Bandits Corrupted by Nature: Lower Bounds on Regret and Robust Optimistic Algorithms

Anonymous authors
Paper under double-blind review

Abstract

We study the corrupted bandit problem, i.e. a stochastic multi-armed bandit problem with k unknown reward distributions, which are heavy-tailed and corrupted by a history-independent adversary or Nature. To be specific, the reward obtained by playing an arm comes from corresponding heavy-tailed reward distribution with probability $1-\varepsilon \in (0.5, 1]$ and an arbitrary corruption distribution of unbounded support with probability $\varepsilon \in [0, 0.5)$. First, we provide *a problem-dependent lower bound on the regret* of any corrupted bandit algorithm. The lower bounds indicate that the corrupted bandit problem is harder than the classical stochastic bandit problem with subGaussian or heavy-tail rewards. Following that, we propose a novel UCB-type algorithm for corrupted bandits, namely `HuberUCB`, that builds on Huber's estimator for robust mean estimation. Leveraging a novel concentration inequality of Huber's estimator, we prove that `HuberUCB` achieves a near-optimal regret upper bound. Since computing Huber's estimator has quadratic complexity, we further introduce a sequential version of Huber's estimator that exhibits linear complexity. We leverage this sequential estimator to design `SeqHuberUCB` that enjoys similar regret guarantees while reducing the computational burden. Finally, we experimentally illustrate the efficiency of `HuberUCB` and `SeqHuberUCB` in solving corrupted bandits for different reward distributions and different levels of corruptions.

1 Introduction

Multi-armed bandit problem is an archetypal setting to study sequential decision-making under incomplete information (Lattimore & Szepesvári, 2018). In the classical setting of stochastic multi-armed bandits, the decision maker or agent has access to $k \in \mathbb{N}$ unknown reward distributions or arms. At every step, the agent plays an arm and obtains a reward. The goal of the agent is to maximize the expected total reward accumulated by a given horizon $T \in \mathbb{N}$.

In this paper, we are interested in a challenging extension of the classical multi-armed bandit problem, where the reward at each step is corrupted by Nature, which is a stationary mechanism independent of the agent's decisions and observations. This setting is often referred as the *Corrupted Bandits*. Specifically, we extend the existing studies of corrupted bandits (Lykouris et al., 2018; Bogunovic et al., 2020; Kapoor et al., 2019) to the more general case, where the 'true' reward distribution might be heavy-tailed and the corruption can be unbounded.

A Motivating Example: Treatments of Varroa Mites. Though this article focuses on the theoretical aspects of this problem, we hereby illustrate a case study with roots in agriculture that motivates us. A bee-keeper has to choose between a set of treatments to save her bees from varroa mites. Every year, the bee-keeper must rotate between the treatments as the varroa mites develop resistance to a given treatment (Rinkevich, 2020; Kamler et al., 2016). The goal of the bee-keeper is to choose a sequence of treatments over the years that eliminate as many number of varroa mites as possible. The reward of a treatment is measured by the number of fallen varroa mites due to it. This reward function is heavy-tailed. As the number of fallen varroa mites is counted manually by the bee-keeper, this process is prone to human error. The reward of a treatment is further corrupted due to plethora of confounding variables, e.g. the weather, the way to administer the treatment, the state of the hive etc. (Semkiw et al., 2013), which are hard to model. The corruption in particular has been witnessed empirically while plotting the number of fallen mites for a given treatment (ref. Fig. 1 (Semkiw et al., 2013)). Interestingly, in this problem, corruptions in the measured rewards are natural and non-adversarial but possibly unbounded. The heavy-tailed and corrupted nature of the problem resists application of the non-robust bandit algorithms, such as UCB (Auer et al., 2002a), and motivates us to introduce the setting of *Bandits corrupted by Nature.*

Bandits corrupted by Nature. Motivated by the aforementioned example, we model a corrupted reward distribution as $(1-\varepsilon)P+\varepsilon H$, where P is the distribution of inliers with a finite variance, H is the distributions of outliers with probably unbounded support, and $\varepsilon \in [0, 1/2)$ is the proportion of outliers. Thus, in the corresponding stochastic bandit setting, an agent has access to k arms of corrupted reward distributions $\{(1-\varepsilon)P_i+\varepsilon H_i\}_{i=1}^{k}$. Here, P_i's are uncorrupted reward distributions with heavy-tails and bounded variances, and H_i's are corruption distributions with probably unbounded corruptions. The goal of the agent is to maximize the expected total reward accumulated oblivious to the corruptions. This is equivalent to considering a setting where at every step Nature flips a coin with success probability ε. The agent obtains a corrupted reward if Nature obtains 1 and otherwise, an uncorrupted reward. We call this setting *Bandits corrupted by Nature* as the corruption introduced in each step does not depend on the present or previous choices of arms and observed rewards. Our setting encompasses both heavy-tailed rewards and unbounded corruptions. We formally define the setting and corresponding regret definition in Section 3.

Bandits corrupted by Nature is different from the adversarial bandit setting (Auer et al., 2002b). The adversarial bandit assumes existence of a non-stochastic adversary that can return at each step the worst-case reward to the agent depending on its history of choices. Incorporating corruptions in this setting, Lykouris et al. (2018); Bogunovic et al. (2020) consider settings where the rewards can be corrupted by a history-dependent adversary but the total amount of corruption and also the corruptions at each step are bounded. However, we encounter problems in ecology and agronomy, such as treatments against varroa mites, where the corruptions are not adversarial, and are independent of the previous history of decisions. Thus, in contrast to the adversarial corruption setting in literature, we consider a non-adversarial proportion of corruptions ($\varepsilon \in [0, 1/2)$) at each step, which are stochastically generated from unbounded corruption distributions $\left(\{H_i\}_{i=1}^{k}\right)$. To the best of our knowledge, only Kapoor et al. (2019) have studied similar non-adversarial corruption setting with a history-independent proportion of corruption at each step. But they assume that the probable corruptions at each step are bounded, and the uncorrupted rewards are sub-Gaussian. Hence, we observe that *there is a gap in the literature in studying unbounded stochastic corruption for bandits with probably heavy-tailed rewards and this article aims to fill this gap.* Specifically, we aim to deal with unbounded corruption and heavy-tails simultaneously, which requires us to develop a novel sensitivity analysis of the robust estimator in lieu of a worst-case (adversarial bandits) analysis.

Our Contributions. Specifically, in this paper, we aim to investigate three main questions:

1. Is the setting of bandits corrupted by Nature with unbounded corruptions and heavy tails fundamentally harder (in terms of the regret lower bound) than the classical sub-Gaussian and uncorrupted bandit setting?
2. Is it possible to design an *efficient and robust algorithm* that achieves an order-optimal performance (*logarithmic* regret) in the corrupted by Nature setting?
3. Are robust bandit algorithms *efficient in practice*?

These questions have led us to the following contributions:

1. Hardness of bandits corrupted by Nature with unbounded corruptions and heavy tails. In order to understand the fundamental hardness of the proposed setting, we use a suitable notion of regret, denoted by $\mathfrak{R}_n$, (Equation (Corrupted regret), (Kapoor et al., 2019)) that extends the traditional pseudo-regret (Lattimore & Szepesvári, 2018) to the corrupted setting. Then, in Section 4, we derive lower bounds on regret that reveal increased difficulties of corrupted bandits with heavy tails in comparison with the classical non-corrupted and light-tailed Bandits. (a) In the Heavy-tailed regime (3), we show that even when the suboptimality gap Δ_i[1] is large, the regret increase with Δ_i because of the difficulty to distinguish between two arms when the rewards of Heavy-tailed. (b) Our lower bounds indicate that when Δ_i is large, the logarithmic regret is asymptotically achievable, but the hardness depends on the corruption proportion ε, variance of P_i, i.e. σ_i, and the suboptimality gap Δ_i. Specifically, if $\frac{\Delta_i}{\sigma_i}$'s are small, i.e. we are in low distinguishability/high variance regime, the hardness is dictated by $\frac{\sigma_i^2}{\overline{\Delta}_{i,\varepsilon}^2}$. Here, $\overline{\Delta}_{i,\varepsilon} \triangleq \Delta_i(1-\varepsilon)-2\varepsilon\sigma_i$ is the '*corrupted suboptimality gap*' that replaces the traditional suboptimality gap Δ_i in the lower bound of non-corrupted and light-tailed bandits (Lai & Robbins, 1985). Since $\overline{\Delta}_{i,\varepsilon} \leq \Delta_i$, it is harder to distinguish the optimal and suboptimal arms in the corrupted settings. They are the same when the corruption proportion $\varepsilon = 0$.

Additionally, our analysis addresses an open problem in heavy-tailed bandits. Works on heavy-tailed bandits (Bubeck et al., 2013; Agrawal et al., 2021) rely on the assumption that a bound on the $(1+\varepsilon)$-moment, i.e. $\mathbb{E}[|X|^{1+\varepsilon}]$, is known for some $\varepsilon > 0$. We do not assume such a restrictive bound as knowing a bound on $\mathbb{E}[|X|^{1+\varepsilon}]$ implies the knowledge of a bound on the sub-optimality gap Δ. Instead, we assume that the centered moment, specifically the variance, is bounded by a known constant. Thus, we address the open problem mentioned in (Agrawal et al., 2021) by relaxing the classical bounded $(1+\varepsilon)$-moment assumption with the bounded centered moment one.

2. Robust and Efficient Algorithm Design. In Section 5, we propose a robust algorithm, called `HuberUCB`, that leverages the Huber's estimator for robust mean estimation. We derive a novel concentration inequality on the deviation of empirical Huber's estimate that allows us to design robust and tight confidence intervals for `HuberUCB`. In Theorem 3, we show that `HuberUCB` achieves the logarithmic regret, and also the optimal rate when the sub-optimality gap Δ is not too large. We show that for `HuberUCB`, $\mathfrak{R}_n$ can be decomposed according to the respective values of Δ_i and σ_i:

$$\mathfrak{R}_n \leq \underbrace{\mathcal{O}\left(\sum_{i:\Delta_i>\sigma_i} \log(n)\sigma_i\right)}_{\text{Error due to Heavy-tail}} + \underbrace{\mathcal{O}\left(\sum_{i:\Delta_i\leq\sigma_i} \log(n)\Delta_i \frac{\sigma_i^2}{\overline{\Delta}_{i,\varepsilon}^2}\right)}_{\sigma^2/\Delta \text{ error with corrupted suboptimality gaps}}.$$

Thus, our upper bound allows us to segregate the errors due to heavy-tail, corruption, and corruption-correction with heavy tails. The error incurred by `HuberUCB` can be directly compared to the lower bounds obtained in Section 4 and interpreted in both the high distinguishability regime and the low distinguishability regime as previously mentioned.

3. Empirically Efficient and Robust Performance. To the best of our knowledge, we present the first robust mean estimator that can be computed in a linear time in a sequential setting (Section 6). Existing robust mean estimators, such as Huber's estimator, need to be recomputed at each iteration using all the data, which implies a quadratic complexity. Our proposal recomputes Huber's estimator only when the iteration number is a power of 2 and computes a sequential approximation on the other iterations. We use the Sequential Huber's estimator to propose `SeqHuberUCB`. We theoretically show that `SeqHuberUCB` achieves similar order of regret as `HuberUCB`, while being computationally efficient. In Section 7, we also experimentally illustrate that `HuberUCB` and `SeqHuberUCB` achieve the claimed performances for corrupted Gaussian and Pareto environments.

We further elaborate on the novelty of our results and position them in the existing literature in Section 2. For brevity, we defer the detailed proofs and the parameter tuning to Appendix.

2 Related Work

Due to the generality of our setting, this work either extends or relates to the existing approaches in both the heavy-tailed and corrupted bandits literature. While designing the algorithm, we further leverage the literature of robust mean estimation. In this section, we connect to these three streams of literature. Table 1 summarises the previous works and posits our work in lieu.

---

[1] The suboptimality gap of an arm is the difference in mean rewards of an optimal arm and that arm.

Algorithms	Settings	Corruption	Type of outliers	Heavy-tailed	Adversarial/ Stochastic
Our work	MAB	Yes	Unbounded	Yes	Stochastic
Bubeck et al. (2013); Agrawal et al. (2021); Lee et al. (2020)	MAB	No	x	Yes	Stochastic
Lykouris et al. (2018)	MAB	Yes	Bounded	No	Stochastic
Bogunovic et al. (2020)	GP Bandits	Yes	Bounded	No	Adversarial
Kapoor et al. (2019)	MAB & Linear Bandits	Yes	Bounded	No	Stochastic
Medina & Yang (2016); Shao et al. (2018)	Linear Bandits	No	x	Yes	Stochastic
Bouneffouf (2021)	Contextual Bandits	context only	Unbounded	No	Stochastic
Agarwal et al. (2019)	Control	Yes	Bounded	x	Adversarial
Hajiesmaili et al. (2020); Auer et al. (2002b); Pogodin & Lattimore (2020)	MAB	Yes	Bounded	x	Adversarial

Table 1: Comparison of existing results on Corrupted and Heavy-tailed Bandits.

Heavy-tailed bandits. Bubeck et al. (2013) are one of the first to study robustness in multi-armed bandits by studying the heavy-tailed rewards. They use robust mean estimator to propose the RobustUCB algorithms. They show that under assumptions on the raw moments of the reward distributions, a logarithmic regret is achievable. It sprouted research works leading to either tighter rates of convergence (Lee et al., 2020; Agrawal et al., 2021), or algorithms for structured environments (Medina & Yang, 2016; Shao et al., 2018). Our article uses Huber's estimator which was already discussed in (Bubeck et al., 2013). However, the chosen parameters in (Bubeck et al., 2013) were suited for heavy-tailed distributions, and thus, *render their proposed estimator non-robust to corruption. We address this gap in this work.*

Corrupted bandits. The existing works on Corrupted Bandits (Lykouris et al., 2018; Bogunovic et al., 2020; Kapoor et al., 2019) are restricted to bounded corruption. When dealing with bounded corruption, one can use techniques similar to adversarial bandits Auer et al. (2002b) to deal with an adversary that can't corrupt an arm too much. The algorithms and proof techniques are fundamentally different in our article because the stochastic (or non-adversarial) corruption by Nature allows us to learn about the inlier distribution on the condition that corresponding estimators are robust. Thus, *our bounds retain the problem-dependent regret, while successfully handling probably unbounded corruptions with robust estimators.*

Robust mean estimation. Our algorithm design leverages the rich literature of robust mean estimation, specifically the influence function representation of Huber's estimator. The problem of robust mean estimation in a corrupted and heavy-tailed setting stems from the work of Huber (Huber, 1964; 2004). Recently, in tandem with machine learning, there have been numerous advances both in the heavy-tailed (Devroye et al., 2016; Catoni, 2012; Minsker, 2019), and in the corrupted settings (Lecué & Lerasle, 2020; Minsker & Ndaoud, 2021; Prasad et al., 2019; 2020; Depersin & Lecué, 2019; Lerasle et al., 2019; Lecué & Lerasle, 2020). Our work, specifically the novel concentration inequality for Huber's estimator, enriches this line of work with a result of parallel interest. We introduce a sequential version of Huber's estimator achieving linear complexity.

3 Bandits corrupted by Nature: Problem formulation

In this section, we present the corrupted bandits setting that we study, together with the corresponding notion of regret. Similarly to the classical bandit setup, the regret decomposition lemma allows us to focus on the expected number of pulls of a suboptimal arm as the central quantity to control algorithmic standpoint.

Notations. We denote by $\mathcal{P}$ the set of probability distributions on the real line $\mathbb{R}$ and by $\mathcal{P}_{[q]} \triangleq \{P \in \mathcal{P} : \mathbb{E}_P[|X|^q] < \infty\}$ the set of distributions with at least $q \geq 1$ finite moments. $\mathbf{1}\{A\}$ is the indicator function for the event A being true. We denote the mean of a distribution P_i as $\mu_i \triangleq \mathbb{E}_{P_i}[X]$. For any $\mathcal{D} \subset \mathcal{P}$, we denote $\mathcal{D}(\varepsilon) \triangleq \{(1-\varepsilon)P + \varepsilon H : P \in \mathcal{D}, H \in \mathcal{P}\}$ the set of corrupted distributions from $\mathcal{D}$.

Problem Formulation. In the setting of *Bandits corrupted by Nature*, a bandit algorithm faces an environment with $k \in \mathbb{N}$ many reward distributions in the form $\{(1-\varepsilon)P_i + \varepsilon H_i\}_{i=1}^k$. Here P_i, H_i are real-valued distributions and ε is a mixture parameter assumed to be in $[0, 1/2)$, that is P_i is given more weights than H_i in the mixture of arm i. For this reason the $\{P_i\}_{i=1}^k$ are called the *inlier* distributions and the $\{H_i\}_{i=1}^k$ the *outlier* distributions. We assume the inlier distributions have at least 2 finite moments that is $P_1, \ldots, P_k \in \mathcal{P}_{[2]}$, while no restriction is put on the outlier distributions, that is $H_1, \ldots, H_k \in \mathcal{P}$. For this reason, we also refer to the outlier distributions as the *corrupted* distributions, and to the inlier distributions as the *non-corrupted* ones. ε is called the level of corruption. We write ν^ε the law of the corrupted environment, and we refer to that of the non-corrupted environment ν^0 by ν.

The game proceed as follows: At each step $t \in \{0, \ldots, n\}$, the agent policy π interacts with the corrupted environment by choosing an arm A_t and obtaining a reward corrupted by Nature. To generate this reward, Nature first draws a random variable $C_t \in \{0, 1\}$ from a Bernoulli distribution with mean $\varepsilon \in [0, 1/2)$. If $C_t = 1$, it generates a corrupted reward Z_t from distribution H_{A_t} corresponding to the chosen arm $A_t \in \{1, \ldots, k\}$. Otherwise, it generates a non-corrupted X'_t from distribution P_{A_t}. More formally, Nature generates reward $X_t = X'_t \mathbf{1}\{C_t = 0\} + Z_t \mathbf{1}\{C_t = 1\}$ which the learner observes. The learner leverages this observation to choose another arm at the next step in order to maximize the total cumulative reward obtained after n steps. In Algorithm 1, we outline a pseudocode of this framework.

Algorithm 1 Bandits corrupted by Nature

```
Require: $\varepsilon \in [0, 1/2)$ and $q \geq 2$
Input: $P_1, \ldots, P_k \in \mathcal{P}_{[q]}$ be the uncorrupted reward distributions and $H_1, \ldots, H_k \in \mathcal{P}$ be the corrupted reward distributions.
for $t = 1, \ldots, n$ do
    Player plays an arm $A_t \in \{1, \ldots, k\}$
    Nature draws a Bernoulli $C_t \sim Ber(\varepsilon)$
    Generate a corrupted reward $Z_t \sim H_{A_t}$ and an uncorrupted reward $X'_t \sim P_{A_t}$
    Player observe the reward $X_t = X'_t \mathbf{1}\{C_t = 0\} + Z_t \mathbf{1}\{C_t = 1\}$
end for
```

Remark 1 (Non-adversarial corruption.) *In the setting of Bandits corrupted by Nature, we consider that the reward received by the learner is corrupted when $C_t = 1$ and non-corrupted otherwise. Since the law of C_t is a Bernoulli $Ber(\varepsilon)$, the corruption is stochastic, and independent on other variables. This is in contrast with* adversarial *setups, where corruption is typically chosen by an opponent and possibly depending on other variables. Assuming a non-adversarial behavior of the Nature seem more justified than assuming an adversarial setup in applications, such as agriculture where corruption is often due to external disturbances, such as pests appearance or weather hazards, whose occurrence are typically non-adversarial. Now when corruption happens, we do not put restriction on the level of corruption. For example, we can imagine a pest outburst or hail, that may have huge impact on a crop but does not occur adversarially.*

Remark 2 (Weak assumption on inliers) *Let us highlight that we do not assume sub-Gaussian behavior for the inlier distributions P_i. Instead, we consider only a weak moment assumption, i.e.* the inlier distributions P_i have a finite variance. *Thus, our setting is capable of modelling both the heavy-tailed and corrupted settings. We highlight this generality in the regret lower bounds and empirical performance analysis in Section 4 and 7.*

Corrupted regret. In this setting, we observe that a corrupted reward distribution $((1-\varepsilon)P_i + \varepsilon H_i)$ might not have finite mean, unlike the true P_i's. Thus, the regret with respect to the corrupted reward distributions might fail to quantify the goodness of the policy and its immunity to corruption while learning.

In this setup, the natural notion of expected regret is measured with respect to the mean of the non-corrupted environment ν specified by $\{P_i\}_{i=1}^k$. We define the regret of learning algorithm playing strategy π after n

steps of interaction with the environment ν^ε as

$$\mathfrak{R}_n(\pi, \nu^\varepsilon) \triangleq n \max_i \mathbb{E}_{P_i}[X'] - \mathbb{E}\left[\sum_{t=1}^{n} X'_t\right]. \tag{Corrupted regret}$$

The expectation is crucially taken on $X'_i \sim P_i$ and $X'_t \sim P_{A_t}$ but not on X_i and X_t. The expectation on the right also incorporates possible randomization from the learner. Thus, (Corrupted regret) quantifies the loss in the rewards accumulated by policy π from the inliers while learning only from the *corrupted rewards* and also not knowing the arm with the best *true reward* distribution. Thus, this definition of corrupted regret quantifies the rate of learning of a bandit algorithm as regret does for non-corrupted bandits. A similar notion of regret is considered in (Kapoor et al., 2019) that deals with bounded stochastic corruptions.

Due to the non-adversarial nature of the corruption, the regret can be decomposed, as in classical stochastic bandits, to make appear the expected number of pulls of suboptimal arms $\mathbb{E}_{\nu^\varepsilon}[T_i(n)]$, which allow us to focus the regret analysis on bounding these terms.

Lemma 1 (Decomposition of corrupted regret) *In a corrupted environment ν^ε, the regret writes*

$$\mathfrak{R}_n(\pi, \nu^\varepsilon) = \sum_{i=1}^{k} \Delta_i \mathbb{E}_{\nu^\varepsilon}\left[T_i(n)\right],$$

where $T_i(n) \triangleq \sum_{t=1}^{n} \mathbf{1}\{A_t = i\}$ denotes the number of pulls of arm i until time n and the problem-dependent quantity $\Delta_i \triangleq \max_j \mu_j - \mu_i$ is called the suboptimality gap of arm i.

4 Lower bounds for uniformly good policies under heavy-tails and corruptions

In order to derive the lower bounds, it is classical to consider *uniformly good* policies on some family of environments, Lai & Robbins (1985). We introduce below the corresponding notion for corrupted environments with the set of laws $\mathfrak{D}^{\otimes k} = \mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_k$, where $\mathcal{D}_i \subset \mathcal{P}$ for each $i \in \{1, \dots, k\}$.

Definition 1 (Robust uniformly good policies) *Let $\mathfrak{D}^{\otimes k}(\varepsilon) = \mathcal{D}_1(\varepsilon) \otimes \cdots \otimes \mathcal{D}_k(\varepsilon)$ be a family of corrupted bandit environments on $\mathbb{R}$. For a corrupted environment $\nu^\varepsilon \in \mathfrak{D}^{\otimes k}(\varepsilon)$ with corresponding uncorrupted environment ν, let $\mu_i(\nu)$ denote the mean reward of arm i in the uncorrupted setting and $\mu_\star(\nu) \triangleq \max_a \mu_i(\nu)$ denote the maximum mean reward. A policy π is uniformly good on $\mathfrak{D}^{\otimes k}(\varepsilon)$ if for any $\alpha \in (0, 1]$,*

$$\forall \nu \in \mathfrak{D}^{\otimes k}(\varepsilon), \forall i \in \{1, \dots, k\}, \mu_i(\nu) < \mu_\star(\nu) \Rightarrow \quad \mathbb{E}_{\nu^\varepsilon}[T_i(n)] = o(n^\alpha).$$

Since the corrupted setup is a special case of stochastic bandits, a lower bound can be immediately recovered with classical results, such as Lemma 2 below, that is a version of the change of measure argument (Burnetas & Katehakis, 1997), and can be found in (Maillard, 2019, Lemma 3.4).

Lemma 2 (Lower bound for uniformly good policies) *Let $\mathfrak{D}^{\otimes k} = \mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_k$, where $\mathcal{D}_i \subset \mathcal{P}$ for each $i \in \{1, \dots, k\}$ and let $\nu \in \mathfrak{D}^{\otimes k}$. Then, any uniformly good policy on $\mathfrak{D}^{\otimes k}$ must pull arms such that for any $P_i \in \mathcal{D}_i$, $i \in \{1, \dots, k\}$,*

$$\forall i \in \{1, \dots, k\},\ \mu_i \leq \mu_\star(\nu) \quad \Rightarrow \quad \liminf_{n \to \infty} \frac{\mathbb{E}_\nu[T_i(n)]}{\log(n)} \geq \frac{1}{\mathcal{K}_i(P_i, \mu(P^*))}.$$

where $\mathcal{K}_i(P_i, \mu(P^*)) = \inf\{D_{\mathrm{KL}}(P_i, \nu) : \nu_i \in \mathcal{D}_i, \mu(\nu_i) \geq \mu(P^*)\}$.

Lemma 2 is used in the traditional bandit literature to obtain lower bound on the regret using the decomposition of regret from Lemma 1. In our setting however, the lower bound is more complex as it involves optimization on the non-convex set $\mathcal{P}_{[2]}$ of distributions with a bounded variance. It also involves an optimization in both the first and second term of the KL because we consider the worst-case corruption in both the optimal arm P^* and non-optimal arm P_i. In this section, we do not solve these problems, but we propose lower bounds derived from the study of a specific class of heavy-tailed distributions on one hand (Lemma 3) and the study of a specific class of corrupted (but not heavy-tailed) distributions on the other hand (Lemma 4).

Using the fact that $\mathcal{K}_i(P_i, \mu(P^*))$ is an infimum that is smaller than the D_{KL} for the choice $\nu = P^*$, Lemma 2 induces the following weaker lower-bound:

$$\forall i \in \{1, \dots, k\},\, \mu_i \leq \mu_\star(\nu) \quad \Rightarrow \quad \liminf_{n\to\infty} \frac{\mathbb{E}_\nu[T_i(n)]}{\log(n)} \geq \frac{1}{D_{\mathrm{KL}}(P_i, P^*)}. \tag{1}$$

Equation (1) shows that *it is sufficient to have an upper bound on the* D_{KL}*-divergence of the reward distributions interacting with the policy to get a lower bound on the number of pulls of a sub-optimal arm.*

In order to bound the D_{KL}-divergence, we separately focus on two families of reward distributions, namely Student's distribution without corruption and corrupted Bernoulli distribution, that reflect the hardness due to heavy-tails and corruptions, respectively.

Student's distribution without corruption. To obtain a lower bound in the heavy-tailed case we use Student distributions. Student distribution are well adapted because they exhibit a finite number of finite moment which makes them heavy-tailed, and we can easily change the mean and variances of Student distribution without changing its shape parameter d. We denote by $\mathcal{T}_d$ the set of Student distributions with d degrees of freedom,

$$\mathcal{T}_d = \left\{ P \in \mathcal{P},\, P \text{ has distribution defined for } t \in \mathbb{R} \text{ by } p(t) = \frac{\Gamma(\frac{d+1}{2})}{\Gamma(d/2)\sqrt{d\pi}} \left(1 + \frac{t^2}{d}\right)^{-\frac{d+1}{2}} \right\}.$$

Lemma 3 (Control of KL-divergence for Heavy-tails) *Let* P_1, P_2 *be two Student distributions with* $d > 1$ *degrees of freedom with* $\mathbb{E}_{P_1}[X] = 0$ *and* $\mathbb{E}_{P_2}[X] = \Delta$*. Then,*

$$D_{\mathrm{KL}}(P_1, P_2) \leq \begin{cases} \frac{3^{d-1}(d+1)^2\Delta^2}{5\sqrt{d}} & \text{if } \Delta \leq 1\,, \\ (d+1)\log(\Delta) + \log\left(3^d \frac{(d+1)^2}{5\sqrt{d}}\right) & \text{if } \Delta > 1\,. \end{cases} \tag{2}$$

Corrupted Bernoulli distributions. Now, we study the cost of corruption using the corrupted Bernoulli distributions. Let P_0, P_1 be two Bernoulli distributions on $\{0, 1\}$ such that $\mathbb{P}_{P_0}(1) = \mathbb{P}_{P_1}(0) > P_{P_0}(0) = P_{P_1}(1)$. We corrupt both P_0 and P_1 with a proportion $\varepsilon > 0$ to get $Q_0 \triangleq (1-\varepsilon)P_0 + \varepsilon\delta_c$ and $Q_1 \triangleq (1-\varepsilon)P_1 + \varepsilon\delta_0$. We obtain Lemma 4 that illustrates three bounds on $D_{\mathrm{KL}}(Q_0, Q_1)$ as functions of the suboptimality gap $\Delta \triangleq \mathbb{E}_{P_0}[X] - \mathbb{E}_{P_1}[X]$, variance $\sigma^2 \triangleq \mathrm{Var}_{P_0}(X) = \mathrm{Var}_{P_1}(X)$, and corruption proportion ε.

Lemma 4 (Control of KL-divergence for Corruptions) *There exists* P_0, P_1 *two Bernoulli probability distribution with* $\Delta = \mathbb{E}_{P_0}[X] - \mathbb{E}_{P_1}[X]$ *and* $\sigma^2 = \mathrm{Var}_{P_0}(X) = \mathrm{Var}_{P_1}(X)$ *for which there exists* Q_0 *and* Q_1 *some* ε*-corruptions of* P_0 *and* P_1 *respectively, that have shifted suboptimality gap given by* $\overline{\Delta}_\varepsilon = \mathbb{E}_{Q_0}[X] - \mathbb{E}_{Q_1}[X] = \Delta(1-\varepsilon) - 2\varepsilon\sigma$*. Furthermore, they can be chosen so as to satisfy*

- ***Uniform Bound.*** *For any* Δ, σ*, we have*

$$D_{\mathrm{KL}}(Q_0, Q_1) \leq (1-2\varepsilon)\log\left(1 + \frac{1-2\varepsilon}{\varepsilon}\right). \tag{3}$$

- ***High Distinguishability/Low Variance Regime.*** *If* $2\sigma\frac{\varepsilon}{\sqrt{1-2\varepsilon}} < \Delta < 2\sigma$*, we get*

$$D_{\mathrm{KL}}(Q_0, Q_1) \leq \frac{\overline{\Delta}_\varepsilon}{2\sigma} \log\left(1 + \frac{\overline{\Delta}_\varepsilon}{2\sigma - \overline{\Delta}_\varepsilon}\right). \tag{4}$$

- ***Low Distinguishability/High Variance Regime.*** *If* $\Delta \leq 2\sigma\frac{\varepsilon}{\sqrt{1-2\varepsilon}}$*, there exists* $\varepsilon' \leq \varepsilon$ *and* Q_0', Q_1' *some* ε'*- versions of* P_0 *and* P_1 *such that* $D_{\mathrm{KL}}(Q_0', Q_1') = 0$*.*

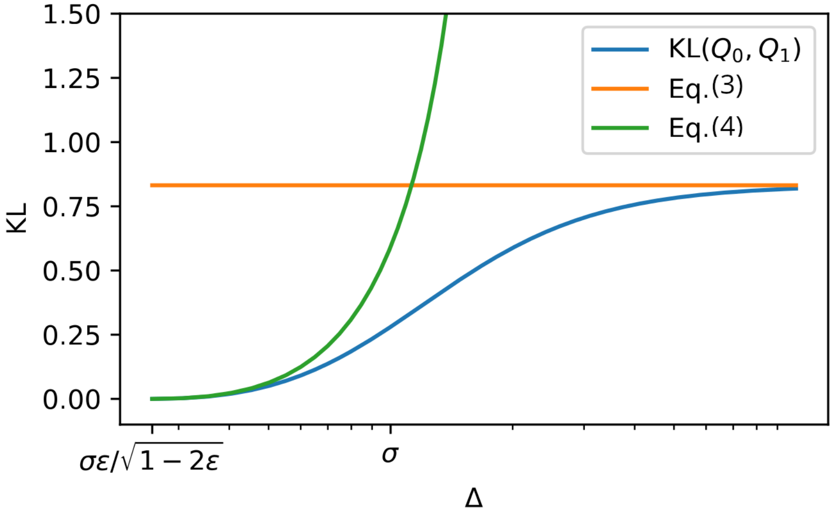

Figure 1: Visualizing the KL and the corresponding bounds in Lemma 4 for $\sigma = 1$ and $\varepsilon = 0.2$ (x axis is in log scale).

Consequences of Lemma 4. We illustrate the bounds of Lemma 4 in Figure 1. The three upper bounds on the KL-divergence of corrupted Bernoullis provide us some insights regarding the impact of corruption.

1. *Three Regimes of Corruption:* We observe that depending on Δ/σ, we can categorize the corrupted environment in three categories. For $\Delta/\sigma \in [2, +\infty)$, we observe that the KL-divergence between corrupted distributions Q_0 and Q_1 is upper bounded by a function of only corruption proportion ε and is independent of the uncorrupted distributions. Whereas for $\Delta/\sigma \in (2\varepsilon/\sqrt{1-2\varepsilon}, 2)$, the distinguishability of corrupted distributions depend on the distinguishibility of uncorrupted distributions and also the corruption level. We call this the High Distinguishability/Low Variance Regime. For $\Delta/\sigma \in [0, 2\varepsilon/\sqrt{1-2\varepsilon}]$, we observe that the KL-divergence can always go to zero. We refer to this setting as the Low Distinguishability/High Variance Regime.

2. *High Distinguishability/Low Variance Regime:* In Lemma 4, we observe that the effective gap to distinguish the optimal arm to the closest suboptimal arm that dictates hardness of a bandit instance has shifted from the uncorrupted gap Δ to a *corrupted suboptimality gap*: $\overline{\Delta}_\varepsilon \triangleq \Delta(1-\varepsilon) - 2\varepsilon\sigma$.

3. *Low Distinguishability/High Variance Regime:* We notice also that there is a limit for Δ below which the corruption can make the two distributions Q_0 and Q_1 indistinguishable, this is a general phenomenon in the setting of testing in corruption neighborhoods (Huber, 1965).

From KL Upper bounds to Regret Lower Bounds. Substituting the results of Lemma 3 and 4 in Equation (1) yield the lower bounds on regret of any uniformly good policy in heavy-tailed and corrupted settings, where reward distributions either belong to the class of corrupted student distributions or the class of corrupted Bernoulli distributions, respectively. We denote

$$\mathfrak{D}_{\mathcal{T}_2}^{\otimes k} \triangleq \mathcal{T}_2 \otimes \cdots \otimes \mathcal{T}_2,$$

where $\mathcal{T}_2$ is the set of Student distributions with more than 2 degrees of freedoms. We also define

$$\mathfrak{D}_{\mathcal{B}(\varepsilon)}^{\otimes k} \triangleq \mathcal{B}(\varepsilon) \otimes \cdots \otimes \mathcal{B}(\varepsilon),$$

where $\mathcal{B}(\varepsilon) = \{(1-\varepsilon)P + \varepsilon H;\ H \sim Ber(p) \text{ and } P \sim Ber(p'), p, p' \in [0,1]\}$ is the set of corrupted Bernoulli distributions.

Theorem 1 (Lower bound for heavy-tailed and corrupted bandit) *Let i be a suboptimal arm such that $\mathbb{E}_{P_i}[X] \leq \max_a \mathbb{E}_{P_a}[X]$ and denote $\Delta_i \triangleq \mathbb{E}_{P_i}[X] - \max_a \mathbb{E}_{P_a}[X]$ and $\overline{\Delta}_{i,\varepsilon} \triangleq \Delta_i(1-\varepsilon) - 2\varepsilon\sigma_i$.*

Student's distributions. *Suppose that the arms are pulled according to a policy that is uniformly good on* $\mathfrak{D}_{\mathcal{T}_2}^{\otimes k}$*. Then,*

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\nu^\varepsilon}[T_i(n)]}{\log(n)} \geq \frac{\sigma_i^2}{51\Delta_i^2} \vee \frac{1}{4\log(\Delta_i/\sigma_i)+22}. \tag{5}$$

Corrupted Bernoulli distributions*: Suppose that the arms are pulled according to a policy that is uniformly good on* $\mathfrak{D}_{\mathcal{B}(\varepsilon)}^{\otimes k}$*. Then, we have for* $2\sigma_i \frac{\varepsilon}{\sqrt{1-2\varepsilon}} < \Delta_i < 2\sigma_i$*, then*

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\nu^\varepsilon}[T_i(n)]}{\log(n)} \geq \frac{2\sigma_i}{\overline{\Delta_{i,\varepsilon}} \log\left(1 + \frac{\overline{\Delta_{i,\varepsilon}}}{2\sigma_i - \overline{\Delta_{i,\varepsilon}}}\right)}, \tag{6}$$

and for $\Delta_i > 2\sigma_i$*,*

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\nu^\varepsilon}[T_i(n)]}{\log(n)} \geq \frac{1}{(1-2\varepsilon)\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}. \tag{7}$$

For brevity, the detailed proof is deferred to Appendix A.1.

Small gap versus large gap regimes. Due to the restriction in the family of distributions considered in Theorem 1, the lower bounds are not tight and may not exhibit the correct rate of convergence for all families of distributions. However, this theorem provide some insights about the difficulties that one may encounter in corrupted and heavy-tail bandits problems, including the logarithmic dependence on n.

In Theorem 1, if Δ_i is small, we see that in the heavy-tailed case (Student's distribution), we recover a term very similar to the lower bound when the arms are from a Gaussian distribution. Now in the case where Δ_i is large, the number of suboptimal pulls in the heavy-tail setting is $\Omega\left(1/\log\left(\frac{\Delta_i}{\sigma_i}\right)\right)$. This is the price to pay for heavy-tails.

If we are in the high distiguishability/low variance regime, i.e. $\frac{\overline{\Delta_{i,\varepsilon}}}{2\sigma_i} \in (\frac{\varepsilon}{\sqrt{1-2\varepsilon}}, 1)$, we recover a logarithmic lower bound which depends on a *corrupted gap between means* $\overline{\Delta_{i,\varepsilon}} = \Delta_i(1-\varepsilon) - 2\varepsilon\sigma_i$. Since the corrupted gap is always smaller than the true gap Δ_i, this indicates that a corrupted bandit ($\varepsilon > 0$) must incur higher regret than a uncorrupted one ($\varepsilon = 0$). For $\varepsilon = 0$, this lower bound coincides with the lower bound for Gaussians with uncorrupted gap of means Δ_i and variance σ_i^2. On the other hand, if $\frac{\overline{\Delta_{i,\varepsilon}}}{2\sigma_i}$ is larger than 1, we observe that we can still achieve logarithmic regret but the hardness depends on only the corruption level ε, specifically $\frac{1}{(1-2\varepsilon)\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}$.

5 Robust bandit algorithm: Huber's estimator and upper bound on the regret

In this section, we propose an UCB-type algorithm, namely `HuberUCB`, addressing the Bandits corrupted by Nature setting (Algorithm 2). This algorithm uses primarily a robust mean estimator called Huber's estimator (Section 5.1) and corresponding confidence bound to develop `HuberUCB` (Section 5.2). We further provide a theoretical analysis in Theorem 3 leading to upper bound on regret of `HuberUCB`. We observe that the proposed upper bound matches the lower bound in Theorem 1 under some settings.

5.1 Robust mean estimation and Huber's estimator

We begin with a presentation of the Huber's estimator of mean (Huber, 1964).

As we aim to design a UCB-type algorithm, the main focus is to obtain an empirical estimate of the mean rewards. Since the rewards are heavy-tailed and corrupted in this setting, we have to use a robust estimator of mean. We choose to use Huber's estimator (Huber, 1964), an M-estimator that is known for its robustness properties and have been extensively studied (e.g. the concentration properties (Catoni, 2012)).

Huber's estimator is an M-estimator, which means that it can be derived as a minimizer of some loss function. Given access to n i.i.d. random variables $X_1^n \triangleq \{X_1, \ldots, X_n\}$, we define Huber's estimator as

$$\mathrm{Hub}(X_1^n) \in \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \rho(X_i - \theta), \tag{8}$$

where ρ is Huber's loss function with parameter $\beta > 0$. ρ is a loss function that is quadratic near 0 and linear near infinity, with β thresholding between the quadratic and linear behaviors.

In the rest of the paper, rather than using the aforementioned definition, we represent the Huber's estimator as a root of the following equation (Mathieu, 2021):

$$\sum_{i=1}^{n} \psi\left(X_i - \mathrm{Hub}(X_1^n)\right) = 0. \tag{9}$$

Here, $\psi(x) \triangleq x\mathbf{1}\{|x| \leq \beta\} + \beta\,\mathrm{sign}(x)\mathbf{1}\{|x| > \beta\}$ is called the influence function. Though the representations in Equation (8) and (9) are equivalent, we prefer to use representation Equation (9) as we prove the properties of Huber's estimator using those of ψ.

β plays the role of a scaling parameter. Depending on β, Huber's estimator exhibits a trade-off between the efficiency of the minimizer of the square loss, i.e. the empirical mean, and the robustness of the minimizer of the absolute loss, i.e. the empirical median.

5.2 Concentration of Huber's estimator in corrupted setting

Let use denote the true Huber mean for a distribution P as $\mathrm{Hub}(P)$. This means that for a random variable Y with law P, $\mathrm{Hub}(P)$ satisfies $\mathbb{E}[\psi(Y - \mathrm{Hub}(P))] = 0$.

We now state our first key result on the concentration of Huber's estimator around $\mathrm{Hub}(P)$ in a corrupted and Heavy-tailed setting.

Theorem 2 (Concentration of Empirical Huber's estimator) *Suppose that $X_1, \ldots, X_n$ are i.i.d with law $(1-\varepsilon)P + \varepsilon H$ for some $P, H \in \mathcal{P}$ and proportion of outliers $\varepsilon \in (0, 1/2)$, and P has a finite variance σ^2. Then, with probability larger than $1 - 5\delta$,*

$$|\mathrm{Hub}(X_1^n) - \mathrm{Hub}(P)| \leq \frac{\sigma\sqrt{\frac{2\ln(1/\delta)}{n}} + \beta\frac{\ln(1/\delta)}{3n} + 2\beta\bar{\varepsilon}\sqrt{\frac{\ln(1/\delta)}{n}} + 2\beta\varepsilon}{\left(p - \sqrt{\frac{\ln(1/\delta)}{2n}} - \varepsilon\right)_+}.$$

Here, $p = \mathbb{P}_P(|Y - \mathbb{E}_P[Y]| \leq \beta/2)$ with $p > 5\varepsilon$, $\beta > 4\sigma$, $\bar{\varepsilon} = \sqrt{\frac{(1-2\varepsilon)}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}$, and $\delta \geq \exp\left(-n\frac{128(p-5\varepsilon)^2}{49\left(1+2\bar{\varepsilon}\sqrt{2}\right)^2}\right)$.

Theorem 2 gives us the concentration of $\mathrm{Hub}(X_1^n)$ around $\mathrm{Hub}(P)$, i.e. the Huber functional of the *inlier* distribution P. This theorem will allow us to construct a UCB-type algorithm to solve the Bandits corrupted by Nature.

For convenience of notation, hereafter, we denote the rate of convergence of $\mathrm{Hub}(X_1^n)$ to $\mathrm{Hub}(P)$ as

$$r_n(\delta) \triangleq \frac{\sigma\sqrt{\frac{2\ln(1/\delta)}{n}} + \beta\frac{\ln(1/\delta)}{3n} + 2\beta\bar{\varepsilon}\sqrt{\frac{\ln(1/\delta)}{n}} + 2\beta\varepsilon}{\left(p - \sqrt{\frac{\ln(1/\delta)}{2n}} - \varepsilon\right)_+}. \tag{10}$$

Discussion. Now, we provide a brief discussion on the implications of Theorem 2.

1. Value of p: For most laws that exhibit concentration properties, the constant p is close to 1 as $\beta \geq 4\sigma$. One might also use Markov inequality to lower bound p, depending on the number of finite moments P has.

Bounding p then becomes a trade-off on the value of β, where large values of β implies that p is close to 1. But larger β also leads to a less robust estimator, since the error bound in Theorem 2 increases with β.

2. Tightness of constants: If there are no outliers ($\varepsilon = 0$), the optimal rate of convergence in such a setting is at least of order $\sigma\sqrt{2\ln(1/\delta)/n}$ due to the central limit theorem. Theorem 2 shows that we are very close to attaining this optimal constant in the leading $1/\sqrt{n}$ term. This result for Huber's estimator echoes the one presented in (Catoni, 2012).

3. Value of β: β is a parameter that achieve a trade-off between accuracy in the light-tailed uncorrupted setting and robustness. For our result, β must be at least of the order of 4σ. We provide a detailed discussion on the choice of β in Section 5.4.

4. Restriction on the values of δ: In Theorem 2, δ must be at least of order e^{-n}. This restriction may seem arbitrary but it is in fact unavoidable as shown in (Devroye et al., 2016, Theorem 4.3). This is a limitation of robust mean estimation that enforces our algorithm to perform a forced exploration in the beginning.

5. Restriction on the values of ε: In Theorem 2, ε can be at most $p/5$, which implies that it is smaller than 1/5. This restriction is common in robustness literature. In particular, in Kapoor et al. (2019), ε is supposed smaller than Δ/σ. In robustness literature, Lecué & Lerasle (2020) and Dalalyan & Thompson (2019) assumed that $\varepsilon \leq 1/768$ and 1/400 respectively. In contrast, our analysis can handle ε up to 0.2, which is signifcantly higher than the existing restrictions.

Bias of Huber's Estimate. If P is symmetric, we have $\mathrm{Hub}(P) = \mathbb{E}[X]$. When P is non-symmetric, we need to control the distance of the Huber's estimate from the true mean, i.e. $|\mathrm{Hub}(P) - \mathbb{E}[X]|$. We call it the bias of Huber's estimate. We need to bound this biad to get a concentration of the empirical Huber's estimate $\mathrm{Hub}(X_1^n)$ around the true mean $\mathbb{E}[X]$. We control the bias using the following lemma, which is a direct consequence of (Mathieu, 2021, Lemma 4).

Lemma 5 (Bias of Huber's estimator) *Let Y be a random variable with $\mathbb{E}[|Y|^q] < \infty$ for $q \geq 2$ and suppose that $\beta^2 \geq 9\mathrm{Var}(Y)$. Then*

$$|\mathbb{E}[Y] - \mathrm{Hub}(P)| \leq \frac{2\mathbb{E}[|Y - \mathbb{E}[Y]|^q]}{(q-1)\beta^{q-1}}.$$

Using Lemma 5 and Theorem 2, we can control the deviations of $\mathrm{Hub}(X_1^n)$ from $\mathbb{E}[X]$. This allows us to formulate an index-based algorithm (UCB-type algorithm) for corrupted Bandits. We present this algorithm in Section 5.3.

5.3 `HuberUCB`: Algorithm and regret bound

In this section, we describe a robust, UCB-type algorithm called `HuberUCB`. We denote μ_i as the mean of arm i and its variance as σ_i^2. We assume that we know the variances of the reward distributions. We refer to Section 5.4 for a discussion on the choice of the parameters when the reward distributions are unknown.

`HuberUCB`: The algorithm. In order to deploy the Huber's estimator in the multi-armed bandits setting, we need to estimate the mean of the rewards of each arm separately. We do that by defining a parameter β_i for each arm and estimating separately each μ_i using

$$\mathrm{Hub}_{i,s} = \mathrm{Hub}\left(X_t, \quad 1 \leq t \leq s \quad \text{such that} \quad A_t = i, \right).$$

Now, at each step t, we define a confidence bound for arm i with s number of pulls as

$$B_i(s,t) \triangleq \begin{cases} r_s(1/t^2) + b_i & \text{if } s \geq s_{lim}(t) \\ \infty & \text{if } s < s_{lim}(t) \end{cases}, \tag{11}$$

where $r_s(1/t^2)$ is defined by Equation (10), $s_{lim}(t) = \log(t)\frac{98}{128(p-5\varepsilon)^2}\left(1 + 2\sqrt{2}\left(\overline{\varepsilon} \vee \frac{9}{14\sqrt{2}}\right)\right)^2$, $\overline{\varepsilon} = \sqrt{\frac{(1-2\varepsilon)}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}$, and b_i is a bound on the bias $|\mathbb{E}[X] - \mathrm{Hub}(P_i)|$. b_i is zero if P_i is symmetric and controlled by Lemma 5 otherwise. For example, one can assign $b_i = 2\sigma_i^2/\beta_i$ by imposing $q = 2$, i.e. finite second moment, in Lemma 5.

Now, we propose `HuberUCB` that selects an arm a_t at step t based on the index

$$I_i^{\texttt{HuberUCB}}(t) = \mathrm{Hub}_{i,T_i(t-1)} + B_i(T_i(t-1), t). \tag{12}$$

The index of `HuberUCB` together with the confidence bound defined in Equation (11) dictates that if an arm is less explored, i.e. $T_i(t-1) < s_{lim}(t)$, we choose that arm, and if multiple arms satisfy this, we break the tie randomly. As t grows and for all the arms $T_i(t-1) \geq s_{lim}(t)$ is satisfied, we choose the arms according to the adaptive bonus. Thus, `HuberUCB` induces an initial forced exploration to obtain confident-enough robust estimates followed by a time-adaptive selection of arms. We present a pseudocode of `HuberUCB` in Algorithm 2.

Algorithm 2 `HuberUCB`

```
Require: $\varepsilon \in [0, 1/2)$ and $\beta > 0$
for $t = 1, \dots, n$ do
    Compute index $I_i^{\texttt{HuberUCB}}(t)$ (Equation (12)) for $i \in \{1, \dots, k\}$ using $X_1, \dots, X_{t-1}$.
    Choose arm $a_t \in \arg\max_i I_i(t)$.
    Observe a reward $X_t$.
end for
```

Regret Analysis. Now, we provide a regret upper bound for `HuberUCB`.

Theorem 3 (Upper Bound on number of pulls of suboptimal arms with `HuberUCB`) *Suppose that for all i, we have $P_i \in \mathcal{P}_{[2]}$, i.e. a reward distribution with finite variance σ_i^2. We assign $\beta \geq 4\sigma_i$ and $p = \inf_{1 \leq i \leq k} \mathbb{P}_{P_i}(|X - \mathbb{E}_{P_i}[X]| \leq \beta_i/2)$ such that $p > 5\varepsilon$ and $\varepsilon < 1/5$. We denote $\widetilde{\Delta}_{i,\varepsilon} = (\Delta_i - 2b_i)(p - \varepsilon) - 8\beta_i\varepsilon > 0$ and $\sqrt{\frac{(1-2\varepsilon)}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}} \leq \overline{\varepsilon}$.*

- *If $\widetilde{\Delta}_{i,\varepsilon} > 12\frac{\sigma_i^2}{\beta_i}\left(\sqrt{2} + 2\frac{\beta_i}{\sigma_i}\overline{\varepsilon}\right)^2$, then*

$$\mathbb{E}[T_i(n)] \leq \log(n) \max\left(\frac{32\beta_i}{3\widetilde{\Delta}_{i,\varepsilon}}, \frac{4}{(p-5\varepsilon)^2}\left(1 + 2\sqrt{2}\left(\overline{\varepsilon} \vee \frac{9}{14\sqrt{2}}\right)\right)^2\right) + 10(\log(n)+1)$$

- *If $\widetilde{\Delta}_{i,\varepsilon} \leq 12\frac{\sigma_i^2}{\beta_i}\left(\sqrt{2} + 2\frac{\beta_i}{\sigma_i}\overline{\varepsilon}\right)^2$, then*

$$\mathbb{E}[T_i(n)] \leq \log(n) \max\left(\frac{50\sigma_i^2}{9\widetilde{\Delta}_{i,\varepsilon}^2}\left(\sqrt{2}+2\frac{\beta_i}{\sigma_i}\overline{\varepsilon}\right)^2, \frac{4}{(p-5\varepsilon)^2}\left(1+2\sqrt{2}\left(\overline{\varepsilon} \vee \frac{9}{14\sqrt{2}}\right)\right)^2\right) + 10(\log(n)+1).$$

Using Theorem 3 and Lemma 1, a bound on the corrupted regret of `HuberUCB` follows immediately.

We now state a simplified version of Theorem 3 with worse but explicit constants for easier comprehension. Let us fix $\beta_i^2 = 16\sigma_i^2$ and $\varepsilon \leq 1/10$ such that $\overline{\varepsilon} = 4/(5\sqrt{\ln(9)}) \simeq 0.54$, and $p \geq 1 - \frac{4\sigma_i^2}{\beta_i^2} \geq \frac{3}{4} \geq 5\varepsilon + \frac{1}{4}$. Now, if we further assume that P_i symmetric leading to $b_i = 0$, it yields the following upper bounds.

Corollary 1 (Simplified version of Theorem 3) *Suppose that for all i, P_i is a symmetric distribution with finite variance σ_i^2. Let also denote $\widetilde{\Delta}_{i,\varepsilon} \triangleq \Delta_i\,(p - \varepsilon) - 32\sigma_i\varepsilon$ for $\varepsilon < 1/10$.*

- *If $\widetilde{\Delta}_{i,\varepsilon} > 6\sigma_i\left(1 + 4\sqrt{2}\overline{\varepsilon}\right)^2$, then*

$$\mathbb{E}[T_i(n)] \leq 43\log(n)\max\left(\frac{\sigma_i}{\widetilde{\Delta}_{i,\varepsilon}}, 10\right) + 10(\log(n) + 1).$$

- *If* $\widetilde{\Delta}_{i,\varepsilon} \leq 6\sigma_i \left(1 + 4\sqrt{2}\overline{\varepsilon}\right)^2$, *then*

$$\mathbb{E}[T_i(n)] \leq 23 \log(n) \max\left(\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2}\left(1 + 32\overline{\varepsilon}^2\right), 18\right) + 10(\log(n) + 1).$$

Remark that in this corollary, we replaced some occurrences of $\overline{\varepsilon}$ by its upper bound, which is also an upper bound on ε. Thus, the presented result is loose up to constants but lend itself to easier comprehension.

Discussions on the Upper Bound. Here, we discuss how this proposed upper bound of `HuberUCB` matches and mismatches with the lower bounds in Theorem 1.

1. *Order-optimality of Upper Bound.* `HuberUCB` achieves the logarithmic regret prescribed by the lower bound (Theorem 1) plus some additive error due to the fact that this is a UCB-type algorithm. Thus, `HuberUCB` is order optimal with respect to n.

2. *Two Regimes of Upper Bound.* When Δ_i is small compared to σ_i, we obtain an upper bound $\mathbb{E}[T_i(n)] \underset{n\to\infty}{=} \mathcal{O}\left(\log(n)\left(\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2}\overline{\varepsilon}^2\right)\right)$ from Corollary 1. $\overline{\varepsilon}^2$ is of the same order of magnitude as Equation (7) because we take ε strictly smaller than 1/2. $\overline{\varepsilon}^2$ acts as an indicator of the corruption level. The term $\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2}$ indicates the hardness due to the corrupted gaps $\widetilde{\Delta}_{i,\varepsilon}$ and echoes the hardness term $\frac{\sigma_i^2}{\Delta_i^2}$ that appears in regret upper bound of UCB for uncorrupted bandits. The hardness term $\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2}$ also appears in the corrupted lower bound (Equation (6)) as well as the heavy-tailed lower bound (Equation (5)) for $\Delta_i \ll {\sigma_i}^2$.

On the other hand, if Δ_i is larger than σ_i, we get that $\mathbb{E}[T_i(n)] = O\left(\log(n)\left(\frac{\sigma_i}{\widetilde{\Delta}_{i,\varepsilon}} \vee \overline{\varepsilon}^2 \vee 1\right)\right)$. This upper bound reflects the lower bound in Equation (7) that holds for $\Delta_i > 2\sigma_i$. This reinstates the fact that for large enough suboptimality gaps, the regret of `HuberUCB` depends solely on the corruption level than the suboptimality gap.

3. *Deviation from the Lower Bound.* The two regimes defined in the upper bound does not follow the exact distinctions made in the lower bounds. We observe that in upper bound, the distinction between regimes depend on a shifted suboptimality gap $\widetilde{\Delta}_{i,\varepsilon} \triangleq \Delta_i (p - \varepsilon) - 32\sigma_i\varepsilon$, while the lower bound depends on the corrupted suboptimality gap $\overline{\Delta}_{i,\varepsilon} \triangleq \Delta_i (1 - \varepsilon) - 2\sigma_i\varepsilon$. This difference in constants hinder the hardness regimes and corresponding constants in upper and lower bounds to match for all Δ_i, σ_i, and ε. This deviation also comes from the fact that the lower bounds proposed in Theorem 1 consider effects of heavy-tails and corruptions separately, while the upper bound of `HuberUCB` consider them in a coupled manner.

Additionally, we observe that regret of `HuberUCB` is suboptimal due to the constant additive error, which appears due to the initial forced exploration of `HuberUCB` up to $s_{lim}(t)$. Our concentration bounds and corresponding regret analysis shows that this forced exploration phase is unavoidable in order to be able to handle the case $\Delta_i \leq \sigma_i$ with `HuberUCB`. Removing this discrepancy between the lower and upper bounds would constitute an interesting future work.

5.4 Computational Details

Here, we discuss the three hyperparameters that `HuberUCB` depends on and also its computational cost.

Choice of σ and ε. In Theorem 3, we assume to know the σ and ε. In practice, these are unknown and we estimate σ^2 with a robust estimator of the variance, such as the median absolute deviation. In contrast, estimating ε is hard. We refer to Appendix C.1 for an ablation study on the choice of ε.

---

[2] We observe that the lower bound in Equation (5) depends on $\frac{\sigma_i^2}{\overline{\Delta_{i,\varepsilon}}^2}$ for $\Delta_i \ll \sigma_i$, since the first order approximation of $\log(1+x)$ is x as $x \to 0$.

Choice of β*.* Ideally, β should be larger than $\max_i\{4\sigma_i\}$. We recommend using the estimator of σ to estimate a good value of β. The choice of β reflects the difference between heavy-tailed bandits and corrupted bandits. When the data are heavy-tailed but not corrupted, Catoni (2012) shows that $\beta \simeq \sigma\sqrt{n}$ is a good choice for the scaling parameter. However, this choice is not robust to outliers and yields a linear regret in our setup (see Section 7). When there is corruption, β must remains bounded even when the sample size goes to infinity in order to retain robustness. In Appendix C.1, we present an ablation study on the choice of ε.

Computational Cost. Huber's estimator has linear complexity due to the involved Iterated Re-weighting Least Squares algorithm, which is not sequential. We have to do this at every iteration, which leads `HuberUCB` to have a quadratic time complexity. This is the computational cost of using a robust mean estimator, i.e. the Huber's estimator.

6 `SeqHuberUCB`: A Faster Robust Bandit Algorithm

In this section, we present a sequential approximation of the Huber's estimator, and we leverage it further to create a robust bandit algorithm with linear-time complexity algorithm. Here, we describe the algorithm (`SeqHuberUCB`) and its theoretical properties.

A sequential approximation of Huber's estimator. The central idea is to compute the Huber's estimator using the full historical data only in logarithmic number of steps than at every step, and in between two of these re-computations, update the estimator using only the samples observed at that step. This allows us to propose a sequential approximation of Huber's estimator, i.e. SeqHub_t, with lower computational complexity.

By fixing the update step $P_2(t) = 2^{\left\lfloor \frac{\log(t)}{\log(2)} \right\rfloor}$ before a given step $t > 0$, we define the estimator SeqHub_t by $\mathrm{SeqHub}_0 = 0$ and

$$\mathrm{SeqHub}_t = \begin{cases} H_t & \text{if } t = P_2(t), \\ H_t + \frac{\sum_{i=P_2(t)}^{t} \psi(X_i - H_t)}{\sum_{i=1}^{t} \psi'(X_i - H_t)} & \text{otherwise.} \end{cases} \tag{13}$$

Here, $H_t \triangleq \mathrm{Hub}(X_1^{P_2(t)})$ and ψ is the influence function defined in Equation (9). SeqHub_t can be conceptualized as a first order Taylor approximation of $\mathrm{Hub}(X_1^t)$ around $\mathrm{Hub}(X_1^{P_2(t)})$.

One might argue that SeqHub_t is not fully sequential rather a phased estimator as we still recompute the Huber's estimator following a geometric schedule. Thus, we still need to keep all the data in memory, leading to linear space complexity as the non-sequential Huber's estimator. But it features the good property of having a linear time complexity when computed using the prescribed geometric schedule. This implies that the `SeqHuberUCB` algorithm leveraging the sequential Huber's estimator achieves a linear time complexity.

Concentration Properties of SeqHub**.** Now, in order to propose `SeqHuberUCB` we first aim to derive the rate of convergence of SeqHub_t towards tge true Huber's mean $\mathrm{Hub}(P)$.

Theorem 4 *If the assumptions of Theorem 2 hold true, with probability larger than* $1 - 14\delta$*, we have*

$$|\mathrm{SeqHub}_t - \mathrm{Hub}(P)| \leq r_t(\delta) + \left(\frac{1}{p - \sqrt{\frac{\log(1/\delta)}{2t}} - \varepsilon} - 1 \right) r_{P_2(t)}(\delta) \tag{14}$$

for any $t > 0$*, and* $\delta \geq \exp\left(-P_2(t)\frac{128(p-5\varepsilon)^2}{49\left(1+2\overline{\varepsilon}\sqrt{2}\right)^2}\right)$*. Here,* $r_t(\delta)$ *is defined as in Equation (10).*

We observe that the confidence bound of SeqHub_t includes the confidence bound of Hub_t, i.e. $r_t(\delta)$, and an additive term proportional to $r_{P_2(t)}(\delta)$. Since $r_{P_2(t)}(\delta) \geq r_t(\delta)$ for $t \geq P_2(t)$, we can show that $|\mathrm{SeqHub}_t - \mathrm{Hub}(P)| \leq \left(p - \sqrt{\frac{\log(1/\delta)}{2t}} - \varepsilon\right)^{-1} r_{P_2(t)}(\delta)$. Thus, we obtain larger confidence bounds for SeqHub than that of Hub, and they differ approximately by a multiplicative constant $(p - \varepsilon)^{-1}$ as $t \to \infty$.

`SeqHuberUCB`: The algorithm. Now, we plug-in the sequential Huber's estimator, SeqHub, and the corresponding confidence bound (Equation (14)), instead of the Huber's estimator and the corresponding confidence bound in the `HuberUCB` algorithm. This allows us to construct the `SeqHuberUCB` algorithm that we present hereafter.

Specifically, we define the index of `SeqHuberUCB` as

$$I_i^{\texttt{SeqHuberUCB}}(t) = \mathrm{SeqHub}_{i,T_i(t-1)} + B_i^{\texttt{SeqHuberUCB}}(T_i(t-1), t). \tag{15}$$

where

$$\mathrm{SeqHub}_{i,s} = \mathrm{SeqHub}\left(X_t, \quad 1 \leq t \leq s \quad \text{such that} \quad A_t = i, \right),$$

and a confidence bound for arm i with s number of pulls is

$$B_i^{\texttt{SeqHuberUCB}}(s,t) \triangleq \begin{cases} r_s(1/t^2) + \left(\frac{1}{p - \sqrt{\frac{\log(1/\delta)}{2s}} - \varepsilon} - 1\right) r_{P_2(s)}(1/t^2) + b_i & \text{if } P_2(s) \geq s_{lim}(t) \\ \infty & \text{if } P_2(s) < s_{lim}(t). \end{cases}$$

Here, $s_{lim}(t)$, $\overline{\varepsilon}$ and b_i are same as defined for `HuberUCB`.

Similar to Corollary 1, we now present a simplified regret upper bound for `SeqHuberUCB`. Retaining the setting of Corollary 1, we assume that $\beta_i^2 = 16\sigma_i^2$, $\varepsilon \leq 1/10$ implying $\overline{\varepsilon} = 4/(5\sqrt{\ln(9)}) \simeq 0.54$, $p \geq 1 - \frac{4\sigma_i^2}{\beta_i^2} \geq \frac{3}{4} \geq 5\varepsilon + \frac{1}{4}$, and P_i symmetric so that $b_i = 0$. Further simplifying the constants yields the following regret upper bound for `SeqHuberUCB`.

Lemma 6 (Simplified Upper Bound on Regret of `SeqHuberUCB`) *Suppose that for all i, P_i is a distribution with finite variance σ_i^2. Let us also denote $\widetilde{\Delta}_{i,\varepsilon} = \Delta_i(p-\varepsilon) - 32\sigma_i\varepsilon$,*

- *If $\widetilde{\Delta}_{i,\varepsilon} > 18\sigma_i\left(1 + 4\sqrt{2\overline{\varepsilon}}\right)^2$, then*

$$\mathbb{E}[T_i(n)] \leq 128\log(n)\max\left(\frac{\sigma_i}{\widetilde{\Delta}_{i,\varepsilon}}, 2\right) + 28(\log(n)+1).$$

- *If $\widetilde{\Delta}_{i,\varepsilon} \leq 18\sigma_i\left(1 + 4\sqrt{2\overline{\varepsilon}}\right)^2$, then*

$$\mathbb{E}[T_i(n)] \leq 80\log(n)\max\left(\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2}\left(1 + 32\overline{\varepsilon}^2\right), 3\right) + 28(\log(n)+1).$$

Comparison between Regrets of `HuberUCB` and `SeqHuberUCB`. Lemma 6 yields similar regret bounds for `SeqHuberUCB` as the ones obtained for `HuberUCB` in Corollary 1. We observe that the regrets of these two algorithms only differ in n-independent constants. Specifically, regret of `SeqHuberUCB` can be approximately 3−4 times higher than that of `HuberUCB`. For simplicity of exposition, we present approximate constants in our results. A more careful analysis might yield more fine-tuned constants. Theorem 4 and experimental results (Figure 2) indicate that it is possible to have very close performances with `SeqHuberUCB` and `HuberUCB`.

7 Experimental Evaluation

In this section, we assess the experimental efficiency of `HuberUCB` and `SeqHuberUCB` by plotting the empirical regret. Contrary to the uncorrupted case, we cannot really estimate the corrupted regret in (Corrupted regret) only using the observed rewards. Instead, we use the true uncorrupted gaps that we know because we are in a simulated environment, and we estimate the corrupted regret R_n using $\sum_{i=1}^k \Delta_i \widehat{T_i}(n)$, where $\widehat{T_i}(n) = \frac{1}{M}\sum_{m=1}^M (T_i(n))_m$ is a Monte-Carlo estimation of $\mathbb{E}_{\nu^\varepsilon}[T_i(n)]$ over M experiments. We use rlberry library (Domingues et al., 2021) and Python3 for the experiments. We run the experiments on an 8 core Intel(R) Core(TM) i7-8665U CPU@1.90GHz. For each algorithm, we perform each experiment 100 times to get a Monte-Carlo estimate of regret.

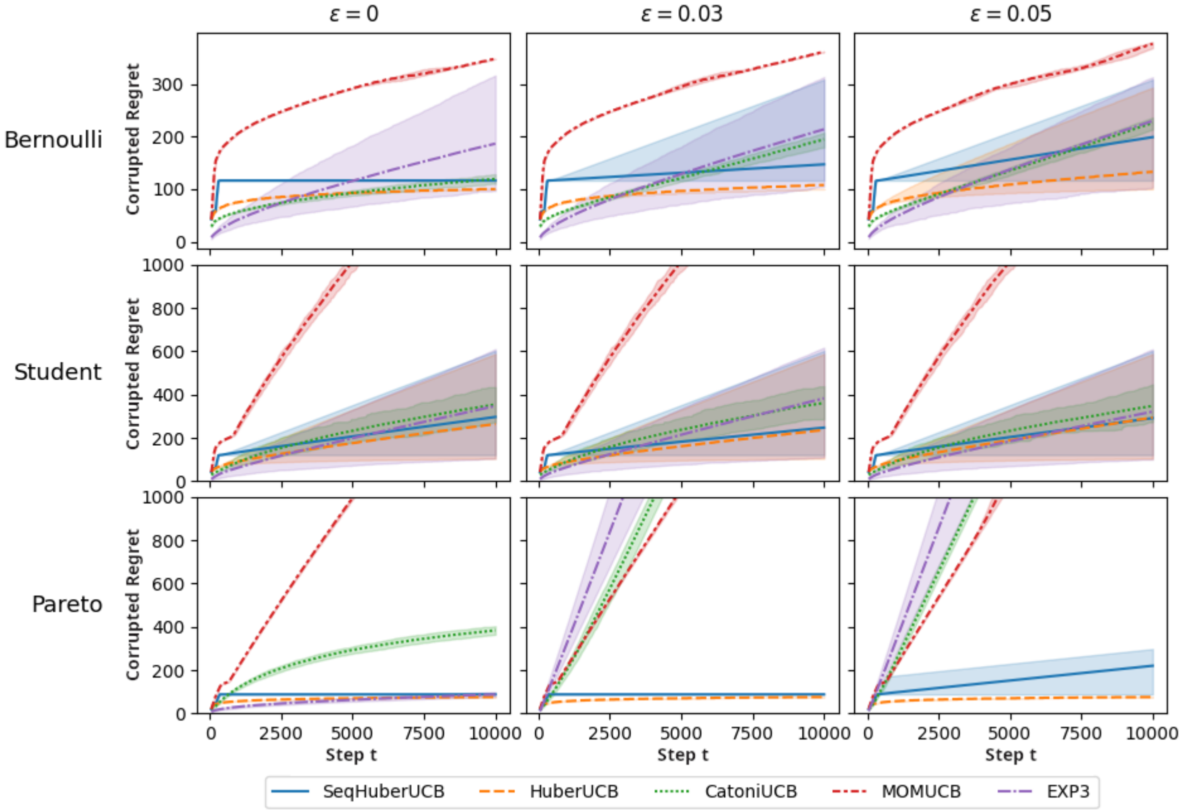

Figure 2: Cumulative regret plot of the algorithms on a corrupted Bernoulli (above), Student's (middle) and Pareto (below) reward distributions with various corruption levels ε. Lower corrupted regret indicates better performance for an algorithm.

Comparison with Bandit Algorithms for Heavy-tailed and Adversarial Settings. To the best of our knowledge, there is no existing bandit algorithm for handling unbounded stochastic corruption prior to this work. Hence, we focus on comparing ourselves to the closest settings, i.e. bandits in heavy-tailed setting and adversarial bandit algorithms. We empirically and competitively study five different algorithms: `HuberUCB`, `SeqHuberUCB`, two RobustUCB algorithms with Catoni-Huber estimator and Median of Means (MOM) (Bubeck et al., 2013), and and adversarial bandit algorithm: Exp3.

`HuberUCB` is closely related to the RobustUCB with Catoni Huber estimator, which also uses Huber's estimator but with another set of parameters and confidence intervals. The RobustUCB algorithms are tuned for uncorrupted heavy-tails. Hence, they incur linear regret in a corrupted setting. This is reflected in the experiments. *We also improve upon (Bubeck et al., 2013) as we can handle arm-dependent variances.* Exp3 is an algorithm designed for bounded Adversarial corruption, and thus, fails as the corruption is too severe.

Corrupted Bernoulli setting: In Figure 2 (above), we study a 3-armed bandits with corrupted Bernoulli distributions with means $0.1, 0.97, 0.99$. The corruption applied to this bandit problem are Bernoulli distributions with means $0.999, 0.999, 0.001$, respectively. For `HuberUCB` and `SeqHuberUCB`, we choose to use $\beta_i = 0.1\sigma_i$, which seems to work better despite the theory presented before. We plot the mean plus/minus the standard error of the result in Figure 2. We do that for the three corruption proportions ε equal to 0%, 3% and 5%. We notice that there is a short linear regret phase at the beginning due to the forced exploration performed by the algorithms. Followed by that, `HuberUCB` and `SeqHuberUCB` incur logarithmic regret. On the other hand, Exp3, Catoni Huber Agent and MOM Agent incur logarithmic regret only in the uncorrupted setting. When the data are corrupted, i.e. $\varepsilon > 0$, their regret grow linearly.

Corrupted Student setting: In Figure 2 (middle), we study a 3-armed bandits with corrupted Student's distributions with 3 degrees of freedom (finite second moment) and with means $0.1, 0.95, 1$. The corruption applied to this bandit problem are Gaussians with variance 1, and means $100, 100, -1000$ respectively. For `HuberUCB` and `SeqHuberUCB`, we choose to use $\beta_i = \sigma_i$. The results echo the observations for the Bernoulli case except that the corruption is more drastic and affect the performance even more.

Corrupted Pareto setting: In Figure 2 (bottom), we illustrate the results for a 3-armed bandits with corrupted Pareto distributions having shape parameters $3, 3, 2.1$ (i.e. they have finite second moments), and scale parameters $0.1, 0.2, 0.3$ respectively. Thus, the corresponding means are $0.15, 0.3$ and 0.57 and the standard deviations are $0.09, 0.17, 1.25$, respectively. The corruption applied to this bandit problem are Gaussians with variance 1, and centered at $100, 100, -1000$ respectively. For `HuberUCB` and `SeqHuberUCB`, we choose to use $\beta = 1.5\sigma_i$ and we also bound the bias b_i by σ_i^2/β_i. The results echo the observations for the Student's distributions.

Thus, we conclude that `HuberUCB` incur the lowest regret among the competing algorithms in the Bandits Corrupted by Nature setting, specially for higher corruption levels ε. Also, performances of `SeqHuberUCB` and `HuberUCB` are very close except for the Pareto distributions with high corruption level.

8 Conclusion

In this paper, we study the setting of Bandits corrupted by Nature that encompasses both the heavy-tailed rewards with bounded variance and unbounded corruptions in rewards. In this setting, we prove lower bounds on the regret that shows the heavy-tail bandits and corrupted bandits are strictly harder than the usual sub-Gaussian bandits. Specifically, in this setting, the hardness depends on the suboptimality gap/variance regimes. If the suboptimality gap is small, the hardness is dictated by $\sigma_i^2/\overline{\Delta}_{i,\varepsilon}^2$. Here, $\overline{\Delta}_{i,\varepsilon}$ is the corrupted sub-optimality gap, which is smaller than the uncorrupted gap Δ and thus, harder to distinguish. To complement the lower bounds, we design a robust algorithm `HuberUCB` that uses Huber's estimator for robust mean estimation and a novel concentration bound on this estimator to create tight confidence intervals. `HuberUCB` achieves logarithmic regret that matches the lower bound for low suboptimality gap/high variance regime. We also present a sequential Huber estimator that could be of independent interest and we use it to state a linear-time robust bandit algorithm, `SeqHuberUCB`, that presents the same efficiency as `HuberUCB`. Unlike existing literature, we do not need any assumption on a known bound on corruption and a known bound on the $(1+\varepsilon)$-uncentered moment, which was posed as an open problem in (Agrawal et al., 2021).

Since our upper and lower bounds disagree in the high gap/low variance regime, it will be interesting to investigate this regime further. From multi-armed bandits, we know that the tightest lower and upper bounds depend on the KL-divergence between optimal and suboptimal reward distributions. Thus, it would be imperative to study KL-divergence with corrupted distributions to better understand the Bandits corrupted by Nature problem. Also, following the reinforcement learning literature, it will be natural to extend `HuberUCB` to contextual and linear bandit settings with corruptions and heavy-tails. This will facilitate its applicability to practical problems, such as choosing treatments against pests.

References

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 111–119. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/agarwal19c.html`.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1071–1081. PMLR, 2020.

Djallel Bouneffouf. Corrupted contextual bandits: Online learning with corrupted context. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3145–3149. IEEE, 2021.

Hippolyte Bourel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Tightening Exploration in Upper Confidence Reinforcement Learning. In *International Conference on Machine Learning*, Vienna, Austria, July 2020. URL `https://hal.archives-ouvertes.fr/hal-03000664`.

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

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

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

Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using l1-penalized huber's m-estimator. *Advances in neural information processing systems*, 32, 2019.

Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Omar Darwiche Domingues, Yannis Flet-Berliac, Edouard Leurent, Pierre Ménard, Xuedong Shang, and Michal Valko. rlberry - A Reinforcement Learning Library for Research and Education, 10 2021. URL `https://github.com/rlberry-py/rlberry`.

Mohammad Hajiesmaili, Mohammad Sadegh Talebi, John Lui, Wing Shing Wong, et al. Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm. *Advances in Neural Information Processing Systems*, 33:19943–19952, 2020.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:492–518, 1964.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Martin Kamler, Marta Nesvorna, Jitka Stara, Tomas Erban, and Jan Hubert. Comparison of tau-fluvalinate, acrinathrin, and amitraz effects on susceptible and resistant populations of varroa destructor in a vial test. *Experimental and applied acarology*, 69(1):1–9, 2016.

Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL `https://www.sciencedirect.com/science/article/pii/0196885885900028`.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462, 2020.

Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, and Guillaume Lecué. Monk outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*, pp. 3782–3793. PMLR, 2019.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptioreferences 1ns. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Odalric-Ambrym Maillard. *Mathematics of Statistical Sequential Decision Making*. Habilitation à diriger des recherches, Université de Lille Nord de France, February 2019. URL `https://hal.archives-ouvertes.fr/tel-02077035`.

Timothée Mathieu. Concentration study of m-estimators using the influence function, 2021.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pp. 1642–1650. PMLR, 2016.

Stanislav Minsker. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019.

Stanislav Minsker and Mohamed Ndaoud. Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15(2):6036–6070, 2021.

Roman Pogodin and Tor Lattimore. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Uncertainty in Artificial Intelligence*, pp. 894–904. PMLR, 2020.

Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A robust univariate mean estimator is all you need. In *International Conference on Artificial Intelligence and Statistics*, pp. 4034–4044. PMLR, 2020.

Frank D Rinkevich. Detection of amitraz resistance and reduced treatment efficacy in the varroa mite, varroa destructor, within commercial beekeeping operations. *PloS one*, 15(1):e0227264, 2020.

Piotr Semkiw, Piotr Skubida, and Krystyna Pohorecka. The amitraz strips efficacy in control of varroa destructor after many years application of amitraz in apiaries. *Journal of Apicultural Science*, 57:107–121, 06 2013. doi: 10.2478/jas-2013-0012.

Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

James G. Wendel. Note on the gamma function. *American Mathematical Monthly*, 55:563, 1948.