the distribution corresponding to the selected arm and independently from past choices and rewards, given that action. Moreover, we have that I_t should only depend on the history $H_{t-1} = (I_1, X_1, \ldots, I_{t-1}, X_{t-1})$, since the learner cannot peek into the future when choosing their actions. A learner adopts a policy to interact with an environment, with the most common objective of choosing actions that lead to the largest possible cumulative reward over all T rounds, which is given by $\sum_{t=1}^{T} X_t$.

In their original formulation [84], Stochastic Multi-Armed Bandits were defined with probability distributions on $[0, 1]$, but the setting could be more general [12].

Definition 2.1 (Stochastic MAB). *A Stochastic Multi-Armed bandit problem is a collection of probability distributions $\nu = (\nu_i : i \in \mathcal{A})$, where $\mathcal{A}$ is the finite set of available actions.*

Remark 1 (Notation).

- *Let $K = |\mathcal{A}| < \infty$ be the number of arms: it is known, as the number of rounds T, with $T \geq K \geq 2$ ("multi-armed").*
- *Let $\nu_1, \ldots, \nu_K$ be the K probability distributions which, on the other side, are un-*

[1] We use $[T]$ to denote the set $\{1, \ldots, T\}$.

known to the learner. If they were known, one would always pull the arm with the highest mean reward in order to maximize the cumulative rewards.

As expected, the fundamental challenge in bandit problems is that the environment is generally unknown to the learner. The only partial information about the bandit instance $\nu = (\nu_i : i \in \mathcal{A})$ is that $\nu \in \mathcal{E}$, i.e. the true environment lies in the environment class $\mathcal{E}$.

We focus here on unstructured environment classes $\mathcal{E}$, set of bandits where $\mathcal{A}$ is finite and there exist sets of distributions $\mathcal{M}_i$ for each $i \in \mathcal{A}$ such that:

$$\mathcal{E} = \{\nu = (\nu_i : i \in \mathcal{A}) : \nu_i \in \mathcal{M}_i \text{ for all } i \in \mathcal{A}\},$$

or, in short, $\mathcal{E} = \times_{i \in \mathcal{A}} \mathcal{M}_i$. The product structure means that, by playing action i, the learner cannot deduce anything about the distributions of actions $j \neq i$. Some typical choices of unstructured bandits are listed in Table 2.1.

Name	Symbol	Definition
Bernoulli	$\mathcal{E}_{\mathcal{B}}^K$	$\left\{(\mathcal{B}(\mu_i))_i : \boldsymbol{\mu} \in [0,1]^K\right\}$
Uniform	$\mathcal{E}_{\mathcal{U}}^K$	$\left\{(\mathcal{U}(a_i, b_i))_i : \mathbf{a}, \mathbf{b} \in \mathbb{R}^K \text{ with } a_i \leq b_i \text{ for all } i\right\}$
Gaussian (known var.)	$\mathcal{E}_{\mathcal{N}}^K(\sigma^2)$	$\left\{(\mathcal{N}(\mu_i, \sigma^2))_i : \boldsymbol{\mu} \in \mathbb{R}^K\right\}$
Gaussian (unknown var.)	$\mathcal{E}_{\mathcal{N}}^K$	$\left\{(\mathcal{N}(\mu_i, \sigma_i^2))_i : \boldsymbol{\mu} \in \mathbb{R}^K \text{ and } \boldsymbol{\sigma}^2 \in [0,\infty)^K\right\}$
Finite variance	$\mathcal{E}_{\mathbb{V}}^K(\sigma^2)$	$\{(P_i)_i : \mathrm{Var}_{X \sim P_i}(X) \leq \sigma^2 \text{ for all } i\}$
Bounded support	$\mathcal{E}_{[a,b]}^K$	$\{(P_i)_i : \mathrm{Supp}(P_i) \subseteq [a,b]\}$
Subgaussian	$\mathcal{E}_{\mathrm{SG}}^K(\sigma^2)$	$\{(P_i)_i : P_i \text{ is } \sigma\text{-subgaussian for all } i\}$
Heavy-tail	$\mathcal{E}_{\mathrm{HT}}^K(\epsilon, u)$	$\{(P_i)_i : \mathbb{E}_{X \sim P_i}[\lvert X\rvert^{1+\epsilon}] \leq u\}$

Table 2.1: Typical environment classes for stochastic bandits, from [62]. For subgaussian distributions see Section 2.3

.

Let now ν_i be the probability distribution associated to arm i, for $i \in [K]$. The random variable $X_{i,t}$ is the payoff (or reward) of arm i when this arm is pulled at time t. Independence also holds for rewards across the different arms.

In the thesis, we consider stochastic MAB in a stationary setting only, where the distributions do not change over time, such that the rewards of the arms can be modeled with i.i.d. random variables. We have:

$$\mu_{i,t} = \mu_{i,t+1} =: \mu_i, \ \forall i \in \mathcal{A}, \ \forall t \in [T],$$

where $\mu_{i,t} = \mathbb{E}[X_{i,t}] = \mathbb{E}[X_{i,1}]$ for all t and μ_i is the expected reward of arm $i \in [K]$.

Nevertheless, many problems can be modeled through non-stationary MAB [39], since distributions may change over time. Few examples include parameters control [32], investments [87], the dynamic spectrum access problem [79] and evolving diseases in clinical trials [94].

Due to its high generality, there are different variants of the MAB model. For some applications, the assumption that the rewards are stochastic and stationary may be too restrictive. What if the stochastic assumptions fail to hold? What if they are violated for a single round? Or just for one action, at some rounds? Will the algorithms developed be robust to smaller or larger deviations from the modelling assumptions? An extreme idea is to drop all assumptions on how the rewards are generated, except that they are chosen without knowledge of the learner's actions and lie in a bounded set. If these are the only assumptions, we get what is called the setting of *adversarial bandits*.

We will not tackle any aspect of these problems, but it is for sure worth to mention them, so that the reader can have broader overview of the topic. In an adversarial multi-armed bandit [11, 13] the rewards are generated by a process that cannot be treated as a stochastic distribution. They are given by an adversary, who may take advantage of the corner cases in which a bandit algorithm performs badly. As a consequence, adversarial algorithms must be robust to adversary choice of the rewards (although under some specific, weaker, definitions of optimality).

2.2. Learning Objective and Regret

In a general stationary stochastic bandit model, the *optimal* arm is the arm the largest expected reward. We let i^* denote the index of such an arm, and we define the optimal expected reward as $\mu_{i^*} = \max_{i \in [K]} \mu_i$.

Further, let I_t denote the arm played at time t and $N_i(t)$ denote the number of times arm i is chosen by the policy during the first t plays:

$$N_i(t) = \sum_{s=1}^{t} \mathbb{1}_{I_s = i},$$

where $\mathbb{1}_{I_s=i}$ represents the indicator function, equal to 1 if $I_s = i$, and 0 otherwise. In general, $N_i(t)$ is random, because for all rounds t except for the first, the action I_t depends on the rewards observed in rounds $1, 2, \ldots, t-1$, which are random, hence I_t will also inherit their randomness.

We recall that a policy is an algorithm able to choose I_t given the past history of observations and actions, and that the goal of a bandit algorithm is to maximize the sum of cumulative rewards. Therefore, to quantify the performance of a policy we introduce the concept of *regret*.

Definition 2.2. *[21] The (cumulative) regret of a policy operating in the context of stochastic MAB setting, after T plays, is defined as:*

$$\tilde{R}_T := \max_{i\in[K]} \sum_{t=1}^{T} X_{i,t} - \sum_{t=1}^{T} X_{I_t,t}.$$

It measures the loss due to the fact that the policy does not always play the best arm.

The target of the agent is to minimize its *regret* after T rounds, which is equivalent to minimizing the loss incurred during the learning process.

Since both the rewards and the player's actions are stochastic, we introduce the following form of expected regret.

Definition 2.3. *[21] The expected (cumulative) regret of a policy operating in the context of stochastic MAB setting, after T plays, is defined as:*

$$R_T = \mathbb{E}[\tilde{R}_T] = T \max_{i\in[K]} \mu_i - \mathbb{E}\left[\sum_{t=1}^{T} X_{i,t}\right] = T\mu_{i^*} - \sum_{t=1}^{T} \mu_{i_t},$$

where i_t is the realization of random variable I_t, and the expectation is taken with respect to the randomness both in the algorithm and the environment.

It represents the average regret of the forecaster with respect to the best arm on average.

Clearly, the expected-regret is a weaker form of regret as it takes as optimum the action which is optimal only in expectation. However, this form of regret is more suitable for the purpose of our analysis, therefore in subsequent discussion, unless otherwise specified, we only consider expected-regret, dropping, for the sake of simplicity, the term "expected".

The setting depicted so far corresponds to the frequentist approach in which the expected mean rewards of all the arms are considered as unknown deterministic quantities and the goal is to achieve the best parameter-dependent performance, which corresponds to the regret defined in Definition 2.3.

Let us now explicitly mention another way to express the regret in 2.3 which will be useful in consequent analysis. Let us define the suboptimality gap for arm i as the expected loss

of playing the arm:

$$\Delta_i \coloneqq \mu_{i^*} - \mu_i, \ \forall i \in \mathcal{A}. \tag{2.1}$$

Lemma 2.4 (Regret Decomposition Lemma [62]). *For any policy and stochastic bandit environment ν, with $\mathcal{A}$ finite and horizon $T \in \mathbb{N}$, the regret R_T of the policy in ν satisfies:*

$$R_T = \sum_{i \in \mathcal{A}} \Delta_i \mathbb{E}\left[N_i(T)\right]. \tag{2.2}$$

Hence, to keep the regret small, the learner should try to minimise the weighted sum of expected action counts. An algorithm that aims at minimizing the expected regret should minimize the expected sampling times of sub-optimal arms.

2.3. Stochastic Heavy-Tailed MAB

Most of the existing research on bandits problem assume that, given a stochastic bandit $\nu = (\nu_i : i \in \mathcal{A})$, the unknown probability distributions ν_i are *subgaussian*.

Assumption 1 (Subgaussianity). *A random variable X drawn according to the distribution ν_i is σ-subgaussian if, for all $\lambda \in \mathbb{R}$, it holds that:*

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X])] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad \textit{and} \quad \mathbb{E}[\exp(\lambda(\mathbb{E}[X] - X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right),$$

where $M_X(\lambda) = \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X])]$ is the moment generating function of ν_i, which is a function $M_X : \mathbb{R} \to \mathbb{R}$. Moreover, σ^2 is the so-called "variance factor", a parameter that is usually assumed to be known.

Under Assumption 1, the tails of the distribution present a strong decay, implying that every moment of finite order is finite. In particular, the tails of a σ-subgaussian random variable decay approximately as fast as that of a Gaussian with zero mean and the same variance.

While this assumption enables for strong theoretical tools, it is often limiting in many practical scenarios we presented in Section 1.2. In settings where uncertainty is very strong, *heavy-tailed distributions* naturally arise: the tails decay slower than a Gaussian, and the moment generating function is no longer assumed to be finite.

Definition 2.5 (Heavy-Tailed Random Variable). *A random variable X is heavy-tailed if $M_X(\lambda) = \infty$ for all $\lambda > 0$. Otherwise it is light-tailed.*

In this thesis, we investigate the bandits with heavy tail setting, introduced in the seminal work [22], in which only moments of order up to $1+\epsilon$, with $\epsilon \in (0,1]$, are assumed to be finite and uniformly bounded by a constant u.

Assumption 2 (Bandits with Heavy-Tailed Rewards)**.** *Given a stochastic bandit instance $\nu = (\nu_i)_{i\in[K]}$, we assume it is* heavy-tailed *if, for each arm i generating a reward X, it holds:*

$$\exists\ \epsilon \in (0,1],\ u < \infty \quad s.t. \quad \mathbb{E}_{\nu_i}[|X|^{1+\epsilon}] \leq u \quad \forall i \in [K], \tag{2.3}$$

where each moment of higher order than $(1+\epsilon)$ is infinite for at least one arm (otherwise, in principle, this assumption holds also for a bandit with light-tailed distributions).

This is a standard assumption in the heavy tail literature, noting that the upper bound on the $(1+\epsilon)$-th order moment u is assumed to be common for all $(\nu_i)_{i\in[K]}$, without loss of generality.

In the heavy-tailed bandit problem, it is common to assume the knowledge of both ϵ and u. This work aims to bring novelty providing a new regret minimization strategy for this bandit problem that *does not require any prior knowledge on ϵ nor u*, but still achieves comparable performances to other approaches knowing them.

Thus, we can wrap up our final stochastic heavy-tailed multi-armed bandit setting assuming that, from now on, to each arm $i \in [K]$, we associate a probability distribution ν_i satisfying Assumption 2. Distributions with infinite variance are then allowed in this problem formulation. In all the following results presented in Chapter 5, we will consider both quantities to be unknown to the agent. To better clarify what we have anticipated in Section 1.2, from now on, we will refer to any algorithm operating without the knowledge of either ϵ or u as *adaptive w.r.t. ϵ and/or u*, depending on which one is unknown (possibly both).

2.4. Relevant Concentration Inequalities

We present here some concentration inequalities which will be needed to prove the results in following chapters. In particular, we do not highlight here the well-known standard concentration inequalities, but we try to present them in their revisited way to be suited to our applications. If not differently specified, the starting point of the next propositions is inspired from relevant literature on this field, as [17, 74].

For future use, we derive our results for the general case where, given the sample $\mathbf{X} = (X_1, \dots, X_n)$, the random variables X_i are independent, but not necessarily identically

distributed.

2.4.1. Notation

For the purpose of the next sections, we employ the following notation. If X is a real valued random variable we use $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ to denote its expectation and variance, respectively. Moreover, given $X_1, \ldots, X_n$ sequence of independent random variables, we let:

$$P_n(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \mu = \mathbb{E}[P_n(\mathbf{X})] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i]$$

and

$$V_n(\mathbf{X}) = \frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{(X_i - X_j)^2}{2}, \quad \mathrm{Var}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right),$$

where $P_n(X)$ is the unbiased sample mean estimator, $V_n(\mathbf{X})$ is the unbiased sample variance estimator and μ and $\mathrm{Var}(\mathbf{X})$ their respective expected values.

2.4.2. Bernstein's Inequalities

We start presenting some Bernstein-type inequalities for bounded random variables. These are slight variations of the most common standard Bernstein's inequality, which gives bounds on the probability that the sum of random variables deviates from its mean.

Proposition 2.1 (Bernstein's Inequality for Independent Random Variables)**.** *Let $X_1, \ldots, X_n$ be a sequence of independent random variables with $X_t - \mathbb{E}\left[X_t\right] \leq b$ almost surely $\forall t \in [n]$. Then, for every $\epsilon \geq 0$, we have:*

$$P\left(P_n(\mathbf{X}) \geq \mu + \epsilon\right) \leq \exp\left(-\frac{n^2\epsilon^2}{2\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) + \frac{2nb\epsilon}{3}}\right) = \exp\left(-\frac{n\epsilon^2}{2\,\mathrm{Var}(\mathbf{X}) + \frac{2\epsilon b}{3}}\right).$$

The proposition showed above assumes to know the real variance of the random variables, but knowing a bound is enough since if:

$$\mathrm{Var}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) \leq v,$$

then:

$$P\left(P_n(\mathbf{X}) \geq \mu + \epsilon\right) = \exp\left(-\frac{n\epsilon^2}{2\,\mathrm{Var}(\mathbf{X}) + \frac{2\epsilon b}{3}}\right) \leq \exp\left(-\frac{n\epsilon^2}{2v + \frac{2\epsilon b}{3}}\right). \tag{2.4}$$

We can re-write the inequality exploiting the equivalence between

$$\mathbb{P}\left(P_n(\mathbf{X}) - \mu \geq \epsilon\right) \leq \delta(\epsilon) \quad \text{and} \quad \mathbb{P}\left(P_n(\mathbf{X}) - \mu \leq \epsilon(\delta)\right) \geq 1 - \delta,$$

where $\epsilon(\delta)$ is the inverse function of $\delta(\epsilon)$.

In this way, if we set:

$$\delta = \exp\left(-\frac{n\epsilon^2}{2v + \frac{2\epsilon b}{3}}\right),$$

we can use that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ holds for positive x, y, to show:

$$\epsilon = \sqrt{\frac{2v\log(\delta^{-1})}{n} + \frac{b^2\log^2(\delta^{-1})}{9n^2}} + \frac{b\log(\delta^{-1})}{3n} \leq \sqrt{\frac{2v\log(\delta^{-1})}{n}} + \frac{2b\log(\delta^{-1})}{3n}.$$

That is, we have our final useful form of Bernstein's inequality:

$$\mathbb{P}\left(P_n(\mathbf{X}) - \mu \leq \sqrt{\frac{2v\log(\delta^{-1})}{n}} + \frac{2b\log(\delta^{-1})}{3n}\right) \geq 1 - \delta. \tag{2.5}$$

Let us now introduce another useful concentration inequality for self-bounding random variables (Theorem 13 in [72]):

Theorem 2.6. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in some set $\mathcal{X}$. For $1 \leq k \leq n$ and $y \in \mathcal{X}$, we use $\mathbf{X}_{y,k}$ to denote the vector obtained from $\mathbf{X}$ by replacing X_k by y. Suppose that $a \geq 1$ and that $Z = Z(\mathbf{X})$ satisfies the inequalities*

$$Z(\mathbf{X}) - \inf_{y\in\mathcal{X}} Z\left(\mathbf{X}_{y,k}\right) \leq 1 \ \forall k, \tag{2.6}$$

$$\sum_{k=1}^{n}\left(Z(\mathbf{X}) - \inf_{y\in\mathcal{X}} Z\left(\mathbf{X}_{y,k}\right)\right)^2 \leq aZ(\mathbf{X}), \tag{2.7}$$

almost surely. Then, for $t > 0$,

$$\mathbb{P}\left(\mathbb{E}[Z] - Z > t\right) \leq \exp\left(\frac{-t^2}{2a\mathbb{E}[Z]}\right). \tag{2.8}$$

If Z satisfies only the self-boundedness condition (2.7), we have:

$$\mathbb{P}\left(Z - \mathbb{E}[Z] > t\right) \leq \exp\left(\frac{-t^2}{2a\mathbb{E}[Z] + at}\right). \tag{2.9}$$

2.4.3. Empirical Bernstein's Bound

From [73], let us highlight a relevant variance sensitive confidence bound:

Theorem 2.7 (Empirical Bernstein's). *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a vector of independent random variables with values in $[0,1]$. Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$ in $\mathbf{X}$, we have:*

$$\mathbb{E}\left[P_n(\mathbf{X})\right] \leq P_n(\mathbf{X}) + \sqrt{\frac{2V_n(\mathbf{X})\log\left(2\delta^{-1}\right)}{n}} + \frac{7\log\left(2\delta^{-1}\right)}{3(n-1)}. \tag{2.10}$$

To allow this inequality to hold in a more general case, we extend this result to a vector of independent random variables in $[a,b]$.

Let us suppose we have $\mathbf{Z} = (Z_1, \dots, Z_n)$ vector of independent random variables in $[a,b]$. We can normalize them and obtain $\mathbf{X} = (X_1, \dots, X_n)$, with:

$$X_i = \frac{Z_i - a}{b-a} \in [0,1] \quad \forall i \in [n].$$

We thus get:

$$P_n(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i - a}{b-a} = \frac{P_n(\mathbf{Z})}{b-a} - \frac{a}{b-a} \tag{2.11}$$

$$\Rightarrow \mathbb{E}\left[P_n(\mathbf{X})\right] = \frac{\mathbb{E}\left[P_n(\mathbf{Z})\right]}{b-a} - \frac{a}{b-a}; \tag{2.12}$$

$$V_n(\mathbf{X}) = \frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{(X_i - X_j)^2}{2} = \frac{1}{(b-a)^2}V_n(\mathbf{Z}). \tag{2.13}$$

Since $\mathbf{X}$ satisfies the hypothesis of Theorem 2.7, we get Equation (2.10). If we substitute there the results of Equations (2.11), (2.12) and (2.13) we obtain, for $\delta \in (0,1)$, a more general concentration result:

$$\mathbb{P}\left(\frac{\mathbb{E}\left[P_n(\mathbf{Z})\right]}{b-a} \leq \frac{P_n(\mathbf{Z})}{b-a} + \frac{1}{b-a}\sqrt{\frac{2V_n(\mathbf{Z})\log\left(2\delta^{-1}\right)}{n}} + \frac{7\log\left(2\delta^{-1}\right)}{3(n-1)}\right) \geq 1-\delta$$

$$\Rightarrow \mathbb{P}\left(\mathbb{E}\left[P_n(\mathbf{Z})\right] \leq P_n(\mathbf{Z}) + \sqrt{\frac{2V_n(\mathbf{Z})\log\left(2\delta^{-1}\right)}{n}} + (b-a)\frac{7\log\left(2\delta^{-1}\right)}{3(n-1)}\right) \geq 1-\delta. \tag{2.14}$$

2.4.4. Concentration Result for Sample Variance

For the purpose of Algorithm 5.1, we need to establish confidence bounds for the deviation of the estimated standard deviation (sample estimate) from the actual one [73]:

Theorem 2.8. *Let $n \geq 2$ and $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in $[0,1]$. Then for $\delta > 0$ we have, writing $\mathbb{E}[V_n(\mathbf{X})]$ for $\mathbb{E}_{\mathbf{X}}[V_n(\mathbf{X})]$,*

$$\mathbb{P}\left(\sqrt{\mathbb{E}[V_n(\mathbf{X})]} > \sqrt{V_n(\mathbf{X})} + \sqrt{\frac{2\log(\delta^{-1})}{n-1}}\right) \leq \delta, \tag{2.15}$$

$$\mathbb{P}\left(\sqrt{V_n(\mathbf{X})} > \sqrt{\mathbb{E}[V_n(\mathbf{X})]} + \sqrt{\frac{2\log(\delta^{-1})}{n-1}}\right) \leq \delta, \tag{2.16}$$

where $\mathbb{E}[V_n(\mathbf{X})]$ is the unknown variance of the vector, since $V_n(\mathbf{X})$ is the unbiased sample variance.

We now extend these inequalities such that they hold for vector of independent and bounded random variables, not necessarily in $[0,1]$.

Using the same perspective of the generalization performed in Section 2.4.3, we can compute

$$\mathbb{E}[(X_i - X_j)^2] = \mathbb{E}\left[\left(\frac{Z_i - Z_j}{b-a}\right)^2\right] = \frac{1}{(b-a)^2}\mathbb{E}\left[(Z_i - Z_j)^2\right] \quad \forall i, j \in [n].$$

$$\begin{aligned}\mathbb{E}\left[V_n(\mathbf{X})\right] = \sigma_n^2(\mathbf{X}) &= \frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{\mathbb{E}\left[(X_i - X_j)^2\right]}{2} \\ &= \frac{1}{(b-a)^2}\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{\mathbb{E}\left[(Z_i - Z_j)^2\right]}{2} = \frac{1}{(b-a)^2}\mathbb{E}\left[V_n(\mathbf{Z})\right].\end{aligned} \tag{2.17}$$

At this point, since $\mathbf{X}$ satisfies the hypotheses of Theorem 2.8, we get both (2.15) and (2.16). If we substitute there the results of Equations (2.17) and (2.13), we obtain the more general concentration results for the variance of bounded $\mathbf{Z}$. For $\delta \in (0,1)$:

$$\mathbb{P}\left(\sqrt{\mathbb{E}[V_n(\mathbf{Z})]} \leq \sqrt{V_n(\mathbf{Z})} + (b-a)\sqrt{\frac{2\log(\delta^{-1})}{n-1}}\right) \geq 1 - \delta, \tag{2.18}$$

$$\mathbb{P}\left(\sqrt{V_n(\mathbf{Z})} \leq \sqrt{\mathbb{E}[V_n(\mathbf{Z})]} + (b-a)\sqrt{\frac{2\log(\delta^{-1})}{n-1}}\right) \geq 1 - \delta. \tag{2.19}$$

2.5. Lower Bounds for Bandits with Finitely Many Arms

Most of the theoretical notions and mathematical tools presented so far were preliminaries for the fully adaptive algorithm presented in Chapter 5. The focus of the study will be on achieving the same order of performance of other less general approaches, where the performance is evaluated through the upper bound on regret in Definition 2.3.

As satisfying as the upper bounds on the regret might be, the real truth of a problem is usually found in the lower bounds. An upper bound does not indeed tell much about what an approach could be missing out on. The only way to demonstrate that an algorithm really is (close to) optimal is to prove a lower bound showing that no algorithm can do better. Moreover, thinking about lower bounds forces to understand what is hard about the problem.

History shows that it usually turns out to be easier to get the lower bound, and then the challenge lies in improving algorithm guarantees until eventually the upper bound matches some known lower bound. That's why we decided here to provide mathematical results that will help us justifying the analysis on lower bounds at chapter 4, before entering in the discussion of the algorithm itself.

So what is the form of a typical lower bound?

The first example is the *worst-case lower bound*, which is suitable for understanding the robustness of a policy and corresponds to a claim of the form following form:
"For any policy, there exists an instance of a bandit problem ν on which the regret is at least L".
In our next analyses, we will focus mostly on finite time worst-case lower bounds, also called *minimax lower bounds*, but for the sake of completeness we mention briefly also the second type, the *instance-dependent lower bounds*. They provide a lower bounds on the regret of an algorithm for specific instances, and they have a different form that usually reads like the following:
"For any reasonable policy, then its regret on any instance ν is at least $L(\nu)$".
The statement only holds for some policies - the "reasonable" ones, whatever that means. But the guarantee is also more refined because bound controls the regret for these policies on every instance by a function that depends on this instance. The inclusion of the word "reasonable" is unfortunately necessary. For every bandit instance ν there is a policy that just chooses the optimal action in ν. Such policies are not reasonable because they have linear regret for bandits with a different optimal arm.

While minimax lower bounds serve as a useful measure of the robustness of a policy, they are often excessively conservative. On the other side, instance-dependent lower bounds try to capture the optimal performance of a policy on a specific bandit instance.

We now introduce the formal definition of minimax regret and lower bounds, discussing the line of attack for proving them.

2.5.1. Minimax Lower Bounds

Definition 2.9 (Minimax Regret, [62]). *The worst-case regret of a policy π on a set of stochastic bandit environments $\mathcal{E}$ is:*

$$R_T(\pi, \mathcal{E}) = \sup_{\nu \in \mathcal{E}} R_T(\pi, \nu).$$

Let Π be the set of all policies. The minimax regret is:

$$R_T^*(\mathcal{E}) = \inf_{\pi \in \Pi} R_T(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}} R_T(\pi, \nu).$$

Thus, a policy is called minimax optimal for $\mathcal{E}$ if $R_T(\pi, \mathcal{E}) = R_T^*(\mathcal{E})$, where $R_T^*(\mathcal{E})$ is indeed the minimax lower bound on the regret for the stochastic MAB setting.

Minimax optimality is not a property of a policy alone. It is a property of a policy together with a set of environments $\mathcal{E}$ and a horizon $T \in \mathbb{N}$. Finding a minimax policy is generally too computationally expensive to be practical. For this reason, we almost always settle for a policy that is nearly minimax optimal (see the beginning of Chapter 4).

In the context of minimax theory, the goal is to find a strategy (or policy) that minimizes the maximum possible loss (or regret) against an adversary. The adversary, in turn, tries to choose the worst-case scenario (an instance of the bandit problem) to maximize the regret. See Figure 2.1 for a simplifying representation of this idea.

The term minimax will be used in the following also for $R_T(\pi, \mathcal{E})$ if, except for constant factors, this worst-case bound cannot be improved on by any algorithm.

The high-level idea is to select two bandit problem instances in such a way that the following two conditions hold simultaneously:

1. Competition: An action, or, more generally, a sequence of actions that is good for one bandit is not good for the other.

2. Similarity: The instances are 'close' enough that the policy interacting with either

of the two instances cannot statistically identify the true bandit with reasonable statistical accuracy.

The problem described here is called *hypothesis testing*, and hence the two requirements are clearly conflicting. The first makes us want to choose instances with means $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathbb{R}^K$ that are far from each other, while the second requirement makes us want to choose them to be close to each other. The lower bound will follow by optimising this trade-off.

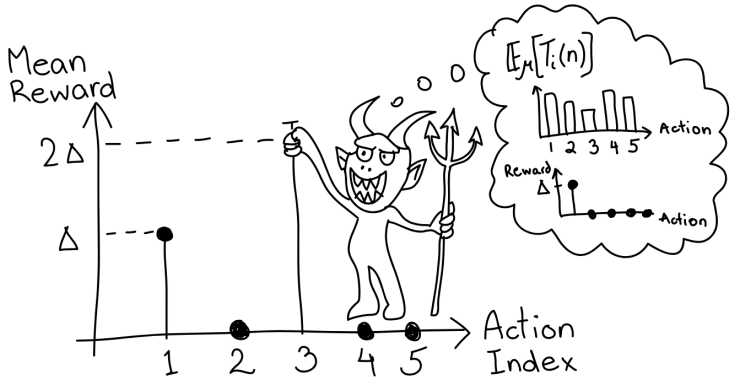


Figure 2.1: The idea of the minimax lower bound, from [62]: given a policy and one environment, the evil antagonist picks another environment so that the policy will suffer a large regret in at least one environment.

In order to prove a lower bound, it suffices to show that, for every strategy π, there exists a choice of $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ such that:

$$\max\left\{\frac{R_T(\pi,\nu)}{f(K,T)}, \frac{R_T(\pi,\nu')}{f(K,T)}\right\} \geq C,$$

where $C > 0$ is a universal constant and the function $f(K,T)$ depends on the bandit setting we are dealing with.

Practically, since we want ν and ν' to be hard to distinguish and yet have different optimal actions, we should make $\boldsymbol{\mu}'$ as close to $\boldsymbol{\mu}$ except in the arm where π expects to explore the least. And this is the approach followed in Chapter 4 to tackle the proofs on lower bounds.

From now on, we will refer to a policy π as *minimax optimal up to constant factors* for a given set of bandit problems $\mathcal{E}^K$ with K arms, if there exists a constant $C > 0$, such that:

$$\frac{R_T\left(\pi, \mathcal{E}^K\right)}{R_T^*\left(\mathcal{E}^K\right)} \leq C \text{ for all } K \text{ and } T.$$

Otherwise, we say that a policy π is *minimax optimal up to logarithmic factors* for the environment class $\mathcal{E}^K$ if:

$$\frac{R_T\left(\pi, \mathcal{E}^K\right)}{R_T^*\left(\mathcal{E}^K\right)} \leq C(T, K) \text{ for all } K \text{ and } T,$$

where $C(T, K)$ is logarithmic in T and K.

2.5.2. Entropy and Information Theory Inequalities

To make the arguments in the previous section rigorous and generalizable to many settings, we need some mathematical tools from information theory and statistics. This subsection takes us on a brief excursion into information theory, with the main references that has to be found in [30, 70].

The most important of these notions is the *relative entropy*, also known as the *Kullback–Leibler divergence* (KL divergence) [56]. We first introduce a real analysis notion to then define the relative entropy between probability measures P and Q on arbitrary measurable spaces.

Definition 2.10 (Absolutely Continuous Measures)**.** *Let P and Q be measures (not necessarily probability measures) on arbitrary measurable space $(\Omega, \mathcal{F})$. We say that P* is absolutely continuous with respect to *Q, denoted as $P \ll Q$, if:*

$$Q(A) = 0 \Longrightarrow P(A) = 0 \text{ for all } A \in \mathcal{F}.$$

Theorem 2.11 (Relative Entropy or Kullback-Leibler Divergence)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, and let P and Q be measures on this space. Then,*

$$\mathrm{D}_{KL}(P\|Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}(\omega)\right) dP(\omega), & \text{if } P \ll Q \\ +\infty, & \text{otherwise} \end{cases}.$$

Note that, in this general case, the relative entropy between P and Q can still be infinite even when $P \ll Q$.

The KL divergence is a measure of difference between two distributions, it is non-negative and assumes its global minimum $\mathrm{D}_{KL} = 0$ when they coincide.

In the case of discrete measures the above expression reduces to the following:

Corollary 2.12 (KL Divergence for Discrete Distributions). *We let P and Q be two discrete probability distributions defined on the same state space Ω, with Ω discrete such that $p_i = P(X = i)$ and $q_i = Q(X = i)$ $\forall i \in \Omega$. The relative entropy from Q to P is defined as:*

$$\mathrm{D}_{KL}(P\|Q) = \sum_{i\in\Omega:p_i>0} p_i \log\left(\frac{1}{q_i}\right) - \sum_{i\in\Omega:p_i>0} p_i \log\left(\frac{1}{p_i}\right) = \sum_{i\in\Omega:p_i>0} p_i \log\left(\frac{p_i}{q_i}\right). \tag{2.20}$$

In other words, $\mathrm{D}_{KL}(P\|Q)$ equals the expectation of the logarithmic difference between the distributions P and Q, where the expectation is taken using P. We see that the sufficient and necessary condition for $\mathrm{D}_{KL}(P\|Q) < +\infty$ is that for each i with $q_i = 0$, we also have $p_i = 0$. This condition is equivalent to say that P is absolutely continuous with respect to Q (see Definition 2.10). If not, $\mathrm{D}_{KL}(P\|Q) = \infty$. But still, absolute continuity implies a finite relative entropy when $X \sim P$ takes finitely many values only.

We also highlight that the KL-divergence is not a distance since it does not satisfy the triangle inequality, nor is symmetric.

It remains now to state the key lemma that connects the relative entropy to the hardness of hypothesis testing.

Theorem 2.13 (Bretagnolle-Huber Inequality, [18]). *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary measurable event. Then,*

$$P(A) + Q\left(A^c\right) \geq \frac{1}{2}\exp(-\mathrm{D}_{KL}(P\|Q)), \tag{2.21}$$

where $A^c = \Omega\backslash A$ is the complement of A.

Let us emphasize a simple interpretation of this statement. Suppose that $\mathrm{D}_{KL}(P\|Q)$ is small, then P is close to Q in some sense. Since P is a probability measure, we have $P(A)+P\left(A^c\right) = 1$. If Q is close to P, then we might expect that $P(A)+Q\left(A^c\right)$ should be large. The purpose of the theorem is to quantify just how large. Also note that the result is symmetric. We could replace $\mathrm{D}_{KL}(P\|Q)$ with $\mathrm{D}_{KL}(Q\|P)$, which sometimes leads to a stronger result because the relative entropy is not symmetric.

For what is next, we assume to be in the K-armed stochastic bandits framework, with horizon $T > 0$ and the number of actions $K > 1$. In particular, we show an exact calculation of the relative entropy between measures in a bandit model, considering a fixed policy and different instances.

Lemma 2.14 (Divergence Decomposition, [41]). *Let $\nu = (P_1, \dots, P_K)$ be the reward distributions associated with one K-armed bandit instance, and let $\nu' = (P'_1, \dots, P'_K)$ be the reward distributions associated with another K-armed bandit instance. Fix some policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the probability measures induced by the T-round interconnection of π and ν (respectively, π and ν'). Then:*

$$\mathrm{D}\left(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}\right) = \sum_{i=1}^{K} \mathbb{E}_\nu\left[N_i(T)\right] \mathrm{D}_{KL}\left(P_i \| P'_i\right). \tag{2.22}$$

According to [63], all these theoretical guarantees will be used to prove the novelties presented in Chapter 4.

3 | Related Works

3.1. Stochastic Bandits

Fundamental algorithms for unstructured stochastic bandits with finitely many actions have been widely studied in literature. Firstly, their simplicity makes them relatively easy to analyse and allows a deep understanding of the trade-off between exploration and exploitation. Secondly, many of the algorithms designed for finite-armed bandits, with the principle underlying them, can be generalised to other settings.

A good strategy to tackle the exploration-exploitation dilemma for the learner is to, in a certain sense, simultaneously perform exploration and exploitation. A simple heuristic principle for that is the so-called *"optimism in face of uncertainty"* [59], which states that one should act as if the environment is as nice as plausibly possible. In this context, taking an optimistic view of an unknown choice leads to exploration, while a pessimistic view would discourage it.

The idea is very general and applies to many sequential decision making problems in uncertain environments. Assume that the forecaster has accumulated some data on the environment and must decide how to act next. First, a set of "plausible" environments which are "consistent" with the data (typically, through concentration inequalities) is constructed. Then, the most "favorable" environment is identified in this set. Based on that, the heuristic prescribes that the decision which is optimal in this most favorable and plausible environment should be made. This principle gives simple and yet almost optimal algorithms for the stochastic multi-armed bandit problem.

Practically, the optimism principle means using the data observed so far to assign to each arm an index, called the *Upper Confidence Bound (UCB)*, that with a high probability of at least $1 - \delta$, given $\delta \in (0, 1)$, is an overestimate of the unknown arm expected reward. The index is usually composed by the empirical estimates of the arm expected reward and the uncertainty about the estimate. This way guarantees that every arm will be selected for a sufficient number of times, while also exploiting the currently believed best arm in the meanwhile.

Lai et al. [59] also proved that the regret of any stochastic MAB policy is at least logarithmic in time, and the intuitive reason behind a sublinear regret is quite simple. Assuming the upper confidence bound assigned to the optimal arm is indeed an overestimate, then another arm can only be played if its upper confidence bound is larger than that of the optimal arm, which, in turn, is larger than the mean of the optimal arm. And yet this cannot happen too often because the additional data provided by playing a suboptimal arm implies that the upper confidence bound for this arm will eventually fall below that of the optimal arm.

Discussing further the "optimism in face of uncertainty", it is fair to wonder how much optimistic we should be. This is a difficult decision, which has been tackled in the literature since [58], with the first version of an UCB algorithm. Choosing the confidence level $1-\delta$ and quantifying the degree of certainty is, indeed, challenging. It should be high enough to ensure optimism with high probability, but not so high that suboptimal arms are explored excessively.

The algorithm we will present in Chapter 5 is also well-founded on this principle, thus it is worth mentioning how much it has been studied in literature throughout the years. Other early works in the nineties as [1, 47, 50] dealt with UCB approaches under parametric assumptions, but all these policies were either computationally unfeasible to implement or lack of a finite time theoretical analysis of their regret.

The policy proposed in [12], under the name of `UCB1`, is the first computationally efficient policy for which a logarithmic regret is guaranteed uniformly over time. For this reason, let us briefly recap this algorithm for the nowadays usual case of subgaussian random variables (see Section 2.3), even if in [12] the payoffs are confined in $[0,1]$ interval. Prior MAB literature, indeed, mostly studies settings where the loss distributions are supported on a bounded interval I (e.g., $I=[0,1]$) known to the agent before-hand.

Let now $(X_t)_{t=1}^T$ be a sequence of independent σ-subgaussian random variables, with σ known positive parameter, and $\hat{\mu}$ be the sample mean estimator such that $\hat{\mu} = \frac{1}{T}\sum_{t=1}^T X_t$.

Then a reasonable candidate for a bound on the unknown mean of arm i which is "as large as plausibly possible" equals to:

$$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} +\infty, & \text{if } N_i(t-1)=0 \\ \hat{\mu}_i(t-1) + \sigma\sqrt{\frac{2\log(\delta^{-1})}{N_i(t-1)}}, & \text{otherwise} \end{cases},$$

where we recall that $N_i(t-1) = \sum_{s=1}^{t-1} \mathbb{1}_{I_s=i}$ denotes the number of times arm i is pulled during the first $t-1$ rounds. Consequently, the pseudo-code of this policy is reported in

Algorithm 3.1.

Algorithm 3.1 `UCB1`

1: Input number of arms K, error probability $\delta \in (0, 1)$ and variance factor σ^2.
2: **for** $t \in [T]$ **do**
3: Choose arm $I_t = \underset{i \in [K]}{\operatorname{argmax}} \operatorname{UCB}_i(t-1, \delta)$.
4: Observe reward X_t from the arm selected and update upper confidence bounds.
5: **end for**

The following result due to [62] provides an upper bound on regret relative to `UCB1` policy:

Theorem 3.1 (Upper Bound on the Regret for `UCB1` Algorithm). *Suppose that $\sigma^2 > 0$ is a known constant, and consider `UCB1` as shown in Algorithm 3.1 on a stochastic K-armed σ-subgaussian bandit problem, i.e. any $\nu \in \mathcal{E}_{\mathrm{SG}}^K(1)$ environment. For any horizon T, if $\delta = 1/T^2$, then the regret is bounded by:*

$$R_T \leq \mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\sigma^2 \log(T)}{\Delta_i} + \sum_{i=1}^{K} \Delta_i\right), \tag{3.1}$$

$$R_T \leq \mathcal{O}\left(\sigma\sqrt{TK\log(T)} + \sum_{i=1}^{K} \Delta_i\right), \tag{3.2}$$

where the $\mathcal{O}$ notation is defined in Appendix A, Equation (A.1).

Note that a regret of at least $\sum_{i=1}^{K} \Delta_i$ is suffered by any strategy that pulls each arm at least once.

Since [12] was published, a huge amount of works have been extending reward assumptions on the same setting with a frequentist approach, leading to UCB-V [10], MOSS [7] and KL-UCB [38], for which finite time regret bounds are also provided. Stochastic bandit problems were also studied under Bayesian assumptions and reasoning [2, 42, 51], as well as with other specific structures and characteristics [53, 76, 92], for which sub-linear regret bounds have been provided.

Despite all the different facets, most of the research for stochastic MABs has been investigated under the sub-Gaussian assumption on reward distributions, which have the exponential-decaying behavior. This assumption, indeed, includes many well-known distributions, as Gaussian, Bernoulli [52], and any bounded variable in general. We refer the reader to [21] for a survey of the extensive literature of this problem and its variations.

Moreover, there have been several works exploring the case of unknown subgaussian constant σ, as [8, 10, 31]. These are the first examples of finding optimal algorithms that extend previous results which require the knowledge of the parameters characterizing the light-tailed distributions. Our contribution is on the same fashion of these ones but for different environments, i.e., the ones of stochastic bandits with heavy-tailed reward distributions.

3.1.1. On Finite Time Instance-Independent Lower Bounds

The first known work on lower bounds for stochastic MABs was presented in [95], with a precise minimax analysis of two-armed Bernoulli bandits. The cases of setting with Bernoulli rewards have been highly studied in literature [11, 25, 59], but still we would like to briefly discuss the stochastic subgaussian bandit setting to understand the optimality of results presented in Section 3.1, and to easily extend the reasoning to heavy-tailed bandits.

Recall that $\mathcal{E}^K_{\mathcal{SG}}(\sigma)$ is the class of subgaussian bandits with variance factor σ^2 and K arms, which can be parameterised by the mean vector $\boldsymbol{\mu} \in \mathbb{R}^K$. We let $\nu_{\boldsymbol{\mu}}$ be the subgaussian bandit instance for which the arm i has reward distribution ν_i.

Theorem 3.2 (Minimax Lower Bound on Regret for Stochastic σ-Subgaussian Bandit, [13, 27])**.** *Let $K > 1$ and $T \geq K - 1$. Then, for any policy π, there exists a mean vector $\boldsymbol{\mu} \in \mathbb{R}^K$ such that:*

$$R_T(\pi, \nu_{\boldsymbol{\mu}}) \geq \Omega\left(\sigma\sqrt{KT}\right),$$

where the Ω notation is introduced in Appendix A, Equation (A.2)*.*

Since $\nu_{\boldsymbol{\mu}} \in \mathcal{E}^K_{\mathcal{SG}}(\sigma)$, it follows that the minimax regret for $\mathcal{E}^K_{\mathcal{SG}}(\sigma)$ is lower-bounded by the right-hand side of the above display as soon as $T \geq K - 1$:

$$R^*_T\left(\mathcal{E}^K_{\mathcal{SG}}(\sigma)\right) \geq \Omega\left(\sigma\sqrt{KT}\right). \tag{3.3}$$

The theorem above, together with Equation (3.2), shows that Algorithm 3.1 is minimax optimal up to logarithmic factors in T for σ-subgaussian bandits with suboptimality gaps in $[0, 1]$, or, more generically, in $\mathbb{R}$.

3.2. Bandits with Heavy Tails

Early researches for stochastic MABs have been investigated under the subgaussian assumption on reward distributions. However, there remains a large class of distributions which are not covered by the subgaussianity, e.g., distributions with heavy tails. This setting naturally extends classical MAB settings, including subgaussian-reward MAB and bounded-reward MAB. We recall that the heavy-tailed bandit problem is defined under Assumption 2, with the key parameters ϵ and u, such that $1+\epsilon$ is the maximum order of finite moments for the reward distributions, and u is a uniform bound on these moments.

While various researches have investigated heavy-tailed reward setting, they focused on variants of the MAB such as linear bandit [75], contextual bandit [86], Lipschitz bandit [68], or ϵ-contaminated bandit [81] (none of these algorithms removes the dependency on parameter ϵ). On the other side, the stochastic HT bandit model was first introduced by [22]. Here, Bubeck et al. [2013], have proposed `RobustUCB` by employing the "optimism in face of uncertainty" principle with a confidence bound on a class of robust estimators (see Section 3.2.3).

Instance-dependent lower-bounds and *instance-independent* ones (also known as *minimax*) were given in their paper, and reported below:

Theorem 3.3 (Non-Adaptive Lower Bounds for Stochastic HT Bandit, [22]). *For any algorithm and for any fixed T, there exists a set of K distributions satisfying Assumption 2 with $u = 1$, such that:*

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

where Δ_i refers to the suboptimality gap defined in Equation (2.1).

In particular, re-adapting theorem 3.3 to the general case of bound $u \neq 1$, the result easily extends as:

$$R_T \geq \Omega\left(\sum_{i:\Delta_i>0} \left(\frac{u}{\Delta_i}\right)^{\frac{1}{\epsilon}} \log T\right), \tag{3.4}$$
$$R_T \geq \Omega\left((uT)^{\frac{1}{1+\epsilon}} K^{\frac{\epsilon}{1+\epsilon}}\right). \tag{3.5}$$

Equation (3.5) is independent of the problem instance and shows how the dependency on T deteriorates as $\epsilon \to 0$. In the particular scenario in which variance is finite, i.e., $\epsilon = 1$, the minimax lower bound achieves the same order as the one for classic stochastic

multi-armed bandit problems [62], reported in Equation (3.3).

We focus here on a specific type of the `RobustUCB` algorithm, i.e., the one that considers the trimmed mean estimator in Definition 3.4 below. This choice is due to the fact that our algorithm proposed in Chapter 5 will consider an estimator which is its adaptive extension.

Definition 3.4 (Truncated Mean Robust Estimator, by [22])**.** *Let $\delta \in (0,1), \varepsilon \in (0,1]$, and $u > 0$. Suppose $X_1, \dots, X_s$ are i.i.d random variables with finite mean, then the truncated empirical mean estimator $\widehat{\mu}^T$ defined as*

$$\widehat{\mu}^T_{s,\delta} = \frac{1}{s}\sum_{j=1}^{s} X_j \mathbb{1}_{\left\{|X_j| \leq \left(\frac{uj}{\log(\delta^{-1})}\right)^{\frac{1}{1+\varepsilon}}\right\}}. \tag{3.6}$$

We show in Algorithm 3.2 the pseudo-code of the trimmed mean version of `RobustUCB`.

Algorithm 3.2 `RobustUCB`

Input number of arms K, error probability $\delta = t^{-4}$, heavy tails parameters ϵ and u.
Initialize $s_i \leftarrow 0, \quad \mathbf{X}_i \leftarrow \emptyset, \quad \hat{\mu}_{i,0,1} \leftarrow +\infty \quad \forall i \in [K]$.
for $t \in [T]$ **do**
  for $i \in [K]$ **do**
    Compute trimmed mean estimator:

$$\widehat{\mu}^T_{i,s,t}(\mathbf{X}_i) = \begin{cases} +\infty, & s_i = 0 \\ \frac{1}{s_i}\sum_{j=1}^{s_i} X_{i,j}\mathbb{1}_{\left\{|X_{i,j}| \leq \left(\frac{uj}{\log(t^4)}\right)^{\frac{1}{1+\varepsilon}}\right\}}, & s_i \neq 0 \end{cases}$$

  end for
  Select an action:

$$i_t \in \underset{i\in[K]}{\operatorname{argmax}} \left\{ \widehat{\mu}^T_{i,s,t}(\mathbf{X}_i) + 4u^{\frac{1}{1+\epsilon}} \left(\frac{\log(t^4)}{s_i}\right)^{\frac{\epsilon}{1+\epsilon}} \right\}.$$

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

We state now the upper bound on regret suffered by `RobustUCB` policy:

Theorem 3.5 (Upper Bound on Regret for `RobustUCB`, [19]). *Suppose to be in a stochastic K-armed heavy-tailed bandit setting, i.e., any $\nu \in \mathcal{E}_{\mathrm{HT}}^{K}(\epsilon, u)$ environment, with $\varepsilon \in (0, 1]$, $u < +\infty$ such that the rewards distributions satisfy Assumption 2. For any horizon T, the regret of `RobustUCB` policy reported in Algorithm 3.2 satisfies:*

$$R_T \leq \sum_{i:\Delta_i>0} \mathcal{O}\left(\left(\frac{u}{\Delta_i}\right)^{\frac{1}{\varepsilon}} \log T + \sum_{i:\Delta_i>0} \Delta_i\right). \tag{3.7}$$

Also, with a suitable T big enough, then

$$R_T \leq \mathcal{O}\left(u^{\frac{1}{1+\varepsilon}} (K \log T)^{\frac{\varepsilon}{1+\varepsilon}} T^{\frac{1}{1+\varepsilon}}\right). \tag{3.8}$$

We remark that, even if the variance is finite, i.e., $\epsilon = 1$, but the higher order moments are not, the same guarantees of order $\sum_{\Delta_i>0} \log T/\Delta_i$ attained in the classic stochastic bandit setting [12] can be achieved in the heavy-tailed bandit problem. However, if $\epsilon < 1$, then the dependency on the suboptimality gaps Δ_i deteriorates and the upper bound on regret is of order $\sum_{\Delta_i>0} \log T/\Delta_i^{\frac{1}{\epsilon}}$. Moreover, the dependency on $\Delta_i^{\frac{1}{\epsilon}}$ is unavoidable.

Comparing the results of Theorem 3.5 with Equations (3.4) and (3.5), `RobustUCB` policy is optimal for the gap-dependent case (up to constant factors), while it achieves upper bounds matching minimax lower bounds up to a logarithmic factor in T for the gap-independent case.

For the analysis of Theorem 3.5, ϵ and u are both known to the agent, and this knowledge was used both for the estimator and for the algorithm itself. This is a common assumption in the HT bandit literature 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. That is a huge drawback, since prior knowledge on these two quantities is hardly available for practical problems. We then aim to design an algorithm based on optimism with a practical usefulness that requires less prior knowledge about rewards yet achieves an optimal efficiency.

Let us now understand the current state-of-the-art optimal results achieved with adaptive algorithms for the stochastic HT bandit problem, even if in a conventional regret minimization MAB setting only few methods have handled heavy-tailed distributions.

A lot of works still assume the knowledge of both ϵ and u. For instance, [93] derived a logarithmic upper-bound with ϵ, u presented to the agent, while requiring also the gap

information to balance the exploration and exploitation. These features make the the policy proposed by Vakili et al. [2013], namely *DSEE*, impractical, since information about the bound or the gap is not accessible in general.

Remark 2 (Parameters knowledge required for estimator and algorithm)**.** *In all the works tackling the heavy-tailed bandit setting, we see some differences in how the approaches presented depend on ϵ and u. In particular, the knowledge of the two parameters can serve as input either for both the robust estimator and the algorithm, or only for the algorithm, with the estimator not depending on them. The latter case has, of course, weaker assumptions, but still can not be considered adaptive since requires anyway a distributional parameter as input. The fully-adaptive case is the one where both the parameters are not required, neither to compute the estimator, nor to run the algorithm.*

All the contributions presented in the following paragraphs will be collected in Table 3.1 at the end of the chapter, which allows a more straightforward comparison, distinguishing the cases where the parameters need to be known:

- by both the estimator and the algorithm;
- by the algorithm only;
- by any of the two (adaptive approach).

3.2.1. Adaptive Approaches in u, with ϵ known

The dependence on u was first removed in [28] for both the estimator and the algorithm, but assuming $\epsilon = 1$. Cesa-Bianchi et al. [2017], instead of using a confidence bound, employed the *Boltzmann-Gumbel exploration* (BGE) with a robust estimator.

Also [64] got rid of the requirement of u, yielding near-optimal regret bounds with an algorithm built on the UCB framework, that uses a novel p-robust estimator and requires a prior knowledge on ϵ only. In this way, it performs an *adaptively perturbed exploration* (APE^2), proving that, only for small suboptimality gaps, the perturbation method outperforms `RobustUCB`.

The knowledge of the bound u on the moments of rewards was not required even in [15], where it was adapted in a data-driven manner, through a best arm identification algorithm in a fixed confidence setting. Anyway, Bhatt et al. [2022], were still requiring the knowledge of ϵ for both the estimator and algorithm. Their algorithm is based on *Lepski's method* [65], which is an adaptation method that has been employed for many different tasks, and might be of inspiration to extend fully adaptive approaches [46].

3.2.2. Fully Adaptive Approaches, with u and ϵ not known

There is a rich literature in deriving algorithms adaptive to the loss sequences, for either full information setting [69, 82], stochastic bandits [38, 60, 61] or adversarial bandits [24, 97]. There are also many algorithms adaptive to the loss range when not known, e.g., the so-called "scale-free" MAB [35, 83].

However, as mentioned above, to our knowledge, our work is the first regret minimization approach to adapt fully and optimally to heavy-tail parameters u and ϵ.

The dependency on both u and $\epsilon \in (0, 1]$ was attempted to be removed firstly in [48], by proposing a *generalized successive rejects* (GSR) method. While GSR does not depend on any prior knowledge of the reward distributions, however, it only focuses on identifying the optimal arm, also known as pure exploration [20], rather than minimizing the cumulative regret. Hence, GSR loses much reward during the learning process.

The requirement of ϵ for both the estimator and the algorithm was relaxed also in [6], proposing a distribution oblivious algorithm, which requires no prior information about the parameters of the arm distributions. In particular, [6] refers to algorithm, called *R-UCB-G*, which suffers a regret slightly super-logarithmic. Nevertheless, it requires anyway the knowledge of a scaling function as input, since it uses a truncation-based estimator in conjunction with a robust scaling of the confidence bound. This work proved also that, if a specific UCB algorithm designed for $\sigma-$subgaussian distributions is used in a subgaussian setting with a mismatched variance parameter, the learning performance could be inconsistent. This is of practical concern because the parameters that define the space of arm distributions (usually in the form of support/moment bounds) are often estimated from limited data samples, and are therefore prone to errors. This statement encouraged our research towards a fully adaptive algorithm that could still have a logarithmic performance matching the lower bound in Equation (3.4).

Eventually, [45] proposes a fully adaptive algorithm, namely *AdaTINF*, with minimax optimal regret $\mathcal{O}\left(u^{\frac{1}{1+\epsilon}} K^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} + \sqrt{KT}\right)$, under specific assumptions on the losses (e.g. "truncated non-negativity"), and with ϵ, u both unknown. This shows that the existing lower bound for the HT setting $\Omega\left(u^{\frac{1}{1+\epsilon}} K^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}\right)$ is tight even when all prior knowledge on ϵ, u is absent. This work looks promising, but still the algorithm is applied to an adversarial setting employing the Follow-The-Regularized-Leader (FTRL) technique [43], and allowing the adversary to choose the distributions of losses and not directly the single losses. Indeed, to our knowledge, the theoretical novelty introduced in Chapter 5 has not been achieved yet by UCB algorithms in stochastic setting with expected regret minimization. Moreover, for the stochastic setting, [45] presented another fully adaptive

algorithm, namely *Optimistic HTINF*, which instead is not optimal, giving a sub-optimal instance-dependent regret of order:

$$\mathcal{O}\left(\sum_{i:\Delta_i>0}\left(\frac{u^2}{\Delta_i^{2\epsilon}}\right)^{\frac{1}{\epsilon}}\log T\right).$$

From all these arguments, we can highlight once again that, to our knowledge, there is not yet any fully adaptive algorithm for the stochastic HT problem giving optimal instance-dependent regret and, simultaneously, an instance-independent one optimal at most up to logarithmic terms.

3.2.3. Robust Estimators

The main issue for the extension of traditional MAB algorithms to the stochastic HT setting lies in estimators theory. It is reflected through the incapacity of the sample mean in providing a reliable estimate of the mean of a heavy-tailed distributed random variable.

This is why a broad variants of estimators has been employed in the heavy-tailed bandit literature to overcome the issue of robustness in the estimates.

Theorem 3.5 shows that the subgaussian assumption can be relaxed to require only finite variance at the price of constant factors. This result is only possible by replacing the standard empirical estimator with something more robust as employed in `RobustUCB` algorithms [19], e.g., truncated mean, median of means [4] for $\epsilon \in (0, 1]$ or Catoni's M estimator for $\epsilon = 1$ [26]. In particular, under Assumption 2 on reward distributions, the disadvantage of truncated mean estimator and Catoni's estimator is that they give a regret bound requiring the knowledge of both ϵ and a bound u on the moments. Choosing the location of truncation requires prior knowledge about the approximate location of the mean. On the contrary, median-of-mean estimators is, in some sense, more flexible since it does not depend on u. Yet another idea of robust approaches would be to minimize the Huber loss [89]. Sun et al. [2020] focuse on linear models, but the results still apply in one dimension.

A lot of extensions of basic robust estimators have been tried out. Cesa-Bianchi et al. [28] have proposed a robust estimator by modifying the Catoni's M estimator, providing a weak tail bound that allows error probability to decay slower than that of Catoni's. Another extension of Catoni's estimator has been provided in [15], allowing its validity for the case of $\epsilon \in (0, 1]$ and not just $\epsilon = 1$ as in [22].

As said, trimmed mean is a common estimator in the heavy-tailed statistics literature. It

is, indeed, very robust to outlying values, which are usually cut out and not included in the computation of averages. This guideline of intuition has been followed in [45], which presents adaptive approaches that rely on the idea of estimators with truncated losses. This is worth to mention since it is very relevant also for our research. In Chapter 5, we will extend these concepts, introducing a new trimmed robust estimator not requiring any prior knowledge on the parameters of reward distributions.

3.2.4. On Regret Definition

We recall that the (cumulative) regret as defined in Definition 2.2 is a random variable. In all our research, we are considering its expected value (Definition 2.3), depending on the arm means, as measure of performance for a policy. Since we highlighted many times that, in a heavy-tailed setting, the empirical mean does not provide robust estimates, the reader might wonder about the reason of this choice. The expected regret is indeed a custom performance measure that is coherent with classical multi-armed bandit problems literature, where the expected value of an arm is used as a metric to evaluate its goodness.

However, the expected value is a risk-neutral metric, such that, in specific applications, different approaches should be considered. For instance, in many domains like finance, one is interested in balancing the expected return of an arm (or portfolio) with the risk associated with that return. Coherently, in [48], the target problem is the one of selecting the arm that optimizes a linear combination of the expected reward and the associated *Conditional Value at Risk (CVaR)* [5]. Along the same idea, [71] considers CVaR for stochastic MAB, with an approach where the regret itself is redefined.

This analysis reflect a recent interests in risk-aware multi-armed bandit problems. On this wave, the optimization of *Value at Risk* (VaR) measure has often been considered instead of the usual regret [33, 34]. VaR was also taken into account in the context of a specific bandit policy family by [8, 10], but CVar is still preferred because it is a more coherent risk measure.

Despite all the recent research, the analysis of CVaR measure is, on a technical level, similar to the one of the expected regret. This is why, for our work, we keep assessing performance in stochastic HT bandit setting being aligned with the most used approach in literature, i.e. using the expected regret as in Definition 2.3.

To close this chapter, let us report Table 3.1 below to recap the main adaptive results investigated so far in literature.

bound u		**maximal order** $\epsilon \in (0, 1]$
Estimator	**Algorithm**	known
known	known	RobustUCB - trimmed mean, Catoni's ($\epsilon = 1$), [22] DSEE [93]
unknown	known	RobustUCB (median of means), [22]
unknown	unknown	Variant BGE ($\epsilon = 1$), [28] Adaptively Perturbed Exploration Method (APE^2), [64] Adaptive Best Arm Identif. - Generalized Catoni's, [15]
Estimator	**Algorithm**	unknown
known	unknown	
unknown	known	
unknown	unknown	GSR - Pure Exploration, [48] Robust-UCB-G, [6] Optimistic HTINF - AdaTINF, [45] Our contribution: `Adaptive Robust UCB`

Table 3.1: Review of the main adaptive Heavy-Tailed bandit algorithms presented in literature, given their different "degree of knowledge" on parameters ϵ and u.

4 | Regret Lower Bounds for Adaptive Heavy-Tailed Bandits

In this chapter, we show that, in a general stochastic heavy-tailed bandit problem with no additional assumption on the distributions of the arms, it is not possible to derive an algorithm that is adaptive in either ϵ or u, matching the lower bound on the regret stated in [22]. This means that any algorithm unaware of these two quantities cannot achieve the same regret order as the one stated in Theorem 3.3. We recall that in this theorem we provided non-adaptive lower bounds when ϵ and u are known.

We now prove a lower bound on the expected regret that any adaptive policy or algorithm (with respect to either u or ϵ) can achieve in this setting.

In the following, we will refer to any algorithm with an upper bound on the regret matching Theorem 3.3 minimax lower bound as a *matching algorithm*. More formally, a matching algorithm provides a regret upper bound UB_T at time T, such that:

$$\lim_{T\to+\infty} \frac{UB_T}{LB_T} = c \in \mathbb{R},$$

where LB_T is the lower bound on regret at time T. In this case, we say that the algorithm is *tight*, since upper bound and worst-case lower bound have the same order up to constants.

4.1. Non-Existence of a Matching Algorithm u-Adaptive

We start by stating the regret lower bound for any algorithm adaptive w.r.t. u, the value (or the lowest upper bound) of the finite moment of maximum order. In particular, we adopt here a non-standard procedure that consists in constructing lower bounds on $\frac{R_T}{(uT)^{\frac{1}{1+\epsilon}}}$. With a fair amount of intuition, we decided to focus on this quantity since the standard HT lower bound in Equation (3.5) has the same dependencies.

Theorem 4.1 (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 Assumption 2 with u and u' respectively (assume $u' > u$ without loss of generality), such that:*

$$\max\left\{\frac{R_T}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq C_1 \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(1+\epsilon)^2}}, \tag{4.1}$$

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 such that no algorithm can match on both the lower bound on regret, presented in Equation (3.5), and, instead, some regret is accrued in a way that is proportional to the ratio of the two $(1+\epsilon)$-th order moments of those instances. In our construction (more details can be found in Section 4.1.1), the ratio between u' and u can be taken arbitrarily large, and thus the regret gap with the non-adaptive lower bound presented in Equation (4.1) can be arbitrarily large. In particular, this result shows that is not possible to be adaptive in u without the risk of incurring in an arbitrarily large regret bound.

Remark 3. *In the lower bound in Equation* (3.5) *we see a dependency in $K^{\frac{\epsilon}{1+\epsilon}}$. The reader might wonder why this dependency is not included in the discussion here. This choice is for a matter of simplifying the computations. Without loss of generality, our focus in only on the dependency on* u, u' *and* T *in the right-hand side part, since, finding any of these terms, would be enough to show the non-existence of a matching u-adaptive algorithm.*

To show the result of Theorem 4.1, we start from the construction of [19, 22].

4.1.1. Step 1: Instance Construction

Let $u < u'$, we construct the two instances each made of just two arms.

Base instance

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

where δ_x denotes the Dirac delta measure centered in x and Δ is such that $\Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \in (0,1) \implies 0 < \Delta < u^{\frac{1}{1+\epsilon}}$.

We have:

$$\mu_1 = 0, \quad \mu_2 = \Delta,$$
$$\mathbb{E}_{\nu_1}[|X|^{1+\epsilon}] = 0, \quad \mathbb{E}_{\nu_2}[|X|^{1+\epsilon}] = u.$$

The optimal arm is arm 2.

Alternative instance

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

for Δ such that $(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}} \in (0,1) \implies 0 < \Delta < \frac{1}{2}(u')^{\frac{1}{1+\epsilon}}$.

We have:

$$\mu_1' = 2\Delta, \quad \mu_2' = \Delta,$$
$$\mathbb{E}_{\nu_1'}[|X|^{1+\epsilon}] = u', \quad \mathbb{E}_{\nu_2'}[|X|^{1+\epsilon}] = u.$$

The optimal arm is arm 1.

Remark 4. *These two instances belong to the heavy-tailed bandit problem since the reward distributions satisfy Assumption 2. Anyway, we can notice that these distributions in general have finite second order moments, because they are bounded with support in* $[0, u^{\frac{1}{\epsilon}}\Delta^{-\frac{1}{\epsilon}}]$ *and* $[0, (u')^{\frac{1}{\epsilon}}(2\Delta)^{-\frac{1}{\epsilon}}]$*, respectively. This might seem a simplification of our heavy-tailed framework, but [22] shows that these instances are difficult enough to be suitable for our context, meaning that* `UCB1` *algorithm (see Section 3.1) is not tight on them, it does not provide a regret upper bound matching the lower bounds.*

4.1.2. Step 2: Lower Bounding the "Adaptive" Regret

Suppose by contradiction that a matching adaptive algorithm in u exists. In such a case (see Equation 3.5), we will have that the expected regret of the base instance R_T is of order $(uT)^{\frac{1}{1+\epsilon}}$ while the regret of the alternative instance R_T' is of order $(u'T)^{\frac{1}{1+\epsilon'}}$, apart from constants not dependening on u, u' and T. Thus, it should hold that:

$$\max\left\{\frac{R_T}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R_T'}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \leq c, \tag{4.2}$$

where c is a constant that does not depend on T.

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

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

being f a function increasing in T. This suffices to show the non-existence of an algorithm adaptive in u matching the minimax lower bound in Theorem 3.3.

The proof is quite technical and merges the approach of [23] with that of [62].

First, we observe that:

$$\max\left\{\frac{R_T}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq \frac{R_T}{(uT)^{\frac{1}{1+\epsilon}}} \overset{(2.2)}{=} \frac{\Delta \mathbb{E}[N_1(T)]}{(uT)^{\frac{1}{1+\epsilon}}}, \tag{4.3}$$

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

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

$$\begin{aligned}
&\max\left\{\frac{R_T}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \\
&\qquad \overset{(2.2)}{\geq} (u'T)^{-\frac{1}{\epsilon+1}} \max\left\{\frac{\Delta T}{2}\mathbb{P}\left(N_1(T) \geq T/2\right), \frac{\Delta T}{2}\mathbb{P}'\left(N_1(T) < T/2\right)\right\} \\
&\qquad \geq \frac{\Delta}{4}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \left(\mathbb{P}\left(N_1(T) \geq T/2\right) + \mathbb{P}'\left(N_1(T) < T/2\right)\right) \\
&\qquad \geq \frac{\Delta}{8}(u')^{-\frac{1}{\epsilon+1}} T^{\frac{\epsilon}{\epsilon+1}} \exp\left(-\mathbb{E}[N_1(T)] D_{KL}(\nu_1 \| \nu'_1)\right),
\end{aligned} \tag{4.4}$$

where we used Bretagnolle-Huber inequality (2.21) and divergence decomposition (2.22), together with $\max\{a, b\} \geq \frac{1}{2}(a+b)$ for $a, b \geq 0$. Let us now compute the KL-divergence using Equation (2.20) and noting that $\nu_1 \ll \nu'_1$:

$$\begin{aligned}
D_{KL}(\nu_1 \| \nu'_1) &= \nu_1(0) \log \frac{\nu_1(0)}{\nu'_1(0)} \\
&= \log \frac{1}{1 - (2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}}} \leq 2(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}},
\end{aligned} \tag{4.5}$$

for $0 < \Delta < u^{\frac{1}{1+\epsilon}}$ such that $(2\Delta)^{1+\frac{1}{\epsilon}}(u')^{-\frac{1}{\epsilon}} \in (0, 1/2)$:

$$\implies 0 < \Delta < \min\left\{u^{\frac{1}{1+\epsilon}}, \left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}} (u')^{\frac{1}{1+\epsilon}}\right\}.$$

Combining together Equations (4.3), (4.4) and (4.5), we have:

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

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

$$\begin{aligned}
&x^* \text{ s.t. } \frac{1}{(uT)^{\frac{1}{1+\epsilon}}} - \frac{1}{4}(u')^{-\left(\frac{1}{\epsilon+1}+\frac{1}{\epsilon}\right)}T^{\frac{\epsilon}{\epsilon+1}}(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\exp\left(-2x^*(2\Delta)^{\frac{1+\epsilon}{\epsilon}}(u')^{-\frac{1}{\epsilon}}\right) = 0 \\
&\implies 2x^*(2\Delta)^{\frac{1+\epsilon}{\epsilon}}(u')^{-\frac{1}{\epsilon}} = -\log\left(\frac{1}{(uT)^{\frac{1}{1+\epsilon}}}\frac{1}{\frac{1}{4}(u')^{-\left(\frac{1}{\epsilon+1}+\frac{1}{\epsilon}\right)}T^{\frac{\epsilon}{\epsilon+1}}(2\Delta)^{\frac{1+\epsilon}{\epsilon}}}\right) \\
&\implies x^* = \frac{1}{2}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\log\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{4(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right) \\
&\implies g(x^*) = \frac{\Delta}{4}(uT)^{-\frac{1}{\epsilon+1}}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\log\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{4(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right) \\
&\qquad\qquad + \frac{\Delta}{4}\frac{1}{8}(u')^{-\frac{1}{\epsilon+1}}T^{\frac{\epsilon}{\epsilon+1}}\frac{8(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}{Tu^{\frac{1}{\epsilon+1}}}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}} \\
&\qquad = \frac{\Delta}{4}(uT)^{-\frac{1}{\epsilon+1}}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\left[\log\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{4(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right)+1\right] \\
&\qquad = \frac{\Delta}{4}(uT)^{-\frac{1}{\epsilon+1}}(2\Delta)^{-\frac{1+\epsilon}{\epsilon}}(u')^{\frac{1}{\epsilon}}\log\left(\frac{Tu^{\frac{1}{\epsilon+1}}}{4(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}e(2\Delta)^{\frac{1+\epsilon}{\epsilon}}\right).
\end{aligned}$$

We take Δ:

$$\begin{aligned}
&\frac{Tu^{\frac{1}{\epsilon+1}}}{4(u')^{\frac{1}{\epsilon}+\frac{1}{\epsilon+1}}}(2\Delta)^{\frac{1+\epsilon}{\epsilon}} = e^{\epsilon} \\
&\implies (2\Delta)^{-\frac{1+\epsilon}{\epsilon}} = \frac{T}{4}e^{-\epsilon}u^{\frac{1}{\epsilon+1}}(u')^{-\frac{1+2\epsilon}{\epsilon^2+\epsilon}}, \\
&\qquad \Delta = 2^{\frac{\epsilon-1}{1+\epsilon}}e^{\frac{\epsilon^2}{1+\epsilon}}T^{-\frac{\epsilon}{\epsilon+1}}u^{-\frac{\epsilon}{(\epsilon+1)^2}}(u')^{\frac{1+2\epsilon}{(\epsilon+1)^2}},
\end{aligned}$$

which is reasonable since $\Delta < \min\left\{u^{\frac{1}{1+\epsilon}}, \left(\frac{1}{2}\right)^{\frac{2\epsilon+1}{1+\epsilon}} (u')^{\frac{1}{1+\epsilon}}\right\}$ for sufficiently large T.

This implies that:

$$g(x^*) = 2^{\left(\frac{\epsilon-1}{1+\epsilon}-2-2\right)} u^{\left(-\frac{\epsilon}{(\epsilon+1)^2}+\frac{1}{\epsilon+1}-\frac{1}{\epsilon+1}\right)} (u')^{\left(\frac{1}{\epsilon}-\frac{1+2\epsilon}{\epsilon^2+\epsilon}+\frac{1+2\epsilon}{(\epsilon+1)^2}\right)} e^{\left(\frac{\epsilon^2}{1+\epsilon}-\epsilon\right)} T^{\left(1-\frac{1}{\epsilon+1}-\frac{\epsilon}{\epsilon+1}\right)} (1+\epsilon).$$

Ending with the calculations, we get:

$$g(x^*) = 2^{-\frac{3\epsilon+5}{\epsilon+1}} (1+\epsilon) e^{-\frac{\epsilon}{\epsilon+1}} u^{-\frac{\epsilon}{(\epsilon+1)^2}} (u')^{\frac{\epsilon}{(\epsilon+1)^2}} \geq C_1 \cdot \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}},$$

where $C_1 = 2^{-\frac{3\epsilon+5}{\epsilon+1}} (1+\epsilon) e^{-\frac{\epsilon}{\epsilon+1}}$.

Thus, we have

$$\max\left\{\frac{R_T}{(uT)^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{(u'T)^{\frac{1}{1+\epsilon}}}\right\} \geq C_1 \cdot \left(\frac{u'}{u}\right)^{\frac{\epsilon}{(\epsilon+1)^2}}, \tag{4.6}$$

proving Equation (4.1). Since $u' > u$ can be taken arbitrarily large, we have that gap can be arbitrarily large. Note that, when $\epsilon = 0$, the gap vanishes.

4.2. Non-Existence of a Matching Algorithm ϵ-Adaptive

We present a new result concerning adaptivity with respect to the parameter ϵ characterizing the maximum order $(1+\epsilon)$ of finite moments for reward distributions.

Theorem 4.2 (Lower Bound on Regret for Stochastic Adaptive Heavy-Tailed Bandit, unknown ϵ)**.** *For any algorithm adaptive w.r.t. to ϵ, with the maximum order finite moment u known, and for any fixed T, there exist two stochastic heavy-tailed bandit instances satisfying (2.3) with ϵ and ϵ' respectively (assume $\epsilon' < \epsilon$ without loss of generality), such that:*

$$\max\left\{\frac{R_T}{T^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq C_2 T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon)(1+\epsilon')^2}}, \tag{4.7}$$

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.

This theorem works independently from the value of u, as soon as Assumption 2 is fulfilled. It only accounts for the fact that, since we are studying the adaptivity with respect to ϵ-like parameter only, u is assumed to be known given ϵ and ϵ'. Indeed, u is never mentioned in the proof below, simply considering that it satisfies the following:

$$u := \max\left\{\mathbb{E}_\nu\left[|X|^{1+\epsilon}\right], \mathbb{E}_{\nu'}[|X|^{1+\epsilon'}]\right\}.$$

Differently from Theorem 4.1, where u and u' could take arbitrarily high values on the positive semi-axis of real numbers, the values of ϵ and ϵ' are known to belong to the set $(0, 1]$ and thus, for any fixed T, the term on the right-hand side of (4.7) cannot grow arbitrarily. Modern statistical literature presents methods to adapt to any unknown quantity for which lower and upper bounds are known while controlling finite-time convergence [65], thus, to be adaptive w.r.t. unknown ϵ is an easier task than adapting to an unknown u. For instance, we can observe that by searching for the maximum value of the right-hand side of (4.7), we get that for $\epsilon = 1$ and $\epsilon' = \frac{1}{3}$ the gap's order is $\approx T^{\frac{1}{16}}$.

To show Theorem 4.2 we will start from the construction of [19, 22].

4.2.1. Step 1: Instance Construction

Let $0 < \epsilon' < \epsilon < 1$, we construct the two instances each made of just two arms:

Base instance

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

where $\gamma = (2\Delta)^{\frac{1}{\epsilon}}$, for $\Delta \in [0, 1/2]$. Note that $\Delta \in [0, 1/2]$ is chosen such that the probability density functions are well defined. We have:

$$\begin{aligned} &\mu_1 = 0, \quad \mu_2 = \Delta, \\ &\mathbb{E}_{\nu_1}[|x|^\alpha] = 0, \quad \mathbb{E}_{\nu_2}[|x|^\alpha] = 2^{\frac{1-\alpha}{\epsilon}} \Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}, \end{aligned}$$

which are guaranteed to be bounded by constant (since we will make $\Delta \to 0$ in the construction) only if $\alpha \leq \epsilon + 1$. Thus, this bandit admits moments finite up to order $\epsilon + 1$. The optimal arm is arm 2.

Alternative instance

$$\begin{cases} \nu'_1 = (1 - (\gamma')^{1+\epsilon'})\delta_0 + (\gamma')^{1+\epsilon'}\delta_{1/\gamma'} \\ \nu'_2 = \nu_2, \end{cases}$$

where $\gamma' = (2\Delta)^{\frac{1}{\epsilon'}}$, for $\Delta \in [0, 1/2]$. We have:

$$\mu'_1 = 2\Delta, \quad \mu'_2 = \Delta, \tag{4.8}$$

$$\mathbb{E}_{\nu'_1}[|x|^\alpha] = (2\Delta)^{\frac{1+\epsilon'-\alpha}{\epsilon'}}, \quad \mathbb{E}_{\nu'_2}[|x|^\alpha] = 2^{\frac{1-\alpha}{\epsilon}} \Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}, \tag{4.9}$$

which are guaranteed to be bounded by constant only if $\alpha \leq 1 + \epsilon'$ (for the same reason as before, recalling that $\epsilon' < \epsilon$ too). Thus, this bandit admits moments finite up to order $1 + \epsilon'$. The optimal arm is arm 1.

4.2.2. Step 2: Lower Bounding the "Adaptive" Regret

Suppose by contradiction that an adaptive algorithm in ϵ exists. In such a case, we will have that the expected regret of the base instance R_T is of order $T^{\frac{1}{1+\epsilon}}$, while the regret of the alternative instance R'_T is of order $T^{\frac{1}{1+\epsilon'}}$, apart from constants independent of T.

Thus, it should hold that:

$$\max\left\{\frac{R_T}{T^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{T^{\frac{1}{1+\epsilon'}}}\right\} \leq c, \tag{4.10}$$

where c is a constant that does not depend on T. We will prove that this is not the case and, specifically that for any algorithm:

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

being f a function increasing in T. This suffices to show the non-existence of an algorithm adaptive in ϵ matching the minimax lower bound in Theorem 3.3.

As previously done, we highlight that also here, with respect to the lower bound in Theorem 3.3, we consider, for simplicity of calculations, only the term of order $\Omega\left(T^{\frac{1}{1+\epsilon}}\right)$. Finding a dependency on T in the right hand-side of Equation (4.10) is, indeed, enough to reach our conclusion of non-existence.

The proof is quite technical and emulates the analyses and steps performed at the previous Section 4.1.

First, we observe that:

$$\max\left\{\frac{R_T}{T^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq \frac{R_T}{T^{\frac{1}{1+\epsilon}}} \overset{(2.2)}{=} \frac{\Delta \mathbb{E}[N_1(T)]}{T^{\frac{1}{1+\epsilon}}}, \tag{4.11}$$

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

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

have:

$$\begin{aligned}
&\max\left\{\frac{R_T}{T^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{T^{\frac{1}{1+\epsilon'}}}\right\} \\
&\quad \overset{(2.2)}{\geq} T^{-\frac{1}{\epsilon'+1}} \max\left\{\frac{\Delta T}{2}\mathbb{P}\left(N_1(T) \geq \frac{T}{2}\right), \frac{\Delta T}{2}\mathbb{P}'\left(N_1(T) < \frac{T}{2}\right)\right\} \\
&\quad \geq \frac{\Delta}{4} T^{\frac{\epsilon'}{\epsilon'+1}} \left(\mathbb{P}\left(N_1(T) \geq \frac{T}{2}\right) + \mathbb{P}'\left(N_1(T) < \frac{T}{2}\right)\right) \\
&\quad \geq \frac{\Delta}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \exp\left(-\mathbb{E}[N_1(T)] D_{KL}(\nu_1 \| \nu'_1)\right),
\end{aligned} \tag{4.12}$$

where we used Bretagnolle-Huber inequality (2.21) and divergence decomposition (2.22), together with $\max\{a, b\} \geq \frac{1}{2}(a+b)$ for $a, b \geq 0$. Let us now compute the KL-divergence using (2.20) and noting that $\nu_1 \ll \nu'_1$:

$$\begin{aligned}
D_{KL}(\nu_1 \| \nu'_1) &= \nu_1(0) \log \frac{\nu_1(0)}{\nu'_1(0)} \\
&= \log \frac{1}{1 - (2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}} \leq 2(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}},
\end{aligned} \tag{4.13}$$

for $\Delta \in [0, 1/4]$.

Putting together Equations (4.11), (4.12) and (4.13), we have:

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

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

$$x^* \text{ s.t. } \frac{1}{T^{\frac{1}{1+\epsilon}}} - \frac{1}{4} T^{\frac{\epsilon'}{\epsilon'+1}} (2\Delta)^{\frac{1+\epsilon'}{\epsilon'}} \exp\left(-2x^*(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right) = 0$$

$$\implies 2x^*(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}} = -\log\left(\frac{1}{T^{\frac{1}{1+\epsilon}}} \frac{1}{\frac{1}{4} T^{\frac{\epsilon'}{\epsilon'+1}} (2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}}\right)$$

$$\implies x^* = \frac{1}{2}(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \log\left(\frac{T^{\frac{1}{\epsilon+1} + \frac{\epsilon'}{1+\epsilon'}}}{4} (2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)$$

Substituting in $g(x)$, we then get:

$$\begin{aligned} g(x^*) &= \frac{\Delta}{4} T^{-\frac{1}{\epsilon+1}} (2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \log\left(\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{4}(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right) + \\ &\quad + \frac{\Delta}{2}\frac{1}{8} T^{\frac{\epsilon'}{\epsilon'+1}} \left[\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{4}(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right]^{-1} \\ &= \frac{\Delta}{4} T^{-\frac{1}{\epsilon+1}} (2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \left[\log\left(\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{4}(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right) + 1\right] \\ &= \frac{\Delta}{4} T^{-\frac{1}{\epsilon+1}} (2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} \log\left(\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{4} e (2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right) \end{aligned}$$

We take Δ such that:

$$\begin{aligned} &\frac{T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}}{4}(2\Delta)^{\frac{1+\epsilon'}{\epsilon'}} = 1 \\ \implies &(2\Delta)^{-\frac{1+\epsilon'}{\epsilon'}} = \frac{1}{4} T^{\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}}, \\ &\Delta = 2^{\frac{\epsilon'-1}{1+\epsilon'}} T^{-\frac{\epsilon'}{1+\epsilon'}\left(\frac{1}{\epsilon+1}+\frac{\epsilon'}{1+\epsilon'}\right)}, \end{aligned}$$

verifying that $\Delta < 1/4$ holds for sufficiently large T.

This implies:

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

where $C_2 = 2^{\frac{-3\epsilon'-5}{1+\epsilon'}}$.

Thus, we have that:

$$\max\left\{\frac{R_T}{T^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq C_2 \cdot T^{\frac{\epsilon'(\epsilon-\epsilon')}{(1+\epsilon')^2(1+\epsilon)}}, \tag{4.14}$$

eventually proving Equation (4.7). Notice that the dependence on T vanishes when $\epsilon' = 0$. This is correct since, even when knowing that $\epsilon' = 0$, the regret lower bound is linear, so that we can just focus on ϵ.

To conclude, in this chapter we have shown how any algorithm adaptive with respect to 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 here refers 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 | Adaptive Robust UCB Algorithm

In this Chapter, we finally answer our original research question, i.e. whether there is an algorithm adaptive w.r.t. both ϵ and u matching the standard heavy-tailed setting's lower bound stated in Theorem 3.3. In Chapter 4, we already showed how adaptivity has a cost, and, thus, the lower bound presented in Theorem 3.3 is not achievable by any algorithm unaware of at least one of these quantities. Luckily, it is possible to restrict the set of heavy-tailed bandit problem instances under analysis to a special set, that will be defined in Section 5.2, on which our algorithm `Adaptive Robust UCB` (shortly `AdaR-UCB`), which is unaware of parameters ϵ and u, is able to achieve a regret order matching the lower bound for the standard heavy-tailed bandit problem.

5.1. New Robust Estimator Independent of ϵ and u

Trimmed mean is a common estimator in the heavy-tailed statistics literature, where observations are averaged while cutting-off values outside of a limited and bounded set of the form $[-M, M]$, thus, being more robust to extreme values than the empirical mean estimator.

Our goal is to seek for a proper value of M giving concentration results which are powerful enough for our algorithm `AdaR-UCB` to achieve an upper bound on regret matching the lower bound of Equation (3.4), without requiring the knowledge of u nor ϵ for the threshold's construction.

As already mentioned in Section 3.2.3, the most common paper in the stochastic HT literature [22] presents the `RobustUCB` algorithm, where the trimmed mean estimator replaces sample average in a standard optimism in the face of uncertainty strategy. Here, the truncated mean estimator for the mean of a set of independent observations $\mathbf{X} = (X_1, \dots, X_s)$ is defined with the following threshold, depending on both the quantities ε and u:

$$M^*_{j,s} = \left(\frac{uj}{\log\left(\delta^{-1}\right)} \right)^{\frac{1}{1+\epsilon}}, \ \forall j \in [s], \tag{5.1}$$

where $\delta \in (0,1)$, and $\epsilon \in (0,1], u < +\infty$ such that $\mathbb{E}[|X|^{1+\epsilon}] \leq u$.

We now present our extension of this robust estimator, which is similar to the "winsorized mean" and "trimmed mean" of Tukey [16], and which does not require the knowledge of the aforementioned parameters.

Definition 5.1. *Given $s \in \mathbb{N}$ and $X_1, \ldots, X_s \overset{i.i.d.}{\sim} X$ random variables distributed according to the same probability distribution, we define the robust estimator of $\mathbb{E}[X]$ as:*

$$\widehat{\mu}_s = \frac{1}{s} \sum_{j=1}^{s} X_j \mathbb{1}_{|X_j| \leq M_s}, \tag{5.2}$$

where the threshold M_s is a positive random variable defined as the solution of the following equation, given $\delta \in (0,1)$:

$$f(M) = \frac{1}{s} \sum_{i=1}^{s} \min\left\{X_i^2, M^2\right\} - \frac{25M^2 \log(\delta^{-1})}{s} = 0 \tag{5.3}$$

$$\Rightarrow \frac{1}{s} \sum_{i=1}^{s} \frac{\min\left\{X_i^2, M^2\right\}}{M^2} - \frac{25 \log(\delta^{-1})}{s} = 0 \tag{5.4}$$

Under Definition 5.1, for all $s \in \mathbb{N}$, the threshold M_s depends on none of the two parameters ϵ and u, and so does the estimator $\widehat{\mu}_s$. Another relevant difference of our estimator with respect to the trimmed estimator of [19] is that Bubeck et al. [2013] assumed the threshold $M_{j,s}^*$ in Equation (5.1) to be a real number and not a random variable.

5.1.1. Uniqueness of Estimator

The reader might now be wondering whether Equation (5.3) has a unique solution M_s. We notice that it has a trivial solution $M_s = 0$, but since to obtain Equation (5.4) we divided by M, from that point on we started assuming M_s different from 0, looking for a positive threshold. Is this problem well-posed?

Proposition 5.1 (Uniqueness of positive solution M_s, [96]). *Provided $\delta \in (0,1), s \in \mathbb{N}$ and $(X_i)_{i \in [s]}$ such that*

$$0 < 25 \log\left(\delta^{-1}\right) < \sum_{i=1}^{s} \mathbb{1}_{|X_i| > 0}, \tag{5.5}$$

then Equation (5.4) admits a unique positive solution.

This proposition allows us to give a theoretical foundation to all our next results ensur-

ing that, under mild assumptions, we have uniqueness of our threshold for the trimmed estimator in Equation (5.2).

Throughout, denote M_s as the solution to Equation (5.4), which is unique and positive whenever $0 < 25\log(\delta^{-1}) < \sum_{i=1}^{s} \mathbb{1}_{|X_i|>0}$. For completeness, we set $M_s = 0$ on $\{25\log(\delta^{-1}) \geq \sum_{i=1}^{s} \mathbb{1}_{|X_i|>0}\}$.
If $\mathbb{P}(X = 0) = 0$ and $0 < 25\log(\delta^{-1}) < s$, then $M_s > 0$ with probability one. With M_s well defined, we investigate its properties below.

5.1.2. Properties of Solution

Let us algebraically analyze the stochastic Equation (5.4) to find out relevant properties of its random solution M_s.

We firstly assume M in Equation (5.4) as positive real number, defining the following:

$$U_M = \min\left\{\left(\frac{X}{M}\right)^2, 1\right\} \in [0,1]. \tag{5.6}$$

While, given $s \in \mathbb{N}$, we let:

$$U_M(\mathbf{X}) = \frac{1}{s} Z_M(\mathbf{X}), \quad \text{with} \quad Z_M(\mathbf{X}) = \sum_{i=1}^{s} U_M^{(i)} = \sum_{i=1}^{s} \min\left\{\left(\frac{X_i}{M}\right)^2, 1\right\} \in [0,s], \tag{5.7}$$

where $X_1, \ldots, X_s \sim X$ are sampled independently, implying $U_M^{(i)}$ to be independent random variables, and M a fixed parameter. For the sake of completeness, we will define

$$\mathbf{U_M} := \left(U_M^{(i)}\right)_{i\in[s]}.$$

It can be easily noticed that $\mathbb{E}_X[U_M(\mathbf{X})] = \mathbb{E}_X[U_M]$. In particular,

$$\begin{aligned}\mathbb{E}_X\left[U_M(\mathbf{X})\right] = \mathbb{E}_X\left[\min\left\{\left(\frac{X}{M}\right)^2, 1\right\}\right] &= \frac{1}{M^2}\mathbb{E}_X[\min\{X^2, M^2\}] \\ &\leq \frac{1}{M^2}\mathbb{E}_X[|X|^{1+\epsilon}M^{1-\epsilon}] = \frac{u}{M^{1+\epsilon}}.\end{aligned}$$

On the other side,

$$\mathbb{E}_X\left[U_M(\mathbf{X})\right] = \mathbb{E}_X\left[\min\left\{\left(\frac{X}{M}\right)^2, 1\right\}\right] \geq \mathbb{E}_X[\mathbb{1}_{|X|>M}] = \mathbb{P}(|X| > M).$$

Now, we take advantage of this notation to state the following proposition, which gives a concentration interval in high probability for the sample mean estimator $U_M(\mathbf{X})$.

Proposition 5.2. *Assuming $U_M(\mathbf{X})$ as defined in Equation (5.7), we have that, for every $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\left|\sqrt{U_M(\mathbf{X})} - \sqrt{\mathbb{E}\left[U_M(\mathbf{X})\right]}\right| \leq 2\sqrt{\frac{2\log(2\delta^{-1})}{s}}. \tag{5.8}$$

Proof. To prove the result, we need to resort to Theorem 2.6 for self-bounding random variables. Let us first verify that its two assumptions (2.6) and (2.7) indeed hold for $Z_M(\mathbf{X})$ as defined in Equation (5.7).

Fixing some $k = 1, \ldots, s$ and choosing any $y \in \mathcal{X}$, with $\mathcal{X}$ identifying the support set of X, we have that:

$$Z_M(\mathbf{X}) - Z_M(\mathbf{X}_{y,k}) = \min\left\{\left(\frac{X_k}{M}\right)^2, 1\right\} - \min\left\{\left(\frac{y}{M}\right)^2, 1\right\} \leq 1 \ \forall M > 0.$$

Equation (2.6) then follows straightforward as:

$$Z_M(\mathbf{X}) - \inf_{y\in\mathcal{X}} Z_M(\mathbf{X}_{y,k}) \leq 1.$$

Following a similar reasoning, we also have that:

$$\begin{aligned}
&\sum_{k=1}^{n}\left(Z_M(\mathbf{X}) - \inf_{y\in\mathcal{X}} Z_M(\mathbf{X}_{y,k})\right)^2 \\
&\quad= \sum_{k=1}^{s}\left(\min\left\{\left(\frac{X_k}{M}\right)^2, 1\right\} - \inf_{y\in\mathbb{X}} \min\left\{\left(\frac{y}{M}\right)^2, 1\right\}\right)^2 \\
&\quad\leq \sum_{k=1}^{s}\left(\min\left\{\left(\frac{X_k}{M}\right)^2, 1\right\} + \min\left\{\left(\frac{X_k}{M}\right)^2, 1\right\}\right)^2 \\
&\quad\leq 4\sum_{k=1}^{s}\min\left\{\left(\frac{X_k}{M}\right)^2, 1\right\} = 4Z_M(\mathbf{X}),
\end{aligned}$$

almost surely, since:

$$\left(\min\left\{\left(\frac{X_k}{M}\right)^2, 1\right\}\right)^2 \leq \min\left\{\left(\frac{X_k}{M}\right)^2, 1\right\} \in [0,1].$$

Therefore, we also proved the second assumption (2.7), giving the self-boundedness of $Z_M(\mathbf{X})$ with $a = 4$. Applying Theorem 2.6, we get Equation (2.8) and we can now state that, for $\epsilon > 0$,

$$\mathbb{P}\left(\mathbb{E}[Z_M(\mathbf{X})] - Z_M(\mathbf{X}) > s\epsilon\right) \leq \exp\left(\frac{-\epsilon^2 s^2}{8\mathbb{E}[Z_M(\mathbf{X})]}\right)$$

$$\implies \mathbb{P}\left(\mathbb{E}[U_M(\mathbf{X})] - U_M(\mathbf{X}) > \epsilon\right) \leq \exp\left(\frac{-\epsilon^2 s^2}{8s\mathbb{E}[U_M(\mathbf{X})]}\right) = \exp\left(\frac{-\epsilon^2 s}{8\mathbb{E}[U_M(\mathbf{X})]}\right).$$

Let us now define:

$$\begin{aligned}
&\delta(s) = \exp\left(\frac{-\epsilon^2 s}{8\mathbb{E}[U_M(\mathbf{X})]}\right) \\
&\implies \frac{\epsilon^2 s}{8\mathbb{E}[U_M(\mathbf{X})]} = \log(\delta^{-1}) \\
&\implies \epsilon = 2\sqrt{\frac{2\mathbb{E}[U_M(\mathbf{X})]\log(\delta^{-1})}{s}}.
\end{aligned}$$

We get that, with probability at most δ:

$$\begin{aligned}
&U_M(\mathbf{X}) < \mathbb{E}[U_M(\mathbf{X})] - 2\sqrt{\frac{2\mathbb{E}[U_M(\mathbf{X})]\log(\delta^{-1})}{s}} \\
&\Leftrightarrow \mathbb{E}[U_M(\mathbf{X})] - 2\sqrt{\frac{2\log(\delta^{-1})}{s}}\sqrt{\mathbb{E}[U_M(\mathbf{X})]} - U_M(\mathbf{X}) > 0 \\
&\Leftrightarrow \sqrt{\mathbb{E}[U_M(\mathbf{X})]} > \sqrt{\frac{2\log(\delta^{-1})}{s}} + \sqrt{\frac{2\log(\delta^{-1})}{s} + U_M(\mathbf{X})} \\
&\qquad \bigcup \sqrt{\mathbb{E}[U_M(\mathbf{X})]} < \sqrt{\frac{2\log(\delta^{-1})}{s}} - \sqrt{\frac{2\log(\delta^{-1})}{s} + U_M(\mathbf{X})}.
\end{aligned}$$

The event:

$$\sqrt{\mathbb{E}[U_M(\mathbf{X})]} < \sqrt{\frac{2\log(\delta^{-1})}{s}} - \sqrt{\frac{2\log(\delta^{-1})}{s} + U_M(\mathbf{X})}$$

has null probability since $U_M(\mathbf{X}) \in [0, 1]$, such that:

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

Using a root inequality that reads as $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for a, b positive, we can rewrite

the above as:

$$\mathbb{P}\left(\sqrt{\mathbb{E}[U_M(\mathbf{X})]} > 2\sqrt{\frac{2\log(\delta^{-1})}{s}} + \sqrt{U_M(\mathbf{X})}\right) \leq \delta$$
$$\implies \mathbb{P}\left(\sqrt{\mathbb{E}[U_M(\mathbf{X})]} - \sqrt{U_M(\mathbf{X})} \leq 2\sqrt{\frac{2\log(\delta^{-1})}{s}}\right) \geq 1-\delta. \tag{5.9}$$

For the left tail, a similar reasoning holds. We can indeed exploit Theorem 2.6 to get Equation (2.9), such that, for $\epsilon > 0$:

$$\mathbb{P}\left(Z_M(\mathbf{X}) - \mathbb{E}[Z_M(\mathbf{X})] > s\epsilon\right) \leq \exp\left(\frac{-\epsilon^2 s^2}{8\mathbb{E}[Z_M(\mathbf{X})] + 4s\epsilon}\right)$$
$$\implies \mathbb{P}\left(U_M(\mathbf{X}) - \mathbb{E}[U_M(\mathbf{X})] > \epsilon\right) \leq \exp\left(\frac{-\epsilon^2 s^2}{8s\mathbb{E}[U_M(\mathbf{X})] + 4s\epsilon}\right)$$
$$= \exp\left(\frac{-\epsilon^2 s}{8\mathbb{E}[U_M(\mathbf{X})] + 4\epsilon}\right).$$

With similar passages to the above ones, let us define:

$$\delta(s) = \exp\left(\frac{-\epsilon^2 s}{8\mathbb{E}[U_M(\mathbf{X})] + 4\epsilon}\right)$$
$$\implies \frac{\epsilon^2 s}{8\mathbb{E}[U_M(\mathbf{X})] + 4\epsilon} = \log(\delta^{-1})$$
$$\implies \epsilon^2 s - 4\log(\delta^{-1})\epsilon - 8\mathbb{E}[U_M(\mathbf{X})]\log(\delta^{-1}) = 0$$
$$\implies \epsilon_{1,2} = \frac{4\log(\delta^{-1}) \pm \sqrt{16\log(\delta^{-1})^2 + 32s\mathbb{E}[U_M(\mathbf{X})]\log(\delta^{-1})}}{2s}.$$

Since ϵ needs to be positive:

$$\epsilon = \frac{2\log(\delta^{-1})}{s} + \frac{2\sqrt{\log(\delta^{-1})^2 + 2s\mathbb{E}[U_M(\mathbf{X})]\log(\delta^{-1})}}{s}.$$

We get that, with probability at least $1-\delta$, and using the root inequality:

$$U_M(\mathbf{X}) \leq \mathbb{E}[U_M(\mathbf{X})] + \frac{2\log(\delta^{-1})}{s} + \frac{2\sqrt{\log(\delta^{-1})^2 + 2s\mathbb{E}[U_M(\mathbf{X})]\log(\delta^{-1})}}{s}$$
$$\implies \mathbb{P}\left(U_M(\mathbf{X}) \leq \mathbb{E}[U_M(\mathbf{X})] + \frac{4\log(\delta^{-1})}{s} + 2\sqrt{\frac{2\mathbb{E}[U_M(\mathbf{X})]\log(\delta^{-1})}{s}}\right) \geq 1-\delta.$$

This is equivalent to:

$$
\begin{aligned}
U_M(\mathbf{X}) &\overset{1-\delta}{\leq} \left(\sqrt{\mathbb{E}[U_M(\mathbf{X})]} + \sqrt{\frac{2\log(\delta^{-1})}{s}}\right)^2 + \frac{2\log(\delta^{-1})}{s} \\
\Leftrightarrow \sqrt{U_M(\mathbf{X})} &\overset{1-\delta}{\leq} \sqrt{\left(\sqrt{\mathbb{E}[U_M(\mathbf{X})]} + \sqrt{\frac{2\log(\delta^{-1})}{s}}\right)^2 + \frac{2\log(\delta^{-1})}{s}} \\
\implies & \mathbb{P}\left(\sqrt{U_M(\mathbf{X})} \leq \left(\sqrt{\mathbb{E}[U_M(\mathbf{X})]} + \sqrt{\frac{2\log(\delta^{-1})}{s}}\right) + \sqrt{\frac{2\log(\delta^{-1})}{s}}\right) \geq 1-\delta \\
\Leftrightarrow & \mathbb{P}\left(\sqrt{U_M(\mathbf{X})} - \sqrt{\mathbb{E}[U_M(\mathbf{X})]} \leq 2\sqrt{\frac{2\log(\delta^{-1})}{s}}\right) \geq 1-\delta
\end{aligned}
$$

The two inequalities proved for right and left tails are such that:

$$
\mathbb{P}\left(\sqrt{\mathbb{E}[U_M(\mathbf{X})]} - \sqrt{U_M(\mathbf{X})} > 2\sqrt{\frac{2\log(2\delta^{-1})}{s}}\right) < \frac{\delta}{2}. \tag{5.10}
$$

$$
\mathbb{P}\left(\sqrt{U_M(\mathbf{X})} - \sqrt{\mathbb{E}[U_M(\mathbf{X})]} > 2\sqrt{\frac{2\log(2\delta^{-1})}{s}}\right) < \frac{\delta}{2} \tag{5.11}
$$

By using the union bound $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, we have:

$$
\begin{aligned}
&\mathbb{P}\left(\left|\sqrt{U_M(\mathbf{X})} - \sqrt{\mathbb{E}[U_M(\mathbf{X})]}\right| > 2\sqrt{\frac{2\log(2\delta^{-1})}{s}}\right) < \delta \\
&\Leftrightarrow \mathbb{P}\left(\left|\sqrt{U_M(\mathbf{X})} - \sqrt{\mathbb{E}[U_M(\mathbf{X})]}\right| \leq 2\sqrt{\frac{2\log(2\delta^{-1})}{s}}\right) \geq 1-\delta
\end{aligned}
$$

and Equation (5.8) is now proved. ■

For the next theoretical results we will assume, without loss of generality, δ such that:

$$
\delta^{-1} > 2 \implies \log\left(2\delta^{-1}\right) = \log(2) + \log\left(\delta^{-1}\right) \leq 2\log\left(\delta^{-1}\right), \tag{5.12}
$$

so that Equation (5.8) reduces to:

$$
\mathbb{P}\left(\left|\sqrt{U_M(\mathbf{X})} - \sqrt{\mathbb{E}[U_M(\mathbf{X})]}\right| \leq 4\sqrt{\frac{\log(\delta^{-1})}{s}}\right) \geq 1-\delta \tag{5.13}
$$

From now on, if we want to choose the value of M satisfying Equation (5.4), we need to

start assuming its solution as a random variable, due to the randomness in the equation $f(M) = 0$.

Let us now define $M = M_s$ random variable such that, with c positive constant:

$$\frac{c\log(\delta^{-1})}{s} = U_{M_s}(\mathbf{X}). \tag{5.14}$$

We use Proposition 5.2 and Equation (5.13) to show that:

$$\begin{aligned}
\sqrt{\frac{c\log(\delta^{-1})}{s}} = \sqrt{U_{M_s}(\mathbf{X})} &\overset{1-\delta}{\geq} \sqrt{\mathbb{E}\left[U_{M_s}(\mathbf{X})\right]} - 4\sqrt{\frac{\log(\delta^{-1})}{s}} \\
&\overset{1-\delta}{\geq} \sqrt{\mathbb{P}\left(|X| > M_s\right)} - 4\sqrt{\frac{\log(\delta^{-1})}{s}}; \\
\Leftrightarrow (\sqrt{c}+4)\sqrt{\frac{\log(\delta^{-1})}{s}} &\overset{1-\delta}{\geq} \sqrt{\mathbb{P}\left(|X| > M_s\right)}. \\
\Leftrightarrow \mathbb{P}\left(|X| > M_s\right) &\overset{1-\delta}{\leq} (\sqrt{c}+4)^2\frac{\log(\delta^{-1})}{s}.
\end{aligned}$$

On the other side,

$$\begin{aligned}
\sqrt{\frac{c\log(\delta^{-1})}{s}} = \sqrt{U_{M_s}(\mathbf{X})} &\overset{1-\delta}{\leq} \sqrt{\mathbb{E}\left[U_{M_s}(\mathbf{X})\right]} + 4\sqrt{\frac{\log(\delta^{-1})}{s}} \\
&\overset{1-\delta}{\leq} \sqrt{\frac{u}{M_s^{1+\epsilon}}} + 4\sqrt{\frac{\log(\delta^{-1})}{s}}; \\
\Leftrightarrow (\sqrt{c}-4)\sqrt{\frac{\log(\delta^{-1})}{s}} &\overset{1-\delta}{\leq} \sqrt{\frac{u}{M_s^{1+\epsilon}}} \\
\Leftrightarrow (\sqrt{c}-4)^2\frac{\log(\delta^{-1})}{s} &\overset{1-\delta}{\leq} \frac{u}{M_s^{1+\epsilon}} \\
\Leftrightarrow M_s^{1+\epsilon} &\overset{1-\delta}{\leq} \frac{us}{(\sqrt{c}-4)^2\log(\delta^{-1})}.
\end{aligned}$$

We choose $c = 25$ to be consistent with the calculus above, since $(\sqrt{c}-4)$ has to be greater than 0. Then, we get that M_s is indeed the solution of Equation (5.4) and the following bounds in high probability hold:

Corollary 5.2 (High Probability Bounds for Random Threshold M_s)**.** *For $\delta \in (0,1)$ and M_s random variable solution of Equation* (5.4)*, we have that the following two events hold*

together with probability at least $1-\delta$*:*

$$\mathbb{P}\left(|X| > M_s\right) \leq \frac{81\log(\delta^{-1})}{s}, \tag{5.15}$$

and

$$M_s \leq \left(\frac{us}{\log(\delta^{-1})}\right)^{\frac{1}{1+\epsilon}}. \tag{5.16}$$

Inequality (5.16) makes sense since for $s \to \infty$ necessarily $M_s \to \infty$.

5.2. The Truncated Non-Positivity Assumption

In this section, we state a key assumption for our work, namely the *truncated non-positivity assumption*.

Assumption 3 (Truncated Non-Positivity Assumption)**.** *Given a set of* K *distributions satisfying Assumption* (2)*, let* ν_1 *be the distribution of the unique optimal arm, namely* $\mu_1 > \mu_i$ *for all* $i \geq 1$*, then:*

$$\mathbb{E}_{\nu_1}[X\mathbb{1}_{|X|>M}] \leq 0, \quad \forall M \geq 0. \tag{5.17}$$

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 $\mathbb{R}$.
We highlight that this assumption only needs to hold for the *optimal arm*.

A similar version of this assumption, called *truncated non-negativity*, appeared in [45] in the context of heavy-tailed bandits, using losses instead of rewards. In that work, authors discuss the weak nature of this assumption comparing it to more stronger assumptions that are common in the literature. The two lower bounds stated in Equations (4.1) and (4.7) have respectively been obtained by introducing four instances which violate this assumption (see proofs in Sections 4.1.1 and 4.2.1). We, therefore, grasp that 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 Chapter 4. At this point, one may wonder whether enforcing our new assumption leads to a strictly smaller lower bound on the regret.

5.2.1. Minimax Lower Bound for Truncated Non-Positive Bandit Instances

We show now that forcing the truncated non-positivity assumption does not result in an improvement of the lower bound in Theorem 3.3, which we recall below:
For any algorithm and for any fixed T, there exists a set of K distributions satisfying Assumption 2, so that we have:

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

At this point, we employ here the same construction as in [62], but replace the Gaussian distributions with the mixtures of Dirac deltas. We define:

$$\nu_\Delta = \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}} \delta_0 + \left(1 - \Delta^{1+\frac{1}{\epsilon}} u^{-\frac{1}{\epsilon}}\right) \delta_{-u^{\frac{1}{\epsilon}} \Delta^{-\frac{1}{\epsilon}}}, \tag{5.18}$$

The two instances are constructed by the means of Equation (5.18).

Base Instance ν

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

Alternative Instance ν'

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

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

Both instances satisfy Assumption 3, i.e. they are truncated non-positive.

We have:

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

Using Bretagnolle-Huber inequality (2.21) we get:

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

We then develop the Kullback-Leibler divergence between the two instances, relying on

the divergence decomposition in Equation (2.22):

$$\begin{aligned} D_{KL}(\nu||\nu') &= \sum_{j=1}^{K} \mathbb{E}[N_j(T)] D_{KL}(\nu_j||\nu'_j) \\ &= \mathbb{E}[N_i(T)] D_{KL}(\nu_i||\nu'_i) \\ &\overset{(*)}{\leq} \frac{T}{K-1} D_{KL}(\nu_i||\nu'_i), \end{aligned}$$

where the step marked by $(*)$ follows from the fact that i is the least pulled arm in instance ν. Proceeding with the computations, we employ Equation (2.20) to obtain:

$$\begin{aligned} D_{KL}(\nu||\nu') &\leq \frac{T}{K-1}\Bigg[(3\Delta)^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}} \log\left(\frac{(3\Delta)^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}{\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}\right) + \\ &\qquad + (1-(3\Delta)^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}) \log\left(\frac{1-(3\Delta)^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}{1-\Delta^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}}}\right)\Bigg] \\ &\leq \frac{T}{K-1}(3\Delta)^{1+\frac{1}{\epsilon}}u^{-\frac{1}{\epsilon}} \log(3^{1+\frac{1}{\epsilon}}). \end{aligned}$$

Plugging this result, we finally get:

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

We conclude the proof by noting that $\max\{x,y\} > \frac{1}{2}(x+y)$ and setting Δ as:

$$\Delta = \frac{1}{3}\left(\frac{K-1}{T}\frac{u^{\frac{1}{\epsilon}}}{\log\left(3^{1+\frac{1}{\epsilon}}\right)}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

Finally, we have:

$$\max\{R_T, R'_T\} \geq c(uT)^{\frac{1}{1+\epsilon}}K^{\frac{\epsilon}{\epsilon+1}},$$

for some constant c independent of T and u.

This proves that even under Assumption 3 is not possible to further improve the regret lower bound from [22] (see Theorem 3.3).

In general, introducing an additional assumption as *truncated non-positivity*, we might

expect a minimax lower bound that is lower than the one in [22], obtained for the heavy-tailed bandit setting without further assumptions. In this section, for the two instances considered, we obtained a regret lower bound of order $\Omega\left((uT)^{\frac{1}{\epsilon+1}}K^{\frac{\epsilon}{\epsilon+1}}\right)$, implying that, even under Assumption 3, it is not possible to further improve the regret lower bound from [22]. Since `RobustUCB` algorithm is tight, with its upper bound matching the instance-independent lower bound up to logarithmic terms, the result of this section is promising: if we find an adaptive algorithm with the same order of performance of `RobustUCB`, we are sure that it is tight.

We are now ready to show how, under the same assumption, it is possible to be adaptive with respect to both ϵ and u while attaining the best regret order achievable in the heavy-tailed stochastic bandit problem.

5.3. A Fully Adaptive Algorithm: AdaR-UCB

We are now ready to introduce Algorithm 5.1, namely AdaR-UCB, which is based on optimism and able to operate in the heavy-tailed bandit problem *without any prior knowledge on ϵ nor u.*

Algorithm 5.1 AdaR-UCB

Initialize $s_i \leftarrow 0, \quad \mathbf{X}_i \leftarrow \emptyset, \quad \mathbf{X}'_i \leftarrow \emptyset, \quad \hat{\mu}_{i,0,1} \leftarrow +\infty \quad \forall i \in [K]$.

for $t \in [\lfloor \frac{T}{2} \rfloor]$ **do**

  for $i \in [K]$ **do**

    Compute threshold $\widehat{M}_{i,s_i,t}$ solving

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

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

$$\mathbf{Y_{i,t}} \leftarrow \{X_{i,1} \mathbb{1}_{\{|X_{i,1}| \leq \widehat{M}_{i,s_i,t}\}}, \ldots, X_{i,s_i} \mathbb{1}_{\{|X_{i,s_i}| \leq \widehat{M}_{i,s_i,t}\}}\}.$$

    Compute trimmed mean estimator $\hat{\mu}_{i,s_i,t}(\mathbf{X}_i) \leftarrow \frac{1}{s_i} \sum_{j \in [s_i]} Y_{i,j,t}$

    Compute sample variance of trimmed observations

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

  end for

  Select an action

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

  Play action i_t and receive an observation X_t

  Update samples $\mathbf{X}_{i_t} \leftarrow \mathbf{X}_{i_t} \cup \{X_t\}$

  Play action i_t and receive an observation X'_t

  Update samples $\mathbf{X'}_{i_t} \leftarrow \mathbf{X}'_{i_t} \cup \{X'_t\}$

  Update number of pulls $s_{i_t} \leftarrow s_{i_t} + 1$

end for

In RobustUCB 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, it is built under the same principle and strategy, but while in RobustUCB the threshold choice is driven by the values of ϵ and u, AdaR-UCB computes a proxy threshold $\widehat{M}$ without resorting to either ϵ or u (or any estimation of them).

AdaR-UCB operates over T rounds, however in Algorithm 5.1 we presented an interaction with the environment lasting only $\lfloor \frac{T}{2} \rfloor$ rounds. Indeed, for each round t, AdaR-UCB chooses a single arm, but collects two rewards from it instead of one, and this is the reason why two different sets of collected rewards have been introduced: $\mathbf{X}_i$ and $\mathbf{X}'_i$ for each arm $i \in [K]$. The reason behind this choice lies in the fact that threshold $\widehat{M}_{i,s_i,t}$ and trimmed mean estimator $\widehat{\mu}_{i,s_i,t}$ need to be computed from independent samples of data. This design choice will ensure that the concentration inequalities built on both $\widehat{M}_{i,s_i,t}$ and $\widehat{\mu}_{i,s_i,t}$ hold properly at the cost of a 2 factor in the final regret of the algorithm.

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

Theorem 5.3 (Upper Bound on Regret for AdaR-UCB)**.** *Given a heavy-tailed bandit problem instance satisfying Assumption 3, the regret of* ***AdaR-UCB*** *at time horizon T then satisfies:*

$$R_T \leq \sum_{i:\Delta_i>0} \left(160 \left(\frac{40u}{\Delta_i} \right)^{\frac{1}{\varepsilon}} \log T + 7\Delta_i \right). \tag{5.19}$$

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.

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

Theorem 5.4 (Instance-Independent Upper Bound on Regret for AdaR-UCB)**.** *Given any heavy-tailed bandit problem instance with K arms that satisfies Assumption 3, 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:*

$$R_T \leq T^{\frac{1}{1+\epsilon}} (320K \log T)^{\frac{\epsilon}{1+\epsilon}} (40u)^{\frac{1}{1+\epsilon}}. \tag{5.20}$$

Remark 5 (Well-posedness of threshold). *As can be noticed, since the algorithm performs differently on different instances, we are not guaranteed to always have, for each arm i, the parameters s and t satisfying Equation (5.5), i.e. $0 < \log(t^4) < s$. In this unsuccessful case, we can not find any unique positive threshold $\widehat{M}_{i,s_i,t}$ solving equation in line 4 of the pseudo-code, such that, without loss of generality, we set both the threshold and the estimator to $+\infty$ at instant t.*

Remark 6. *The estimator $\widehat{\mu}_s$ in Definition 5.1 is given for a single process generating random variables $X_1, \dots, X_s$, so it depends only on the number of pulls s. In Algorithm 5.1 we extend this Definition considering each estimator associated to a given arm i, such that the rewards X_i considered are sampled only from the selected distribution.*

Now, before proving the key results of Theorems 5.3 and 5.4, we need to introduce new forms of concentration inequalities for the trimmed mean estimator, discussing the role of Assumption 3 on them.

5.4. Derivations of New Concentration Inequalities

We first state a concentration inequality for the trimmed mean estimator in Equation (5.2), which is explicitly dependent on the threshold value M_s. This concentration result is a pivotal ingredient to prove the theoretical performances of our approach.

Theorem 5.5 (Concentration Inequality for Trimmed-Mean Estimator). *Given a set of i.i.d. observations $\mathbf{X} = \{X_1, \dots, X_s\}$, and given a threshold $M_s > 0$, under Assumption 3 we get that for any given $\delta \in (0, 1)$:*

$$\mathbb{P}\left(\mu - \widehat{\mu}_s \leq \sqrt{\frac{2V_s(\mathbf{Y})\log(2\delta^{-1})}{s}} + \frac{14M_s\log(2\delta^{-1})}{3(s-1)}\right) \geq 1 - \delta \tag{5.21}$$

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

The result above can be obtained by decomposing the gap between the true mean and the estimator in a bias-variance fashion by the means of the trimmed variable $\mathbf{Y}$. The bias can be neglected under Assumption 3, while the variance of $\mathbf{Y}$ (which is bounded by construction) is controlled using the well-known Empirical Bernstein bound [73].

Proof. With probability at least $1-\delta$:

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

with

$$V_s(\mathbf{Y}) = \frac{1}{s(s-1)}\sum_{i,j=1}^{s}\frac{(Y_i - Y_j)^2}{2},$$

sample variance estimator of $\mathbf{Y} = (Y_t)_{t\in[s]}$, given:

$$Y_t = X_t \mathbf{1}_{|X_t|\leq M_s} \ \forall t \in [s].$$

Note that in step (I) we used *truncated non-positivity* (Assumption 3) to bound the bias with 0 from above. In step (II), instead, we used Equation (2.14) for $Y_t = X_t \mathbf{1}_{|X_t|\leq M_s}$ $\forall t \in [s]$ independent random variables, with $Y_t \in [-M_s, +M_s]$ $\forall t \in [s]$ almost surely. ■

Given Equation (5.21), we can further investigate on the concentration bound in high probability for $\mu - \widehat{\mu}_s$ retrieving the assumption on δ made in (5.12). It leads to have $\log(2\delta^{-1}) \leq 2\log(\delta^{-1})$, giving:

$$\begin{aligned}
&\sqrt{\frac{2V_s(\mathbf{Y})\log(2\delta^{-1})}{s}} + \frac{14M_s\log(2\delta^{-1})}{3(s-1)} \\
&\leq 2\sqrt{\frac{V_s(\mathbf{Y})\log(\delta^{-1})}{s}} + \frac{28}{3}\frac{M_s\log(\delta^{-1})}{s-1}.
\end{aligned}$$

Moreover, we use that $\frac{1}{s-1} \leq \frac{2}{s}$, for $s \geq 2$, to obtain the final form of the concentration inequality:

$$\implies \mathbb{P}\left(\mu - \widehat{\mu}_s \leq 2\sqrt{\frac{V_s(\mathbf{Y})\log(\delta^{-1})}{s}} + 19\frac{M_s\log(\delta^{-1})}{s}\right) \geq 1-\delta. \tag{5.22}$$

Theorem 5.5, together with (5.22), shows that trimmed mean estimator achieves a sub-Gaussian type concentration rate in function of M. This bound is used at line 9 of Algorithm 5.1 to compute the upper confidence bounds over the trimmed mean estimators, allowing `AdaR-UCB` to choose an action following the optimism in the face of uncertainty paradigm.

Before stating other concentration rules, given M_s random variable solution of Equation (5.4), we let ξ_δ be the event:

$$\xi_\delta = \left\{ M_s \leq \left(\frac{us}{\log(\delta^{-1})} \right)^{\frac{1}{1+\epsilon}} \cap \ \mathbb{P}(|X_t| > M_s) \leq \frac{81 \log(\delta^{-1})}{s} \right\}, \tag{5.23}$$

for $\delta \in (0,1)$. Retrieving the results of Corollary 5.2, we have that $\mathbb{P}_{M_s}(\xi_\delta) \geq 1 - \delta$.

We are now ready to show another bound in high probability for $\widehat{\mu}_s$, which now depends on the true values of parameters ϵ and u, and not on M_s anymore. This happens because we will use here Bernstein's inequality for bounded random variables (Equation 2.5), instead of Empirical Bernstein's.

Theorem 5.6 (Threshold-Independent Concentration Inequality for Trimmed-Mean Estimator)**.** *Let $\mathbf{X} = (X_1, \ldots, X_s)$ be a set of i.i.d. observations with $X_i \sim X$ for all $i \in [s]$. If X satisfies Equation (2.3), we have:*

$$\mathbb{P}\left(\widehat{\mu}_s - \mu \leq 11 u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \cap \xi_\delta \right) \geq 1 - 2\delta \tag{5.24}$$

Proof. Following the same computations as for the proof of Theorem 5.5, we can first compute

$$\mathbb{P}\left(\widehat{\mu}_s - \mu \leq 11 u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \,\middle|\, \xi_\delta \right).$$

We have that, with probability at least $1 - \delta$:

$$\begin{aligned}
\widehat{\mu}_s - \mu =& \frac{1}{s} \sum_{t=1}^{s} X_t \mathbf{1}_{|X_t| \leq M_s} - \mathbb{E}[X] \\
=& \frac{1}{s} \sum_{t=1}^{s} \left(X_t \mathbf{1}_{|X_t| \leq M_s} - \mathbb{E}\left[X_t \mathbf{1}_{|X_t| \leq M_s} \right] \right) + \frac{1}{s} \sum_{t=1}^{s} \left(\mathbb{E}\left[X_t \mathbf{1}_{|X_t| \leq M_s} \right] - \mathbb{E}[X_t] \right) \\
=& \frac{1}{s} \sum_{t=1}^{s} \left(X_t \mathbf{1}_{|X_t| \leq M_s} - \mathbb{E}\left[X_t \mathbf{1}_{|X_t| \leq M_s} \right] \right) + \frac{1}{s} \sum_{t=1}^{s} \mathbb{E}[-X_t \mathbf{1}_{|X_t| > M_s}] \leq
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{s}\sum_{t=1}^{s}\left(X_t\mathbf{1}_{|X_t|\leq M_s}-\mathbb{E}\left[X_t\mathbf{1}_{|X_t|\leq M_s}\right]\right)+\frac{1}{s}\sum_{t=1}^{s}\mathbb{E}[|X_t|\mathbf{1}_{|X_t|>M_s}]\\
&\overset{(*)}{\leq} \frac{1}{s}\sum_{t=1}^{s}\left(X_t\mathbf{1}_{|X_t|\leq M_s}-\mathbb{E}\left[X_t\mathbf{1}_{|X_t|\leq M_s}\right]\right)+\\
&\qquad+\frac{1}{s}\sum_{t=1}^{s}\left(\mathbb{E}\left[|X_t|^{1+\varepsilon}\right]^{\frac{1}{1+\varepsilon}}\right)\left(\mathbb{E}\left[\left(\mathbf{1}_{|X_t|>M_s}\right)^{\frac{1+\varepsilon}{\varepsilon}}\right]^{\frac{\epsilon}{1+\varepsilon}}\right)\\
&\overset{(**)}{\leq} \sqrt{\frac{2M_s^{1-\epsilon}u\log\left(\delta^{-1}\right)}{s}}+\frac{M_s\log\left(\delta^{-1}\right)}{3s}+\frac{1}{s}\sum_{t=1}^{s}\left(u^{\frac{1}{1+\varepsilon}}\right)\left(\mathbb{E}\left[\mathbf{1}_{|X_t|>M_s}\right]^{\frac{\epsilon}{1+\varepsilon}}\right)\\
&\leq \sqrt{\frac{2M_s^{1-\epsilon}u\log\left(\delta^{-1}\right)}{s}}+\frac{M_n\log\left(\delta^{-1}\right)}{3s}+u^{\frac{1}{1+\varepsilon}}\left(\frac{1}{n}\sum_{t=1}^{s}\mathbb{P}\left[|X_t|>M_s\right]^{\frac{\epsilon}{1+\varepsilon}}\right),
\end{aligned}
$$

where in step $(*)$ we applied repetitively Hölder's inequality in (A.4).

In this specific case $q=\frac{1+\varepsilon}{\varepsilon}$, $p=1+\varepsilon$, $X=|X_t|$ and $Y=\mathbf{1}_{|X_t|>M_s}$ for all $t\in[s]$.

Moreover, Bernstein's inequality (2.5) is applied in step $(**)$ since we have that:

$$
\frac{1}{s}\sum_{t=1}^{s}\left(X_t\mathbf{1}_{|X_t|\leq M_s}-\mathbb{E}\left[X_t\mathbf{1}_{|X_t|\leq M_s}\right]\right)=\widehat{\mu}_s-\mu,
$$

for a sequence of independent random variables $Y_1,...,Y_s$ with $Y_t=X_t\mathbf{1}_{|X_t|\leq M_s}$ for all $t\in[s]$. In this context, the random variables have bounded support and bounded second moments; the second moments are not known, only a bound is:

$$
\begin{aligned}
&Y_t\in[-M_s,M_s]\ \forall t\in[s];\\
&\mathbb{E}[Y_t^2]=\mathbb{E}\left(X_t^2\mathbf{1}_{|X_t|\leq M_s}\right)=\mathbb{E}\left(|X_t|^{1+\epsilon}|X_t|^{1-\epsilon}\mathbf{1}_{|X_t|\leq M_s}\right)\\
&\qquad\leq\mathbb{E}\left(|X|^{1+\epsilon}M_s^{1-\varepsilon}\right)\leq uM_s^{1-\epsilon}\ \forall t\in[s];\\
&\mathrm{Var}(Y_t)=\mathbb{E}[(Y_t-\mathbb{E}[Y_t])^2]=\mathbb{E}[Y_t^2]-(\mathbb{E}[Y_t])^2\\
&\qquad\leq\mathbb{E}[Y_t^2]\leq uM_s^{1+\epsilon}\ \forall t\in[s],
\end{aligned}
\tag{5.25}
$$

where in Equation (5.25) the random variable $|X|^{1-\epsilon}\mathbb{1}_{|X_t|\leq M_s}$ was bounded with $M_s^{1-\epsilon}$ inside the expected value.

To proceed further, we recall that we are computing our probability conditioned to the

event ξ_t (see Equation 5.23). In this way, we get, with probability at least $1-\delta$:

$$\begin{aligned}
\widehat{\mu}_s - \mu &\leq \sqrt{\frac{2M_s^{1-\epsilon} u \log\left(\delta^{-1}\right)}{s}} + \frac{M_s \log\left(\delta^{-1}\right)}{3s} + u^{\frac{1}{1+\varepsilon}} \left(\frac{1}{s} \sum_{t=1}^{s} \mathbb{P}\left[|X_t| > M_s\right]^{\frac{\epsilon}{1+\varepsilon}} \right) \\
&\leq \left[\frac{2\left(\frac{us}{\log(\delta^{-1})}\right)^{\frac{1-\epsilon}{1+\epsilon}} u \log\left(\delta^{-1}\right)}{s} \right]^{\frac{1}{2}} + \frac{\left(\frac{us}{\log(\delta^{-1})}\right)^{\frac{1}{1+\epsilon}} \log\left(\delta^{-1}\right)}{3s} + \\
&\qquad + \frac{u^{\frac{1}{1+\varepsilon}}}{s} \sum_{t=1}^{s} \left(\frac{81 \log\left(\delta^{-1}\right)}{s} \right)^{\frac{\epsilon}{1+\varepsilon}} \\
&\leq u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \left(\sqrt{2} + \frac{1}{3} \right) + u^{\frac{1}{1+\epsilon}} \left[\frac{81 \log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \\
&\leq u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \left(\sqrt{2} + \frac{1}{3} + 9 \right) \\
&\implies \mathbb{P}\left(\widehat{\mu}_s - \mu \leq 11 u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \,\middle|\, \xi_\delta \right) \geq 1 - \delta.
\end{aligned}$$

Using Bayes Rule, we complete the proof computing:

$$\begin{aligned}
&\mathbb{P}\left(\widehat{\mu}_s - \mu \leq 11 u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \cap \xi_\delta \right) \\
&\qquad = \mathbb{P}\left(\widehat{\mu}_s - \mu \leq 11 u^{\frac{1}{1+\epsilon}} \left[\frac{\log(\delta^{-1})}{s} \right]^{\frac{\epsilon}{1+\epsilon}} \,\middle|\, \xi_\delta \right) \mathbb{P}\left(\xi_\delta\right) \geq (1-\delta)^2 \geq 1 - 2\delta.
\end{aligned}$$

■

We need just one last theoretical result to recall before proving Theorem 5.3. Taking advantage of Theorem 2.8 and Equation (2.19), we can state the following:

Proposition 5.3. *Let $s \geq 2$, M_s positive threshold and $\mathbf{Y} = (Y_1, \ldots, Y_s)$ be a vector of independent random variables with values in $[-M_s, M_s]$. Then for $\delta > 0$ we have, writing $\mathbb{E}[V_s(\mathbf{Y})]$ for $\mathbb{E}_{\mathbf{Y}}\left[V_s(\mathbf{Y})\right]$,*

$$\mathbb{P}\left(\sqrt{V_s(\mathbf{Y})} \leq \sqrt{\mathbb{E}[V_s(\mathbf{Y})]} + 2M_s \sqrt{\frac{2\log(\delta^{-1})}{s-1}} \right) \geq 1 - \delta,$$

with

$$V_s(\mathbf{Y}) = \frac{1}{s(s-1)} \sum_{i,j=1}^{s} \frac{(Y_i - Y_j)^2}{2}.$$

In particular, we can retrieve the same reasoning done for Equation (5.22), noting that $\frac{1}{s-1} \leq \frac{2}{s}$, for $s \geq 2$. We thus obtain:

$$\mathbb{P}\left(\sqrt{V_n(\mathbf{Y})} \leq \sqrt{\mathbb{E}[V_s(\mathbf{Y})]} + 4M_s\sqrt{\frac{\log(\delta^{-1})}{s}}\right) \geq 1-\delta. \tag{5.26}$$

Eventually, let us note that all the concentration inequalities proved will be used in AdaR-UCB with the choice of $\delta = t^{-4} \in (0,1]$ for all t.

We are now ready to prove the upper bounds on regret for AdaR-UCB under Assumption 3.

5.5. Proof of AdaR-UCB Upper Bound on the Regret

The proof of Theorem 5.3 follows similar steps to the result provided by [22] concerning the upper bound on regret for RobustUCB. Nevertheless, being adaptive w.r.t. both ϵ and u brings additional difficulties in constructing the algorithm and proving its theoretical guarantees.

Proof. We first introduce

$$v := \left\lceil 160\frac{(40u)^{\frac{1}{\epsilon}}}{\Delta_i^{\frac{1+\epsilon}{\epsilon}}}\log T\right\rceil \tag{5.27}$$

and define

$$B_{i,N_i(t-1),t} = \widehat{\mu}_{i,N_i(t-1),t} + 2\sqrt{\frac{V_{N_i(t-1)}(\mathbf{Y})\log(t^4)}{N_i(t-1)}} + 19\frac{M_{N_i(t-1)}\log(t^4)}{N_i(t-1)}$$

where, to not overwhelm the notation, we will use $V_{N_i(t-1)}(\mathbf{Y})$ for $V_{i,N_i(t-1),t}(\mathbf{Y_{i,t}})$ and $M_{N_i(t-1)}$ for $M_{i,N_i(t-1),t}$.

For what follows, we will assume to reason under the assumption that event ξ_t, described in Equation (5.23), holds. This allows to state that:

$$M_{N_i(t-1)} \leq \left(\frac{uN_i(t-1)}{\log(t^4)}\right)^{\frac{1}{1+\epsilon}} \tag{5.28}$$

is true.

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

$$\text{either } B_{i^*,N_{i^*}(t-1),t} \leq \mu^*, \tag{5.29}$$

$$\text{or} \quad \widehat{\mu}_{i,N_i(t-1),t} > \mu_i + 11u^{\frac{1}{1+\epsilon}} \left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}}, \tag{5.30}$$

$$\text{or } N_i(t-1) < 160\frac{(40u)^{\frac{1}{\varepsilon}}}{\Delta_i^{\frac{1+\varepsilon}{\epsilon}}}\log t, \tag{5.31}$$

$$\text{or } \sqrt{V_{N_i(t-1)}(\mathbf{Y})} > \sqrt{\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y})]} + 4M_{N_i(t-1)}\sqrt{\frac{\log(t^4)}{N_i(t-1)}}. \tag{5.32}$$

Indeed, assume that all four inequalities are false.

$$\begin{aligned}
B_{i^*,T_{i^*}(t-1),t} &\overset{(5.29)}{>} \mu^* = \mu_i + \Delta_i \\
&\overset{(5.30)}{\geq} \widehat{\mu}_{i,N_i(t-1),t} - 11u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + \Delta_i \\
&\overset{(*)}{\geq} \widehat{\mu}_{i,N_i(t-1),t} + 2\sqrt{\frac{V_{i,N_i(t-1),t}(\mathbf{Y_{i,t}})\log(t^4)}{N_i(t-1)}} + 19\frac{M_{i,N_i(t-1),t}\log(t^4)}{N_i(t-1)} \\
&= B_{i,N_i(t-1),t}
\end{aligned}$$

The step marked with $(*)$ is a consequence of the fact that both (5.31) and (5.32) are false.

In particular, we need to show that

$$\Delta_i \geq 11u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{V_{N_i(t-1)}(\mathbf{Y})\log(t^4)}{N_i(t-1)}} + 19\frac{M_{N_i(t-1)}\log(t^4)}{N_i(t-1)}. \tag{$*$}$$

We will resort also to the validity of the following inequality:

$$\begin{aligned}
\mathbb{E}\left[V_{N_i(t-1)}(\mathbf{Y})\right] &= \sigma^2_{N_i(t-1)}(\mathbf{Y}) = \sigma^2_{N_i(t-1)}\left(\left(Y_1,\ldots,Y_{N_i(t-1)}\right)\right) = \\
&= \frac{1}{N_i(t-1)[N_i(t-1)-1]}\sum_{l,j=1}^{N_i(t-1)}\frac{\mathbb{E}\left[(Y_l-Y_j)^2\right]}{2} \leq uM^{1-\epsilon}_{N_i(t-1)},
\end{aligned} \tag{5.33}$$

since

$$\begin{aligned}
\sigma^2(Y_t) &= \mathbb{E}\left[Y_t^2\right] - \mathbb{E}\left[Y_t\right]^2 \leq \mathbb{E}\left(X_t^2\mathbf{1}_{|X_t|\leq M_t}\right) \\
&= \mathbb{E}\left(|X_t|^{1+\epsilon}|X_t|^{1-\epsilon}\mathbf{1}_{|X_t|\leq M_t}\right) \leq \mathbb{E}\left(|X_t|^{1+\epsilon}M_t^{1-\varepsilon}\right) \leq uM_t^{1-\epsilon} \quad \forall t.
\end{aligned}$$

Now, we make use of the fact that Equations (5.31) and (5.32) are false, together with

inequalities (5.28) and (5.33), to show that:

$$N_i(t-1) \overset{(5.31)}{\geq} 160\frac{(40u)^{\frac{1}{\epsilon}}}{\Delta_i^{\frac{1+\epsilon}{\epsilon}}}\log t \implies \Delta_i^{\frac{1+\epsilon}{\epsilon}} \geq 160\frac{(40u)^{\frac{1}{\epsilon}}}{N_i(t-1)}\log t = 40\frac{(40u)^{\frac{1}{\epsilon}}}{N_i(t-1)}\log\left(t^4\right).$$

Exploiting the following inequality:

$$40^{\frac{\epsilon}{1+\epsilon}}40^{\frac{1}{1+\epsilon}} \geq 40, \text{ for all } \epsilon \in (0,1],$$

we get:

$$\begin{aligned}
\implies \Delta_i &\geq 40u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} \\
&= (11+2+27)u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} \\
&= 11u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\left(\frac{\log(t^4)u\left(\frac{uN_i(t-1)}{\log(t^4)}\right)^{\frac{1-\epsilon}{1+\epsilon}}}{N_i(t-1)}\right)^{\frac{1}{2}} + \\
&\qquad + 27\frac{\left(\frac{uN_i(t-1)}{\log(t^4)}\right)^{\frac{1}{1+\epsilon}}\log(t^4)}{N_i(t-1)} \\
&\overset{(5.28)}{\geq} 11u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\left(\frac{\log(t^4)uM_{N_i(t-1)}^{1-\epsilon}}{N_i(t-1)}\right)^{\frac{1}{2}} + 27\frac{M_{N_i(t-1)}\log(t^4)}{N_i(t-1)} \\
&\overset{(5.33)}{\geq} 11u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y_i})]\log(t^4)}{N_i(t-1)}} + 27\frac{M_{N_i(t-1)}\log(t^4)}{N_i(t-1)} \\
&= 11u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{\log(t^4)}{N_i(t-1)}}\left[\sqrt{\mathbb{E}[V_{N_i(t-1)}(\mathbf{Y})]}+\right. \\
&\qquad \left. + 4M_{N_i(t-1)}\sqrt{\frac{\log(t^4)}{N_i(t-1)}}\right] + 19\frac{M_{N_i(t-1)}\log(t^4)}{N_i(t-1)} \\
&\overset{(5.32)}{\geq} 11u^{\frac{1}{1+\epsilon}}\left[\frac{\log(t^4)}{N_i(t-1)}\right]^{\frac{\epsilon}{1+\epsilon}} + 2\sqrt{\frac{V_{N_i(t-1)}(\mathbf{Y}_i)\log(t^4)}{N_i(t-1)}} + 19\frac{M_{N_i(t-1)}\log(t^4)}{N_i(t-1)},
\end{aligned}$$

which proves Equation (∗).

Thus, we obtain the following contradiction:

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

implying, in particular, that $I_t \neq i$.

Now we bound the probability that ξ_t holds and at least one of Equations (5.29), (5.30) or (5.32) is true.

By (5.22), (5.24) and (5.26), as well as an union bound over the value of $N_{i^*}(t-1)$ and $N_i(t-1)$, we obtain:

$$\mathbb{P}\left([\text{(5.29) or (5.30) or (5.32) is true}] \text{ and } \xi_t\right) \leq 4\sum_{s=1}^{t}\frac{1}{t^4} = \frac{4}{t^3}$$

At this point, using v as defined in Equation (5.27), we obtain:

$$\begin{aligned}
\mathbb{E}[N_i(T)] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{I_t=i}\right] &\leq v + \mathbb{E}\left[\sum_{t=v+1}^{T}\mathbb{1}_{\{I_t=i \text{ and (5.31) is false }\}}\right] \\
&= v + \mathbb{E}\Bigg[\sum_{t=v+1}^{T}\Big(\mathbb{1}_{\{I_t=i \text{ and (5.31) is false and } \xi_t\}} + \\
&\qquad + \mathbb{1}_{\{I_t=i \text{ and (5.31) is false and } \xi_t^C\}}\Big)\Bigg] \\
&\leq v + \mathbb{E}\Bigg[\sum_{t=v+1}^{T}\Big(\mathbb{1}_{\{I_t=i \text{ and [(5.29) or (5.30) or (5.32) is true] and } \xi_t\}} + \\
&\qquad + \mathbb{1}_{\{\xi_t^C\}}\Big)\Bigg] \\
&\leq v + \sum_{t=v+1}^{T}\left(\frac{4}{t^3} + \frac{1}{t^4}\right) \\
&\leq v + 6,
\end{aligned}$$

where ξ_t^C defines the complementary set of ξ_t.

This proves that

$$\mathbb{E}[N_i(T)] \leq 160\frac{(40u)^{\frac{1}{\varepsilon}}}{\Delta_i^{\frac{1+\varepsilon}{\epsilon}}}\log T + 7, \tag{5.34}$$

for any i not optimal arm.

Using that $R_T = \sum_{i=1}^{K}\Delta_i\mathbb{E}[N_i(T)]$ and (5.34), we directly obtain an *instance-dependent*

bound on the regret which proves Theorem 5.3:

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

■

On the other hand, for the sake of completeness, we can also prove Theorem 5.4 to show an *instance-independent bound* on the regret, assuming T is such that:

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

Proof. We have:

$$\begin{aligned}
R_T &= \sum_{i:\Delta_i>0} \Delta_i \left(\mathbb{E}[N_i(T)]\right)^{\frac{\epsilon}{1+\epsilon}} \left(\mathbb{E}[N_i(T)]\right)^{\frac{1}{1+\varepsilon}} \\
&\leq \sum_{i:\Delta_i>0} \Delta_i \left(\mathbb{E}[N_i(T)]\right)^{\frac{1}{1+\epsilon}} \left(160 \frac{(40u)^{\frac{1}{\epsilon}}}{\Delta_i^{\frac{1+\epsilon}{\epsilon}}} \log T + 7 \right)^{\frac{\epsilon}{1+\epsilon}} \\
&\overset{(i)}{\leq} \sum_{i:\Delta_i>0} \Delta_i \left(\mathbb{E}[N_i(T)]\right)^{\frac{1}{1+\epsilon}} \left(320 \frac{(40u)^{\frac{1}{\epsilon}}}{\Delta_i^{\frac{1+\epsilon}{\epsilon}}} \log T \right)^{\frac{\epsilon}{1+\epsilon}} \\
&= \left[\sum_{i:\Delta_i>0} \left(\mathbb{E}[N_i(T)]\right)^{\frac{1}{1+\epsilon}} \right] \left(320(40u)^{\frac{1}{\epsilon}} \log T \right)^{\frac{\epsilon}{1+\epsilon}} \\
&\overset{(ii)}{\leq} K^{\frac{\epsilon}{1+\epsilon}} \left(\sum_{i:\Delta_i>0} \mathbb{E}[N_i(T)] \right)^{\frac{1}{1+\epsilon}} 320^{\frac{\epsilon}{1+\epsilon}} (40u)^{\frac{1}{1+\epsilon}} (\log T)^{\frac{\epsilon}{1+\epsilon}} \\
&\leq T^{\frac{1}{1+\epsilon}} (320K \log T)^{\frac{\epsilon}{1+\epsilon}} (40u)^{\frac{1}{1+\epsilon}},
\end{aligned}$$

which proves Equation (5.20).

Step (i) above is given by assumption on T, while step (ii) is an application of Hölder's inequality in Appendix A. More specifically, we chose:

$$p = \frac{1+\varepsilon}{\varepsilon},\ q = 1+\varepsilon,\ x_i = 1 \ \text{ and } y_i = \left(\mathbb{E}[N_i(T)]\right)^{\frac{1}{1+\varepsilon}} \ \text{ for all } i = 1, \ldots, K$$

to then apply Equation (A.3). ■

6 | Numerical Simulations

In this chapter, we present numerical results to illustrate the performance of `AdaR-UCB` algorithm presented in Chapter 5. We empirically validate the theoretical improvements achieved in using our proposed algorithm based on a fully adaptive approach, by comparing it with some state of the art regret minimization algorithms, i.e., `UCB1` and `RobustUCB`.

We start outlining below the theory behind the reward distributions of classic benchmark heavy-tailed problems that we adopt to compare the aforementioned algorithms. Afterwards, we report the full results of the experiments.

6.1. Experimental Setting: Pareto Distributions

To make an experimental validation of the theoretical novelties introduced in the previous chapters, we use *Generalized Pareto distributed rewards*, which represent the prototype of distributions to adopt in a standard heavy-tailed modeling framework.

A Pareto-distributed random variable X [77] is characterized by a *scale parameter* x_m and a *shape parameter* α, which is known as the tail index; the first one is the (necessarily positive) minimum possible value of the support of X, and the second one is a positive parameter characterizing the heaviness of the distribution's tail.

The Probability Density Function (PDF) of X is given by:

$$f_X(x) = \begin{cases} \frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}}, & x \geq x_{\mathrm{m}} \\ 0, & x < x_{\mathrm{m}} \end{cases}.$$

When plotted on canonical axes, the distribution assumes the familiar J-shaped curve which approaches the horizontal axis asymptotically. The Cumulative Distribution Function (CDF) of X can be consequently computed as:

$$F_X(x) = 1 - \mathbb{P}(X > x) = 1 - \int_x^{+\infty} f_X(t)\, dt = 1 - \left(\frac{x_m}{x}\right)^{\alpha}, \text{ for } x \geq x_m.$$

In particular, we have that:

$$\mathbb{P}(X > x) \overset{x\to\infty}{\sim} x^{-\alpha}, \alpha > 0,$$

for large x. Thus, tail index α gives an inference about the orders of finite moments.

Properties on the moments:

- The expected value of a random variable X following a Pareto distribution is:

$$\mathrm{E}[X] = \begin{cases} \infty & \alpha \leq 1 \\ \frac{\alpha x_{\mathrm{m}}}{\alpha - 1} & \alpha > 1 \end{cases}$$

- The variance of a random variable X following a Pareto distribution is:

$$\mathrm{Var}(X) = \begin{cases} +\infty & \alpha \in (1, 2] \\ \left(\frac{x_{\mathrm{m}}}{\alpha-1}\right)^2 \frac{\alpha}{\alpha-2} & \alpha > 2 \end{cases}.$$

  If $\alpha \leq 1$, the variance does not exist.

- Let us assume $X \sim Pareto(\alpha, x_m)$, with shape parameter $\alpha \in (1, 2]$ such that the second moment is infinite, and $x_m > 0$. We now compute the *non-centered moment* of order $1 + \epsilon$, with $\epsilon \in (0, 1]$:

$$\begin{aligned} \mathbb{E}[|X|^{1+\epsilon}] &= \int_{x_m}^{+\infty} |x|^{1+\epsilon} f_X(x)\, dx \\ &= \int_{x_m}^{+\infty} x^{1+\epsilon} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}\, dx \\ &= \alpha x_m^\alpha \int_{x_m}^{+\infty} \frac{1}{x^{\alpha-\epsilon}}, \end{aligned}$$

  which converges when $\alpha - \epsilon > 1 \Leftrightarrow \alpha > 1 + \epsilon$.

To recap the final point, we can state the following:

Proposition 6.1. *A Pareto distributed random variable X with shape parameter α and scale parameter x_m has finite moments of order $1 + \epsilon$ with $\epsilon \in (0, 1]$ if and only if the shape parameter α is greater than $1 + \epsilon$.*

Moreover, if $\alpha > 1 + \epsilon$, we can also compute analytically the value of the moment of order

$1+\epsilon$, since the integral can be computed in closed form:

$$\begin{aligned}\mathbb{E}[|X|^{1+\epsilon}] &= \alpha x_m^\alpha \int_{x_m}^{+\infty} \frac{1}{x^{\alpha-\epsilon}} \\ &= \frac{\alpha x_m^\alpha}{\epsilon-\alpha+1}\left[x^{-(\alpha-(1+\epsilon))}\right]_{x_m}^{+\infty} \\ &= \frac{\alpha x_m^\alpha}{\epsilon-\alpha+1}(-x_m^{\epsilon-\alpha+1}) \\ &= \frac{\alpha x_m^{1+\epsilon}}{\alpha-(1+\epsilon)}.\end{aligned} \tag{6.1}$$

To understand the possible theoretical differencies, let us now compute the *centered moment* of order $1+\epsilon$, with $\epsilon \in (0,1]$. Given $\mu = \mathbb{E}[X]$:

$$\begin{aligned}\mathbb{E}[|X-\mu|^{1+\epsilon}] &= \int_{x_m}^{+\infty} |x-\mu|^{1+\epsilon} f_X(x)\, dx \\ &= \alpha x_m^\alpha \int_{x_m}^{+\infty} (x-\mu)^{1+\epsilon} \frac{1}{x^{\alpha+1}}\, dx.\end{aligned}$$

The last integral cannot be solved analytically, but we have that, for $x \to \infty$,

$$\int_{x_m}^{+\infty} \frac{(x-\mu)^{1+\epsilon}}{x^{\alpha+1}}\, dx \sim \int_{x_m}^{+\infty} \frac{x^{1+\epsilon}}{x^{\alpha+1}}\, dx,$$

which convergence properties are analogous to the ones of non-centered moments.

In conclusion, for a given α and ϵ, the non-centered moment of order $1+\epsilon$ exists if and only if the centered one exists, but the expression of the latter cannot be computed analytically, and in general the two do not coincide.

To test the performance of `AdaR-UCB` algorithm, we will consider bandit instances with arms Pareto-distributed, introducing some generalizations to allow more flexibility in the chosen instances. In particular, we will consider rewards distributed as:

- *Positive Pareto*: heavy-tailed Pareto distribution with standard parameters $\alpha \in (1,2]$ and $x_m > 0$, allowing only for a positive tail on the support $S = [x_m, +\infty)$.
- *Negative Pareto*: heavy-tailed Pareto distribution which has a symmetric probability density function with respect to the Positive one. It has parameters $\alpha \in (1,2]$ and $x_m > 0$, with only a negative tail on the support $S = (-\infty, -x_m]$.
- *Double-tailed Pareto*: a random variable which represents the event of sampling with probability $1/2$ from a Positive Pareto, $X_1 \sim Pareto(\alpha_1, x_{m1})$, and with probability

1/2 from a negative one, $X_2 \sim Pareto(\alpha_2, x_{m2})$. It has support $(-\infty, -x_{m2}] \cup [x_{m1}, +\infty)$ and probability density function:

$$f_X(x) = \frac{1}{2} f_{X_1}(x) + \frac{1}{2} f_{X_2}(x),$$

where $f_{X_1}(x)$ and $f_{X_2}(x)$ are the probability density functions of X_1 and X_2 respectively.

6.2. Truncated Non-Positive Bandit Instances

We run some experiments on bandits instances which satisfy Assumption 3, i.e. have a *truncated non-positive optimal arm*. We will compare the results of our new Algorithm 5.1 with respect to the other two standard algorithms for multi-armed bandits problems, `UCB1` (Algorithm 3.1) and `RobustUCB` (Algorithm 3.2).

For the instances that follow, we retrieve the result of Proposition 6.1 to compute the value of ϵ given that $1+\epsilon$ is the number just below the minimum of the shape parameters of the distributions. In this way, it indeed represents the maximum order of finite moments. The uniform bound u is the maximum, among all the distributions, of the $(1+\epsilon)$-th moments of each one. It is computed analytically as in Equation (6.1).

We first consider cases where all the arms follow Negative Pareto distributions, including the optimal arm, giving Assumption 3 to be trivially satisfied.

6.2.1. Simulation 1

For the *first simulation*, the four distributions are displayed in Figure 6.1. The optimal arm is the first one (arm 0), since it has the highest mean.

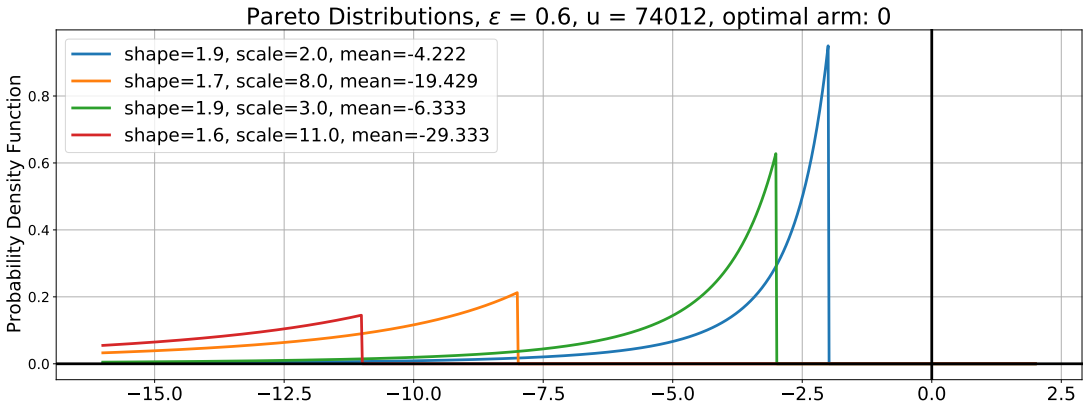


Figure 6.1: Simulation 1 - *Instance a* - Pareto Distributions.

The performances of our three reference algorithms on the instance described are reported in Figure 6.2, where the cumulative regret is computed considering a time horizon $T = 25000$. Moreover, 20 runs are performed, such that we can plot univariate confidence intervals for the regret mean estimates, $\left[\bar{R}_T - \frac{\hat{\sigma}}{\sqrt{20}}, \bar{R}_T + \frac{\hat{\sigma}}{\sqrt{20}}\right]$, where the uncertainty is characterized through the sample standard deviation $\hat{\sigma}$. The intervals are plotted as shaded areas with their upper and lower bounds.

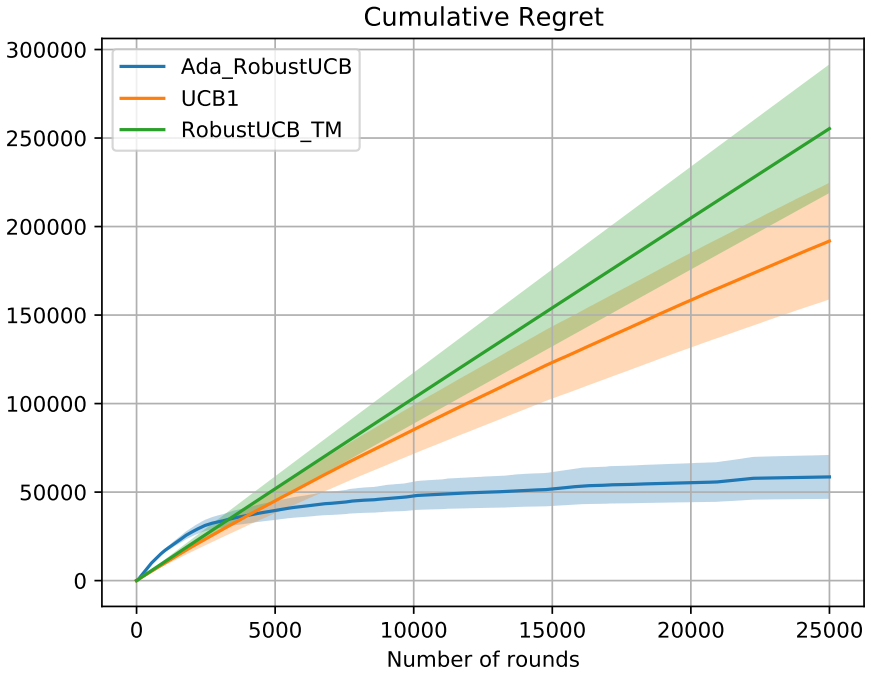


Figure 6.2: Numerical results of simulation 1 - *Instance a* - Baseline comparison of cumulative regret, which is averaged over 20 independent experiments (20 runs, shaded areas are standard deviations.

Comparing the results in Figure 6.2 we see that `AdaR-UCB` performs way better than `UCB1` and `RobustUCB`, with a clearly sub-linear regret that tends to flatten fast. On the other side, `UCB1` and `RobustUCB` algorithms show a regret behaviour only slightly sub-linear, almost linear, with a slower convergence with respect to our algorithm. In particular, applying `UCB1` to a stochastic bandit setting with heavy tails, we do not have any theoretical guarantee on the upper bound on regret, such that empirically a priori we can not expect any specific behaviour.

Ideally, since `UCB1` uses the sample mean estimator that is not robust to extreme outliers, it should perform worse than any robust algorithm suited for heavy-tailed instances. This is not the case in this experiment, and we will show that the empirical evidence of `UCB1` performing better than `RobustUCB` will be recurrent also in the next simulations. The

reason behind this unexpected result lies in the threshold chosen for the truncated mean estimator. Even if we have good convergence result for this robust estimator, the threshold reported in Equation (5.1) is too high with respect to the means of the Pareto, such that it basically never truncates rewards and it takes too long to tighten the upper confidence bound interval. Even if we have theoretical results that guarantee a logarithmic regret for `RobustUCB`, empirically this algorithm is too slow to converge. We will show in the next experiments that this is a frequent pattern, with `RobustUCB` consistently showing poor performances.

To better understand the behaviour of `AdaR-UCB` algorithm, we reported in Figure 6.3 the choice of actions that it performed throughout all the 20 runs (upper plot) and the specific choice of actions performed along time in the last epoch (lower plot).

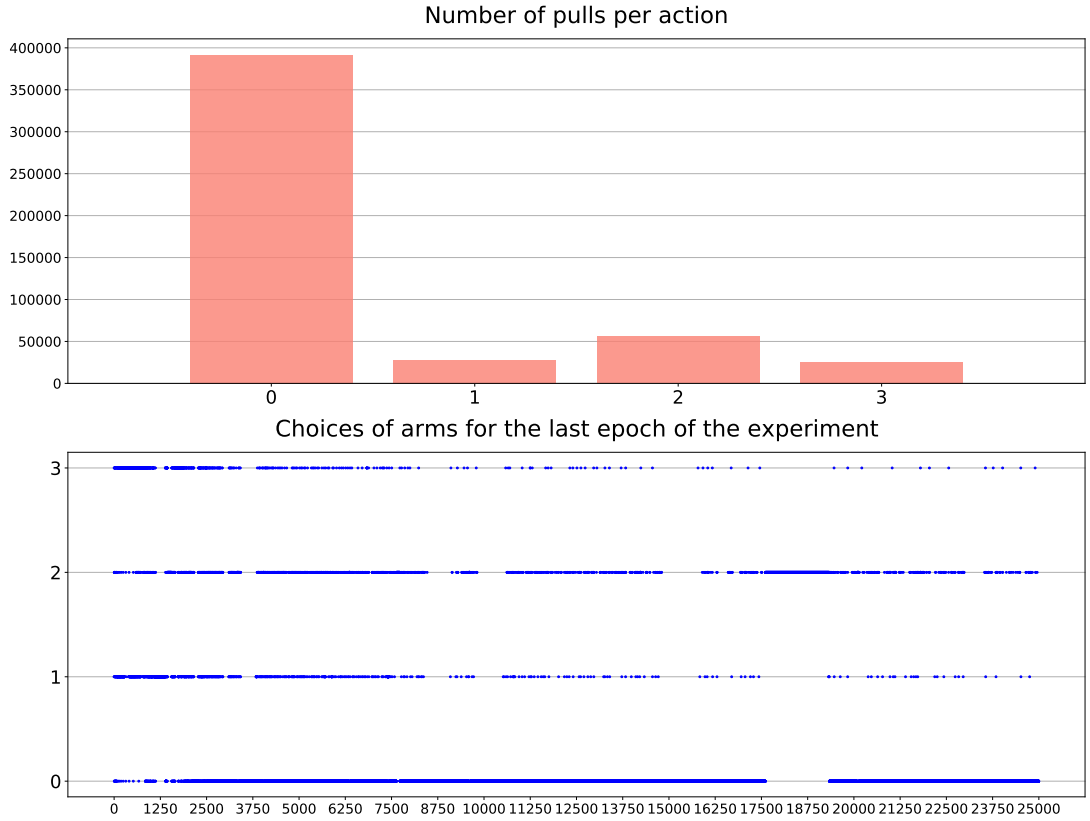


Figure 6.3: Simulation 1 - *Instance a* - Arms pulled by Algorithm `AdaR-UCB`.

We see that, according to regret results, the optimal arm is the most played, but the algorithm keeps anyway exploring occasionally along the time horizon. Thus, `AdaR-UCB` more efficiently explores an optimal action than `RobustUCB` despite of the weakness of the assumptions.

Moreover, doing the last investigations for this instance, we can assess another recurrent

behaviour that repeats along all the possible experiments. The cumulative regret computed by `AdaR-UCB` is in all the cases well below the bounds presented in Theorems 5.3 and 5.4. This shows that regret bounds proved are fairly loose, but this happens usually in literature for all the stochastic bandits algorithms based on UCB approaches [3].

We want now to study how the performances of the algorithms change with respect to parameter u. Taking two bandit instances and keeping the same type of distributions, same ϵ and number of arms, we want to test how the different *scale parameters* influence the behaviour of cumulative regret. Since in the first instance presented we had Negative Pareto's with scale x_m between 2 and 11 (Figure 6.1), we now try an experiment with the same one, but changing the scales to a wider range, e.g. between 40 and 150. The new distributions are displayed in Figure 6.4.

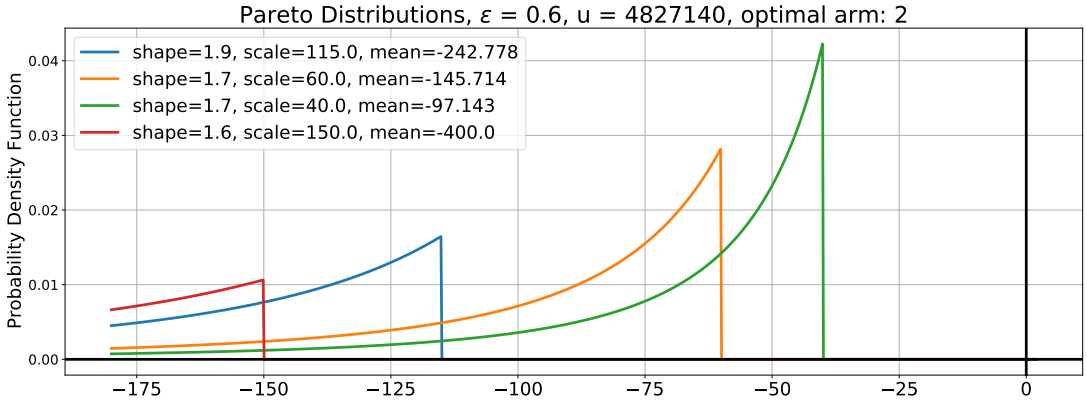


Figure 6.4: Simulation 1 - *Instance b* - Pareto Distributions.

For each algorithm, the mean estimates of cumulative regret over 20 independent runs and along an horizon of $T = 25000$ time instants are reported in Figure 6.5.

Comparing Figures 6.2 and 6.5 in terms of regret trends, we do not see a relevant difference between the two. In both the instances, `AdaR-UCB`, with a clear sub-linear regret behaviour, still outperforms the other two algorithms, that show only a slightly sub-linear behaviour. To compare the absolute values of regret at horizon T, we can refer to Table 6.1.

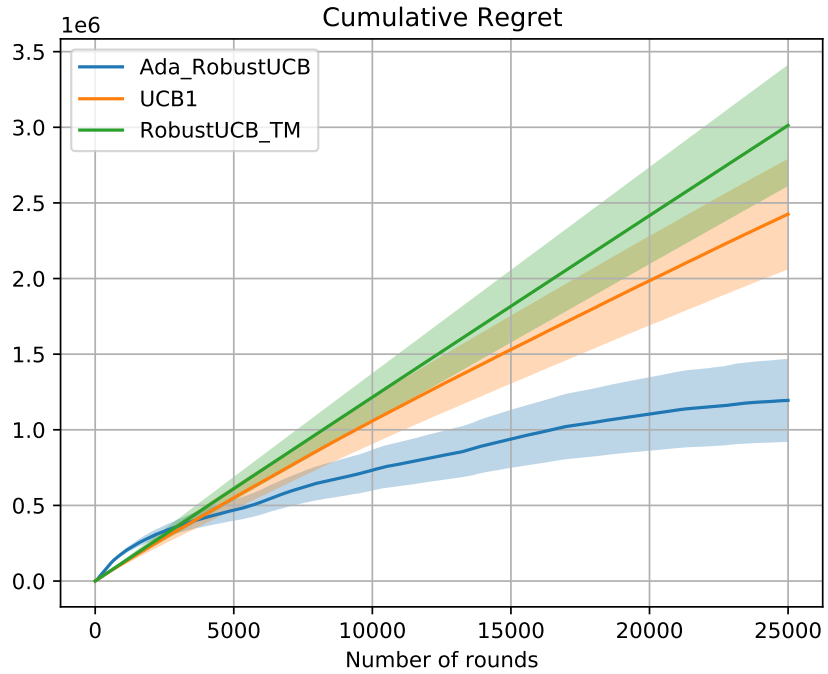


Figure 6.5: Numerical results of simulation 1 - *Instance b* - Baseline comparison of cumulative regret (20 runs, shaded areas are standard deviations).

	Regret R_T	
Algorithm	**Instance a, $u = 7 \cdot 10^4$**	**Instance b, $u = 5 \cdot 10^6$**
`AdaR-UCB`	$6 \cdot 10^4$	$12 \cdot 10^5$
`UCB1`	$19 \cdot 10^4$	$24 \cdot 10^5$
`RobustUCB`	$25 \cdot 10^4$	$30 \cdot 10^5$

Table 6.1: Simulation 1 - Performance comparison of same instances with different u

As expected, we see that, given the same algorithm, cumulative regret increases while increasing u and the suboptimality gaps. On the other hand, with simple computations, we see that the relative behaviour of cumulative regrets of different algorithms in the same bandit instance is not strongly affected from how wider is the difference in scale among Pareto distributions. Considering, for instance, the relative difference in regret between `RobustUCB` and `AdaR-UCB`, and between `RobustUCB` and `UCB1`, we get:

$$\begin{cases} \textbf{Instance a}: \frac{R_T(\texttt{RobustUCB}) - R_T(\texttt{AdaR-UCB})}{R_T(\texttt{RobustUCB})} \approx 0.75; \\ \textbf{Instance b}: \frac{R_T(\texttt{RobustUCB}) - R_T(\texttt{AdaR-UCB})}{R_T(\texttt{RobustUCB})} \approx 0.6. \end{cases}$$

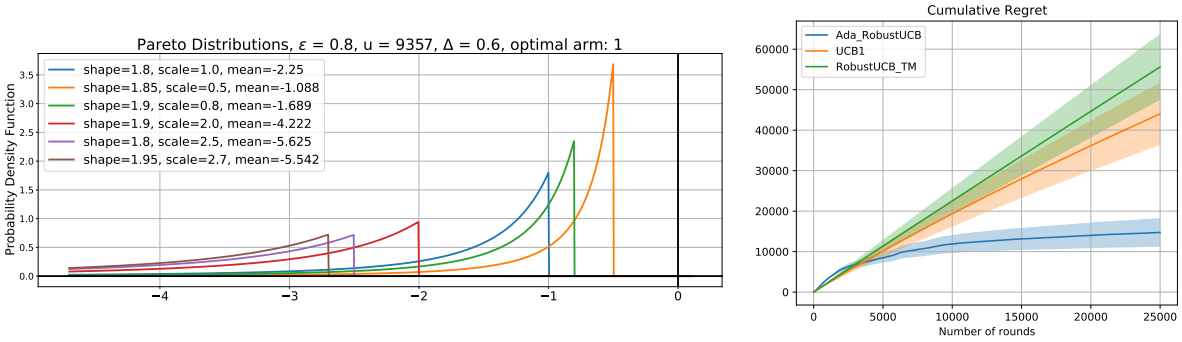
$$\begin{cases} \textbf{Instance a}: \frac{R_T(\texttt{RobustUCB}) - R_T(\texttt{UCB1})}{R_T(\texttt{RobustUCB})} \approx 0.24; \\ \textbf{Instance b}: \frac{R_T(\texttt{RobustUCB}) - R_T(\texttt{UCB1})}{R_T(\texttt{RobustUCB})} \approx 0.2. \end{cases}$$

We can conclude that changing u between same types of instances does not lead to relevant differences in relative performances.

6.2.2. Simulation 2

We want then to study how the performances of the algorithms change with respect to the other key parameter in our stochastic heavy-tailed bandit setting, namely ϵ. For the next analyses, given a bandit instance, we need first to define the parameter Δ as the difference between the mean of the optimal arm and the mean of the second optimal one.

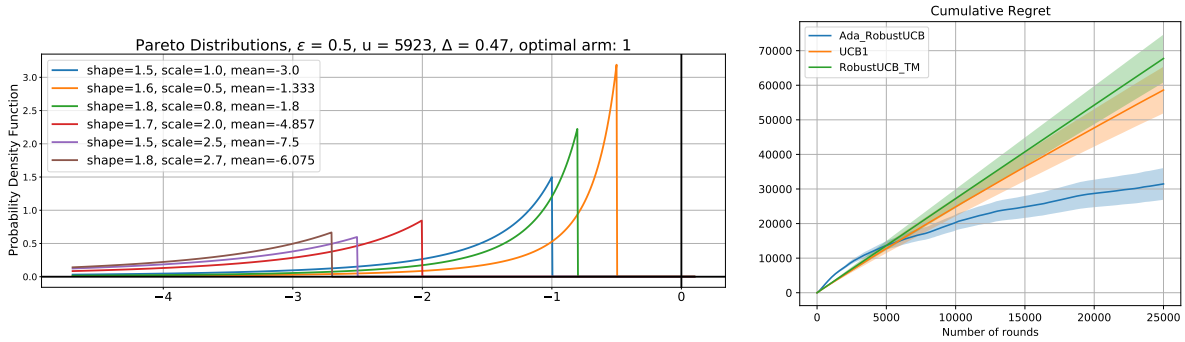
Since the influence of u on the relative performances is negligible, without lack of generality, we can now compare instances with same number of arms, $\Delta \in (0, 1)$ and different $\epsilon \in (0, 1]$.



(a) Pareto Distributions.

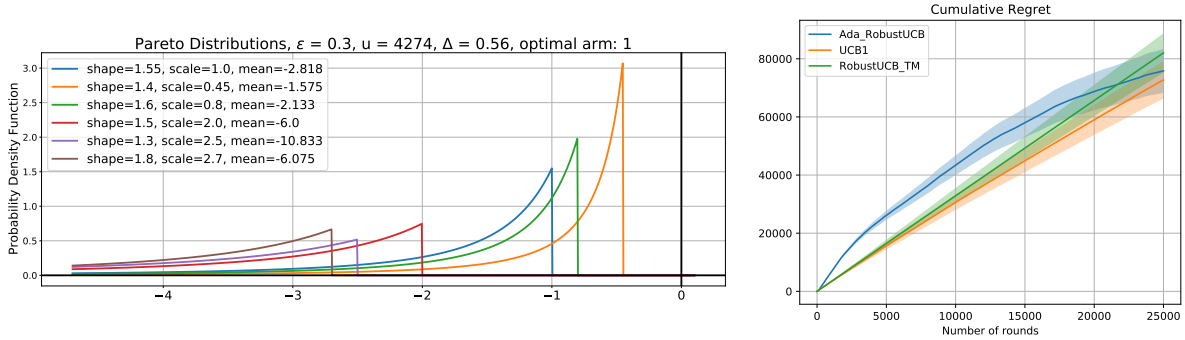
(b) Numerical results (20 runs).

Figure 6.6: Simulation 2 - *Instance a* - Negative Pareto's, 6 arms, $\epsilon = \mathbf{0.8}, \Delta \in (0, 1)$.



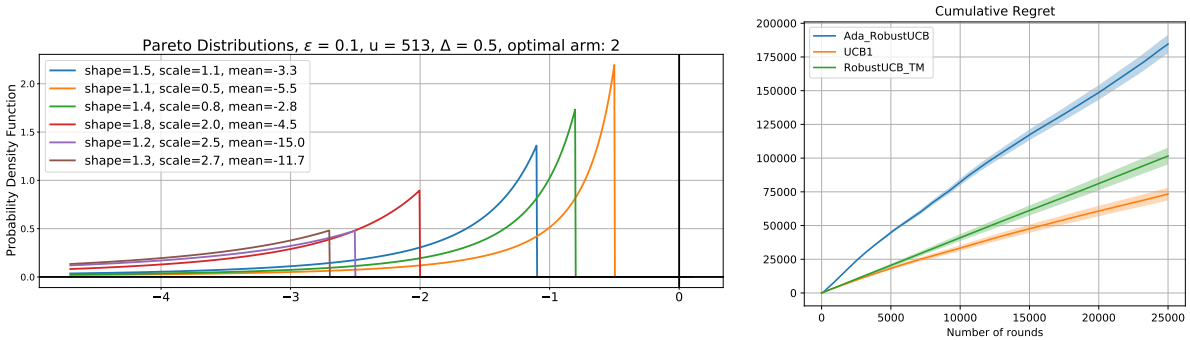
(a) Pareto Distributions. (b) Numerical results (20 runs).

Figure 6.7: Simulation 2 - *Instance b* - Negative Pareto's, 6 arms, $\epsilon = \mathbf{0.5}, \Delta \in (0, 1)$.



(a) Pareto Distributions. (b) Numerical results (20 runs).

Figure 6.8: Simulation 2 - *Instance c* - Negative Pareto's, 6 arms, $\epsilon = \mathbf{0.3}, \Delta \in (0, 1)$.



(a) Pareto Distributions. (b) Numerical results (20 runs).

Figure 6.9: Simulation 2 - *Instance d* - Negative Pareto's, 6 arms, $\epsilon = \mathbf{0.1}, \Delta \in (0, 1)$.

In Table 6.2, we report a summary of regret results at horizon $T = 25000$ for every algorithm, while changing ϵ parameter in each instance.

	Regret R_T			
Algorithm	$\epsilon = 0.8$	$\epsilon = 0.5$	$\epsilon = 0.3$	$\epsilon = 0.1$
`AdaR-UCB`	$15 \cdot 10^3$	$31 \cdot 10^3$	$75 \cdot 10^3$	$180 \cdot 10^3$
`UCB1`	$45 \cdot 10^3$	$58 \cdot 10^3$	$70 \cdot 10^3$	$75 \cdot 10^3$
`RobustUCB`	$55 \cdot 10^3$	$68 \cdot 10^3$	$80 \cdot 10^3$	$100 \cdot 10^3$

Table 6.2: Comparison on the regret performances of the algorithms, given the same type of instances but with different ϵ.

With simple computations, we see that the relative performances between the algorithms change significantly depending on ϵ.

Remark 7. *At a first sight, Figure 6.9b might be misleading. The cumulative regret of* `AdaR-UCB` *seems to have an almost linear behaviour, disagreeing with the theoretical results we showed. Keeping the same Pareto's as in Figure 6.9a, but simply increasing the time horizon to* $T = 100000$*, we see in Figure 6.10 that algorithm* `AdaR-UCB` *coherently shows a sub-linear empirical behaviour, in accordance with the theory.*

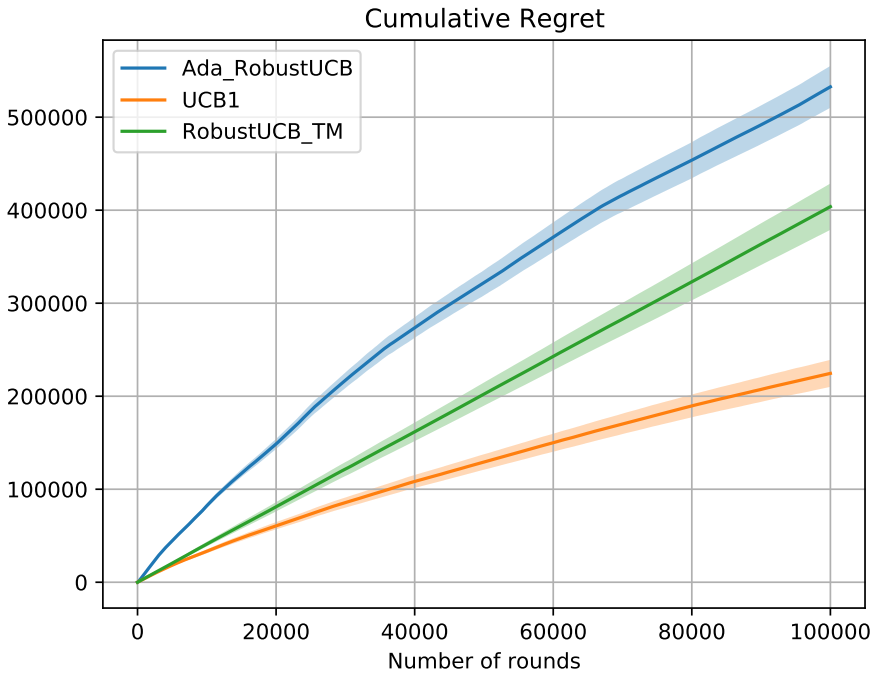


Figure 6.10: Simulation 2 - *Instance d* - Numerical results on regret comparison (20 runs), T = 100000.

From the experiments on these instances we can make a relevant conclusion on the empirical performance of `AdaR-UCB` algorithm: decreasing ϵ parameter makes the slope of cumulative regret decreasing more slowly.

6.2.3. Simulation 3

Let us now investigate a more complicated setting. We keep the optimal action distributed as a negative Pareto distribution, but we consider as possible non-optimal arms also some double-tailed Pareto distributions. The stochastic instance chosen is reported in Figure 6.11. Here Assumption 3 of truncated non-positivity is trivially satisfied since the optimal arm has a negative tail only.

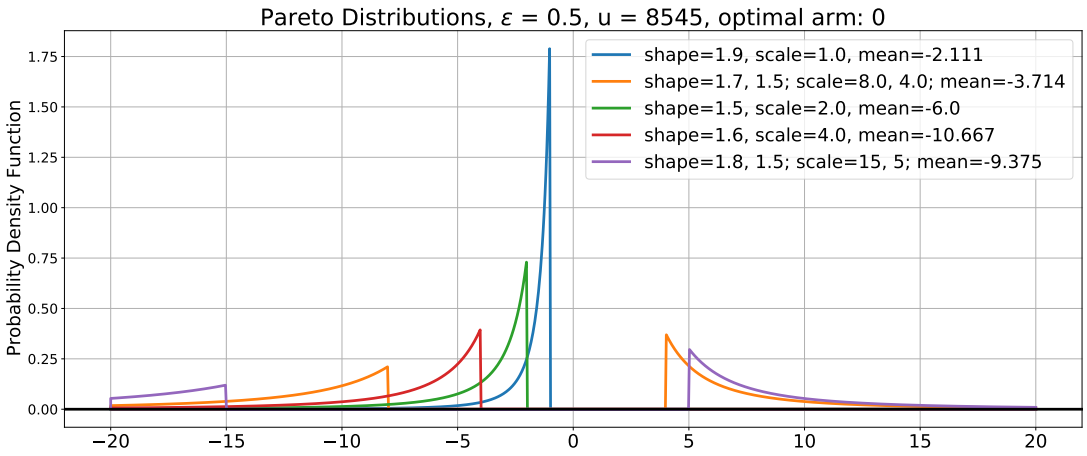


Figure 6.11: Simulation 3 - Pareto Distributions.

Computing the expected regrets over 20 runs and $T = 25000$ time instants, we still see in Figure 6.12 a pattern of regrets similar to the one in Section 6.2.1, with `AdaR-UCB` outperforming the current most common state-of-the-art performances.

6.2.4. Simulation 4

Let us now evaluate empirically the last type of instances we can think of, that still satisfies Assumption 3 of truncated non-positivity. We consider rewards distributed according to negative Pareto's or double-tailed Pareto's, with the optimal arm that follows the latter and observes Assumption 3.

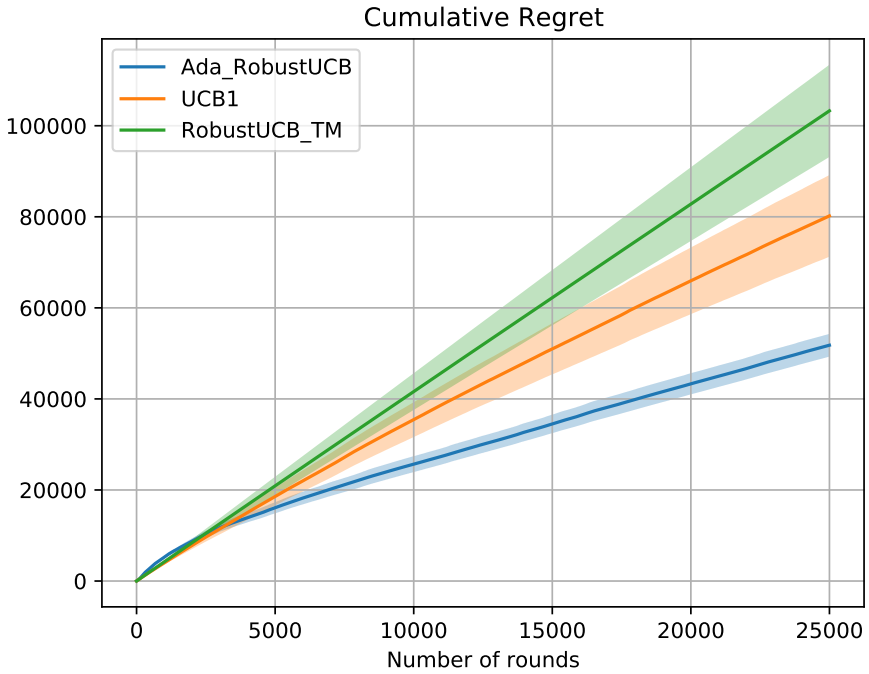


Figure 6.12: Numerical results of simulation 3 - Baseline comparison of cumulative regret (20 runs, shaded areas are standard deviations).

The reader might wonder how to investigate if a double-tailed Pareto random variable X satisfies the *truncated non-positivity assumption*, which we recall here as:

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

Even if, in a Pareto, shape parameter α controls the heaviness of the tails, checking if a double-tailed Pareto satisfies the assumption cannot be done simply comparing the *shape* parameters of the two tails, as one might think straightforward. Since the condition above needs to hold for all $M > 0$, it has to be checked analytically considering both the *shape* and *scale* parameters.

Let us assume X to be a random variable such that with a probability of $\frac{1}{2}$ we sample from a Positive Pareto, $X_1 \sim Pareto(\alpha_1, x_{m1})$, and with probability $\frac{1}{2}$ we sample from a Negative Pareto, $X_2 \sim Pareto(\alpha_2, x_{m2})$.

To check the assumption above we need to show that:

$$\frac{1}{2}\alpha_1 x_{m1}^{\alpha_1} \int_M^{+\infty} \frac{x}{x^{\alpha_1+1}} \mathbb{1}_{x>x_{m1}}\, dx \;\leq\; \frac{1}{2}\alpha_2 x_{m2}^{\alpha_2} \int_M^{+\infty} \frac{x}{x^{\alpha_2+1}} \mathbb{1}_{x>x_{m2}}\, dx \quad \forall M \geq 0. \tag{6.2}$$

Since the support of X_1 is $[x_{m1}, +\infty)$ and the support of X_2 equals to $(-\infty, -x_{m2}]$, the

inequality needs to hold for $M \geq \min(x_{m1}, x_{m2})$. Let us first study the case of $x_{m1} < x_{m2}$. Equation (6.2) becomes:

$$\begin{cases} \alpha_1 x_{m1}^{\alpha_1} \dfrac{M^{1-\alpha_1}}{\alpha_1 - 1} \leq \alpha_2 x_{m2}^{\alpha_2} \dfrac{x_{m2}^{1-\alpha_2}}{\alpha_2 - 1}, & \text{for all } M\text{: } x_{m1} \leq M < x_{m2} \\ \alpha_1 x_{m1}^{\alpha_1} \dfrac{M^{1-\alpha_1}}{\alpha_1 - 1} \leq \alpha_2 x_{m2}^{\alpha_2} \dfrac{M^{1-\alpha_2}}{\alpha_2 - 1}, & \text{for all } M\text{: } x_{m2} \leq M \end{cases}$$

$$\implies \begin{cases} M \geq \left(\dfrac{\alpha_1(\alpha_2 - 1)}{\alpha_2(\alpha_1 - 1)} \dfrac{x_{m1}^{\alpha_1}}{x_{m2}} \right)^{\frac{1}{\alpha_1 - 1}}, & \text{for all M: } x_{m1} \leq M < x_{m2} \\ M \geq \left(\dfrac{\alpha_1(\alpha_2 - 1)}{\alpha_2(\alpha_1 - 1)} \dfrac{x_{m1}^{\alpha_1}}{x_{m2}^{\alpha_2}} \right)^{\frac{1}{\alpha_1 - \alpha_2}}, & \text{for all M: } x_{m2} \leq M \end{cases}. \tag{6.3}$$

While for the case of $x_{m2} < x_{m1}$, Equation (6.2) reads:

$$\begin{cases} \alpha_1 x_{m1}^{\alpha_1} \dfrac{x_{m1}^{1-\alpha_1}}{\alpha_1 - 1} \leq \alpha_2 x_{m2}^{\alpha_2} \dfrac{M^{1-\alpha_2}}{\alpha_2 - 1}, & \text{for all } M\text{: } x_{m2} \leq M < x_{m1} \\ \alpha_1 x_{m1}^{\alpha_1} \dfrac{M^{1-\alpha_1}}{\alpha_1 - 1} \leq \alpha_2 x_{m2}^{\alpha_2} \dfrac{M^{1-\alpha_2}}{\alpha_2 - 1}, & \text{for all } M\text{: } x_{m1} \leq M \end{cases}$$

$$\implies \begin{cases} M \leq \left(\dfrac{\alpha_2(\alpha_1 - 1)}{\alpha_1(\alpha_2 - 1)} \dfrac{x_{m2}^{\alpha_2}}{x_{m1}} \right)^{\frac{1}{\alpha_2 - 1}}, & \text{for all M: } x_{m2} \leq M < x_{m1} \\ M \geq \left(\dfrac{\alpha_1(\alpha_2 - 1)}{\alpha_2(\alpha_1 - 1)} \dfrac{x_{m1}^{\alpha_1}}{x_{m2}^{\alpha_2}} \right)^{\frac{1}{\alpha_1 - \alpha_2}}, & \text{for all M: } x_{m1} \leq M \end{cases}. \tag{6.4}$$

In particular, it is trivial to understand that a double-tailed Pareto random variable X with positive mean for sure does not satisfy the assumption of *truncated non-positivity*. We have, indeed, that $\mathbb{E}_X[X \mathbb{1}_{\{|X|>M\}}] \leq 0$ does not hold for $M = 0$. On the other side, it is not true that the assumption is in general satisfied for every double-tailed Pareto with negative mean, since it has to be checked for all $M \geq 0$.

Let us now consider the instance with rewards of arms distributed as in Figure 6.13.

In particular, the optimal arm is a double-tailed Pareto with the positive tail distributed as $X_1 \sim Pareto(\alpha_1, x_{m1}) = Pareto(1.7, 8)$, and the negative tail as $X_2 \sim Pareto(\alpha_2, x_{m2}) = Pareto(1.3, 10)$. Since $x_{m2} > x_{m1}$, we need to verify Equation (6.3):

$$\begin{cases} M \geq 2.54, & \text{for all M: } 8 \leq M < 10 \\ M \geq 0.10, & \text{for all M: } 10 \leq M \end{cases},$$

so that the truncated non-positivity holds, and our theoretical analysis in Chapter 5 still

ensures a logarithmic regret.

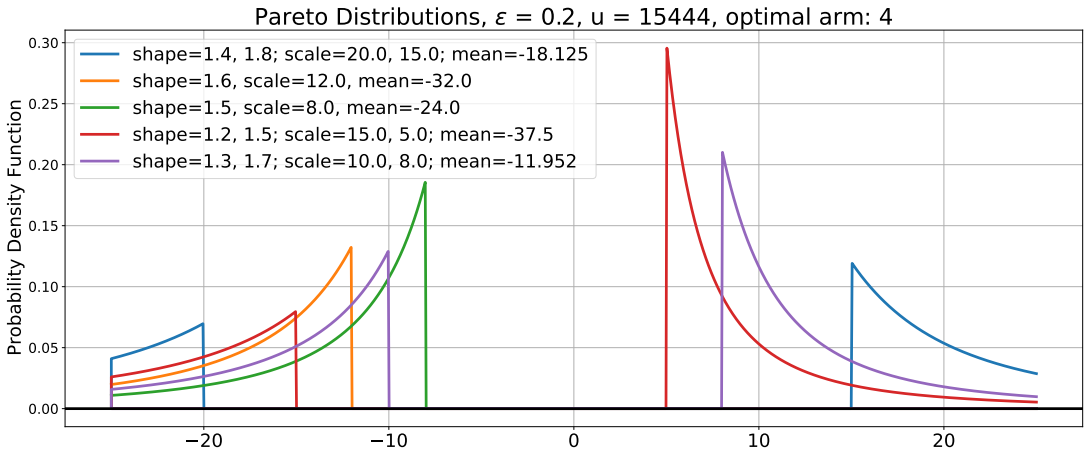


Figure 6.13: Simulation 4 - Pareto Distributions.

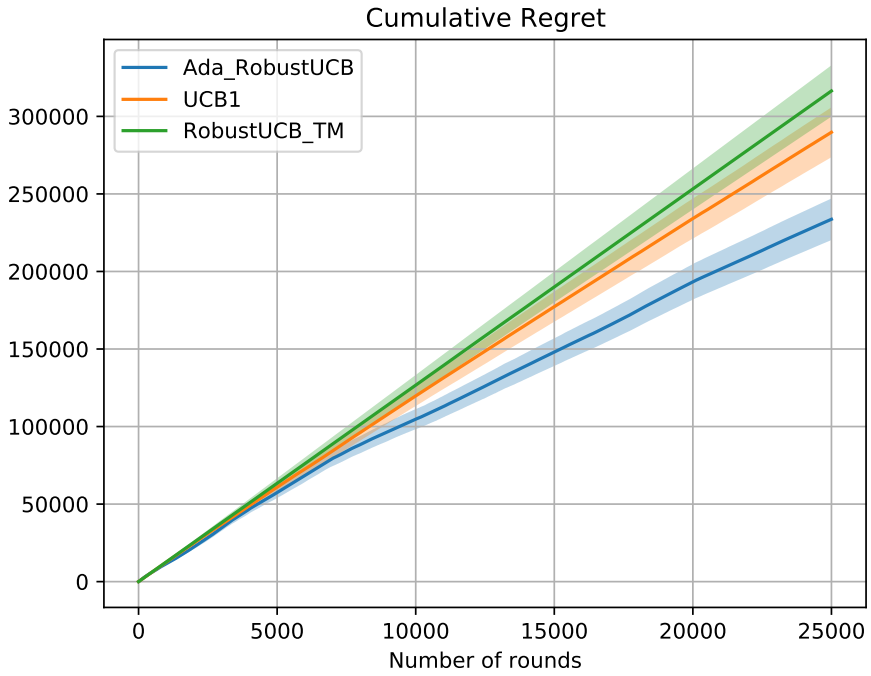


Figure 6.14: Numerical results of simulation 4 - Baseline comparison of cumulative regret (20 runs, shaded areas are standard deviations).

Comparing the regret results in Figure 6.14, we have that `AdaR-UCB` algorithm still performs better empirically than state-of-the-art ones, given a lower cumulative regret with a clear sub-linear behaviour. Nevertheless, the performance of `AdaR-UCB` has degraded a little bit with respect to previous simulations, due to the difficulty of the bandit instance considered.

To conclude our analysis on truncated non-positive bandit instances, we see that `RobustUCB` consistently shows poor performance, while `AdaR-UCB` generally outperforms both `UCB1` and `RobustUCB`, even if its convergence speed decreases significantly for instances with ϵ close to 0. The tightness of `AdaR-UCB` algorithm is more evident for higher values of ϵ and instances with all negative Pareto distributions.

6.3. General Heavy-Tailed Bandit Instances

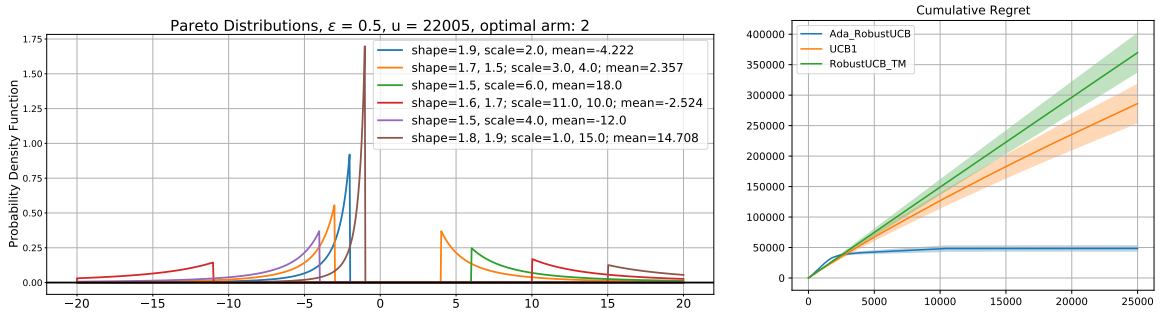
We now want to present other numerical results obtained applying `AdaR-UCB` algorithm on general heavy-tailed bandit instances, not satisfying Assumption 3.

The possible instances belonging to this framework can be the most varied. We can encounter bandits with all the arms distributed as positive Pareto's (see Figure 6.17a), or some arms with the positive tail only and some others double-tailed. In this second case, we might have that the optimal arm is either a positive Pareto or a double-tailed one, but in both cases the truncated non-positivity assumption is not satisfied since the mean is necessarily positive. We might also have all the possible combinations of Pareto distributions in the same bandit instance, including all the three generalized cases presented in Section 6.1 (see Figures 6.15a and 6.16a).

In this case, the theoretical evidencies presented in Chapter 5 for `AdaR-UCB` algorithm do not hold, such that we do not have any guarantee on the behaviour of the cumulative regret.

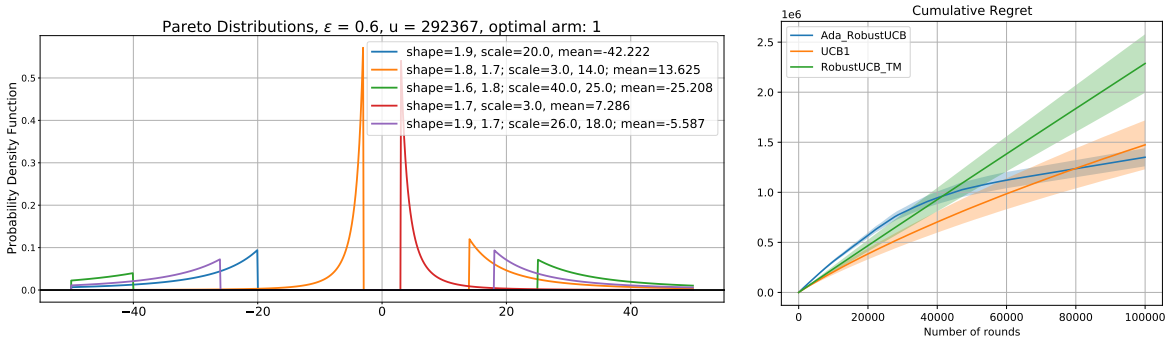
We present here some experiments showing how the performance of algorithm `AdaR-UCB` is not predictable in this setting, with the expected cumulative regrets plotted in Figures 6.15b, 6.16b and 6.17b, considering 20 independent runs and time horizon of either $T = 25000$ or $T = 100000$, depending on how fast the algorithms were converging.

We note that the regret curve of `AdaR-UCB` can grow fast and flatten with time, it can be logarithmic, slightly sub-linear or even linear, coherently with the fact that we have no theoretical result supporting this framework for `AdaR-UCB`. Nevertheless, for general heavy-tailed bandit instances, our algorithm still shows a slope of cumulative regret which decreases faster than `UCB1` and `RobustUCB` algorithms.



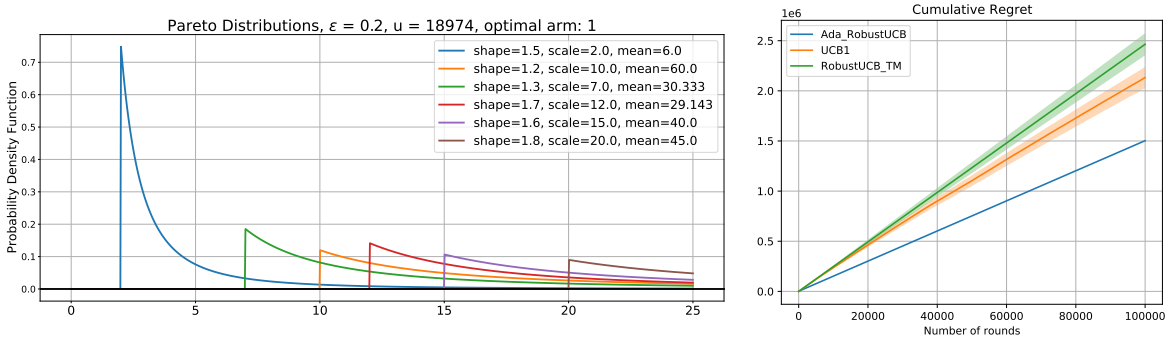
(a) Pareto Distributions. (b) Numerical results on Regret.

Figure 6.15: Simulation 5 - Negative, positive and double-tailed Pareto's, 6 arms, 20 runs, T = 25000, the optimal arm is positive.



(a) Pareto Distributions. (b) Numerical results on Regret

Figure 6.16: Simulation 6 - Negative, positive and double-tailed Pareto's, 5 arms, 20 runs, T=100000, the optimal arm is double-tailed.



(a) Pareto Distributions. (b) Numerical results on Regret

Figure 6.17: Simulation 7 - All positive Pareto's, 6 arms, 20 runs, T=100000.

7 | Conclusions and Future Work

7.1. Conclusions

In this thesis, we studied the *adaptive stochastic heavy-tailed bandit* problem, a variation on the classical stochastic 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 study of heavy-tailed multi-armed bandits is essential for a comprehensive and realistic approach to sequential decision-making in environments with a strong uncertainty and extreme outcomes. Nowadays, many real-world scenarios might present extreme outlying values occurring at high frequencies, and this explains why the research on the heavy-tailed bandit framework experienced a notable advancement over the last decade.

In the literature, given a bandit instance with heavy tailed rewards, it is common to assume the knowledge of the two key parameters characterizing the collection of reward distributions, namely ϵ and u. This is a strong limitation since in practical cases, where real-world samples are collected, these parameters are usually not known. Nevertheless, so far, every regret minimization strategy in literature still requires them as an algorithm's input to achieve optimal instance-dependent efficiency.

Approaching these practical problems in a flexible and efficient way has become crucial. Thus, our focus moved towards an *adaptive* approach, where agents are unaware of the aforementioned quantities, but still achieving comparable performances to non-adaptive approaches.

Our first relevant results concern the intrinsic difficulty of this setting, for which two novel lower bounds on expected regret have been provided in Chapter 4. In particular, we proved that, without any additional assumption, no algorithm can match the performances of a reasonable policy in the non-adaptive setting. This holds because any algorithm adaptive with respect to either u or ϵ has a higher regret lower bound than the one of the non-adaptive heavy-tailed bandit problem. In general, it is not possible to achieve the same order of performance of the state-of-the-art approaches while being unaware of these two

quantities.

We showed how adaptivity comes at a cost, pushing us towards restricting the set of adaptive heavy-tailed bandit problem instances under analysis to a special set, that would allow to accomplish optimality results. In detail, these results hold under a specific distributional assumption over the optimal arm, namely the *truncated non-positivity assumption*.

Finally, we provided a novel algorithm, namely `AdaR-UCB`, that under our assumption is able to achieve the state-of-the-art performances of the standard non-adaptive heavy-tailed bandit problem, i.e., an instance-dependent regret order matching the classical instance-dependent lower bound.
While in the conventional heavy-tailed bandit problem it is not feasible to be adaptive with respect to both ϵ and u while attaining the best regret order achievable, our algorithm shows that, under the aforementioned assumption, it is indeed possible.

Together with theoretical guarantees, we provided an empirical validation of `AdaR-UCB`. We validated the design choices of our solution in a synthetic environment. In each of the simulations we conducted, we observed a clear evidence in favor of `AdaR-UCB`, rather than the other two well-known baselines, namely `RobustUCB` and `UCB1`.

In particular, it is important to note that the theoretical performances in terms of regret upper bound for `AdaR-UCB` and `RobustUCB` coincide when the latter is run with the correct instance-dependent parameters ϵ and u. Nevertheless, even if in the simulations we decided to input to `RobustUCB` the correct values, `AdaR-UCB`, which does not require any prior knowledge but the number of arms, still outperforms `RobustUCB`. This might seem counter-intuitive since the first is an algorithm fully adaptive with less information, but the results were outstanding.

7.2. Future Developments

The main future direction of investigation regards the role of the *truncated non-positivity* assumption. Since with this assumption we selected a subset of the stochastic heavy-tailed instances, we wonder if is it possible to find a weaker one ensuring the kind of performances achieved by `AdaR-UCB` algorithm. This new assumption would of course hold for a set that contains the truncated non-positive instance, but does not contain the bandit instances considered in Chapter 4 to prove the adaptive lower bounds. On the other side, further future inspections should focus on how `AdaR-UCB` algorithm performs on instances not satisfying Assumption 3. We proposed numerical simulations for this case, but no theoretical result or guarantee has been proven yet.

We recall that, under our distributional assumption, `AdaR-UCB` is a tight algorithm for the instance-dependent case, while the instance-independent regret upper bound matches the minimax lower bound up to a logarithmic factor in T. A future work would consist in understanding if it is possible to find an algorithm that, under the truncated non-positive assumption, is tight in both the instance-dependent and instance-independent cases. This property was already achieved from MOSS algorithm in the traditional stochastic setting for subgaussian MAB [9, 60]. Building on UCB, the directly named "*minimax optimal strategy in the stochastic case*" (MOSS) algorithm was the first one that, modifying the confidence levels, managed to remove the $\log T$ factor entirely in instance-independent regret upper bound (see result for `UCB1` algorithm in Theorem 3.1).

Moreover, due to the empirical performances showed in Chapter 6, it would be interesting to better analyze how the non-adaptive `RobustUCB` algorithm behaves when the parameters ϵ and u are incorrectly specified as input.

Lastly, we recall that the two lower bounds introduced in Chapter 4 refer to adaptivity with respect to only one of the two quantities ϵ and u. As a future research development, it could be interesting to investigate if simultaneous adaptivity to both parameters implies an even higher lower bound.

Bibliography

[1] R. Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in applied probability*, 27(4):1054–1078, 1995.

[2] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

[3] S. Agrawal, S. K. Juneja, and W. M. Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR, 2021.

[4] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996.

[5] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[6] K. Ashutosh, J. Nair, A. Kagrecha, and K. Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR, 2021.

[7] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

[8] J.-Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Yheory*, pages 150–165. Springer, 2007.

[9] J.-Y. Audibert, S. Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[10] J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

[11] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[13] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[14] D. A. Berry and B. Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5 (71-87):7–7, 1985.

[15] S. Bhatt, G. Fang, P. Li, and G. Samorodnitsky. Nearly optimal catoni's m-estimator for infinite variance. In *International Conference on Machine Learning*, pages 1925–1944. PMLR, 2022.

[16] P. J. Bickel. On some robust estimates of location. *The Annals of Mathematical Statistics*, 1965.

[17] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5.

[18] J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137, 1979.

[19] S. Bubeck. Bandits games and clustering foundations. *PhD Thesis*, 2010.

[20] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer, 2009.

[21] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

[22] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[23] S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR, 2013.

[24] S. Bubeck, Y. Li, H. Luo, and C.-Y. Wei. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, pages 508–528. PMLR, 2019.

[25] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[26] O. 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.

[27] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[28] N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu. Boltzmann exploration done right. *Advances in Neural Information Processing Systems*, 30, 2017.

[29] Y. Chen, Q. Zhao, V. Krishnamurthy, and D. Djonin. Transmission scheduling for optimizing sensor network lifetime: A stochastic shortest path approach. *IEEE Transactions on Signal Processing*, 55(5):2294–2309, 2007.

[30] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[31] W. Cowan, J. Honda, and M. N. Katehakis. Normal bandits of unknown means and variances. *The Journal of Machine Learning Research*, 18(1):5638–5665, 2017.

[32] L. DaCosta, A. Fialho, M. Schoenauer, and M. Sebag. Adaptive operator selection with dynamic multi-armed bandits. In *Proceedings of the 10th annual Conference on Genetic and Evolutionary Computation*, pages 913–920, 2008.

[33] Y. David and N. Shimkin. Pure exploration for max-quantile bandits. In *Joint European Conference on Machine Learning and Knowledge discovery in databases*, pages 556–571. Springer, 2016.

[34] Y. David, B. Szörényi, M. Ghavamzadeh, S. Mannor, and N. Shimkin. Pac bandits with risk constraints. In *International Symposium on Artificial Intelligence and Mathematics*, 2018.

[35] S. De Rooij, T. Van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1): 1281–1316, 2014.

[36] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley. Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121(2):461–504, 2006.

[37] M. Gagliolo and J. Schmidhuber. Algorithm portfolio selection as a bandit problem with unbounded losses. *Annals of Mathematics and Artificial Intelligence*, 61:49–86, 2011.

[38] A. Garivier and O. 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.

[39] A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

[40] S. Gelly and Y. Wang. Exploration exploitation in Go: UCT for Monte-Carlo Go. In *NIPS: Neural Information Processing Systems Conference On-line trading of Exploration and Exploitation Workshop*, Canada, Dec. 2006.

[41] S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

[42] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[43] G. J. Gordon. Regret bounds for prediction problems. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 29–40, 1999.

[44] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[45] J. Huang, Y. Dai, and L. Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning*, pages 9173–9200. PMLR, 2022.

[46] J.-C. Hütter and C. Mao. Notes on adaptive estimation with lepski's method. 2017.

[47] L. P. Kaelbling. Learning to achieve goals. In *International Joint Conference on Artificial Intelligence*, volume 2, pages 1094–8. Citeseer, 1993.

[48] A. Kagrecha, J. Nair, and K. P. Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems*, pages 11269–11278, 2019.

[49] L. V. Kantorovich and G. P. Akilov. *Functional analysis*. Elsevier, 2016.

[50] M. N. Katehakis and H. Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences*, 92(19):8584–8585, 1995.

[51] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

[52] M. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference. *arXiv preprint arXiv:1301.7392*, 2013.

[53] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.

[54] M. J. Kochenderfer. *Decision making under uncertainty: theory and application*. MIT press, 2015.

[55] L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *European Conference on Machine Learning*, pages 282–293. Springer, 2006.

[56] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[57] A. Lacerda. Multi-objective ranked bandits for recommender systems. *Neurocomputing*, 246:12–24, 2017.

[58] T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

[59] T. L. Lai, H. Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[60] T. Lattimore. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

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

[62] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[63] L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[64] K. Lee, H. Yang, S. Lim, and S. Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462, 2020.

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

[66] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[67] J. Liebeherr, A. Burchard, and F. Ciucu. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2):1010–1024, 2012.

[68] S. Lu, G. Wang, Y. Hu, and L. Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163. PMLR, 2019.

[69] H. Luo and R. E. Schapire. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pages 1286–1304. PMLR, 2015.

[70] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[71] O.-A. Maillard. Robust risk-averse stochastic multi-armed bandits. In *Algorithmic Learning Theory: 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings 24*, pages 218–233. Springer, 2013.

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

[73] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[74] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

[75] A. M. Medina and S. Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650. PMLR, 2016.

[76] P. Ménard and A. Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, 2017.

[77] L. Metcalf and W. Casey. *Cybersecurity and applied mathematics*. Syngress, 2016.

[78] T. M. Mitchell. Machine learning, 1997.

[79] A. Motamedi and A. Bahai. Mac protocol design for spectrum-agile wireless networks: Stochastic control approach. In *2007 2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pages 448–451. IEEE, 2007.

[80] J. Nicolau and P. M. Rodrigues. A new regression-based tail index estimator. *Review of Economics and Statistics*, 101(4):667–680, 2019.

[81] L. Niss and A. Tewari. What you see may not be what you get: Ucb bandit algorithms robust to ε-contamination. In *Conference on Uncertainty in Artificial Intelligence*, pages 450–459. PMLR, 2020.

[82] F. Orabona and D. Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[83] F. Orabona and D. Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018.

[84] H. Robbins. Some aspects of the sequential design of experiments. 1952.

[85] E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

[86] H. Shao, X. Yu, I. King, and M. R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

[87] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha. Portfolio choices with orthogonal bandit learning. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[88] K. Sigman. A primer on heavy-tailed distributions. *Queueing systems*, 33(1-3):261, 1999.

[89] Q. Sun, W.-X. Zhou, and J. Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[90] R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[91] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444.

[92] F. Trovò, S. Paladino, M. Restelli, and N. Gatti. Improving multi-armed bandit algorithms in online pricing settings. *International Journal of Approximate Reasoning*, 98:196–235, 2018.

[93] S. Vakili, K. Liu, and Q. Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):759–767, 2013.

[94] S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[95] W. Vogel. An asymptotic minimax theorem for the two armed bandit problem. *The Annals of Mathematical Statistics*, 31(2):444–451, 1960.

[96] L. Wang, C. Zheng, W. Zhou, and W.-X. Zhou. A new principle for tuning-free huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021.

[97] C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

A | Appendix

A.1. Landau Notation

In the thesis, we make frequent use of the *Bachmann-Landau notation* [62]. Both were nineteenth century mathematicians who could have never expected their notation to be adopted so enthusiastically by computer scientists. Given functions $f, g : \mathbb{N} \to [0, \infty)$, define

$$f(n) = \mathcal{O}(g(n)) \Leftrightarrow \limsup_{n\to\infty} \frac{f(n)}{g(n)} < +\infty, \tag{A.1}$$

$$f(n) = \Omega(g(n)) \Leftrightarrow \liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0. \tag{A.2}$$

We make use of the (Bachmann-)Landau notation to informally describe a result without the clutter of uninteresting constants. For better or worse, this usage is often a little imprecise. For example, we will often write expressions of the form: $R_T \leq \mathcal{O}(u\sqrt{KT})$. Almost always what is meant by this is that there exists a universal constant $c > 0$ (a constant that does not depend on either of the quantities involved) such that $R_T \leq cu\sqrt{KT}$ for all (reasonable) choices of u, K and T. In this context, we are careful not to use Landau notation to hide large lower-order terms.

A.2. Hölder's Inequalities

Theorem A.1 (Hölder's Inequality, [49]). *Let (S, Σ, μ) be a measure space, let $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then for all measurable, real-valued functions f and g on S:*

$$\|fg\|_1 = \|f\|_p \|g\|_q,$$

where the norm here is given by $\|f\|_p = \left(\int_S |f|^p \mid d\mu\right)^{1/p}$.

For both the following specialized cases assume that p and q are in the open interval $(1, \infty)$, with $\frac{1}{p} + \frac{1}{q} = 1$.

Proposition A.1 (μ =Counting measure)**.** *Let $\mathbb{R}^n$ be the n-dimensional Euclidean space, and S be the set $\{1, \ldots, n\}$ with the counting measure, which, for each $A \subseteq S$, computes the cardinality of A. Then, Hölder's inequality reads:*

$$\sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} |y_k|^q \right)^{\frac{1}{q}}, \tag{A.3}$$
$$\text{for all } (x_1, \ldots, x_n), (y_1, \ldots, y_n) \in \mathbb{R}^n.$$

Proposition A.2 (μ = Probability measure)**.** *Let the probability space be $(\Omega, \mathcal{F}, \mathbb{P})$, For real-valued random variables X and Y on Ω, Hölder's inequality reads:*

$$\mathbb{E}[|XY|] \leqslant \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \left(\mathbb{E}\left[|Y|^q\right]\right)^{\frac{1}{q}}. \tag{A.4}$$

B | Appendix

B.1. Further Lower Bound Analysis

To have a complete discussion on the complexity of the adaptive heavy-tailed bandit problem, we further discuss the minimax lower bounds. In Chapter 4, we chose specific instances to state the lower bounds in a general stochastic adaptive HT setting, without additional assumptions. Indeed, we already noticed that these instances do not satisfy Assumption 3 on *truncated non-positivity*.

The reader might wonder if the construction in Sections 4.1.1 and 4.2.1 can be replicated on the negative axis, in order to enforce the truncated non-positive assumption to hold, and still repeat the derivation. Fortunately, this is not the case. Let us focus, for simplicity, on the construction in Section 4.2.1, but the same reasoning holds for the one in the proof of u-adaptive lower bound (Section 4.1.1).

Let us consider the translated **base instance** (with $y \geq 0$):

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

where δ_x denotes the Dirac delta measure centered in x and $\gamma = (2\Delta)^{\frac{1}{\epsilon}}$ for $\Delta \in [0, 1/2]$. The optimal arm is arm 2.

Let us enforce the truncated non-positive assumption on the optimal arm:

$$\mathbb{E}_{X\sim\nu_2}[X\mathbb{1}_{|X|\leq M}] = -y(1 + \Delta\gamma - \gamma^{1+\epsilon})\mathbb{1}_{|-y|\leq M} + (1/\gamma - y)(\gamma^{1+\epsilon} - \Delta\gamma)\mathbb{1}_{|1/\gamma - y|\leq M} \leq 0, \quad \forall M > 0.$$

Clearly, when $y \geq 1/\gamma$, the support contains non-positive points only and thus the assumption holds. Thus, we focus on $y < 1/\gamma$. In such a case, in order for the assumption

to hold it must be that:

$$\begin{cases} -y(1+\Delta\gamma-\gamma^{1+\epsilon})+(1/\gamma-y)(\gamma^{1+\epsilon}-\Delta\gamma)\leq 0 \\ 1/\gamma-y\leq y \end{cases} \implies \begin{cases} y\geq \gamma^{\epsilon}-\Delta \\ y\geq 1/2\gamma^{-1} \end{cases}.$$

Then, a necessary condition for having the property is that $y \geq \frac{1}{2}(2\Delta)^{-\frac{1}{\epsilon}}$.

We immediately observe that:

$$\mathbb{E}_{X\sim\nu_1}[X] = -y \leq -\frac{1}{2}(2\Delta)^{-\frac{1}{\epsilon}},$$

that gets unbounded by making $\Delta \to 0$ (as in the construction). It follows that, since the expectation is unbounded, also the $(1+\epsilon)$-th moments are, and Assumption 2 of heavy-tailed setting does not hold.

Now, starting from the construction in Section 4.2.1, the reader may propose a different construction to force the truncated non-positivity assumption. We consider here two new instances, in which we mirror the two arms for both the base and alternative instance in Section 4.2.1.

Base instance:

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

where $\gamma = (2\Delta)^{\frac{1}{\epsilon}}$ for $\Delta \in [0, 1/2]$. We have:

$$\begin{aligned} &\mu_1 = 0, \quad \mu_2 = -\Delta, \\ &\mathbb{E}_{\nu_1}[|X|^{\alpha}] = 0, \quad \mathbb{E}_{\nu_2}[|X|^{\alpha}] = -2^{\frac{1-\alpha}{\epsilon}}\Delta^{\frac{1+\epsilon-\alpha}{\epsilon}}, \end{aligned}$$

which are guaranteed to be bounded by constant (recall that we will make $\Delta \to 0$ in the construction) only if $\alpha \leq \epsilon+1$. Thus, this bandit admits moments finite up to order $\epsilon+1$. The optimal arm is arm 1.

Alternative instance:

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

where $\gamma' = (2\Delta)^{\frac{1}{\epsilon'}}$ for $\Delta \in [0, 1/2]$. We have:

$$\mu'_1 = -2\Delta, \quad \mu'_2 = -\Delta,$$
$$\mathbb{E}_{\nu'_1}[|X|^\alpha] = -(2\Delta)^{\frac{1+\epsilon'-\alpha}{\epsilon'}}, \quad \mathbb{E}_{\nu'_2}[|X|^\alpha] = -2^{\frac{1-\alpha}{\epsilon}}\Delta^{\frac{1+\epsilon-\alpha}{\epsilon}},$$

which are guaranteed to be bounded by constant only if $\alpha \leq 1 + \epsilon'$ (for the same reason as before, recalling that $\epsilon' < \epsilon$ too). Thus, this bandit admits moments finite up to order $1 + \epsilon'$. The optimal arm is arm 2.

We can attempt to replicate the non-existence result of Section 4.2, but following the same steps as in Section 4.2.2, now we get:

$$\max\left\{\frac{R_T}{T^{\frac{1}{1+\epsilon}}}, \frac{R'_T}{T^{\frac{1}{1+\epsilon'}}}\right\} \geq \max\left\{\frac{\Delta\mathbb{E}[N_2(T)]}{T^{\frac{1}{1+\epsilon}}}, \frac{\Delta}{8}T^{\frac{\epsilon'}{\epsilon'+1}}\exp\left(-c\mathbb{E}[N_1(T)](2\Delta)^{\frac{1+\epsilon'}{\epsilon'}}\right)\right\},$$

that does not allow to conclude the proof.

This shows that the construction in Section 4.2.1 cannot be extended to the case of instances satisfying the truncated non-positive assumption.

Remark 8. *Note that here we did not prove that there are no instances which satisfy Assumption 3 and give a regret lower bound for the adaptive setting of order higher than* $\Omega\left(T^{\frac{1}{1+\epsilon}}\right)$*. Nevertheless, if our theorem about the upper bound on the regret for the fully adaptive algorithm* `AdaR-UCB` *is correct (see Section 5.3), then these instances could not exist, otherwise we would get a contradiction.*

List of Figures

List of Tables

Acknowledgements

Prima di tutto, desidero porre una menzione e ringraziamento particolare per il mio relatore, Prof. Alberto Maria Metelli, e co-relatore, Dott. Gianmarco Genalti. La loro costante presenza, la loro fervida intuizione e il loro prezioso supporto hanno reso questa esperienza di ricerca un'opportunità unica di crescita e apprendimento. Onestamente, non avrei potuto chiedere una conclusione di percorso universitario più appagante di questa.

Ma il mio più grande ringraziamento va a mamma, papà, ai miei nonni, e a tutta la mia famiglia. Siete immensa fonte di ispirazione e tremenda gratitudine.

Una dovuta menzione va anche a chiunque mi sia stato accanto in questi anni, gli amici di una vita, quelli dei banchi dell'università, quelli delle esperienze incredibili all'estero e le persone uniche che hanno illuminato il mio ultimo anno a Milano.

Un grazie speciale, dal profondo del cuore.

Lupo

Milano, 5 Ottobre 2023