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

Debabrota Basu, Odalric-Ambrym Maillard, Timothée Mathieu

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

Debabrota Basu DEBABROTA.BASU@INRIA.FR
Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Timothée Mathieu TIMOTHEE.MATHIEU@INRIA.FR
Université de Lille, Inria, CNRS, Centrale Lille UMR 9189 – CRIStAL, F-59000 Lille, France

Editor: Under Review for COLT 2022

Abstract

In this paper, we study the stochastic bandits problem with k unknown heavy-tailed and corrupted reward distributions or arms with time-invariant corruption distributions. At each iteration, the player chooses an arm. Given the arm, the environment returns an uncorrupted reward with probability $1-\varepsilon$ and an arbitrarily corrupted reward with probability ε. In our setting, the uncorrupted reward might be heavy-tailed and the corrupted reward might be unbounded. We prove a lower bound on the regret indicating that the corrupted and heavy-tailed bandits are strictly harder than uncorrupted or light-tailed bandits. We observe that the environments can be categorised into hardness regimes depending on the suboptimality gap Δ, variance σ, and corruption proportion ϵ. Following this, we design a UCB-type algorithm, namely HuberUCB, that leverages Huber's estimator for robust mean estimation. HuberUCB leads to tight upper bounds on regret in the proposed corrupted and heavy-tailed setting. To derive the upper bound, we prove a novel concentration inequality for Huber's estimator, which might be of independent interest.

Keywords: Unbounded corruption, Heavy-tail distributions, Huber's estimator, Regret bounds

1. Introduction

In this paper, we consider the problem of sequential decision making where the player has incomplete information about a finite number of decisions and the decisions generate corrupted returns. Specifically, we are interested in the problem of Multi-armed Bandits in a corrupted setting, or in short *Corrupted Bandits*, in which the player is presented with $k> 0$ *arms* corresponding to k unknown probability distributions. At each iteration $t \in \{1, 2, \ldots\}$, the player selects an arm and observes a sample, called *reward*, generated independently from the corresponding probability distribution. In Corrupted Bandits, *the observed reward is corrupted by* some unknown adversary or *nature*. The corruption might also differ from arm to arm. In order to model the corrupted rewards, we consider that the rewards of an arm correspond to a possibly *heavy-tailed* (with only a finite variance) and corrupted distribution instead of a well-behaved bounded or sub-Gaussian reward commonly found in literature (cite). By corrupted rewards we mean that during the experiment, the environment may randomly generate a reward that is arbitrarily different from the "true reward", which one would expect to get if the bandit was not corrupted. In such a case, we refer to these arbitrarily different rewards as *outliers*. The goal of the player is to maximize the expected reward obtained oblivious to the corruption.

A Motivating Example. Though this article focuses on the theoretical aspects of this problem, we hereby illustrate a case study with roots in agriculture that motivates us. The Varroa mite is a pest that invades apiaries and causes destruction of bee colonies. There are numerous treatments that

the beekeeper can use against the Varroa, such as applying amitraz or tau-fluvalinate (Kamler et al., 2016) based pesticides, performing organic treatments like thymus or oxalic acid or thymus (Gregorc and Planinc, 2012). Every year, the beekeeper must rotate between treatments as the varroa develops resistance to a given treatment (Rinkevich, 2020; Kamler et al., 2016). The reward of a treatment is dictated by the number of fallen varroa mites due to it. This reward function seems to be heavy-tailed and 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. Interestingly, observed corruptions are natural and non-adversarial but probably unbounded. The heavy-tailed and corrupted nature of the problem resists application of the non-robust bandit algorithms, such as UCB, and motivates us to propose our setting: *Bandits corrupted by Nature*.

Our Contributions. We consider the stochastic bandit with k arms of corrupted rewards $\{(1-\varepsilon)P_i + \varepsilon H_i\}_{i=1}^k$. Here, P_i's are uncorrupted reward distributions with heavy-tails and bounded variances, H_i's are corruption distributions with probably unbounded corruptions, and $\varepsilon \in [0, 1/2)$ is the proportion of corruption. This is equivalent to considering a setting where at every step Nature flips a coin with success probability ε. The player obtains a corrupted reward if Nature obtains 1 and otherwise, an uncorrupted reward. We call this setting 'Bandits corrupted by Nature'. Our setting encompasses both the heavy-tailed reward and unbounded corruptions. We formally define the setting and corresponding regret definition in Section 2.

In order to understand the fundamental hardness of the proposed setting, we derive the lower bounds on regret that illustrates the optimal regret achievable by any algorithm (Section 3). Lower bounds in Theorem 1 confirm that the heavy-tailed and corrupted bandits are harder than light-tailed bandits and there is an unavoidable cost to pay. It also indicates that though the logarithmic regret is asymptotically achievable, the hardness is dictated by the suboptimality gap Δ_i,[1] the variance of reward distributions σ_i and the corruption proportion ϵ. Specifically, when $\frac{\delta_i}{\sigma_i}$'s are large, i.e. the optimal and the suboptimal arms are easy to distinguish, the effect due to this factor is inverse logarithmic and the effect due to corruption is proportional to $(\log(\frac{1-\epsilon}{\epsilon}))^{-1}$. On the other hand, 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, $\overline{\Delta}_{i,\varepsilon} \triangleq \Delta_i(1-\varepsilon) - \varepsilon\sigma_i$ 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 and Robbins, 1985). Since $\overline{\Delta}_{i,\varepsilon} \leq \overline{\Delta}_i$, it shows that in heavy-tailed and corrupt settings, it is harder to distinguish the optimal and suboptimal arms. They are the same when the corruption proportion $\varepsilon = 0$.

Following the lower bounds, in Section 4, we design a robust algorithm, 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. In addition, we demonstrate that the regret of HuberUCB decomposes according to the respective values of Δ_i and σ_i:

$$R_n \asymp \underbrace{\left(\sum_{i:\Delta_i>\sigma_i} \log(n)\sigma_i\right)}_{\text{Error due to Heavy-tail}} + \underbrace{O\left(\sum_{i:\Delta_i\leq\sigma_i} \log(n)\Delta_i\frac{\sigma_i^2}{\overline{\Delta}_{i,\varepsilon}^2}\right)}_{\text{Usual } \sigma^2/\Delta \text{ error with corruption correction}} + \underbrace{O\left(\sum_i \frac{\Delta_i}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}\right)}_{\text{Constant error due to corruption}}.$$

1. Suboptimality gap of an arm is the difference in mean rewards of the optimal arm and that arm.

Thus, our upper bound allows us to segregate the errors due to heavy-tail, corruption, and corruption-correction with heavy tails. Due to the corruption, we incur at least an error $\log(n)\frac{\Delta_i}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}$, and instead of the usual $\log(n)\frac{\sigma_i}{\Delta_i}$, we obtain an error term $\log(n)\Delta_i\frac{\sigma_i^2}{\Delta_{i,\varepsilon}^2}$ with the *corrupted suboptimality gap*. These observations resonate that of the lower bound, i.e. corrupted bandits are harder than uncorrupted ones as the corruption turns distinguishing the optimal arm from the suboptimal ones strictly harder. Due to the heavy-tail behaviour, we have the first term $\sigma_i \log(n)$ that dominates the upper bound for $\Delta_i > \sigma_i$. Thus, the effect of heavy-tailedness separates from that of the corruption for large suboptimality gaps, where the optimal arm is easier to distinguish. In contrast, the effect of heavy-tailedness and corruption entangles and turns the optimal arm harder to detect, if the uncorrupted suboptimality gap is low. Thus, in the spirit of the lower bound, the upper bound of HuberUCB also shows the transition in hardness regime depending on Δ_i/σ_i. In Section 5, we experimentally illustrate the claimed performance of HuberUCB for corrupted Gaussian and Pareto environments. For brevity, we defer the detailed proofs and the parameter tuning to Appendix.

Related Work. Due to the generality of our setting, this work extends the existing methods for both the heavy-tailed and corrupted bandits. While for designing the algorithm, it leverages on the literature of robust mean estimation. Here, we connect to these three streams of literature.

Heavy-tailed bandits. Bubeck et al. (2013) introduced the heavy-tailed bandits problem and uses robust mean estimator to propose RobustUCB algorithms. 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 and Yang, 2016; Shao et al., 2018). These works rely on the assumption that a bound on the $(1+\epsilon)$-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 $\mathbb{E}[|X|^{1+\varepsilon}]$ imply 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 $(1+\epsilon)$-moment assumption with bounded centered moment.

Corrupted bandits. To our knowledge, the Bandits Corrupted by Nature is a novel setting for Bandits. Motivated by the agricultural and biological applications, we consider a non-adversarial proportion ($\varepsilon \in [0, 1/2)$) of corrupted samples with probably unbounded amount of corruptions. This is significantly different than the existing corrupted settings for bandits, which assume that the corruption is bounded and often the bound is known (Lykouris et al., 2018; Bogunovic et al., 2020).

Robust mean estimation. Our algorithm design leverages the rich literature of robust means 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 has been numerous advances bth in the heavy-tailed (Devroye et al., 2016; Catoni, 2012; Minsker, 2019) and in the corrupted settings (Lecué and Lerasle, 2020; Minsker and Ndaoud, 2021; Prasad et al., 2019, 2020; Depersin and Lecué, 2019; Lerasle et al., 2019; Lecué and Lerasle, 2020). Our work, specifically the novel concentration inequality for Huber's estimator, adds a new result in this spirit.

Notations. We denote by $\mathcal{P}$ the set of probabilîty distributions on the real line $\mathbb{R}$ and $\mathcal{P}_{[q]} \triangleq \{P \in \mathcal{P} : \mathbb{E}_P[|X|^q] < \infty\}$ the set of distributions with at least $q \geq 1$ finite moments. $\mathbb{1}\{A\}$ is the indicator function for the event A being true. We denote the mean of a distribution P_i as $\mu_i \triangleq \mathbb{E}_{P_i}[X]$.

2. Bandits corrupted by Nature: Problem setting

In this section, we present the corrupted bandits setting that we study, and introduce the notion of regret decomposition for this setting. Importantly, the regret decomposition allows us to focus on the expected number of pulls of a suboptimal arm as the central quantity to control algorithmically.

2.1. Bandits corrupted by Nature

In the setting of *Bandits corrupted by Nature*, a bandit algorithm, or *policy* π, has access to $k \in \mathbb{N}$ uncorrupted reward distributions $P_1, \dots, P_k \in \mathcal{P}_{[q]}$ and k corrupted reward distributions $H_1, \dots, H_k \in \mathcal{P}$. At each step $t \in \{0, \dots, n\}$, Nature draws a random variable $C_t \in \{0, 1\}$ from a Bernoulli distribution with mean $\varepsilon \in [0, 1/2)$. If $C_t = 1$, the reward will be drawn from the corrupted distribution H_{A_t} corresponding to the chosen arm $A_t \in \{1, \dots, k\}$. Otherwise, it will come from the uncorrupted distribution P_{A_t}. The policy π interacts with these corrupted environment by choosing an arm A_t and obtaining a reward corrupted by Nature X_t. The policy leverages these observations to choose another arm at the next step so that it maximises the total cumulative reward obtained after n steps. In Algorithm 1, we outline a pseudocode of this problem.

Algorithm 1 Bandits corrupted by Nature

Parameters: $\varepsilon \in [0, 1/2)$ and $q \geq 2$

Data: $P_1, \dots, P_k \in \mathcal{P}_{[q]}$ be the uncorrupted reward distributions and $H_1, \dots, H_k \in \mathcal{P}$ be the corrupted reward distributions.

for $t = 1, \dots, n$ **do**

  Player plays an arm $A_t \in \{1, \dots, k\}$

  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 $X_t = X'_t \mathbb{1}\{C_t = 0\} + Z_t \mathbb{1}\{C_t = 1\}$

end

We call ν^ε the law of the corrupted environment and we refer to the uncorrupted environment as ν. In this model of corruption, if $C_t = 1$, the reward will be corrupted and obtaining $C_t = 1$ does not depend on the value of X'_t but rather on the global parameter ε. Thus, this model assumes independence of corruption, i.e. a non-adversarial behaviour of the Nature. Typically, one can think of agricultural applications in which such a setting would make sense because there does not seem to be an adversary that corrupts the data. The corruption is often due to external natural confounders, such as animals or weather, which are non-adversarial.

Remark 1 *Let us highlight that we do not assume sub-Gaussian behaviour for the inlier distributions P_i. Instead, we consider only a weak moment assumption: the inlier distributions P_i have a finite variance. Thus, our setting is capable of modelling both heavy-tailed and corrupted settings. This enables our setting with a generality to develop algorithms for both of the problems. We testify this generality in the regret lower bounds and empirical performance analysis in Sections 3 and 5.*

The *Bandits corrupted by Nature* model is equivalent to the classical finite-armed stochastic bandit setting with distributions $\{(1 - \varepsilon)P_i + \varepsilon H_i\}_{i=1}^k$. The main difference is how we assess the performance of an algorithm in this model: we try to quantify the effect of corruptions on the number

of pulls or arms while ignoring the rewards coming from the H_i's. In the following section, we introduce a regret definition to explicate this connection.

2.2. Regret of corrupted bandits

We define the regret using the *uncorrupted* rewards instead of corrupted ones as otherwise it could always lead to arbitrarily bad regret. In this setting, the regret is the difference between the expected sum of uncorrupted rewards collected from the optimal arm and the expected sum of rewards collected by a bandit algorithm π operating under corrupted setting. Formally,

$$R_n \triangleq n \max_i \mathbb{E}_{P_i}[X'] - \mathbb{E}\left[\sum_{t=1}^{n} X'_t\right]. \tag{1}$$

The randomness in the second expectation is taken with respect to a bandit algorithm π that reacts to the corrupted environment and the reward are those of the uncorrupted environment $(X')_1^n$. The first step is to decompose the regret over the arms as in Lemma 1.

Lemma 1 (Decomposition of corrupted regret) *In a corrupted environment ν^ε, the regret defined in Equation (1) can be decomposed as*

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

where $T_i(n) \triangleq \sum_{t=1}^{n} \mathbb{1}\{A_t = i\}$, i.e. the number of pulls of arm i until time n, $\mathbb{E}_{\nu^\varepsilon}\left[T_i(n)\right]$ is the expected number of pulls of arm i until time n in the corrupted environment, and $\Delta_i \triangleq \max_j \mu_j - \mu_i$, which is called the suboptimality gap of arm i.

Lemma 1 states that the regret is the sum of the gaps in the uncorrupted environment times the expected number of pulls in the corrupted environment. Hence, we will focus on controlling $\mathbb{E}_{\nu^\varepsilon}\left[T_i(n)\right]$, i.e. the expected number of pulls of sub-optimal arms in the corrupted environment.

3. Lower bounds for uniformly good policies

In order to derive the lower bounds, we consider uniformly good policies on some family of environments with the set of laws $\mathcal{D} = \mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_k$, where $\mathcal{D}_i \subset \mathcal{P}$ for each $i \in \{1, \ldots, k\}$.

Definition 2 (Robust uniformly good policies) *Let $\mathcal{D}^\varepsilon = \mathcal{D}_1^\varepsilon \otimes \cdots \otimes \mathcal{D}_k^\varepsilon$ be a family of corrupted bandit environments on $\mathbb{R}$. For a corrupted environment $\nu^\varepsilon \in \mathcal{D}^\varepsilon$ with corresponding uncorrupted environment ν, let $\mu_i(\nu)$ be the mean reward of arm i in the uncorrupted setting and $\mu_*(\nu) \triangleq \max_a \mu_i(\nu)$ be the maximum mean reward. A policy π is uniformly good on $\mathcal{D}^\varepsilon$ if for any $\alpha \in (0, 1]$,*

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

In order to derive the lower bound, we rely on Lemma 2, which is a version of the change of measure argument (Burnetas and Katehakis, 1997), and can be found in (Maillard, 2019, Lemma 3.4).

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

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

Lemma 2 shows that *it is sufficient to have an upper bound on the* 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 the rest of this section, we compute upper bounds on the KL-divergences in some specific cases of heavy-tailed, such as Student's, and corrupted, such as Bernoulli, distributions. We leverage them for deriving the lower bounds on the number of pulls and hence, on the robust regret of a uniformly good policy.

Heavy-tails: Student's Distribution. 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 ν.

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

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

Corruption: Corrupted Bernoullis. We choose the corrupted Bernoulli distributions to get a lower-bound on the number of sub-optimal pulls in corrupted the setting. Let P_0, P_1 be two Bernoulli distributions on $\{0, c\}$ such that $\mathbb{P}_{P_0}(c) = \mathbb{P}_{P_1}(0) > P_{P_0}(0) = P_{P_1}(c)$. 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 $\mathrm{KL}(Q_0, Q_1)$ 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.

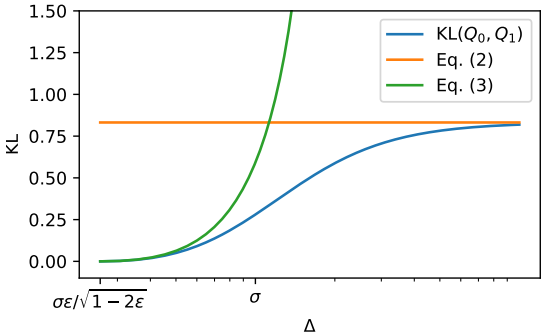


Figure 1: Plot of the KL and the bounds for σ=1 and $\varepsilon = 0.2$ (x axis is in log scale)

Lemma 4 (Control of KL-divergence for Corruptions)
Let P_0, P_1 be two Bernoulli probability distribution with $\Delta = \mathbb{E}_{P_0}[X] - \mathbb{E}_{P_1}[X]$ and $\sigma^2 = \mathrm{Var}_{P_0}(X) = \mathrm{Var}_{P_1}(X)$. There exists Q_0 and Q_1, which are some ε-corruptions of P_0 and P_1 respectively with the shifted suboptiallity gap $\overline{\Delta}_\varepsilon = \Delta(1-\varepsilon) - \varepsilon\sigma$. Given this, we have the following bounds on $\mathrm{KL}(Q_0, Q_1)$.

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

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

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

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

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

Consequences of Lemma 4. We illustrate the bounds 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 categorise the corrupted environment in three categories. For $\Delta/\sigma \in [1, +\infty)$, 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 $\Delta/\sigma \in (\varepsilon/\sqrt{1-2\varepsilon}, 1)$, 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 $\Delta/\sigma \in [0, \varepsilon/\sqrt{1-2\varepsilon}]$, 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*: $\overline{\Delta}_\varepsilon \triangleq \Delta(1-\varepsilon) - \varepsilon\sigma$.
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 neighbourhoods (see (Huber, 1965)). Thus, in Figure 1, we observe that $\mathrm{KL}(Q_0, Q_1)$ and the corresponding upper bound meet at zero.
4. *Boundedness of KL under Corruption:* Contrary to the usual setup, the KL between two corrupted Bernoullis is bounded as the corruption causes Q_0 and Q_1 to have the same support.
5. *Limitations of the Bounds:* Though the bounds are individually tight in their corresponding regimes, Figure 1 indicates looseness of them around σ. In future, it will be interesting to explore how these bounds can be pushed closer to the actual $\mathrm{KL}(Q_0, Q_1)$ for $\frac{\Delta}{\sigma}$ close to 1.

From KL Upper bounds to Regret Lower Bounds. We can substitute the results of Lemma 3 and 4 in Lemma 2 to get the lower bounds on regret of any uniformly good policy in a corrupted and heavy-tailed setting, where reward distributions belong to

$$\mathcal{P}^\varepsilon_{[2]} = \{(1-\varepsilon)P + \varepsilon H : H \in \mathcal{P} \text{ and } \mathbb{E}_P[|X|^2] < \infty\}.$$

Theorem 1 (Lower bound for heavy-tailed and corrupted bandit) *Let* $\mathcal{D}^\varepsilon_{[2]} \triangleq \mathcal{P}^\varepsilon_2 \otimes \cdots \otimes \mathcal{P}^\varepsilon_2$, *and let* $\nu \in \mathcal{D}^\varepsilon_{[2]}$. *Let* i *be a sub-optimal 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* $\overline{\Delta}_{i,\varepsilon} \triangleq \Delta_i(1-\varepsilon) - \varepsilon\sigma_i$.

• *If* $\Delta_i \leq \sigma_i/2$, *any uniformly good policy on* $\mathcal{D}^\varepsilon_{[2]}$ *satisfies*

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

• *If* $\Delta_i \geq \sigma_i$, *any uniformly good policy on* $\mathcal{D}^\varepsilon_{[2]}$ *satisfies*

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\nu^\varepsilon}[T_i(n)]}{\log(n)} \geq \frac{1}{4\log\left(\frac{6\Delta_i}{\sigma_i}\right)} \vee \frac{1}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}.$$

Due to brevity, the detailed proof is deferred to Appendix B. In Theorem 1, if Δ_i is smaller than σ_i, the term $\sigma_i^2/\overline{\Delta}_{i,\varepsilon}^2$ represents the additional error due to corruption as well as the error due to heavy-tailedness. In the case where Δ_i is larger than σ_i, we have a first term $1/4 \log\left(\frac{6\Delta_i}{\sigma_i}\right)$ that is due to heavy-tailedness. In both the cases, we obtain an unavoidable second term, $1/\log\left(\frac{1-\varepsilon}{\varepsilon}\right)$, due to corruption. We also observe that when $\overline{\Delta}_\varepsilon$ is small, we recover a lower bound with the factor $\sigma^2/2\overline{\Delta}_\varepsilon^2$, which is analogous to inverse of the KL between Gaussians with variance σ^2 and a *corrupted gap between means* $\overline{\Delta}_\varepsilon = \Delta(1-\varepsilon) - \varepsilon\sigma$. For $\epsilon = 0$, this exactly leads to the lower bound for Gaussians with uncorrupted gap of means Δ_i and variance σ_i^2. In contrast, for $\Delta_i > \sigma_i$, it is not clear whether the term $1/\log(\Delta_i/\sigma_i)$ is tight because our upper bound do not exhibit this term but a constant error that does not go to zero as Δ_i/σ_i go to infinity.

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

In this section we introduce an UCB-type algorithm adapted to our corrupted and heavy-tailed setup, see Algorithm 2. We further provide its theoretical guarantees in Theorem 3, showing that the rates of Theorem 1 are attained in some settings. Before that, we introduce and discuss Huber'sestimator.

4.1. Robust mean estimation and Huber's estimator

Now, we aim to design a UCB-type algorithm. In UCB, the focus is on mean estimation. Since the rewards are heavy-tailed and corrupted in our 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 robust properties and have been extensively studied, specially the concentration properties (Catoni, 2012).

Huber's estimator is a M-estimator, which means that it can be derived as a minimizer of some loss function. Let $X_1, \dots, X_n$ be i.i.d. random variables and $\beta > 0$, we define Huber's estimator as $\mathrm{Hub}(X_1^n) \in \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^n \rho(X_i - \theta)$, where ρ is Huber's loss function with parameter β and X_1^n is shorthand notation for $(X_1, \dots, X_n)$. ρ is a loss function that is quadratic near 0 and linear near infinity, with β giving the limit between quadratic and linear behavior. In what follows, instead of this definition we will prefer the alternative one as a root of the following equation:

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

where the influence function $\psi(x) \triangleq x\mathbb{1}\{|x| \le \beta\} + \beta\,\mathrm{sign}(x)\mathbb{1}\{|x| > \beta\}$. We prefer this representation as we will show afterwards that the properties of Huber's estimator depend on ψ.

β plays the role of a scaling parameter and depending on β, Huber's estimator is 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).

4.2. Concentration of Huber's estimator in corrupted setting

Let $\mathrm{Hub}(P)$ be the theoretical counterpart of $\mathrm{Hub}(X_1^n)$, defined for Y a random variable with law P by $\mathbb{E}[\psi(Y - \mathrm{Hub}(P))] = 0$.

Theorem 2 (Concentration of Huber's estimator) *We now state our first key result on the concentration of Huber's estimator in a corrupted and Heavy-tailed setting.*

Suppose that $X_1, \dots, X_n$ 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 having a finite variance σ^2. Let $p = \mathbb{P}_P(|Y - \mathbb{E}_P[Y]| \leq \beta/2)$ with $p > 5\varepsilon$, $\beta > 4\sigma$. Let $\overline{\varepsilon} = \sqrt{\frac{(1-2\varepsilon)}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}$ and suppose $\delta \geq \exp\left(-n\frac{128(p-5\varepsilon)^2}{49\left(1+2\overline{\varepsilon}\sqrt{2}\right)^2}\right)$. Then, with probability larger than $1-5\delta$,

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

The Theorem 2 gives us the concentration of $\mathrm{Hub}(X_1^n)$ around $\mathrm{Hub}(P)$, the Huber functional of the *inlier* distribution P, there are a few details that must be explain to understand this Theorem:
1. Value of p: For most laws that exhibit some concentration properties, the constant p is close to 1 as $\beta \geq 4\sigma$. One might also use Markov inequality to lower bound p.
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 was already present 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. See the discussion in Section 4.4.
4. Restriction on value 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 will imply later a forced exploration that we will have to do at the beginning of our algorithm.

When P is non-symmetric, we need to control the distance to the mean $|\mathrm{Hub}(P) - \mathbb{E}[X]|$ (if P is symmetric, we have $\mathrm{Hub}(P) = \mathbb{E}[X]$) to get a concentration of $\mathrm{Hub}(X_1^n)$ around $\mathbb{E}[X]$. We have the following lemma, 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 4.3.

4.3. 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 σ_i^2 its variance. We assume that we know the variances of the reward distributions. We refer to Section 4.4 for a discussion on the choice of the parameters when the reward distributions are unknown.

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

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

Denote for $s \ge s_{lim}(t) = \log(t)\frac{98}{128(p-5\varepsilon)^2}\left(1+2\sqrt{2}\left(\overline{\varepsilon} \vee \frac{9}{14\sqrt{2}}\right)\right)^2$, where $\overline{\varepsilon} = \sqrt{\frac{(1-2\varepsilon)}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}$,

$$B_i(s,t) = \frac{\sigma_i\sqrt{\frac{2\log(t^2)}{s}} + \beta_i\frac{\log(t^2)}{3s} + 2\beta_i\overline{\varepsilon}\sqrt{\frac{\log(t^2)}{s}} + 2\beta_i\varepsilon}{\left(p - \sqrt{\frac{\log(t^2)}{2s}} - \varepsilon\right)} + b_i,$$

and $B_i(s,t) = \infty$ if $s < s_{lim}(t)$, where b_i is a bound on the bias $|\mathbb{E}[X] - \mathrm{Hub}(P_i)|$. This is zero if P_i is symmetric and controlled by Lemma 5 otherwise. For example, one can take $b_i = 2\sigma_i^2/\beta_i^2$.

Algorithm 2 HuberUCB

for $t = 1, \dots, n$ **do**
  Compute $I_i(t)$ 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

Then, we introduce HuberUCB (Algorithm 2), which selects an arm a_t based on the index $I_i(t) = \mathrm{Hub}_{i,T_i(t-1)} + B_i(T_i(t-1), t)$. We now provide the main regret guarantee of this strategy.

Theorem 3 (Upper Bound on Regret of HuberUCB) *Suppose that for all i, P_i is a distribution with finite variance σ_i^2. Suppose $4\sigma_i \le \beta_i$ and $p = \inf_{1\le i\le k} \mathbb{P}_{P_i}(|X - \mathbb{E}_{P_i}[X]| \le \beta_i/2)$ with $p > 5\varepsilon$ (in particular $\varepsilon < 1/5$). Also, $\widetilde{\Delta}_{i,\varepsilon} = (\Delta_i - 2b_i)(p-\varepsilon) - 8\beta_i\varepsilon > 0$ and $\sqrt{\frac{(1-2\varepsilon)}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}} \le \overline{\varepsilon}$.*

- *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*

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

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

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

We now state a simplified version of Theorem 3 with bad but explicit constants for easier understanding. Let $\beta_i^2 = 16\sigma_i^2$, $\varepsilon \le 1/10$ so that $\overline{\varepsilon} = 4/(5\sqrt{\ln(9)}) \simeq 0.54$, $p \ge 1 - \frac{4\sigma_i^2}{\beta_i^2} \ge \frac{3}{4} \ge 5\varepsilon + \frac{1}{4}$ and suppose P_i symmetric so that $b_i = 0$. Further simplifying the constants yields the following.

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

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

$$\mathbb{E}[T_i(n)] \leq 43\log(n)\max\left(\frac{\sigma_i}{\widetilde{\Delta}_{i,\varepsilon}}, 12\overline{\varepsilon}^2+6\right) + 10(\log(n)+1).$$

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

$$\mathbb{E}[T_i(n)] \leq 23\log(n)\max\left(\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2}\left(1+32\overline{\varepsilon}^2\right), 24\overline{\varepsilon}^2+12\right) + 10(\log(n)+1).$$

We see that up to the constants, in the case $\widetilde{\Delta}_{i,\varepsilon}$ small, we recover the rate of convergences of Theorem 1. Indeed, in Theorem 1 and Corollary 1, when Δ_i is small compared to σ_i, we recover the error term $\mathbb{E}[T_i(n)] \asymp \log(n)\left(\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2} \vee \overline{\varepsilon}^2\right)$. On the other hand, if Δ_i is large compared to σ_i, we get that $\mathbb{E}[T_i(n)] \leq O\left(\log(n)\left(\frac{\sigma_i}{\widetilde{\Delta}_{i,\varepsilon}} \vee \overline{\varepsilon}^2 \vee 1\right)\right) \leq O\left(\log(n)\left(1 \vee \overline{\varepsilon}^2\right)\right)$, the rate of convergence of our algorithm is sub-optimal but this was unavoidable due to the forced exploration we have to give to our algorithm (the $s_{lim}(t)$). This forced exploration seems necessary in our approach in order to be able to handle the case $\Delta_i \leq \sigma_i$.

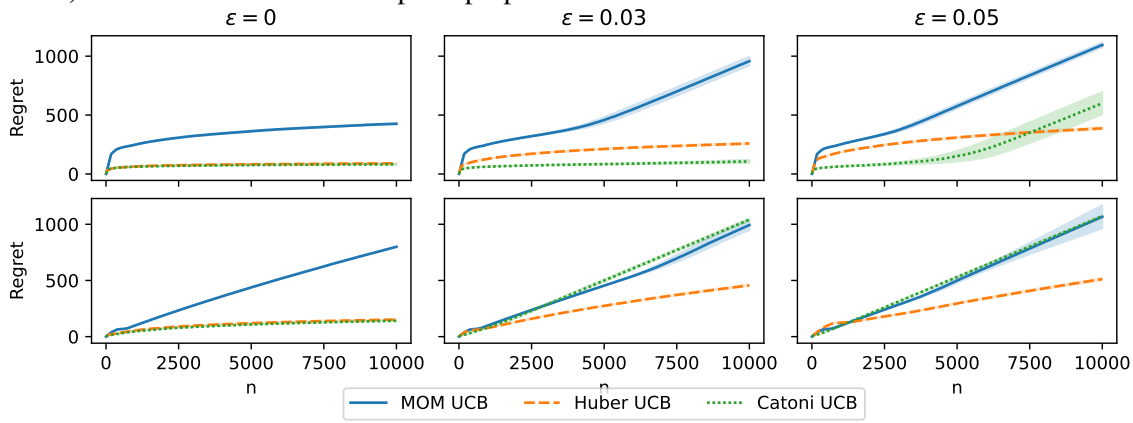
4.4. Discussion

We discuss some properties of HuberUCB and compare it with RobustUCB (Bubeck et al., 2013).
Choice of β, σ and ε. HuberUCB depends on three hyperparameters that we have to choose. 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. Ideally, β should be larger than σ by some constant factor. We recommend to use the estimator of σ to estimate a good value of β. In contrast, estimating ε is hard. However one can use the conservative upper bound $\varepsilon = 0.5$. We refer to Appendix H for an empirical study of the choice of β and ε.
Comparison with Heavy-tail bandits. Linked to the problem of chosing β is 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 5). When there is corruption, β must remains bounded when the sample size goes to infinity in order to stay robust.
Computational cost. Huber's estimator has linear complexity due to the involved Iterated Reweighting Least Squares algorithm, which is not sequential. We have to do this at every iteration, which leads HuberUCB to have quadratic time complexity. This seems to be the price for robustness.

5. Experimental Analysis

In this section, we assess the experimental efficiency of HuberUCB by plotting the empirical regret. Contrary to the uncorrupted case, we cannot really estimate the regret in Equation (1) using the observed regret. Instead, we use the theoretical uncorrupted gaps that we know because we are in a simulated environment and we estimate the regret R_n using Regret $= \sum_{i=1}^k \Delta_i \widehat{T}_i(n)$, where

Figure 2: Cumulative regret plot of the algorithms on a corrupted Gaussian (above) and Pareto (below) datasets with various corruption proportions.



$\widehat{T}_i(n) = \frac{1}{M}\sum_{m=1}^{M}(T_i(n))_m$ is a Monte-Carlo estimation of $\mathbb{E}_{\nu^\varepsilon}[T_i(n)]$ over M experiments. We used rlberry python library Domingues et al. (2021) for the experiments.

Comparison to bandit algorithms for Heavy-tailed setting. There is, to our knowledge, no existing bandit algorithm for the corruption setting prior to this work, hence we focus on comparing ourselves to the closest relatives: bandits in heavy-tailed setting. We empirically and competitively study three different algorithms: HuberUCB and two RobustUCB algorithms with Catoni Huber estimator and Median of Means (MOM) (Bubeck et al., 2013). 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 truly corrupted setting and this is reflected in the experiments. We also improve upon (Bubeck et al., 2013) as we can handle arm-dependent variances.

Corrupted Gaussian setting: In Figure 2 (top), we study a 3-armed bandits with corrupted Gaussian distributions having means $0, 0.9, 1$ and standard deviation 0.1. The corruption applied to this bandit problem are Gaussians with variance 1 and centered in $100, 100$ and -1000 respectively. For HuberUCH, we chose to use $\beta_i = 4\sigma_i$. We perform each experiment 100 times to get a Monte-Carlo error estimation. We plot the mean plus/minus the standard deviation 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 three algorithms. Followed by that, HuberUCB incurs seemingly logarithmic regret. On the other hand, for Catoni Huber Agent and MOM Agent, the regret is logarithmic only in the uncorrupted setting. When the data are corrupted, i.e. $\varepsilon > 0$, the regret becomes linear.

Corrupted pareto setting: In Figure 2 (bottom), we illustrate the results for a 3-armed bandits with corrupted pareto distributions having shape parameters $3, 4$ and 5 (i.e. 2, 3, and 4 finite moments) and scale parameters $0.1, 0.2, 0.3$. Thus, the corresponding means are $0.15, 0.27$ and 0.37 and the standard deviations are $0.3, 0.4, 0.5$, respectively. The corruption applied to this bandit problem are Gaussians with variance 1 and centered in respectively $100, 100$ and -1000 respectively. For HuberUCB, we chose to use $\beta = 3\sigma_i$ and we also bound the bias b_i by σ_i^2/β_i. The results echoes the observations for the Gaussian case except that the learning process takes more time.

6. 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 suboptimality 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. Unlike existing literature, we do not need any assumption on a known bound on corruption and a known bound on the $(1+\epsilon)$-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. Also, following the 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.

Acknowledgments

This work has been supported by the French Ministry of Higher Education and Research, the Hauts-de-France region, Inria, the MEL, the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF, and the Inria A.Ex. SR4SG project.

References

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

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

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, pages 1148–1185, 2012.

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.

Aleš Gregorc and Ivo Planinc. Use of thymol formulations, amitraz, and oxalic acid for the control of the varroa mite in honey bee (apis mellifera carnica) colonies. *Journal of Apicultural Science*, 56(2):61–69, 2012.

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*, pages 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.

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.

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*, pages 3782–3793. PMLR, 2019.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 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*, pages 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.

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*, pages 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.

Appendix A. Proof of Lemma 1: Regret Decomposition

From Equation (1), we have

$$R_n = \sum_{a=1}^{k} \sum_{t=1}^{n} \mathbb{E}\left[(\max_a \mathbb{E}_{P_a}[X'] - X'_t)\mathbb{1}\left\{A_t = a\right\}\right]$$

Then, we condition on A_t

$$\begin{aligned}\mathbb{E}\left[(\max_a \mathbb{E}_{P_a}[X'] - X'_t)\mathbb{1}\left\{A_t = a\right\}|A_t\right] &= \mathbb{1}\{A_t = a\}\mathbb{E}[\max_a \mathbb{E}_{P_a}[X'] - X'_t|A_t] \\ &= \mathbb{1}\{A_t = a\}(\max_a \mathbb{E}_{P_a}[X'] - \mu_{A_t}) \\ &= \mathbb{1}\{A_t = a\}(\max_a \mathbb{E}_{P_a}[X'] - \mu_a) = \mathbb{1}\{A_t = a\}\Delta_a\end{aligned}$$

and this stays true whatever the policy, because the policy at time t use knowledge up to time $t-1$, hence its decision does not depend on X_t. Hence, we have

$$R_n(\pi) = \sum_{a=1}^{k} \Delta_a \mathbb{E}_{\pi(\cdot|X_1^n, A_1^n)}\left[T_a(n)\right]$$

where $T_a(n)$ is with respect to the randomness of π, which is to say that we compute $\mathbb{E}[T_i(n)]$ in the corrupted setting and not in the uncorrupted one.

$$R_n = \sum_{a=1}^{k} \Delta_a \mathbb{E}_{\nu_\varepsilon}\left[T_a(n)\right].$$

Appendix B. Proof of Theorem 1: Regret Lower Bound

From Lemma 2, we have

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\nu[T_i(n)]}{\log(n)} \geq \frac{1}{\mathrm{KL}(P_0, P_1)} \vee \frac{1}{\mathrm{KL}(Q_0, Q_1)} \tag{4}$$

where P_0, P_1 are student distributions with parameter $\nu = 3$ and gap Δ_i as in Lemma 3 renormalized so that the variance is σ_i^2, and Q_0, Q_1 are as in Lemma 4 with gap Δ_i and variance σ_i. From Lemma 3, we get

$$KL(P_1, P_2) \leq \begin{cases} \frac{3^{\nu-1}(\nu+1)^2\Delta_i^2}{\sigma_i^2(\nu-2)} & \text{if } \Delta_i^2 \leq \frac{\nu-2}{\nu}\sigma_i^2 \\ \frac{\nu+1}{2}\log\left(\frac{\nu}{\sigma_i^2(\nu-2)}\Delta_i^2\right) + \log\left(3^\nu \frac{(\nu+1)^2}{\nu}\right) & \text{if } \Delta_i^2 > \frac{\nu-2}{\nu}\sigma_i^2 \end{cases}$$

Hence, with $\nu = 3$,

$$KL(P_1, P_2) \leq \begin{cases} \frac{144\Delta_i^2}{\sigma_i^2} & \text{if } \Delta_i^2 \leq \sigma_i^2/3 \\ 2\log\left(\frac{3\Delta_i^2}{\sigma_i^2}\right) + \log(144) & \text{if } \Delta_i^2 > \sigma_i^2/3 \end{cases} \tag{5}$$

First setting: If $\Delta_i \leq \sigma_i/2$, then $\Delta_i^2 \leq \sigma_i^2/3$ and from Equations (5) and (4) and Lemma 4,

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\nu[T_i(n)]}{\log(n)} \geq \frac{\sigma_i^2}{144\Delta_i^2} \vee \frac{\sigma_i}{\overline{\Delta}_{i,\varepsilon} \log\left(1 + \frac{\overline{\Delta}_{i,\varepsilon}}{\sigma_i - \overline{\Delta}_{i,\varepsilon}}\right)}$$

Then, use that $\overline{\Delta}_{i,\varepsilon} \leq \Delta_i \leq \sigma_i/2$,

$$\log\left(1 + \frac{\overline{\Delta}_{i,\varepsilon}}{\sigma_i - \overline{\Delta}_{i,\varepsilon}}\right) \leq \frac{\overline{\Delta}_{i,\varepsilon}}{\sigma_i - \overline{\Delta}_{i,\varepsilon}} \leq \frac{2\overline{\Delta}_{i,\varepsilon}}{\sigma_i}$$

Hence, considering also the term from Equation (2) in Lemma 4,

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\nu[T_i(n)]}{\log(n)} \geq \frac{\sigma_i^2}{144\Delta_i^2} \vee \frac{\sigma_i^2}{2\overline{\Delta}_{i,\varepsilon}^2} \vee \frac{1}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}$$

We weaken this inequality to the simplified version found in Theorem 1 by dropping the second term on the right-hand side for better interpretability.

Second setting: If $\Delta_i > \sigma_i$, then $\Delta_i^2 > \sigma_i^2/3$ and from Equations (5) and (4) and Lemma 4,

$$\begin{aligned}
\liminf_{n\to\infty} \frac{\mathbb{E}_\nu[T_i(n)]}{\log(n)} &\geq \frac{1}{2\log\left(\frac{3\Delta_i^2}{\sigma_i^2}\right) + \log(144)} \vee \frac{1}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)} \\
&= \frac{1}{4\log\left(\frac{\sqrt{3}\Delta_i}{\sigma_i}\right) + 4\log\left(2\sqrt{3}\right)} \vee \frac{1}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)} \\
&= \frac{1}{4\log\left(\frac{6\Delta_i}{\sigma_i}\right)} \vee \frac{1}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}
\end{aligned}$$

Appendix C. Upper Bounds on KL-divergence: Student's and Corrupted Bernoulli

C.1. Proof of Lemma 3: Student's Distribution

First, we compute the χ^2 divergence between the two laws f_a and f_0. We have, for any $a \in \mathbb{R}$

$$\begin{aligned}
\mathrm{d}_{\chi^2}(f_a, f_0) &= \int \frac{(f_a(x) - f_0(x))^2}{f_0(x)} \mathrm{d}x \\
&= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \int_{\mathbb{R}} \left(\frac{1}{\left(1 + \frac{(x-a)^2}{k}\right)^{\frac{k+1}{2}}} - \frac{1}{\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}} \right)^2 \left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}} \mathrm{d}x \\
&= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \int_{\mathbb{R}} \frac{\left(\left(1 + \frac{(x-a)^2}{k}\right)^{\frac{k+1}{2}} - \left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}\right)^2}{\left(1 + \frac{(x-a)^2}{k}\right)^{k+1} \left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}} \mathrm{d}x \\
&= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \left(\int_{\mathbb{R}} \frac{\mathrm{d}x}{\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}} - 2\int_{\mathbb{R}} \frac{\mathrm{d}x}{\left(1 + \frac{(x-a)^2}{k}\right)^{\frac{k+1}{2}}} + \int_{\mathbb{R}} \frac{\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}}{\left(1 + \frac{(x-a)^2}{k}\right)^{k+1}} \mathrm{d}x \right).
\end{aligned}$$

The first two terms are respectively equal to 1 and -2 using the fact that the student distribution integrate to 1. Then, we do the change of variable $y = x - a$ in the last integral to get

$$\mathrm{d}_{\chi^2}(f_a, f_0) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \int_{\mathbb{R}} \frac{\left(1 + \frac{(y+a)^2}{k}\right)^{\frac{k+1}{2}}}{\left(1 + \frac{y^2}{k}\right)^{k+1}} \mathrm{d}y - 1.$$

this is a polynomial of degree k in the variable a. We have the following Lemma proven in Section G.1.

Lemma 6 *For $a \in \mathbb{R}$ and $k \geq 0$, we have the following algebraic inequality.*

$$\int_{\mathbb{R}} \frac{\left(1 + \frac{(y+a)^2}{k}\right)^{\frac{k+1}{2}}}{\left(1 + \frac{y^2}{k}\right)^{k+1}} \mathrm{d}y \leq 2\frac{a^2}{\sqrt{k}}(k+1)^2 \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1} + \int_{\mathbb{R}} \frac{(1 + y^2/k)^{\frac{k+1}{2}}}{\left(1 + \frac{y^2}{k}\right)^{k+1}} \mathrm{d}y$$

Using this lemma, and because we recognize up to a constant the integral of the student distribution on $\mathbb{R}$ in the right hand side, we have

$$\begin{aligned}
\mathrm{d}_{\chi^2}(f_a, f_0) &= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \left(2\frac{a^2}{\sqrt{k}}(k+1)^2 \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1} + \int_{\mathbb{R}} \frac{(1 + y^2/k)^{\frac{k+1}{2}}}{\left(1 + \frac{y^2}{k}\right)^{k+1}} \mathrm{d}y \right) - 1 \\
&\leq \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} 2\frac{a^2}{\sqrt{k}}(k+1)^2 \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1}
\end{aligned}$$

then, use that for any $k \geq 1$, $\Gamma(\frac{k+1}{2}) \leq \Gamma(\frac{k}{2})\sqrt{k/2}$ from Wendel (1948), hence

$$\mathrm{d}_{\chi^2}(f_a, f_0) \leq \frac{a^2(k+1)^2\sqrt{2}}{k\sqrt{\pi}} \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1} \leq \frac{a^2(k+1)^2}{k} \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1}$$

Then, we use the link between KL divergence and χ^2 divergence to get the result.

$$\begin{aligned}
\mathrm{KL}(f_a, f_0) &\leq \log(1 + \mathrm{d}_{\chi^2}(f_a, f_0)) \\
&\leq \log\left(1 + \frac{a^2(k+1)^2}{k} \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1}\right) \qquad (6)
\end{aligned}$$

Then, use that

$$\log\left(1 + \frac{a^2(k+1)^2}{k} \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1}\right) \leq \begin{cases} \log\left(1 + 3^{k-1}\frac{(k+1)^2}{k}a^2\right) & \text{if } a < 1 \\ \log\left(1 + 3^{k-1}\frac{(k+1)^2}{k}a^{k+1}\right) & \text{if } a \geq 1 \end{cases}$$

hence, using that $1 \leq 3^{k-1}\frac{(k+1)^2}{k}a^{k+1}$

$$\log\left(1 + \frac{a^2(k+1)^2}{k} \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1}\right) \leq \begin{cases} 3^{k-1}\frac{(k+1)^2}{k}a^2 & \text{if } a < 1 \\ (k+1)\log(a) + \log\left(3^k\frac{(k+1)^2}{k}\right) & \text{if } a \geq 1 \end{cases}$$

Inject this in Equation (6) to get the result.

C.2. Proof of Lemma 4: Corrupted Bernoulli Distribution

Let $\alpha \in (0, 1/2)$ and $c > 0$. Define
$P_0 = (1-\alpha)\delta_0 + \alpha\delta_c$,
$P_1 = \alpha\delta_0 + (1-\alpha)\delta_c$,
$Q_0 = (1-\varepsilon)(1-\alpha)\delta_0 + (1-(1-\varepsilon)(1-\alpha))\delta_c$,
$Q_1 = (1-(1-\varepsilon)(1-\alpha))\delta_0 + (1-\varepsilon)(1-\alpha)\delta_c$.

One can check that $Q_0 = (1-\varepsilon)P_0\varepsilon\delta_c$ and $Q_1 = (1-\varepsilon)P_1 + \varepsilon\delta_0$ and hence Q_0 and Q_1 are in the ε-corrupted neighborhood of respectively P_0 and P_1.

We have

$$\begin{aligned}
\mathrm{KL}(Q_0, Q_1) &= \sum_{k\in\{0,c\}} \mathbb{P}_{Q_0}(X=k) \log\left(\frac{\mathbb{P}_{Q_0}(X=k)}{\mathbb{P}_{Q_1}(X=k)}\right) \\
&= (1-\varepsilon)(1-\alpha)\log\left(\frac{(1-\varepsilon)(1-\alpha)}{1-(1-\varepsilon)(1-\alpha)}\right) + (1-(1-\varepsilon)(1-\alpha))\log\left(\frac{1-(1-\varepsilon)(1-\alpha)}{(1-\varepsilon)(1-\alpha)}\right) \\
&= ((1-\varepsilon)(1-\alpha) - (1-(1-\varepsilon)(1-\alpha)))\log\left(\frac{(1-\varepsilon)(1-\alpha)}{1-(1-\varepsilon)(1-\alpha)}\right) \\
&= (1-2\varepsilon-2\alpha+2\varepsilon\alpha)\log\left(1+\frac{1-2\varepsilon-2\alpha+2\varepsilon\alpha}{\varepsilon+\alpha-\varepsilon\alpha}\right)
\end{aligned}$$

Then, note that $\Delta = \mathbb{E}_{P_1}[X] - \mathbb{E}_{P_0}[X] = (1-2\alpha)c$ and $\sigma^2 = \mathrm{Var}_{P_0}(X) = \mathrm{Var}_{P_1}(X) = \alpha(1-\alpha)c^2$. Hence, $c = \sqrt{\Delta^2+\sigma^2}$ and $\alpha = \frac{1}{2}\left(1 - \Delta/\sqrt{\Delta^2+\sigma^2}\right)$.

$$\begin{aligned}
\mathrm{KL}(Q_0, Q_1) &= \left(1-2\varepsilon-\left(1-\frac{\Delta}{\sqrt{\Delta^2+\sigma^2}}\right)(1-\varepsilon)\right)\log\left(1+\frac{1-2\varepsilon-\left(1-\frac{\Delta}{\sqrt{\Delta^2+\sigma^2}}\right)(1-\varepsilon)}{\varepsilon+\frac{1}{2}\left(1-\frac{\Delta}{\sqrt{\Delta^2+\sigma^2}}\right)(1-\varepsilon)}\right) \\
&= \left(\frac{\Delta}{\sqrt{\Delta^2+\sigma^2}}(1-\varepsilon)-\varepsilon\right)\log\left(1+\frac{\frac{\Delta}{\sqrt{\Delta^2+\sigma^2}}(1-\varepsilon)-\varepsilon}{\frac{1}{2}(1+\varepsilon)-\frac{1}{2}\frac{\Delta}{\sqrt{\Delta^2+\sigma^2}}(1-\varepsilon)}\right)
\end{aligned}$$

In the setting $\sigma > \Delta$, we have the bound

$$\mathrm{KL}(Q_0, Q_1) \le \left(\frac{\Delta}{\sigma}(1-\varepsilon)-\varepsilon\right)\log\left(1+2\frac{\frac{\Delta}{\sigma}(1-\varepsilon)-\varepsilon}{1-\left(\frac{\Delta}{\sigma}(1-\varepsilon)-\varepsilon\right)}\right)$$

On the other hand, if $\varepsilon > 0$, we have

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

Appendix D. Regret Upper Bounds for HuberUCB: Proofs of Theorem 3 and Corollary 1

If $A_t = i$ then at least one of the following four inequalities is true:

$$\widehat{\text{Hub}}_{1,T_1(t-1)} + B_1(T_1(t-1), t) \leq \mu_1 \tag{7}$$

or

$$\widehat{\text{Hub}}_{i,T_i(t-1)} \geq \mu_i + B_i(T_i(t-1), t) \tag{8}$$

or

$$\Delta_i < 2B_i(T_i(t-1), t) \tag{9}$$

or

$$T_1(t-1) < s_{lim}(t) = \frac{98\log(t)}{128\,(p-5\varepsilon)^2}\left(1 + 2\sqrt{2}\left(\bar{\varepsilon} \vee \frac{9}{14\sqrt{2}}\right)\right)^2 \tag{10}$$

Indeed, if $T_i(t-1) < s_{lim}(t)$, then $B_i(T_i(t-1), t) = \infty$ and Inequality (9) is true. On the other hand, if $T_i(t-1) \geq s_{lim}(t)$, then we have $B_i(T_i(t-1), t)$ is finite and all four inequalities are false, then,

$$\begin{aligned}
\widehat{\text{Hub}}_{1,T_1(t-1)} + B_1(T_1(t-1), t) &> \mu_1 \\
&= \mu_i + \Delta_i \\
&\geq \mu_i + 2B_i(T_i(t-1), n) \\
&\geq \mu_i + 2B_i(T_i(t-1), t) \\
&\geq \widehat{\text{Hub}}_{i,T_i(t-1)} + B_i(T_i(t-1), t)
\end{aligned}$$

which implies that $A_t \neq i$.

Step 1. We have that $\mathbb{P}\,(7 \text{ is true}) \leq 5/t$.
PROOF:

Then, we have that,

$$\begin{aligned}
\mathbb{P}\left(\widehat{\text{Hub}}_{1,T_1(t-1)} + B_1(T_1(t-1), t) \leq \mu_1\right) &\leq \sum_{s=1}^{t} \mathbb{P}\left(\widehat{\text{Hub}}_{1,s} + B_1(s,t) \leq \mu_1\right) \\
&= \sum_{s=\lceil s_{lim}(t)\rceil}^{t} \mathbb{P}\left(\widehat{\text{Hub}}_{1,s} - \mu_1 \leq -B_1(s,t)\right)
\end{aligned}$$

Then, use Theorem 2, we get

$$\begin{aligned}
\mathbb{P}\left(\widehat{\text{Hub}}_{1,T_1(t-1)} + B_1(T_1(t-1), t) \leq \mu_1\right) &\leq \sum_{s=\lceil s_{lim}(t)\rceil}^{t} 5e^{-\log(t^2)} \\
&\leq \sum_{s=\lceil s_{lim}(t)\rceil}^{t} \frac{5}{t^2} \leq \frac{5}{t}.
\end{aligned}$$

Step 2. Similarly, for arm i, we have

$$\mathbb{P}\left(\widehat{\mathrm{Hub}}_{i,T_i(t-1)} \geq \mu_i + B_i(T_i(t-1),t)\right) \leq \frac{5}{t}$$

PROOF: We have,

$$\begin{aligned}\mathbb{P}\left(\widehat{\mathrm{Hub}}_{i,T_i(t-1)} \geq \mu_i + B_i(T_i(t-1),t)\right) &\leq \sum_{s=\lceil s_{lim}(t)\rceil}^{t} \mathbb{P}\left(\widehat{\mathrm{Hub}}_{i,s} - \mu_i \geq B_i(s,t)\right) \\ &\leq \sum_{s=\lceil s_{lim}(t)\rceil}^{t} 5e^{-\log(t^2)} \leq \frac{5}{t}.\end{aligned}$$

Step 3. Let $v \in \mathbb{N}$. If one of the two following conditions are true, then for all t such that $T_i(t-1) \geq v$, we have $\Delta_i \geq 2B_i(T_i(t-1),t)$ (i.e. Equation (9) is false).
Condition 1: if $\widetilde{\Delta}_{i,\varepsilon} > 12\frac{\sigma_i^2}{\beta_i}\left(\sqrt{2} + 2\frac{\beta_i}{\sigma_i}\overline{\varepsilon}\right)^2$ and $v \leq \log(n)\frac{96\beta_i}{9\widetilde{\Delta}_{i,\varepsilon}}$.
Condition 2: if $\widetilde{\Delta}_{i,\varepsilon} \leq 12\frac{\sigma_i^2}{\beta_i}\left(\sqrt{2} + 2\frac{\beta_i}{\sigma_i}\overline{\varepsilon}\right)^2$ and $v \leq \frac{50}{9\widetilde{\Delta}_{i,\varepsilon}^2}\left(\sigma_i\sqrt{2} + 2\beta_i\overline{\varepsilon}\right)^2\log(n)$.
PROOF: We search for the smallest value $v \geq s_{lim}(n)$ such that Δ_i verifies

$$\Delta_i \geq 2B_i(v,n) = 2\frac{\sigma_i\sqrt{\frac{2\log(n^2)}{v}} + \beta\frac{\log(n^2)}{3v} + 2\overline{\varepsilon}\beta_i\sqrt{\frac{\log(n^2)}{v}} + 2\beta_i\varepsilon}{\left(p - \sqrt{\frac{\log(n^2)}{2v}} - \varepsilon\right)} + 2b_i.$$

First, we simplify the expression, having that $v \geq s_{lim}(n)$, we have

$$\frac{\log(n^2)}{2v} \leq \frac{128(p-5\varepsilon)^2}{98(1+9/7)^2} \leq \frac{(p-\varepsilon)^2}{4},$$

hence we simplify to

$$\Delta_i \geq \frac{4}{(p-\varepsilon)}\left(\sigma_i\sqrt{\frac{2\log(n^2)}{v}} + \beta_i\frac{\log(n^2)}{3v} + 2\beta_i\overline{\varepsilon}\sqrt{\frac{\log(n^2)}{v}} + 2\beta_i\varepsilon\right) + 2b_i$$

let us denote $\widetilde{\Delta}_{i,\varepsilon} = (\Delta_i - 2b_i)(p-\varepsilon) - 8\beta_i\varepsilon$, we are searching for v such that

$$\beta_i\frac{\log(n^2)}{3v} + \sqrt{\frac{\log(n^2)}{v}}\left(\sigma_i\sqrt{2} + 2\beta_i\overline{\varepsilon}\right) - \frac{\widetilde{\Delta}_{i,\varepsilon}}{4} \leq 0$$

This is a second order polynomial in $\sqrt{\log(n^2)/v}$.
If $\widetilde{\Delta}_{i,\varepsilon} > 0$, then the smallest $v > 0$ is

$$\sqrt{\frac{\log(n^2)}{v}} = \frac{3}{2\beta_i}\left(-\left(\sigma_i\sqrt{2} + 2\overline{\varepsilon}\beta_i\right) + \sqrt{\left(\sigma_i\sqrt{2} + 2\beta_i\overline{\varepsilon}\right)^2 + \frac{\widetilde{\Delta}_{i,\varepsilon}\beta_i}{3}}\right).$$

First setting: if $\widetilde{\Delta}_{i,\varepsilon} > 12\frac{\sigma_i^2}{\beta_i}\left(\sqrt{2}+2\frac{\beta_i}{\sigma_i}\overline{\varepsilon}\right)^2$,

In that case, we have

$$\sqrt{\frac{\log(n^2)}{v}} \geq \frac{3}{2\beta_i}\left(-\left(\sigma_i\sqrt{2}+2\beta_i\overline{\varepsilon}\right)+\sqrt{\frac{\beta_i\widetilde{\Delta}_{i,\varepsilon}}{3}}\right) \geq \frac{3}{2\beta_i}\sqrt{\frac{\beta_i\widetilde{\Delta}_{i,\varepsilon}}{12}} = \sqrt{\frac{9\widetilde{\Delta}_{i,\varepsilon}}{48\beta_i}}$$

Hence, $v \leq \log(n)\frac{96\beta_i}{9\widetilde{\Delta}_{i,\varepsilon}}$.

Second setting: 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 we use Lemma 9, using that

$$\frac{\widetilde{\Delta}_{i,\varepsilon}\beta_i}{3\left(\sigma_i\sqrt{2}+2\beta_i\overline{\varepsilon}\right)^2} \leq 4$$

and the fact that $\frac{\sqrt{1+4}-1}{4} \geq \frac{3}{10}$, we get,

$$\sqrt{\frac{\log(n^2)}{v}} \geq \frac{3\widetilde{\Delta}_{i,\varepsilon}}{5\left(\sigma_i\sqrt{2}+2\beta_i\overline{\varepsilon}\right)}$$

Hence,

$$v \leq \frac{50}{9\widetilde{\Delta}_{i,\varepsilon}^2}\left(\sigma_i\sqrt{2}+2\beta_i\overline{\varepsilon}\right)^2\log(n).$$

Step 4. Using All the previous steps, we prove the theorem. PROOF: We have

$$\begin{aligned}
\mathbb{E}[T_i(n)] &= \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{1}\{A_t=i\}\right] \\
&\leq \lfloor\max(v,s_{lim}(n))\rfloor + \mathbb{E}\left[\sum_{t=\lfloor\max(v,s_{lim}(n))\rfloor+1}^{n}\mathbb{1}\{A_t=i \text{ and (9) is false}\}\right] \\
&\leq \lfloor\max(v,s_{lim}(n))\rfloor + \mathbb{E}\left[\sum_{t=\lfloor\max(v,s_{lim}(n))\rfloor+1}^{n}\mathbb{1}\{\text{(7) or (8) or (10) is true}\}\right] \\
&= \lfloor\max(v,s_{lim}(n))\rfloor + \sum_{t=\lfloor\min(v,s_{lim}(n))\rfloor+1}^{n}\mathbb{P}\left(\text{(7) or (8) is true}\right) \\
&\leq \lfloor\max(v,s_{lim}(n))\rfloor + 2\sum_{t=\lfloor\min(v,s_{lim}(n))\rfloor+1}^{n}\frac{5}{t}
\end{aligned}$$

using the harmonic series bound by $\log(n)+1$, we have

$$\mathbb{E}[T_i(n)] \leq \max(v,s_{lim}(n)) + 10(\log(n)+1)$$

Then, we replace the value of v,

First setting: $\widetilde{\Delta}_{i,\varepsilon} > 12\frac{\sigma_i^2}{\beta_i}\left(\sqrt{2}+2\frac{\beta_i}{\sigma_i}\overline{\varepsilon}\right)^2$

$$\mathbb{E}[T_i(n)] \leq \log(n)\max\left(\frac{96\beta_i}{9\widetilde{\Delta}_{i,\varepsilon}}, \frac{4}{(p-5\varepsilon)^2}\left(1+2\sqrt{2}\left(\overline{\varepsilon}\vee\frac{9}{14\sqrt{2}}\right)\right)^2\right)+10(\log(n)+1)$$

Second setting: 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

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

This concludes the proof of Theorem 3.

D.1. Proof of Corollary 1: Simplified Upper Bound of HuberUCB

Replacing β_i by $4\sigma_i$, we have

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

$$\mathbb{E}[T_i(n)] \leq \log(n)\max\left(\frac{128\sigma_i}{3\widetilde{\Delta}_{i,\varepsilon}}, \frac{4}{(p-5\varepsilon)^2}\left(1+2\sqrt{2}\left(\overline{\varepsilon}\vee\frac{9}{14\sqrt{2}}\right)\right)^2\right)+10(\log(n)+1)$$

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

$$\mathbb{E}[T_i(n)] \leq \log(n)\max\left(\frac{50\sigma_i^2}{9\widetilde{\Delta}_{i,\varepsilon}^2}\left(\sqrt{2}+8\overline{\varepsilon}\right)^2, \frac{4}{(p-5\varepsilon)^2}\left(1+2\sqrt{2}\left(\overline{\varepsilon}\vee\frac{9}{14\sqrt{2}}\right)\right)^2\right)+10(\log(n)+1).$$

Then, we use that

$$\begin{aligned}\left(1+2\sqrt{2}\left(\overline{\varepsilon}\vee\frac{9}{14\sqrt{2}}\right)\right)^2 &\leq 2\left(1+\left(2\sqrt{2}\left(\overline{\varepsilon}\vee\frac{9}{14\sqrt{2}}\right)\right)^2\right)\\ &= 2+8\left(\overline{\varepsilon}^2\vee\frac{81}{392}\right) \leq 8\overline{\varepsilon}^2+2+\frac{648}{392} \leq 8\overline{\varepsilon}^2+4\end{aligned}$$

and that $p-5\varepsilon \geq 1/4$, to get

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

$$\begin{aligned}\mathbb{E}[T_i(n)] &\leq \log(n)\max\left(\frac{128\sigma_i}{3\widetilde{\Delta}_{i,\varepsilon}}, 512\overline{\varepsilon}^2+256\right)+10(\log(n)+1)\\ &= \frac{128}{3}\log(n)\max\left(\frac{\sigma_i}{\widetilde{\Delta}_{i,\varepsilon}}, 12\overline{\varepsilon}^2+6\right)+10(\log(n)+1)\\ &\leq 43\log(n)\max\left(\frac{\sigma_i}{\widetilde{\Delta}_{i,\varepsilon}}, 12\overline{\varepsilon}^2+6\right)+10(\log(n)+1)\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[T_i(n)] &\leq \log(n) \max\left(\frac{50\sigma_i^2}{9\widetilde{\Delta}_{i,\varepsilon}^2}\left(\sqrt{2} + 8\overline{\varepsilon}\right)^2, 512\overline{\varepsilon}^2 + 256\right) + 10(\log(n)+1) \\
&\leq \log(n) \max\left(\frac{100\sigma_i^2}{9\widetilde{\Delta}_{i,\varepsilon}^2}\left(2 + 64\overline{\varepsilon}^2\right), 512\overline{\varepsilon}^2 + 256\right) + 10(\log(n)+1) \\
&\leq 23\log(n) \max\left(\frac{\sigma_i^2}{\widetilde{\Delta}_{i,\varepsilon}^2}\left(1 + 32\overline{\varepsilon}^2\right), 24\overline{\varepsilon}^2 + 12\right) + 10(\log(n)+1)
\end{aligned}$$

Appendix E. Proof of Theorem 2: Concentration of Huber's Estimator

First, we control the deviations of Huber's estimator using the deviations of $\psi(X - \mathrm{Hub}(X_1^n))$. We will need the following lemma to control the variance of $\psi(X - \mathrm{Hub}(X_1^n))$, which will in turn allow us to control its deviation with Lemma 8.

Lemma 7 (Controlling Variance of Influence of Huber's Estimator) *Suppose that $Y_1, \ldots, Y_n$ are i.i.d with law P. Then*

$$\mathrm{Var}(\psi(Y - \mathrm{Hub}(P))) \leq \mathrm{Var}(Y) = \sigma^2$$

Lemma 8 (Concentrating Huber's Estimator by Concentrating the Influence) *Suppose that $X_1, \ldots, X_n$ are i.i.d with law $(1-\varepsilon)P + \varepsilon H$ for some $H \in \mathcal{P}$ and proportion of outliers $\varepsilon \in (0, 1/2)$. Then, for any $\eta > 0$ and $\lambda \in (0, \beta/2]$, we have*

$$\mathbb{P}(|\mathrm{Hub}(X_1^n) - \mathrm{Hub}(P)| \geq \lambda) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \psi(X_i - \mathrm{Hub}(P))\right| \geq \lambda\left(p - \eta - \varepsilon\right)_+\right) + 2e^{-2n\eta^2}$$

where $p = \mathbb{P}(|Y - \mathbb{E}[X]| \leq \beta/2)$.

Then, using these Lemmas, we can prove the theorem.

Step 1. For any $\delta \in (0,1)$, with probability larger than $1 - 3\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n \psi(X_i - \mathrm{Hub}(P))\right| \leq \sigma\sqrt{\frac{2\ln(1/\delta)}{n}} + \beta\frac{\ln(1/\delta)}{2n} + 2\beta\varepsilon + 2\beta\sqrt{\frac{\ln(1/\delta)(1-2\varepsilon)}{n\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}. \quad (11)$$

PROOF: Write that $X_i = (1 - W_i)Y_i + W_i Z_i$ where $W_1, \ldots, W_n$ are i.i.d $\{0,1\}$ Bernoulli random variable with mean ε, $Y_1, \ldots, Y_n$ are i.i.d $\sim P$ and $Z_1, \ldots, Z_n$ are i.i.d with law H, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \psi(X_i - \mathrm{Hub}(P))\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^n \psi(Y_i - \mathrm{Hub}(P)) + \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{W_i = 1\}\left(\psi(Z_i - \mathrm{Hub}(P)) - \psi(Y_i - \mathrm{Hub}(P))\right)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \psi(Y_i - \mathrm{Hub}(P))\right| + 2\beta\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{W_i = 1\}
\end{aligned}$$

Remark that by definition of $\mathrm{Hub}(P)$, it is defined as the root of the equation $\mathbb{E}[\psi(Y-\mathrm{Hub}(P))]=0$. From Bernstein's inequality, for any $\delta\in(0,1)$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i-\mathrm{Hub}(P))\right|\geq\sqrt{\frac{2V_\psi\ln(1/\delta)}{n}}+\beta\frac{\ln(1/\delta)}{3n}\right)\leq 2\delta$$

where $V_\psi=\mathrm{Var}(\psi(Y_i-\mathrm{Hub}(P)))$.

Then, using that Bernoulli random variables with mean ε are sub-Gaussian with variance parameter $\frac{1-2\varepsilon}{2\log((1-\varepsilon)/\varepsilon)}$ (see (Bourel et al., 2020, Lemma 6)),

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{W_i=1\}\leq\varepsilon+\sqrt{\frac{\ln(1/\delta)(1-2\varepsilon)}{n\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}\right)\geq 1-\delta.$$

Then, using Lemma 7 we get for any $\delta\in(0,1)$, with probability larger than $1-3\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i-\mathrm{Hub}(P))\right|\leq\sigma\sqrt{\frac{2\ln(1/\delta)}{n}}+\beta\frac{\ln(1/\delta)}{2n}+2\beta\varepsilon+2\beta\sqrt{\frac{\ln(1/\delta)(1-2\varepsilon)}{n\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}.$$

Step 2. Using $\eta=\sqrt{\frac{\ln(1/\delta)}{2n}}$, the hypotheses of Lemma 8 are verified.
PROOF: To apply Lemma 8, it is sufficient that

$$\sigma\sqrt{\frac{2t}{n}}+\beta\frac{\ln(1/\delta)}{3n}+2\beta\sqrt{\frac{\ln(1/\delta)(1-2\varepsilon)}{n\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}\leq\frac{\beta}{2}\left(p-\sqrt{\frac{\ln(1/\delta)}{2n}}-\varepsilon\right)$$

and using that $4\sigma\leq\beta$, we have that it is sufficient that

$$\sqrt{\frac{\ln(1/\delta)}{2n}}+\frac{\ln(1/\delta)}{3n}+2\sqrt{\frac{\ln(1/\delta)(1-2\varepsilon)}{n\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}\leq\frac{1}{2}\left(p-5\varepsilon\right). \tag{12}$$

This is a polynomial in $\sqrt{\ln(1/\delta)/n}$ that we need to solve. We use the following elementary algebra lemma.

Lemma 9 (2nd order polynomial root bound) *let a,b,c be three positive constants and x verify $ax^2+bx-c\leq 0$. Suppose that $\frac{4ac}{b^2}\leq d$, then x verifies*

$$x\geq\frac{2c(\sqrt{d+1}-1)}{db}.$$

Observe that we have

$$\frac{2\left(p-5\varepsilon\right)}{3\left(\frac{1}{\sqrt{2}}+\frac{2\sqrt{1-2\varepsilon}}{\sqrt{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}\right)^2}\leq\frac{4}{3}$$

and $(\sqrt{4/3+1}-1)/(4/3) \geq 8/7$, hence, from Lemma 9, we get the following sufficient condition for Equation (12) to hold:

$$\sqrt{\ln(1/\delta)/n} \leq \frac{8\sqrt{2}\,(p-5\varepsilon)}{7\left(1+\frac{2\sqrt{2(1-2\varepsilon)}}{\sqrt{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}\right)}.$$

Hence, taking this to the square,

$$\ln(1/\delta) \leq n\frac{128\,(p-5\varepsilon)^2}{49\left(1+\frac{2\sqrt{2(1-2\varepsilon)}}{\sqrt{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}}\right)^2}.$$

Step 3. Using Lemma 8 and Step 1 prove that the theorem is true. PROOF: The hypotheses of Lemma 8 are verified and we can use its result and together with Equation (11) we get with probability larger than $1-5\delta$,

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

Appendix F. Controlling Variance and Concentration of Huber's Estimator with Influence Function

F.1. Proof of Lemma 7: Controlling Variance of Influence of Huber's Estimator

Let ρ be Huber's loss function, with $\psi = \rho'$. We have that for any $x > 0$, $\psi(x)^2 \leq 2\rho(x)$. Hence,

$$\mathrm{Var}(\psi(Y - \mathrm{Hub}(P))) = \mathbb{E}[\psi(Y-\mathrm{Hub}(P))^2] \leq 2\mathbb{E}[\rho(Y-\mathrm{Hub}(P))].$$

Then, use that by definition of $\mathrm{Hub}(P)$, $\mathrm{Hub}(P)$ is a minimizer of $\theta \mapsto \mathbb{E}[\rho(Y-\theta)]$, hence,

$$\mathrm{Var}(\psi(Y - \mathrm{Hub}(P))) \leq 2\mathbb{E}[\rho(Y-\mathbb{E}[Y])].$$

and finally, use that $\rho(x) \leq x^2/2$ to conclude.

F.2. Proof of Lemma 8 : Concentrating Huber's Estimator by Concentrating the Influence

For all $n \in \mathbb{N}^*$, $\lambda > 0$, let

$$f_n(\lambda) = \frac{\mathrm{sign}(\Delta_n)}{n}\sum_{i=1}^n \psi(X_i - \mathrm{Hub}(P) - \lambda\,\mathrm{sign}(\Delta_n)),$$

where $\Delta_n = \mathrm{Hub}(P) - \mathrm{Hub}(X_1^n)$.

Step 1. For any $\lambda > 0$, $\mathbb{P}(\exists n \leq N : \quad |\Delta_n| \geq \lambda) \leq \mathbb{P}(\exists n \leq N : \quad f_n(\lambda) \geq 0)$.
PROOF: For all $y \in \mathbb{R}$, let $J_n(y) = \frac{1}{n}\sum_{i=1}^n \rho(X_i - y)$ we have,

$$J_n''(y) = \frac{1}{n}\sum_{i=1}^n \psi'\left(X_i - y\right).$$

In particular, having $f_n(\lambda) = -\operatorname{sign}(\Delta_n) J'(\mathrm{Hub}(P) + \lambda \operatorname{sign}(\Delta_n))$ if we take the derivative of f_n with respect to λ, we have the following equation

$$\begin{aligned}\frac{\partial}{\partial\lambda} f_n(\lambda) &= -\operatorname{sign}(\Delta_n)^2 J_n''(\mathrm{Hub}(P) + \lambda \operatorname{sign}(\Delta_n)) \\ &\leq -\frac{1}{n}\sum_{i=1}^n \psi'(X_i - \mathrm{Hub}(P) - \lambda \operatorname{sign}(\Delta_n)). \end{aligned} \tag{13}$$

Then, because ψ' is non-negative, the function $\lambda \mapsto f_n(\lambda,)$ is non-increasing. Hence, for all $n \in \mathbb{N}^*$ and $\lambda > 0$,

$$|\Delta_n| \geq \lambda \Rightarrow f_n(|\Delta_n|) = 0 \leq f_n(\lambda),$$

Hence,

$$\mathbb{P}(\exists n \leq N : \quad |\Delta_n| \geq \lambda) \leq \mathbb{P}(\exists n \leq N : \quad f_n(\lambda) \geq 0). \tag{14}$$

Step 2. For all $\lambda > 0$,

$$f_n(\lambda) \leq f_n(0) - \lambda \inf_{t\in[0,\lambda]} \left|f_n'(t)\right|.$$

PROOF: We apply Taylor's inequality to the function f_n. As f_n is non-increasing (because its derivative is non-positive, see Equation (13)), we get

$$f_n(\lambda) \leq f_n(0) - \lambda \inf_{t\in[0,\lambda]} \left|f_n'(t)\right|.$$

Step 3. Let $m_n = \mathbb{E}\left[\inf_{t\in[0,\lambda]} \frac{1}{n}\sum_{i=1}^n \psi'(X_i' - \mathrm{Hub}(P) - t)\right]$. With probability larger than $1 - 2e^{-2n\eta^2}$,

$$\inf_{t\in[0,\lambda]} \left|f_n'(t))\right| \geq m_n - 2\eta - \varepsilon,$$

PROOF: Write that $X_i = (1-W_i)Y_i + W_i Z_i$ where $W_1, \ldots, W_n$ are i.i.d Bernoulli random variable with mean ε, $Y_1, \ldots, Y_n$ are i.i.d $\sim P$ and $Z_1, \ldots, Z_n$ are i.i.d with law H.

From equation (13),

$$\begin{aligned}
\left|f_n'(t))\right| \geq & \frac{1}{n}\sum_{i=1}^{n}\psi'(X_i - \mathrm{Hub}(P) - t\,\mathrm{sign}(\Delta)) \\
\geq & \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{W_i = 0\}\psi'(Y_i - \mathrm{Hub}(P) - t\,\mathrm{sign}(\Delta)) && (15)\\
& + \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{W_i = 1\}\psi'(Z_i - \mathrm{Hub}(P) - t\,\mathrm{sign}(\Delta)) && (16)\\
\geq & \frac{1}{n}\sum_{i=1}^{n}\psi'(Y_i - \mathrm{Hub}(P) - t\,\mathrm{sign}(\Delta)) && (17)\\
& + \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{W_i = 1\}\left(\psi'(Z_i - \mathrm{Hub}(P) - t\,\mathrm{sign}(\Delta)) - \psi'(W_i - \mathrm{Hub}(P) - t\,\mathrm{sign}(\Delta))\right) && (18)
\end{aligned}$$

Hence, because $\psi' \in [0,1]$, we have

$$\left|f_n'(t))\right| \geq \frac{1}{n}\sum_{i=1}^{n}\psi'(Y_i - \mathrm{Hub}(P) - t\,\mathrm{sign}(\Delta)) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{W_i = 1\}) \tag{19}$$

The right-hand side depends on the infimum of the mean of n i.i.d random variables in $[0,1]$. Hence, the function

$$Z(X_1^n) \mapsto \sup_{t\in[0,\lambda]}\sum_{i=1}^{n}\psi'(X_i' - \mathrm{Hub}(P) - t)$$

satisfies, by sub-linearity of the supremum operator and triangular inequality, the bounded difference property, with differences bounded by 1. Hence, by Hoeffding's inequality, we get with probability larger than $1 - e^{-2n\eta^2}$,

$$\inf_{t\in[0,\lambda]}\left|f_n'(t))\right| \geq \mathbb{E}\left[\inf_{t\in[0,\lambda]}\frac{1}{n}\sum_{i=1}^{n}\psi'(X_i' - \mathrm{Hub}(P) - t)\right] - \eta - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{W_i = 1\})$$

and using Hoeffding's inequality to control $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{W_i = 1\}$, we have with probability larger than $1 - 2e^{-2\eta^2/n}$,

$$\inf_{t\in[0,\lambda]}\left|f_n'(t))\right| \geq \mathbb{E}\left[\inf_{t\in[0,\lambda]}\frac{1}{n}\sum_{i=1}^{n}\psi'(X_i' - \mathrm{Hub}(P) - t)\right] - 2\eta - \varepsilon$$

Step 4. For $\lambda \in (0, \beta/2)$,

$$\mathbb{P}\left(\quad |\Delta_n| \geq \lambda\right) \leq \mathbb{P}\left(\quad \left|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i - \mathrm{Hub}(P))\right| \geq \lambda\left(m_n - \eta - \varepsilon\right)\right) + 2e^{-2n\eta^2}.$$

PROOF: For any $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{P}(|\Delta_n| \geq \lambda) &\leq \mathbb{P}(f_n(\lambda) \geq 0) && \text{(from Step 1)} \\
&\leq 1 - \mathbb{P}\left(f_n(0) - \lambda \inf_{t\in[0,\lambda]} \left|f_n'(t)\right| \leq 0\right) && \text{(from Step 2)} \\
&\leq 1 - \mathbb{P}\left(f_n(0) \leq \lambda\left(m_n - 2\eta - \varepsilon\right)\right) + 2e^{-2n\eta^2} && \text{(from Step 3)} \\
&= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \mathrm{Hub}(P))\right| \geq \lambda\left(m_n - \eta - \varepsilon\right)\right) + 2e^{-2n\eta^2}. && (20)
\end{aligned}$$

Step 5. We prove that $m_n \geq p$, and hence

$$\mathbb{P}\left(|\Delta_n| \geq \lambda\right) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \mathrm{Hub}(P))\right| \geq \lambda\left(p - \eta - \varepsilon\right)\right) + 2e^{-2n\eta^2}$$

PROOF: For all $\lambda \leq \beta/2$,

$$\begin{aligned}
\mathbb{E}\left[\inf_{t\in[0,\lambda]} \frac{1}{n}\sum_{i=1}^{n} \psi'(X_i' - \mathrm{Hub}(P) - t)\right] &= \mathbb{E}\left[\inf_{t\in[0,\lambda]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{|X_i' - \mathrm{Hub}(P) - t| \leq \beta\}\right] \\
&\geq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{|X_i' - \mathrm{Hub}(P)| \leq \beta - \lambda\}\right] \\
&\geq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{|X_i' - \mathrm{Hub}(P)| \leq \beta/2\}\right] = p
\end{aligned}$$

Then, we plug the bound on m_n found in the previous step in equation (20), we get for any $\eta > 0$ and $\lambda \in (0, \beta/2]$,

$$\mathbb{P}(|\Delta_n| \geq \lambda) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \mathrm{Hub}(P))\right| \geq \lambda\left(p - \eta - \varepsilon\right)\right) + 2e^{-2n\eta^2}$$

Appendix G. Proofs of Auxiliary Lemmas

G.1. Proof of Lemma 6

We have,

$$\begin{aligned}\int_{\mathbb{R}} \frac{\left(1+\frac{(y+a)^2}{k}\right)^{\frac{k+1}{2}}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y &= \int_{\mathbb{R}} \sum_{l=0}^{\frac{k+1}{2}} \binom{\frac{k+1}{2}}{l} \frac{(y+a)^{2l}}{k^l\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y \\ &= \int_{\mathbb{R}} \sum_{l=0}^{\frac{k+1}{2}} \sum_{j=0}^{2l} \binom{\frac{k+1}{2}}{l}\binom{2l}{j} \frac{y^j a^{2l-j}}{k^l\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y \\ &= \sum_{l=0}^{\frac{k+1}{2}} \sum_{j=0}^{2l} \binom{\frac{k+1}{2}}{l}\binom{2l}{j} \int_{\mathbb{R}} \frac{y^j a^{2l-j}}{k^l\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y\end{aligned}$$

Remark that the integral is 0 if j is odd. Hence,

$$\int_{\mathbb{R}} \frac{\left(1+\frac{(y+a)^2}{k}\right)^{\frac{k+1}{2}}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y = \sum_{l=0}^{\frac{k+1}{2}} \sum_{j=1}^{l} \binom{\frac{k+1}{2}}{l}\binom{2l}{2j} \frac{a^{2l-2j}}{k^l} \int_{\mathbb{R}} \frac{y^{2j}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y$$

Then, we compute the integrals. By change of variable $u = y/k$, we have

$$\int_{\mathbb{R}} \frac{y^{2j}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y = k^{j+1/2} \int_{\mathbb{R}} \frac{u^{2j}}{(1+u^2)^{k+1}}\mathrm{d}u \leq 2k^{j+1/2}$$

and for $l = j$,

$$\sum_{l=0}^{\frac{k+1}{2}} \binom{\frac{k+1}{2}}{l} \frac{1}{k^l} \int_{\mathbb{R}} \frac{y^{2l}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y = \int_{\mathbb{R}} \frac{(1+y^2/k)^{\frac{k+1}{2}}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y$$

Hence,

$$\begin{aligned}\int_{\mathbb{R}} \frac{\left(1+\frac{(y+a)^2}{k}\right)^{\frac{k+1}{2}}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y &\leq 2\sum_{l=1}^{\frac{k+1}{2}} \sum_{j=0}^{l-1} \binom{\frac{k+1}{2}}{l}\binom{2l}{2j} \frac{a^{2l-2j}}{k^l} k^{j+1/2} + \int_{\mathbb{R}} \frac{(1+y^2/k)^{\frac{k+1}{2}}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y \\ &= 2\sum_{l=1}^{\frac{k+1}{2}} a^{2l} \sum_{j=0}^{l-1} \binom{\frac{k+1}{2}}{l}\binom{2l}{2j} \frac{a^{-2j}}{k^l} k^{j+1/2} + \int_{\mathbb{R}} \frac{(1+y^2/k)^{\frac{k+1}{2}}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y \\ &\leq 2\sum_{l=1}^{\frac{k+1}{2}} a^{2l} \sum_{j=0}^{l-1} \binom{\frac{k+1}{2}}{l}\binom{2l}{2j} \frac{a^{-2j}}{k^l} k^{j+1/2} + \int_{\mathbb{R}} \frac{(1+y^2/k)^{\frac{k+1}{2}}}{\left(1+\frac{y^2}{k}\right)^{k+1}}\mathrm{d}y \quad (21) \\ &\quad (22)\end{aligned}$$

And,

$$\begin{aligned}
\sum_{l=1}^{\frac{k+1}{2}} a^{2l} \sum_{j=0}^{l-1} \binom{\frac{k+1}{2}}{l} \binom{2l}{2j} \frac{a^{-2j}}{k^l} k^{j+1/2} &= \sqrt{k} \sum_{l=1}^{\frac{k+1}{2}} \sum_{j=0}^{l-1} \binom{\frac{k+1}{2}}{l} \binom{2l}{2j} a^{2(l-j)} k^{j-l} \\
&\leq \sqrt{k} \sum_{l=1}^{\frac{k+1}{2}} \sum_{j=0}^{l-1} \binom{\frac{k+1}{2}}{l} \binom{2l}{2j} a^{2(l-j)} k^{j-l} \\
&\leq \sqrt{k} \sum_{l=1}^{\frac{k+1}{2}} \sum_{j=0}^{l-1} \binom{\frac{k+1}{2}}{l} \binom{2(l-1)}{2j} 4l^2 \left(\frac{a^2}{k}\right)^{l-j} \\
&\leq (k+1)^2 \sqrt{k} \sum_{l=1}^{\frac{k+1}{2}} \binom{\frac{k+1}{2}}{l} \left(\frac{a^2}{k}\right)^l \left(1 + \frac{\sqrt{k}}{a}\right)^{2(l-1)} \\
&= \frac{a^2}{k} (k+1)^2 \sqrt{k} \sum_{l=1}^{\frac{k+1}{2}} \binom{\frac{k+1}{2}}{l} \left(\frac{a}{\sqrt{k}} + 1\right)^{2(l-1)} \\
&\leq \frac{a^2}{\sqrt{k}} (k+1)^2 \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1}.
\end{aligned}$$

Then, inject this in Equation (21) to get

$$\int_{\mathbb{R}} \frac{\left(1 + \frac{(y+a)^2}{k}\right)^{\frac{k+1}{2}}}{\left(1 + \frac{y^2}{k}\right)^{k+1}} \mathrm{d}y \leq 2 \frac{a^2}{\sqrt{k}} (k+1)^2 \left(2 + \frac{a}{\sqrt{k}}\right)^{k-1} + \int_{\mathbb{R}} \frac{(1 + y^2/k)^{\frac{k+1}{2}}}{\left(1 + \frac{y^2}{k}\right)^{k+1}} \mathrm{d}y.$$

G.2. Proof of Lemma 9

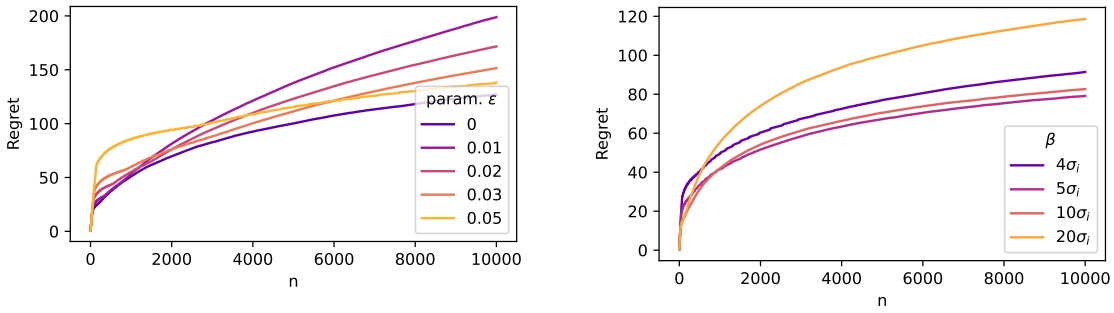
The solutions of the second order polynomial indicate that x must verify

$$x \geq \frac{-b + \sqrt{b^2 + 4ac}}{2a} \geq \frac{b}{2a} \left(-1 + \sqrt{1 + \frac{4ac}{b^2}}\right).$$

Then, use that the function $x \mapsto \sqrt{x+1}$ is concave and hence the graph of $x \mapsto \sqrt{x+1}$ is above its chords and we have for any $x \in [0, d]$, $\sqrt{1+x} \geq 1 + x\frac{\sqrt{d+1}-1}{d}$. Hence,

$$x \geq \frac{b}{2a} \left(\frac{4ac(\sqrt{d+1}-1)}{db^2}\right) = \frac{2c(\sqrt{d+1}-1)}{db}.$$

Figure 3: Cumulative regret plots for different values of the parameters ε and β on a Weibull dataset.



(a) Dependency in ε (b) Dependency in β

Appendix H. Sensitivity to β and ε

In this section we illustrate the impact of the choice of β and ε on the estimation.

Choice of β (Figure 4(*b*)): The choice of β is a trade-off between the bias (distance $|\mathrm{Hub}(P) - \mathbb{E}[X]|$ which decreases as β go to infinity) and robustness (when β goes to 0, $\mathrm{Hub}(P)$ goes to the median). To illustrate this trade-off we use the Weibull distribution for which can be very asymmetric. We use a 3-armed bandit problem with shape parameters $(2, 2, 0.75)$ and scale parameters $(0.5, 0.7, 0.8)$ which implies that the means are approximately $(0.44, 0.62, 0.95)$. These distributions are very asymmetric, hence the bias $|\mathrm{Hub}(P)-\mathbb{E}[X]|$ is high and in fact even though arm 3 has the optimal mean, arm 2 will have the optimal median, the medians are given by $(0.41, 0.58, 0.49)$. In this experiment we don't use any corruption as we don't want to complicate the interpretation. As expected by the theory, we get that β_i should not be too small or too large but it should be around $4\sigma_i$.

Choice of ε (Figure 4(*a*)): To illustrate the dependency in ε, we also use the Weibull distribution to show the dependency in ε with the same parameters as in the previous Weibull example, except that we choose $\beta_i = 5\sigma_i$ which is around the optimum found in the previous experiment and we corrupt with 2% of outliers (this is the true ε while we will make the ε used in the definition of the algorithm vary). The outliers are constructed as in Section 5. The effect of the parameter ε is difficult to assess because ε has an impact on the length of force exploration that we impose at the beginning of our algorithm (the s_{lim}).