What You See May Not Be What You Get: UCB Bandit Algorithms Robust to ε-Contamination

Laura Niss
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Ambuj Tewari
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Abstract

Motivated by applications of bandit algorithms in education, we consider a stochastic multi-armed bandit problem with ε-contaminated rewards. We allow an adversary to give arbitrary unbounded contaminated rewards with full knowledge of the past and future. We impose the constraint that for each time t the proportion of contaminated rewards for any action is less than or equal to ε. We derive concentration inequalities for two robust mean estimators for sub-Gaussian distributions in the ε-contamination context. We define the ε-contaminated stochastic bandit problem and use our robust mean estimators to give two variants of a robust Upper Confidence Bound (UCB) algorithm, crUCB. Using regret derived from only the underlying stochastic rewards, both variants of crUCB achieve $\mathcal{O}(\sqrt{KT\log T})$ regret for small enough contamination proportions. Our simulations assume small horizons, reflecting the newly explored setting of bandits in education. We show that in certain adversarial regimes crUCB not only outperforms algorithms designed for stochastic (UCB1) and adversarial (EXP3) bandits but also those that have "best of both worlds" guarantees (EXP3++ and TsallisInf) even when our constraint on the proportion of contaminated rewards is broken.

1 INTRODUCTION

We first review the problem of stochastic multi-armed bandits (sMAB) with contaminated rewards, or contaminated stochastic bandits (CSB). This scenario assumes that rewards associated with an action are sampled i.i.d. from a fixed distribution and that the learner observes the reward after an adversary has the opportunity to contaminate it. The observed reward can be unrelated to the reward distribution and can be maliciously chosen to fool the learner. An outline for this setup is presented in Section 2.

We are primarily motivated by the use of bandit algorithms in education, where the rewards often come directly from human opinion. Whether responses come from undergraduate students, a community sample, or paid participants on platforms like MTurk, there is always reason to believe some responses are careless or inattentive to the question or could be assisted by bots (Curran, 2016; Necka et al., 2016).

An example in education is a recent paper testing bandit Thompson sampling to identify high quality student generated solution explanations to math problems using MTurk participants (Williams et al., 2016). Using a rating between 1-10 from 150 participants, the results showed that Thompson sampling identified participant generated explanations that when viewed by other participants significantly improved their chance of solving future problems compared to no explanation or "bad" explanations identified by the algorithm. While the proportion of contaminated responses will always depend on the population, recent work suggests even when screening out fraudulent participants, between $2-30\%$ of MTurk participants give low-quality samples (Ahler, Roush, and Sood, 2019; Necka et al., 2016; Ryan, 2018). This is consistent with measurements of careless and inattentive responses seen in survey data, which reports $1-30\%$ with an estimated mode of $8-12\%$, with the conclusion that these responses are generally not a random sample (Curran, 2016). Accounting for these low quality responses is especially relevant in educational setting where the number of iterations an algorithm can run is often significantly smaller than those used by big tech (e.g. advertising).

Recent work in CSB has various assumptions on the ad-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

versary, the contamination, and the reward distributions. Many papers require the rewards and contamination to be bounded (Gupta, Koren, and Talwar, 2019; Kapoor, Patel, and Kar, 2018; Lykouris, Mirrokni, and Leme, 2018). Others do not require boundedness, but do assume that the adversary contaminates uniformly across rewards (Altschuler, Brunel, and Malek, 2019). All works make some assumption on the number of rewards for an action an adversary can contaminate. We discuss previous work more thoroughly in Section 3.

Our work expands on these papers by allowing for a full knowledge adaptive adversary that can give unbounded contamination in any manner. However, there is a trade off when compared to work assuming bounded rewards and contamination: we require an estimate of the upper bound on the reward variance. This can often allow for simpler implementation than some algorithms that require boundedness, as we will discuss in section 4. Our constraint on the adversary is that for some fixed ε, no more than ε proportion of rewards for an action are contaminated. We provide a ε-contamination robust UCB algorithm by first proving concentration inequalities for two robust mean estimators in the ε-contamination context. We are able to show that the regret of our algorithm analyzed on the true reward distributions is $\mathcal{O}(\sqrt{KT\log T})$ provided that the contamination proportion is small enough. Through simulations, we show that with a Bernoulli adversary, our algorithm outperforms algorithms designed for stochastic (UCB1) and adversarial (EXP3) bandits as well as those that have "best of both worlds" guarantees (EXP3++ and TsallisInf) even when our constraint on the adversary is broken.

Though we are motivated by of bandit algorithms applications in education and use this context to determine appropriate parameters in the simulations, we point out opportunities for CSB modeling to arise in other contexts as well.

Human feedback: There is always a chance that human feedback is careless or inattentive, and therefore is not representative of the underlying truth related to an action. This may appear in online surveys that are used for A/B testing, or as is the case above in the explanation generation example. Adaptive surveys, such as choosing question ordering to minimize dropout rates, are also an example where the sample sizes can be small compared to other bandit deployments.

Click fraud: Internet users who wish to preserve privacy can intentionally click on ads to obfuscate their true interests either manually or through browser apps. Similarly, malware can click on ads from one company to falsely indicate high interest, which can cause higher rankings in searches or more frequent use of the ad than it would otherwise merit (Crussell, Stevens, and Chen, 2014; Pearce et al., 2014).

Measurement errors: If rewards are gathered through some process that may occasionally fail or be inaccurate, then the rewards may be contaminated. For example, in health apps that use activity monitors, vigorous movement of the arms may be perceived as running in place (Bai et al., 2018; Feehan et al., 2018).

2 PROBLEM SETTING

Here we specify our notation and present the ε-contaminated stochastic bandit problem. We then argue for a specific notion of regret for CSB. We compare our setting to others current in the field in section 3.

Notation We use $[K]$ to represent $\{1, ..., K\}$ for $K \in \mathbb{R}$ to represent the number of actions and the indicator function $\mathbb{I}\{\cdot\}$ to be 1 if true and 0 otherwise. Let $N_a(t)$ be the number of times action a has been chosen at time t and $\mathbf{x}_a(t) = \{x_a(1), ..., x_a(N_a(t))\}$ to be the vector of all observed rewards for action a at time t. The suboptimality gap for action a is Δ_a and we define $\Delta_{\min} = \min_{a\in[K]} \Delta_a$.

2.1 ε-CONTAMINATED STOCHASTIC BANDITS

A basic parameter in our framework is ε, the fraction of rewards for an action that the adversary is allowed to contaminate. Before play, the environment picks a true reward $r_a(t) \sim D_a$ from fixed distribution D_a for all $a \in [K]$ and $t \in [T]$. The adversary observes these rewards and then play begins. At time $t = 1, 2, ..., T$ the learner chooses an action $A_t \in [K]$. The adversary sees A_t then chooses an observed reward $x_{A_t}(t)$ and then the learner observes only $x_{A_t}(t)$.

We present the contaminated stochastic bandits game in algorithm 1.

Algorithm 1: Contaminated Stochastic Bandits

input: Number of actions K, time horizon T.
fix : $r_a(t)\ \forall a \in [K],\ t \in [T]$.
Adversary observes fixed rewards.
for $t = 1, ..., T$ **do**
  Learner picks action $A_t \in [K]$.
  Adversary observes A_t and chooses $x_{A_t}(t)$.
  Learner observes $x_{A_t}(t)$.
end

We allow the adversary to corrupt in any fashion as long as for every time t there is no more than an ε-fraction of contaminated rewards for any action. That is, we constrain the adversary such that,

$$\forall a \in [K],\ \forall t \in [T],\ \sum_{i=1}^{N_a(t)} \mathbb{I}\{r_a(i) \neq x_a(i)\} \leq \varepsilon \cdot N_a(t).$$

We allow the adversary to give unbounded contamination that can be chosen with full knowledge of the learner's history as well as current and future rewards. This setting allows the adversary to act differently across actions and places no constraints on the contamination itself, but rather the rate of contamination.

2.2 NOTION OF REGRET

A traditional goal in bandit learning is to minimize the observed cumulative regret gained over the total number of plays T. Because the adversary in this model can affect the observed cumulative regret, we argue to instead use a notion of regret that considers only the underlying true rewards. We call this uncontaminated regret and give the definition below for any time T and policy π in terms of the true rewards r,

$$\bar{R}_T(\pi) = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} r_a(t) - \sum_{t=1}^{T} r_{A_t}(t)\right]. \quad (2.1)$$

This definition eq. (2.1) is first mentioned in Kapoor, Patel, and Kar (2018) along with another notion of regret that compares the sum of the observed (possibly contaminated) rewards to the sum of optimal, uncontaminated rewards,

$$\bar{R}_T(\pi) = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} r_a(t) - \sum_{t=1}^{T} x_{A_t}(t)\right]. \quad (2.2)$$

We argue that eq. (2.2) gives little information about the performance of an algorithm. This notion of regret can be negative, and with no bounds on the contamination it can be arbitrarily small and potentially meaningless. We believe that any regret that compares a true component to an observed (possibly contaminated) component is not a useful measure of performance in CSB as it is unclear what regret an optimal strategy should produce.

3 RELATED WORK

We start by briefly addressing why adversarial and "best of both world" algorithms are not optimized for CSB. We then cover relevant work in robust statistics, followed by current work in robust bandits and how our model differs and relates.

3.1 ADVERSARIAL BANDITS

Adversarial bandits with an oblivious environment allows the adversary to first look at the learners policy and then choose all rewards before the game begins. If the learner chooses a deterministic policy, the adversary can choose rewards such that the learner cannot achieve sublinear worst-case regret (Lattimore, 2020). Algorithms such as EXP3 (Auer, Nicolò Cesa-Bianchi, et al., 2002) are thus randomized, but their regret is analysed with respect to the best fixed action where "best" is defined using the *observed* rewards. There are no theoretical guarantees with respect to the uncontaminated regret, so it is not immediately clear how they will perform in a CSB problem. We remark that adversarial analysis assumes uniformly bounded observed rewards whereas we allow observed rewards to be unbounded. Additionally, the general adversarial framework does not take advantage of the structure present in CSB, namely that the adversary can only corrupt a small fraction of rewards, so it is likely that performance improvements can be made.

3.2 BEST OF BOTH WORLDS

A developing line of work is algorithms that enjoy "best of both worlds" guarantees. That is, they perform well in both stochastic and adversarial environments without knowing a priori which environment they will face. Early work in this area (Auer and Chiang, 2016; Bubeck and Slivkins, 2012) started by assuming a stochastic environment and implementing some method to detect a failure of the i.i.d. assumption on rewards, at which point the algorithm switches to an algorithm for the adversarial environment for the remainder of iterations. Further work implements algorithms that can handle an environment that is some mixture of stochastic and adversarial, as in EXP3++ and TsallisInf (Seldin and Slivkins, 2014; Zimmert and Seldin, 2019).

While these algorithms are aimed well for a stochastic environment with some adversarial rewards, they differ from contamination robust algorithms in that all observed rewards are thought to be informative. Their uncontaminated regret has not been analysed and therefore there are no guarantees in the CSB setting.

3.3 CONTAMINATION ROBUST STATISTICS

The ε-contamination model we consider is closely related to the one introduced by Huber in 1964 (Huber, 1964). Their goal was to estimate the mean of a Gaussian mixture model where ε fraction of the sample was not sampled from the main Gaussian component. There has been a recent increase of work using this model, especially in extensions to the high-dimensional case (Di-

akonikolas et al., 2019, Kothari, Steinhardt, and Steurer, 2018, Lai, Rao, and Vempala, 2016, L. Liu, Li, and Caramanis, 2019). These works often keep the assumption of a Gaussian mixture component, though there has been expanding work with non-Gaussian models as well.

3.4 CONTAMINATION ROBUST BANDITS

Some of the first work in CSB started by assuming both rewards and contamination were bounded (Gupta, Koren, and Talwar, 2019; Lykouris, Mirrokni, and Leme, 2018). These works assume an adversary that can contaminate at any time step, but that is constrained in the cumulative contamination. They bound the cumulative max (over actions) absolute difference of the contaminated reward, x, to the true reward, r, $\sum_t \max_a |r_a(t) - x_a(t)| \leq C$. Lykouris, Mirrokni, and Leme (2018) provides a layered UCB-type active arm elimination algorithm. Gupta, Koren, and Talwar (2019) expands on this work to provide an algorithm similar to active arm elimination in spirit, but which never completely eliminates an action, and which has better regret guarantees.

Recent work in implementing a robust UCB replaces the empirical mean with the empirical median, and gives guarantees for the uncontaminated regret with Gaussian rewards (Kapoor, Patel, and Kar, 2018). They consider an adaptive adversary but require the contamination to be bounded, though the bound need not be known. They cite work that can expand their robust UCB to distributions with bounded fourth moments by using the agnostic mean (Lai, Rao, and Vempala, 2016), though give no uncontaminated regret guarantees. In one dimension, the agnostic mean takes the mean of the smallest interval containing $(1 - \alpha)$ fraction of points. This estimator is also known as the α-shorth mean. Our work expands on this model by allowing for unbounded contamination and analysing the uncontaminated regret for sub-Gaussian rewards when implementing a UCB algorithm with the α-shorth mean.

CSB has also been analysed in the best arm identification problem (Altschuler, Brunel, and Malek, 2019). Using a Bernoulli adversary that contaminates any reward with probability ε, Altschuler, Brunel, and Malek (2019) consider three adversaries of increasing power, from the oblivious adversary, which does not know the player's history nor the current action or reward, to a malicious adversary, which can contaminate knowing the player's history and the current action and reward. They give analysis of the probability of best arm selection and sample complexity of an active arm elimination algorithm. While their performance measure is different than ours, we generalize their context to allow an adversary to contaminate in any fashion.

There is also work that explores the impact of an adaptive adversarial contamination on ε-greedy and UCB algorithms (Jun et al., 2018). They give a thorough analysis with both theoretical guarantees and simulations of the effects an adversary can have on these two algorithms when the adversary does not know the optimal action but is otherwise fully adaptive. They show these standard algorithms are susceptible to contamination. Similar work looks at contamination in contextual bandits with a non-adaptive adversary (Ma et al., 2019).

4 MAIN RESULTS

We present concentration bounds for both the α-shorth and α-trimmed mean estimators in the ε-contamination context for sub-Gaussian random variables.

Our contribution to the CSB problem is in providing a contamination robust UCB algorithm that is simple to implement and has theoretical regret guarantees close to those of UCB algorithms in the uncontaminated setting.

4.1 CONTAMINATION ROBUST MEAN ESTIMATORS

The estimators we analyse have been in use for many decades as robust statistics. Our contribution is to analyze them within our ε-contamination model with sub-Gaussian samples and provide simple *finite-sample concentration inequalities* for ease of use in UCB-type algorithms.

4.1.1 Trimmed Mean

Our first estimator suggested for use in the contaminated model is the α-trimmed mean (L. Liu, Li, and Caramanis, 2019).

α-trimmed mean Trim the smallest and largest α-fraction of points from the sample and calculate the mean of the remaining points. This estimator uses $1 - 2\alpha$ fraction of sample points.

Algorithm 2: α-Trimmed Mean

input : $X_n = (x_1, ..., x_n)$, α
output: α-trimmed mean
$(x_{(1)}, ..., x_{(n)})$ = sorted X_n $s.t.$ $x_{(i)} \leq x_{(i+1)}$
cut = $\lceil \alpha * n \rceil$
return *mean*$(x_{(cut)}, ..., x_{(n\text{-}cut)})$

The intuition being if the contamination is large, then it will be removed from the sample. If it is small, it should have little affect on the mean estimate. Next we provide the concentration inequality for the α-trimmed mean.

Theorem 1 (Trimmed mean concentration)**.** *Let G be the set of points $x_1, ...x_n \in \mathbb{R}$ that are drawn from a σ-sub-Gaussian distribution with mean μ. Let S_n be a sample where an ε-fraction of these points are contaminated by an adversary. For $\varepsilon \leq \alpha < 1/2$, $t \geq n$ we have,*

$$|\textit{trMean}_\alpha(S_n) - \mu| \leq \frac{\sigma}{(1-2\alpha)}\left(\sqrt{\frac{4}{n}\log(t)} + 4\alpha\sqrt{6\log(t)}\right)$$

with probability at least $1 - \frac{4}{t^2}$.

Proof follows from L. Liu, Li, and Caramanis (2019) and can be found in the appendix.

4.1.2 Shorth Mean

The agnostic mean from Lai, Rao, and Vempala (2016), which we use the more common term α-shorth mean for, can be considered a variation of the trimmed mean.

α-shorth mean Take the mean of the shortest interval that removes the smallest δ_1 and largest δ_2 fraction of points such that $\delta_1 + \delta_2 = \alpha$, where $\delta_1,\ \delta_2$ are chosen to minimize the interval length of remaining points. Uses $1 - \alpha$ fraction of sample points.

The α-shorth mean is less computationally efficient than the trimmed mean, but may be a better mean estimator when the contaminated points are not large outliers and are skewed in one direction. Intuitively this is because the α-shorth mean can trim off contamination that would require removing most of the sample with the trimmed mean. Next we provide the concentration inequality for the α-shorth mean.

Algorithm 3: α-Shorth Mean

input : $X_n = (x_1, ..., x_n)$, α
output: A mean estimate for the distribution of X
$(x_{(1)}, ..., x_{(n)})$ = sorted X_n *s.t.* $x_{(i)} \leq x_{(i+1)}$
$n_\alpha = \lfloor (1-\alpha) * n \rfloor$
$\mathcal{I} \in \text{argmin}_k\{x_{(k+n_\alpha)} - x_{(k)}\}$
Choose uniformly at random from set $\mathcal{I}$ if there is more than one starting index with the smallest interval length
return $\textit{sMean}(X) \leftarrow \textit{mean}(x_{(\mathcal{I})}, ..., x_{(\mathcal{I}+n_\alpha)})$

Theorem 2 (α-shorth mean concentration)**.** *Let G_n be the set of points $x_1, ...x_n \in \mathbb{R}$ that are drawn from a σ-sub-Gaussian distribution with mean μ. Let S_n be a sample where an ε-fraction of these points are contaminated by an adversary. For $\varepsilon \leq \alpha < 1/3$, $t \geq n$, we have,*

$$|\textit{sMean}_\alpha(S_n) - \mu| \leq \frac{\sigma}{1-2\alpha}\sqrt{\frac{4}{n}\log t} + \frac{(6\alpha - 8\alpha^2)\sigma}{(1-2\alpha)(1-\alpha)}\sqrt{6\log t}$$

with probability at least $1 - \frac{4}{t^2}$.

Proof sketch. Without loss of generality assume $\mu = 0$ for the underlying true distribution. Let $\tilde{G} \subset G_n$ represent the points which are not contaminated and $C \subset G_n$ represent the contaminated points. Then our sample can be represented by the union $S_n = \tilde{G} \cup C$

Let J be the interval that contains the shortest $1 - \alpha$ fraction of S_n, I be the interval that contains $\tilde{G}$ (i.e. the remaining good points after contamination), and T be the interval that contains the points of S_n after trimming the α largest and smallest fraction of points. Use $|I|$ to denote the length of interval I. It must be that $I \cap J \neq \emptyset$ because otherwise the points in $I \cup J$ would contain $2 - 2\alpha > 1$ fraction of S_n. Let c be a point in $I \cap J$ and x be a point in J. Recall that $\text{trMean}_\alpha(S_n)$ is the trimmed mean of the contaminated sample S_n. Then we have,

$$\begin{aligned}|x| &\leq |x - c| + |c - \text{trMean}_\alpha(S_n)| + |\text{trMean}_\alpha(S_n)| \\ &\leq |J| + |I| + |\text{trMean}_\alpha(S_n)| \\ &\leq 2|I| + |\text{trMean}_\alpha(S_n)|\end{aligned}$$

The second step comes from x and c both being in J and because $I \supseteq T$. The third step comes from $|J| \leq |I|$.

To bound the length of I we have,

$$|I| \leq 2 \max_{x \in G_n} |x| \ \text{ w.p. at least } 1 - \delta_2.$$

Finally, since

$$|\text{trMean}_\alpha(S_n)| \leq \frac{1}{(1-2\alpha)}(|\bar{x}_{G_n}| + 4\alpha \max_{x \in G_n} |x|)$$

with probability at least $1 - \delta_1 - \delta_2$, we get that for $x \in J$,

$$\begin{aligned}|x| &\leq 4 \max_{i \in [n]} |x_i| + \frac{1}{(1-2\alpha)}(|\bar{x}_{G_n}| + 4\alpha \max_{x \in G_n} |x|) \\ &= \frac{|\bar{x}_{G_n}|}{1-2\alpha} + \left(4 + \frac{4\alpha}{1-2\alpha}\right) \max_{x \in G_n} |x|.\end{aligned}$$

Now that we have a bound on the contaminated points in J, our analysis follows similarly as the trimmed mean by bounding A_1, A_2, A_3 as defined below.

$$\begin{aligned}&|\text{sMean}_\alpha(S_n)| \\ &\leq \frac{1}{(1-\alpha)n}\left(\underbrace{\left|\sum_{x \in \tilde{G}} x\right|}_{A_1} + \underbrace{\left|\sum_{x \in \tilde{G} \cap \neg J} x\right|}_{A_2} + \underbrace{\left|\sum_{x \in C \cap J} x\right|}_{A_3}\right)\end{aligned}$$

□

The full proof is contained in the appendix and follows a similar approach as for the trimmed mean.

Our methods ensured that the first term in each concentration bound is the same, giving them similar regret guarantees when implemented in a UCB algorithm. We emphasize that the α-shorth mean uses $1-\alpha$ fraction of a sample while the α-trimmed mean uses $1-2\alpha$ fraction of a sample. We remark that if there is no contamination and $\alpha = 0$ then our inequalities reduce to the standard concentration inequality for the empirical mean of samples drawn from a sub-Gaussian distribution.

4.2 CONTAMINATION ROBUST UCB

We present the contamination robust-UCB (crUCB) algorithm for ε-CSB with sub-Gaussian rewards.

Algorithm 4: crUCB

input: number of actions K, time horizon T, upper bound on fraction contamination α, upper bound on sub-Gaussian constant σ_0, mean estimate function (α trimmed or shorth mean) f.

for $t \leq K$ **do**
  Pick action a when $t = a$.
end
for $t > K$ **do**
  for $a \in [K]$ *compute* **do**
    $f(\mathbf{x}_a(t)) \leftarrow$ mean estimate of rewards.
    $N_a(t) \leftarrow$ number of times action has been played.
  end
  Pick action $A_t =$
  $\text{argmax}_{a\in[K]} f(\mathbf{x}_a(t)) + \frac{\sigma_0}{(1-2\alpha)}\left(\sqrt{4\frac{\log(t)}{N_a(t)}}\right)$.
  Observe reward $x_{A_t}(t)$.
end

We provide uncontaminated regret guarantees for crUCB below for both the α-trimmed and the α-shorth mean.

Theorem 3 (α-trimmed mean crUCB uncontaminated regret)**.** *Let $K > 1$ and $T \geq K-1$. Then with algorithm 4 with the α-trimmed mean, σ-sub-Gaussian reward distributions with $\sigma_a \leq \sigma_0$, and contamination rate $\varepsilon \leq \alpha \leq \frac{\Delta_{min}}{4(\Delta_{min}+4\sigma_0\sqrt{6\log T})}$, we have the uncontaminated regret bound,*

$$\bar{R}(UCB) \leq 8\sigma_0\sqrt{KT\log T} + \sum 15\Delta_a.$$

Corollary 1 (α-trimmed mean crUCB uncontaminated regret bounded rewards)**.** *If the rewards are bounded by b, and have contamination rate $\varepsilon \leq \alpha \leq \frac{\Delta_{\min}}{4(\Delta_{\min}+4b)}$, then*

$$\bar{R}_T \leq 8\sigma_0\sqrt{KT\log(T)} + \sum 15\Delta_a.$$

Theorem 4 (α-shorth mean crUCB uncontaminated regret)**.** *Let $K > 1$ and $T \geq K-1$. Then with algorithm 4 with the α-shorth mean, sub-Gaussian reward distributions with $\sigma_a \leq \sigma_0$, and contamination rate $\varepsilon \leq \alpha \leq \frac{\Delta_{min}}{4(\Delta_{min}+9\sigma_0\sqrt{6\log T})}$, we have the uncontaminated regret bound,*

$$\bar{R}(UCB) \leq 8\sigma_0\sqrt{KT\log T} + \sum 15\Delta_a.$$

Corollary 2 (α-shorth mean crUCB uncontaminated regret bounded rewards)**.** *If the rewards are bounded by b, and have contamination rate $\varepsilon \leq \alpha \leq \frac{\Delta_{\min}}{4(\Delta_{\min}+9b)}$, then*

$$\bar{R}_T \leq 8\sigma_0\sqrt{KT\log(T)} + \sum 15\Delta_a.$$

Proofs for theorem 3 and 4 and their corollaries follow standard analysis and are provided in the appendix.

From theorem 3 and 4 we get that crUCB has the same order of regret in the CSB setting as UCB1 has in the standard sMAB setting. The constraint on the magnitude of ε is quite strong, but we show in section 5 that they can be broken and still obtain good empirical performance.

Remark Our bounds above do not allow ε to be too big relative to the minimum suboptimality gap $\Delta_{\min}$. This is natural: if $\varepsilon > \Delta_{\min}$ then no algorithm can get sublinear regret since distinguishing between the top two actions is statistically impossible even with infinite samples. We give a simple example in Appendix B. Furthermore, it is possible to derive a regret bound[1] of $\tilde{O}(\sigma_0\sqrt{KT} + \frac{\alpha\sigma_0}{1-4\alpha}T)$ for any choice of α such that $\varepsilon \leq \alpha < 1/4$. The linear term in regret (which is unavoidable for large ε) may be acceptable if the corruption proportion is not very large.

5 SIMULATIONS

We compare our crUCB algorithms using the trimmed mean (tUCB) and shorth mean (sUCB) against a standard stochastic algorithm (UCB1, Auer and Nicolo Cesa-Bianchi, 2002), a standard adversarial algorithm (EXP3, Auer, Nicolò Cesa-Bianchi, et al., 2002), two "best of both worlds" algorithms (EXP3++, Seldin and Lugosi, 2017, 0.5-TsallisInf, Zimmert and Seldin, 2019), and

[1]The $\tilde{O}(\cdot)$ notation hides constants and logarithmic terms. See Appendix B for details.

another contamination robust algorithm (RUCB-MAB, Kapoor, Patel, and Kar, 2018). Each trial has five actions ($K = 5$), is run for 1000 iterations ($T = 1000$), for $\varepsilon \in \{0.05, 0.1\}$. For sUCB and tUCB, we set $\alpha = \varepsilon$ and $\sigma_0 = \sigma$. The plots are average results over 10 trials with error bars showing the standard deviation.

Our choice of T comes from our motivation to apply contaminated bandits in education, where the sample sizes are often much smaller than for example in advertising. While $T = 1000$ would be considered a large university class, it still allows one to visually see regret for smaller iterations and see how performance stabilizes. We similarly chose number K of arms and proportion contamination ε to be in a realistic range for the application we have in mind. All algorithms use recommended parameter settings given within their respective papers.

Rewards and gaps We chose the reward distribution to be binomial(n=10) to simulate likert scale and because this distribution has bounded rewards and is not symmetric for large p. For the optimal action, $p = .9$ and for suboptimal actions $p = .8$, thus the suboptimality gap is $\Delta = 1$. All non-optimal actions have the same true distribution.

Adversaries We focus on a Bernoulli adversary which gives a contaminated reward at every time step with probability ε. We also implement a cluster adversary which contaminates at the beginning of play to show the weakness of algorithms to this type of attack.

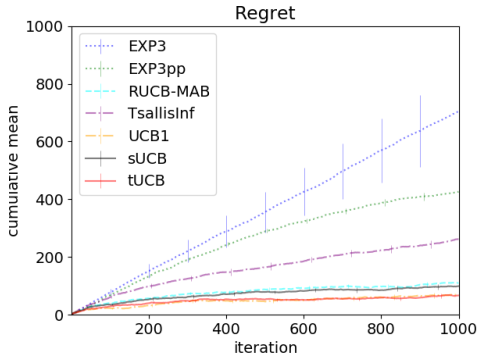
Contamination We use a random malicious contamination scheme which chooses a contaminated reward uniformly from ranges that increase suboptimal action means and decrease the optimal action's mean.

Performance measurement We plot the average regret over 10 trials for 1000 iterations.

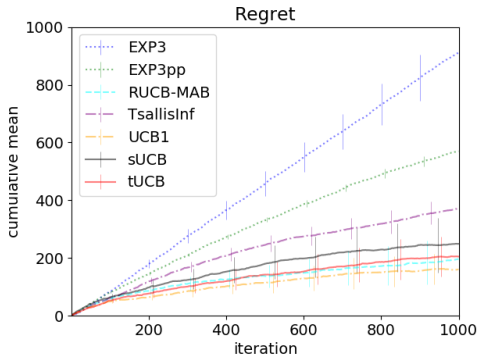
We recommend to view the plots on a color screen.

In Figure 1a we see that the adversarial and best of both worlds algorithms, EXP3, EXP3++, and TsallisInf, perform poorly in the purely stochastic setting compared to the UCB type algorithms. In Figure 1, we see the best of these, TsallisInf, starts to degrade as the proportion of contamination increases while the robust UCB algorithms are only slightly affected. These simulations show a clear performance benefit to using algorithms that specifically account for contaminated rewards.

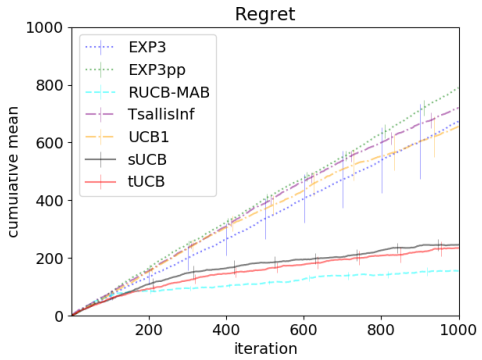
Figure 3 and Figure 4 shows that for both sUCB and tUCB, the choice of α is much less sensitive than choice of σ. Over estimating or slightly underestimating α does not degrade performance significantly. Underestimating



(a) $\varepsilon = 0$



(b) $\varepsilon = 0.05$

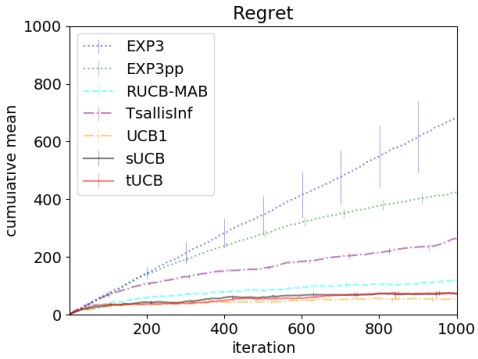


(c) $\varepsilon = 0.1$

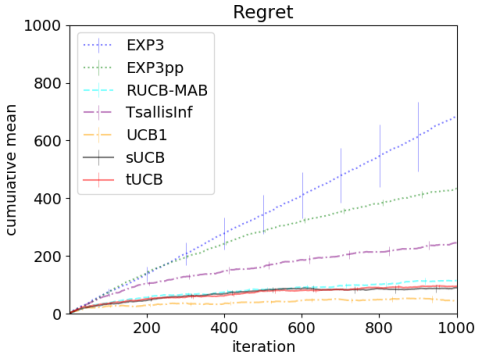
Figure 1: Binomial Rewards With Varying Proportion Of Contamination

σ can give a significant boost to performance while over estimating can degrade it. This is consistent with the performance of UCB algorithms in practice, which often scale the exploration term to improve empirical performance (Y.-E. Liu et al., 2014).

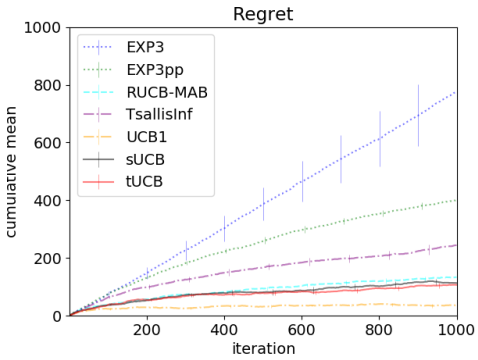
To look at the impact of using a contamination robust algorithm when there is no contamination, we plotted various α values when $\varepsilon = 0$, shown in Figure 2. Assuming small amounts of contamination when there is none only



(a) $\alpha = 0$



(b) $\alpha = 0.05$

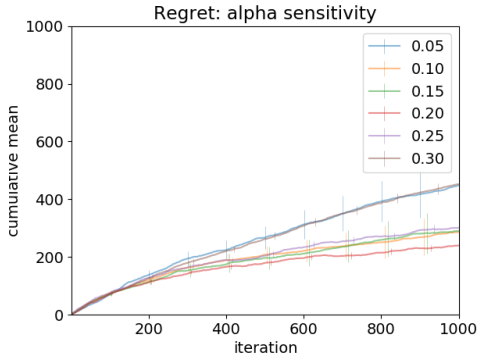


(c) $\alpha = 0.1$

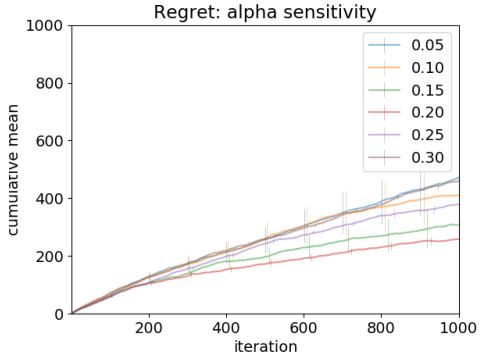
Figure 2: Misspecified α For $\varepsilon = 0$.

has a small impact on performance, suggesting it is permissible to use contamination robust methods when there is uncertainty of contamination. Similarly, small K and large Δ can render bounded contamination impotent and would not require algorithms that account for it.

We have included RUCB-MAB in our simulations because it is simple to implement and can perform similarly well to our algorithms. We note it currently has guarantees only for Gaussian rewards (Kapoor, Patel, and Kar, 2018).



(a) sUCB



(b) tUCB

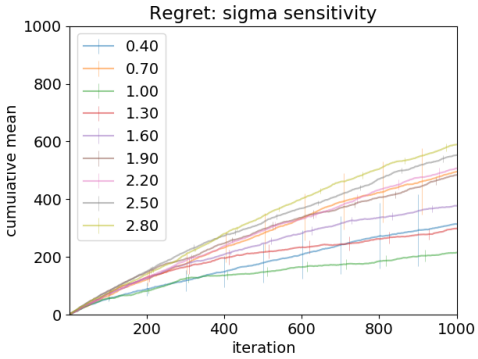
Figure 3: Regret Sensitivity For Various α.

Figure 5 shows the poor performance of all algorithms when the first ε rewards are contaminated. TsallisInf and EXP3++ show some recovery, but it is clear this type of adversary is harmful. This remains an open problem for scenarios with small T.

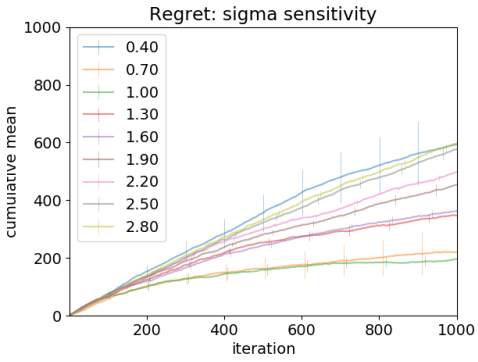
We also considered including the BARBAR algorithm (Gupta, Koren, and Talwar, 2019) whose epoch scheme is the only algorithm we know that accounts for the front cluster attack. We chose against this as for our setting of $T = 1000$ the BARBAR algorithm only has one epoch, and thus does not make any updates to the estimated gaps, resulting in pure random exploration.

6 DISCUSSION

We have presented two variants of an ε-contamination robust UCB algorithm to handle uninformative or malicious rewards in the stochastic bandit setting. As the main contribution, we proved concentration inequalities for the α-trimmed and α-shorth mean in the ε-contamination setting with sub-Gaussian samples and guarantees on the uncontaminated regret of the crUCB algorithms. The regret guarantees are similar to those in



(a) sUCB



(b) tUCB

Figure 4: Regret Sensitivity For Various σ.

the uncontaminated sMAB setting.

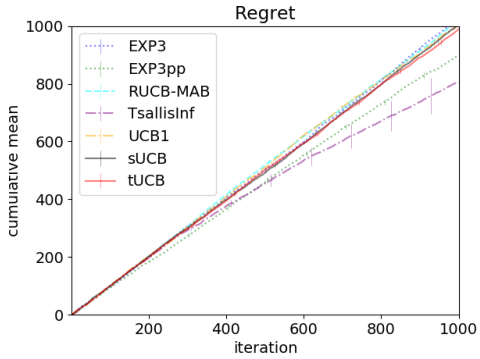
We have shown through simulation that these algorithms can outperform "best of both worlds" algorithms and those for stochastic or adversarial environments when using a small number of iterations and ε chosen to be reasonable when implementing bandits in education.

We highlight that our algorithms are simple to implement. In practice, it is often easy to find upper bounds on the parameters which are robust to underestimation. Our algorithms are numerically stable and have clear intuition to their actions.

A weak point of these algorithms is they require knowledge of α before hand. Choices of α may come from domain knowledge, but could also require a separate study.

In this work we assumed a fully adaptive adversarial contamination, constrained only by the total fraction of contamination at any time step. By making more assumptions about the adversary, it is likely possible to improve uncontaminated regret bounds.

Limitations The adversary used in the simulation is quite simple and does not take full advantage of the



(a) $\varepsilon = .1$

Figure 5: Front Cluster Attack

power we allow in our model. We designed it as a first test of our algorithms and associated theory. In the future, we would like to design simulated adversaries that are modeled on real world contamination. It will also be important to deploy contamination robust algorithms in the real world. This will require thought on how to select various tuning parameters ahead of the deployment.

There remain many open questions in this area. In particular, we think this work could be improved along the following directions.

Randomized algorithms UCB-type algorithms are often outperformed in applications by the randomized Thompson sampling algorithm. Creating a randomized algorithm that accounts for the contamination model would increase the practicality of this line of work.

Contamination correlated with true rewards One possibility is that the contaminated rewards contain information of the true rewards. For example if contamination can be missing data, we know dropout can be correlated with the treatment condition.

Acknowledgements

L.N. acknowledges the support of NSF via grant DMS-1646108 and thanks Joseph Jay Williams for helpful discussions and for inspiring this work. A.T. would like to acknowledge the support of a Sloan Research Fellowship and NSF grant CAREER IIS-1452099.

References

Ahler, Douglas J, Carolyn E Roush, and Gaurav Sood (2019). "The micro-task market for lemons: Data quality on Amazon's Mechanical Turk". In: *Meeting of the Midwest Political Science Association*.

Altschuler, Jason, Victor-Emmanuel Brunel, and Alan Malek (2019). "Best Arm Identification for Contaminated Bandits". In: *Journal of Machine Learning Research* 20, pp. 1–39.

Auer, Peter and Nicolo Cesa-Bianchi (2002). "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine learning,* p. 22.

Auer, Peter, Nicolò Cesa-Bianchi, et al. (Jan. 2002). "The Nonstochastic Multiarmed Bandit Problem". In: *SIAM Journal on Computing* 32.1, pp. 48–77.

Auer, Peter and Chao-Kai Chiang (May 2016). "An Algorithm with Nearly Optimal Pseudo-Regret for Both Stochastic and Adversarial Bandits". In: *Conference on Learning Theory*, pp. 116–120.

Bai, Yang et al. (Aug. 2018). "Comparative Evaluation of Heart Rate-Based Monitors: Apple Watch vs Fitbit Charge HR". In: *Journal of Sports Sciences* 36.15, pp. 1734–1741.

Bubeck, Sebastien and Aleksandrs Slivkins (Feb. 2012). "The Best of Both Worlds: Stochastic and Adversarial Bandits". In: *Conference on Learning Theory*.

Crussell, Jonathan, Ryan Stevens, and Hao Chen (2014). "MAdFraud: Investigating Ad Fraud in Android Applications". In: *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '14*. Bretton Woods, New Hampshire, USA: ACM Press, pp. 123–134.

Curran, Paul G. (Sept. 2016). "Methods for the detection of carelessly invalid responses in survey data". In: *Journal of Experimental Social Psychology* 66, pp. 4–19.

Diakonikolas, Ilias et al. (2019). "Robust Estimators in High Dimensions without the Computational Intractability". In: *SIAM Journal on Computing* 48.2, pp. 742–864.

Feehan, Lynne M et al. (Aug. 2018). "Accuracy of Fitbit Devices: Systematic Review and Narrative Syntheses of Quantitative Data". In: *JMIR mHealth and uHealth* 6.8, e10527.

Gupta, Anupam, Tomer Koren, and Kunal Talwar (2019). "Better Algorithms for Stochastic Bandits with Adversarial Corruptions". In: *Conference on Learning Theory*, pp. 1562–1578.

Huber, Peter (Mar. 1964). "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1, pp. 73–101.

Jun, Kwang-Sung et al. (2018). "Adversarial attacks on stochastic bandits". In: *Advances in Neural Information Processing Systems*, pp. 3640–3649.

Kapoor, Sayash, Kumar Kshitij Patel, and Purushottam Kar (Aug. 2018). "Corruption-Tolerant Bandit Learning". In: *Machine Learning*, pp. 1–29.

Kothari, Pravesh K, Jacob Steinhardt, and David Steurer (2018). "Robust moment estimation and improved clustering via sum of squares". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1035–1046.

Lai, K. A., A. B. Rao, and S. Vempala (Oct. 2016). "Agnostic Estimation of Mean and Covariance". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 665–674.

Lattimore, Szepesvari (2020). *Bandit Algorithms*. Cambridge University Poress.

Liu, Liu, Tianyang Li, and Constantine Caramanis (Jan. 2019). "High Dimensional Robust Estimation of Sparse Models via Trimmed Hard Thresholding". In: *arXiv preprint*.

Liu, Yun-En et al. (2014). "Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits." In: *EDM*, pp. 161–168.

Lykouris, Thodoris, Vahab Mirrokni, and Renato Paes Leme (Mar. 2018). "Stochastic Bandits Robust to Adversarial Corruptions". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pp. 114–122.

Ma, Yuzhe et al. (2019). "Data Poisoning Attacks in Contextual Bandits". In: *International Conference on Decision and Game Theory for Security*.

Necka, Elizabeth A. et al. (June 2016). "Measuring the Prevalence of Problematic Respondent Behaviors among MTurk, Campus, and Community Participants". In: *PLOS ONE* 11.6. Ed. by Jelte M. Wicherts, e0157732.

Pearce, Paul et al. (2014). "Characterizing Large-Scale Click Fraud in ZeroAccess". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*. Scottsdale, Arizona, USA: ACM Press, pp. 141–152.

Ryan, Timothy (2018). *Data Contamination on MTurk — Timothy J. Ryan*. en-US.

Seldin, Yevgeny and Gábor Lugosi (July 2017). "An Improved Parametrization and Analysis of the EXP3++ Algorithm for Stochastic and Adversarial Bandits". In: *Proceedings of the 2017 Conference on Learning Theory*. Vol. 65. PMLR, pp. 1743–1759.

Seldin, Yevgeny and Aleksandrs Slivkins (2014). "One Practical Algorithm for Both Stochastic and Adversarial Bandits." In: *ICML*, pp. 1287–1295.

Williams, Joseph Jay et al. (2016). "AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning". In: *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. Edinburgh, Scotland, UK: ACM Press, pp. 379–388.

Zimmert, Julian and Yevgeny Seldin (2019). "An Optimal Algorithm for Stochastic and Adversarial Bandits". In: *Proceedings of Machine Learning Research*. Vol. 89, pp. 467–475.

A Proofs

A.1 Theorem 1

Theorem 1 (Trimmed mean concentration)**.** *Let G be the set of points $x_1, ...x_n \in \mathbb{R}$ that are drawn from a σ-sub-Gaussian distribution with mean μ. Let S_n be a sample where an ε-fraction of these points are contaminated by an adversary. For $\varepsilon \leq \alpha < 1/2$, $t \geq n$ we have,*

$$|\mathit{trMean}_\alpha(S_n) - \mu| \leq \frac{\sigma}{(1-2\alpha)}\left(\sqrt{\frac{4}{n}\log(t)} + 4\alpha\sqrt{6\log(t)}\right)$$

with probability at least $1 - \frac{4}{t^2}$.

Proof of theorem 1. Without loss of generality assume $\mu = 0$ for the underlying true distribution. For $X \sim \sigma$-sub-Gaussian, by definition, we have:

$$P\Big(|X| \geq \mu + \eta\Big) \leq 2\exp(-\frac{\eta^2}{2\sigma^2})$$
$$P\left(|\bar{x}_n - \mu| \geq \sigma\sqrt{\frac{2}{n}\log\frac{2}{\delta_1}}\right) \leq \delta_1$$

and

$$P\Big(\max_{i\in[n]}|X_i| \geq t\Big) \leq 2n\exp\Big(-\frac{t^2}{2\sigma^2}\Big)$$
$$P\Big(\max_{i\in[n]}|X_i| \geq \sigma\sqrt{2\log\frac{2n}{\delta_2}}\Big) \leq \delta_2.$$

Let $\tilde{G} \subset G_n$ represent the points which are not contaminated and $C \subset G_n$ represent the contaminated points. Then our sample can be represented by the union $S_n = \tilde{G} \cup C$. Let R represent the points that remain after trimming α fraction of the largest and smallest points, and T be the set of points that were trimmed. Then we have that.

$$\begin{aligned}
|\text{trMean}_\alpha(S_n)| &= \left|\frac{1}{(1-2\alpha)n}\sum_{x\in R} x\right| \\
&= \frac{1}{(1-2\alpha)n}\left|\sum_{x\in\tilde{G}\cap R} x + \sum_{x\in C\cap R} x\right| \\
&\leq \frac{1}{(1-2\alpha)n}\left|\underbrace{\sum_{x\in\tilde{G}} x}_{A_1} - \underbrace{\sum_{x\in\tilde{G}\cap T} x}_{A_2} + \underbrace{\sum_{x\in C\cap R} x}_{A_3}\right| \\
&\leq \frac{1}{(1-2\alpha)n}\left(\left|\underbrace{\sum_{x\in\tilde{G}} x}_{A_1}\right| + \left|\underbrace{\sum_{x\in\tilde{G}\cap T} x}_{A_2}\right| + \left|\underbrace{\sum_{x\in C\cap R} x}_{A_3}\right|\right)
\end{aligned}$$

with

$$A_1 = \left|\sum_{x\in G_n} x - \sum_{x\in G_n\setminus\tilde{G}} x\right| \leq \left|\sum_{x\in G_n} x\right| + \left|\sum_{x\in G_n\setminus\tilde{G}} x\right| \leq n|\bar{x}_{G_n}| + \varepsilon n\max_{x\in G_n}|x| \quad \text{w.p. at least } 1-\delta_1-\delta_2,$$
$$A_2 \leq 2\alpha n\max_{x\in G_n}|x| \quad \text{w.p. at least } 1-\delta_2,$$
$$A_3 \leq \varepsilon n\max_{x\in G_n}|x| \quad \text{w.p. at least } 1-\delta_2.$$

Combining we get,

$$\begin{aligned}|\text{trMean}_\alpha(S_n) - \mu| &\leq \frac{1}{(1-2\alpha)}\left(|\bar{x}_{G_n}| + \max_{x\in G_n}|x|(2\varepsilon + 2\alpha)\right)\\ &\leq \frac{1}{(1-2\alpha)}\left(|\bar{x}_{G_n}| + \max_{x\in G_n}|x|(4\alpha)\right)\\ &\leq \frac{\sigma}{(1-2\alpha)}\left(\sqrt{\frac{2}{n}\log\frac{2}{\delta_1}} + 4\alpha\sqrt{2\log\frac{2t}{\delta_2}}\right)\end{aligned}$$

with probability at least $1-\delta_1-\delta_2$. Letting $\delta_1 = \frac{2}{t^2}$ and $\delta_2 = \frac{2}{t^2}$, and assuming $\alpha \geq \varepsilon$, we have,

$$|\text{trMean}_\alpha(S_n) - \mu| \leq \frac{\sigma}{(1-2\alpha)}\left(\sqrt{\frac{4}{n}\log(t)} + 4\alpha\sqrt{6\log(t)}\right)$$

with probability at least $1-\frac{4}{t^2}$. □

A.2 Theorem 2

Theorem 2 (α-shorth mean concentration)**.** *Let G_n be the set of points $x_1, ...x_n \in \mathbb{R}$ that are drawn from a σ-sub-Gaussian distribution with mean μ. Let S_n be a sample where an ε-fraction of these points are contaminated by an adversary. For $\varepsilon \leq \alpha < 1/3$, $t \geq n$, we have,*

$$\begin{aligned}&|\textit{sMean}_\alpha(S_n) - \mu| \leq\\ &\qquad\frac{\sigma}{1-2\alpha}\sqrt{\frac{4}{n}\log t} + \frac{(6\alpha - 8\alpha^2)\sigma}{(1-2\alpha)(1-\alpha)}\sqrt{6\log t}\end{aligned}$$

with probability at least $1-\frac{4}{t^2}$.

Proof of theorem 2. Without loss of generality assume $\mu = 0$ for the underlying true distribution. Let $X \sim \sigma$-sub-Gaussian.

We want to bound the impact of the contaminated points in our interval. Once we have this bound, the proof follows just as in the trimmed mean.

Assume $\alpha < 1/3$ and $\varepsilon \leq \alpha$. Let J be the interval that contains the shortest $1-\alpha$ fraction of S_n, I be the interval that contains $\tilde{G}$ (i.e. the remaining good points after contamination), and T be the interval that contains the points of S_n after trimming the α largest and smallest fraction of points. Use $|I|$ to denote the length of interval I. It must be that $I \cap J \neq \emptyset$ because otherwise the points in $I \cup J$ would contain $2-2\alpha > 1$ fraction of S_n. Let c be a point in $I \cap J$ and x be a point in J. Recall that $\text{trMean}_\alpha(S_n)$ is the trimmed mean of the contaminated sample S_n from above. Then we have,

$$\begin{aligned}|x| &\leq |x-c| + |c - \text{trMean}_\alpha(S_n)| + |\text{trMean}_\alpha(S_n)|\\ &\leq |J| + |I| + |\text{trMean}_\alpha(S_n)|\\ &\leq 2|I| + |\text{trMean}_\alpha(S_n)|\end{aligned}$$

The second step comes from x and c both being in J and because $I \supseteq T$. The third step comes from $|J| \leq |I|$.

To bound the length of I we have,

$$|I| \leq 2\max_{x\in G_n}|x| \text{ w.p. at least } 1-\delta_2.$$

Finally, since

$$|\text{trMean}_\alpha(S_n)| \leq \frac{1}{(1-2\alpha)}(|\bar{x}_{G_n}| + 4\alpha\max_{x\in G_n}|x|)$$

with probability at least $1-\delta_1-\delta_2$, we get that for $x \in J$,

$$\begin{aligned}|x| &\le 4\max_{x\in G_n}|x| + \frac{1}{(1-2\alpha)}(|\bar{x}_{G_n}| + 4\alpha\max_{x\in G_n}|x|) && \text{w.p. at least } 1-\delta_1-\delta_2,\\ &= \frac{|\bar{x}_{G_n}|}{1-2\alpha} + \Big(4+\frac{4\alpha}{1-2\alpha}\Big)\max_{x\in G_n}|x|.\end{aligned}$$

Now that we have a bound on the contaminated points in J, our analysis follows as before,

$$\begin{aligned}&|\text{sMean}_\alpha(S_n)|\\ &\qquad\le \frac{1}{(1-\alpha)n}\Bigg(\Big|\underbrace{\sum_{x\in\tilde{G}} x}_{A_1}\Big| + \Big|\underbrace{\sum_{x\in\tilde{G}\cap\neg J} x}_{A_2}\Big| + \Big|\underbrace{\sum_{x\in C\cap J} x}_{A_3}\Big|\Bigg)\end{aligned}$$

where

$$\begin{aligned}A_1 &\le n|\bar{x}_{G_n}| + \varepsilon n\max_{x\in G_n}|x| && \text{w.p. at least } 1-\delta_1-\delta_2,\\ A_2 &\le \alpha n\max_{x\in G_n}|x| && \text{w.p. at least } 1-\delta_2,\\ A_3 &\le \varepsilon n\left(\frac{|\bar{x}_{G_n}|}{1-2\alpha} + \Big(4+\frac{4\alpha}{1-2\alpha}\Big)\max_{x\in G_n}|x|\right) && \text{w.p. at least } 1-\delta_1-\delta_2.\end{aligned}$$

Combining we get,

$$\begin{aligned}&|\text{sMean}_\alpha(S_n)-\mu|\\ &\le \frac{1}{(1-\alpha)}\left(|\bar{x}_{G_n}|\big(1+\frac{\varepsilon}{1-2\alpha}\big) + \max_{x\in G_n}|x|\big(5\varepsilon+\alpha+\frac{4\alpha\varepsilon}{1-2\alpha}\big)\right)\\ &\le \frac{1}{(1-\alpha)}\left(|\bar{x}_{G_n}|\big(\frac{1-\alpha}{1-2\alpha}\big) + \max_{x\in G_n}|x|\big(6\alpha+\frac{4\alpha^2}{1-2\alpha}\big)\right)\\ &= \frac{1}{1-\alpha}\left(|\bar{x}_{G_n}|\big(\frac{1-\alpha}{1-2\alpha}\big) + \max_{x\in G_n}|x|\frac{6\alpha-8\alpha^2}{1-2\alpha}\right)\\ &\le \frac{\sigma}{1-2\alpha}\sqrt{\frac{2}{n}\log\frac{2}{\delta_1}} + \frac{(6\alpha-8\alpha^2)\sigma}{(1-2\alpha)(1-\alpha)}\sqrt{2\log\frac{2t}{\delta_2}}\end{aligned}$$

With probability at least $1-\delta_1-\delta_2$. Letting $\delta_1 = \frac{2}{t^2}$ and $\delta_2 = \frac{2}{t^2}$, and assuming $\alpha \ge \varepsilon$, we have,

$$\begin{aligned}&|\text{sMean}_\alpha(S_n)-\mu|\\ &\qquad\le \frac{\sigma}{1-2\alpha}\sqrt{\frac{4}{n}\log t} + \frac{(6\alpha-8\alpha^2)\sigma}{(1-2\alpha)(1-\alpha)}\sqrt{6\log t}\end{aligned}$$

With probability at least $1-\frac{4}{t^2}$. □

A.3 Theorem 3

Theorem 3 (α-trimmed mean crUCB uncontaminated regret)**.** *Let $K > 1$ and $T \ge K-1$. Then with algorithm 4 with the α-trimmed mean, σ-sub-Gaussian reward distributions with $\sigma_a \le \sigma_0$, and contamination rate $\varepsilon \le \alpha \le \frac{\Delta_{min}}{4(\Delta_{min}+4\sigma_0\sqrt{6\log T})}$, we have the uncontaminated regret bound,*

$$\bar{R}(UCB) \le 8\sigma_0\sqrt{KT\log T} + \sum 15\Delta_a.$$

Proof of theorem 3. First will show that $\mathbb{E}[N_a(t)] < \infty$ for non-optimal actions. Assume $N_a(t) \geq \frac{64\sigma_0^2 \log(T)}{\Delta_a^2}$.

$$
\begin{aligned}
&\hat{\mu}_a + \frac{\sigma_0}{(1-2\alpha)}\left(\sqrt{\frac{4}{N_a(t)}\log t} + 4\alpha\sqrt{6\log(t)}\right) \\
&\leq \mu_a + \frac{\sigma_i + \sigma_0}{(1-2\alpha)}\left(\sqrt{\frac{4}{N_a(t)}\log t} + 4\alpha\sqrt{6\log(t)}\right) && \text{w.p. at least } 1 - \frac{4}{t^2} \\
&\leq \mu^* - \Delta_a + \frac{2\sigma_0}{(1-2\alpha)}\left(\sqrt{\frac{4}{N_a(t)}\log t} + 4\alpha\sqrt{6\log(t)}\right) \\
&\leq \mu^* - \Delta_a + \frac{\Delta_a}{2(1-2\alpha)} + \frac{2\sigma_0 4\alpha}{(1-2\alpha)}\sqrt{6\log t} && N_a(t) \geq \frac{64\sigma_0^2\log(T)}{\Delta_a^2} \\
&\leq \mu^* && \alpha \leq \frac{\Delta_a}{4(\Delta_a + 4\sigma_0\sqrt{6\log(t))}} \\
&\leq \hat{\mu}^* + \frac{\sigma_{i^*}}{(1-2\alpha)}\left(\sqrt{\frac{4}{N^*(t)}\log t} + 4\alpha\sqrt{6\log(t)}\right) && \text{w.p. at least } 1 - \frac{4}{t^2} \\
&\leq \hat{\mu}^* + \frac{\sigma_0}{(1-2\alpha)}\left(\sqrt{\frac{4}{N^*(t)}\log t} + 4\alpha\sqrt{6\log(t)}\right).
\end{aligned}
$$

Now to find $\mathbb{E}[N_a(T)]$ for non-optimal actions.

$$
\begin{aligned}
\mathbb{E}[N_a(T)] &= 1 + \mathbb{E}\Big[\sum_{t=K+1}^{T} \mathbf{1}\{A_t = a\}\Big] \\
&= 1 + \mathbb{E}\Big[\sum_{t=K+1}^{T} \mathbf{1}\Big\{A_t = a, N_a(t) \leq \frac{64\sigma_0^2\log(T)}{\Delta_a^2}\Big\} + \mathbf{1}\Big\{A_t = a, N_a(t) > \frac{64\sigma_0^2\log(T)}{\Delta_a^2}\Big\}\Big] \\
&\leq 1 + \frac{64\sigma_0^2\log(T)}{\Delta_a^2} + \sum_{t=K+1}^{T} \mathbb{P}\Big[A_t = a, N_a(t) > \frac{64\sigma_0^2\log(T)}{\Delta_a^2}\Big] \\
&= 1 + \frac{64\sigma_0^2\log(T)}{\Delta_a^2} + \sum_{t=K+1}^{T} \mathbb{P}\Big[A_t = a | N_a(t) > \frac{64\sigma_0^2\log(T)}{\Delta_a^2}\Big]\mathbb{P}\Big[N_a(t) > \frac{64\sigma_0^2\log(T)}{\Delta_a^2}\Big] \\
&\leq 1 + \frac{64\sigma_0^2\log(T)}{\Delta_a^2} + \sum_{t=K+1}^{T} \frac{8}{t^2} \\
&\leq \frac{64\sigma_0^2\log(T)}{\Delta_a^2} + 15.
\end{aligned}
$$

Finally, we can find the regret following the standard analysis,

$$
\begin{aligned}
\bar{R} &= \sum_{a=2}^{K} \Delta_a \mathbb{E}[N_a(T)] \\
&= \sum_{\Delta_a < \Delta} \Delta_a \mathbb{E}[N_a(T)] + \sum_{\Delta_a \geq \Delta} \Delta_a \mathbb{E}\Big[N_a(T)\Big] \\
&\leq \Delta T + \sum_{\Delta_a \geq \Delta} \big[\frac{64\sigma_0^2\log(T)}{\Delta_a} + 15\Delta_a\big] && \mathbb{E}[N_a(t)] \leq \frac{64\sigma_0^2\log(T)}{\Delta_a} + 15 \\
&\leq 8\sigma_0\sqrt{KT\log(T)} + \sum 15\Delta_a && \Delta = \sqrt{\frac{64K\sigma_0^2\log(T)}{T}}.
\end{aligned}
$$

□

A.4 Corollary 1

Corollary 1 (α-trimmed mean crUCB uncontaminated regret bounded rewards)**.** *If the rewards are bounded by b, and have contamination rate $\varepsilon \leq \alpha \leq \frac{\Delta_{\min}}{4(\Delta_{\min}+4b)}$, then*

$$\bar{R}_T \leq 8\sigma_0\sqrt{KT\log(T)} + \sum 15\Delta_a.$$

Proof of corollary 1. By replacing the part of the concentration bound for the trimmed mean that is based on the maximum value in the sample with b, we get that,

$$|\text{trMean}_\alpha(S_n) - \mu| \leq \frac{\sigma}{(1-2\alpha)}\sqrt{\frac{4}{n}\log(t)} + \frac{4\alpha}{1-2\alpha}b$$

with probability at least $1 - \frac{4}{t^2}$.

First will show that $\mathbb{E}[N_a(t)] < \infty$ for non-optimal actions. Assume $N_a(t) \geq \frac{64\sigma_0^2\log(T)}{\Delta_a^2}$.

$$\begin{aligned}
&\hat{\mu}_a + \frac{\sigma_0}{(1-2\alpha)}\sqrt{\frac{4}{N_a(t)}\log t} + \frac{4\alpha}{1-2\alpha}b \\
&\leq \mu_a + \frac{\sigma_i + \sigma_0}{(1-2\alpha)}\sqrt{\frac{4}{N_a(t)}\log t} + \frac{8\alpha}{1-2\alpha}b && \text{w.p. at least } 1 - \frac{4}{t^2} \\
&\leq \mu^* - \Delta_a + \frac{2\sigma_0}{(1-2\alpha)}\sqrt{\frac{4}{N_a(t)}\log t} + \frac{8\alpha}{1-2\alpha}b \\
&\leq \mu^* - \Delta_a + \frac{\Delta_a}{2(1-2\alpha)} + \frac{8\alpha}{(1-2\alpha)}b && N_a(t) \geq \frac{64\sigma_0^2\log(T)}{\Delta_a^2} \\
&\leq \mu^* && \alpha \leq \frac{\Delta_a}{4(\Delta_a+4b)} \\
&\leq \hat{\mu}^* + \frac{\sigma_{i^*}}{(1-2\alpha)}\sqrt{\frac{4}{N^*(t)}\log t} + \frac{4\alpha}{1-2\alpha}b && \text{w.p. at least } 1 - \frac{4}{t^2} \\
&\leq \hat{\mu}^* + \frac{\sigma_0}{(1-2\alpha)}\sqrt{\frac{4}{N^*(t)}\log t} + \frac{4\alpha}{1-2\alpha}b.
\end{aligned}$$

Results follow with a similar analysis as above. □

A.5 Theorem 4

Theorem 4 (α-shorth mean crUCB uncontaminated regret)**.** *Let $K > 1$ and $T \geq K - 1$. Then with algorithm 4 with the α-shorth mean, sub-Gaussian reward distributions with $\sigma_a \leq \sigma_0$, and contamination rate $\varepsilon \leq \alpha \leq \frac{\Delta_{min}}{4(\Delta_{min}+9\sigma_0\sqrt{6\log T})}$, we have the uncontaminated regret bound,*

$$\bar{R}(UCB) \leq 8\sigma_0\sqrt{KT\log T} + \sum 15\Delta_a.$$

Proof of theorem 4. The proof for the contamination robust UCB using the α-shorth mean is similar to that of the trimmed mean.

$$
\begin{aligned}
&\hat{\mu}_a + \frac{\sigma_0}{1-2\alpha}\sqrt{\frac{4}{N_a(t)}\log t} + \frac{(6\alpha - 8\alpha^2)\sigma}{(1-2\alpha)(1-\alpha)}\sqrt{6\log t} \\
&\leq \mu^* - \Delta_a + \frac{2\sigma_0}{1-2\alpha}\sqrt{\frac{4}{N_a(t)}\log t} + 2\frac{(6\alpha - 8\alpha^2)\sigma_0}{(1-2\alpha)(1-\alpha)}\sqrt{\log t} && \text{w.p.a.l } 1 - \frac{4}{t^2} \\
&\leq \mu^* - \Delta_a + \frac{\Delta_a}{2(1-2\alpha)} + \frac{18\alpha\sigma_0}{(1-2\alpha)}\sqrt{6\log t} && N_a(t) \geq \frac{64\sigma_0^2\log(t)}{\Delta_a^2},\ \alpha < 1/3 \\
&\leq \mu^* && \alpha \leq \frac{\Delta_a}{4(\Delta_a + 9\sigma_0\sqrt{6\log t})} \\
&\leq \hat{\mu}^* + \frac{\sigma_0}{1-2\alpha}\sqrt{\frac{4}{N^*(t)}\log t} + \frac{6\alpha - 8\alpha^2\sigma}{(1-2\alpha)(1-\alpha)}\sqrt{6\log t}
\end{aligned}
$$

Using the analysis from the trimmed mean regret, we again get,

$$\mathbb{E}[N_a(t)] \leq \frac{64\sigma_0^2\log T}{\Delta_a} + \sum 15\Delta_a$$

Using this value and standard regret analysis yields

$$\bar{R}_T \leq 8\sigma_0\sqrt{KT\log(T)} + \sum 15\Delta_a.$$

□

A.6 Corollary 2

Corollary 2 (α-shorth mean crUCB uncontaminated regret bounded rewards)**.** *If the rewards are bounded by b, and have contamination rate* $\varepsilon \leq \alpha \leq \frac{\Delta_{\min}}{4(\Delta_{\min}+9b)}$*, then*

$$\bar{R}_T \leq 8\sigma_0\sqrt{KT\log(T)} + \sum 15\Delta_a.$$

Proof of corollary 2. By replacing the part of the concentration bound for the trimmed mean that is based on the maximum value in the sample with b, we get that,

$$|\text{sMean}_\alpha(S_n) - \mu| \leq \frac{\sigma}{1-2\alpha}\sqrt{\frac{4}{n}\log t} + \frac{6\alpha - 8\alpha^2}{(1-2\alpha)(1-\alpha)}b$$

With probability at least $1 - \frac{4}{t^2}$.

Follow similar analysis as in section A.4 but setting constraint to be,

$$\varepsilon \leq \alpha \leq \frac{\Delta_{\min}}{4(\Delta_{\min} + 9b)}$$

□

B Relationship of ε and $\Delta_{\min}$

One quick example showing that $\varepsilon > \Delta_{\min}$ can prohibit sublinear regret is to consider the CSB game with two actions and Bernoulli rewards. If $a_1 \sim B(p)$ and $a_2 \sim B(p - \varepsilon)$ then an adversary can choose all the contaminated rewards for a_2 to be 1 making it appear that $a_2 \sim B(p)$. Thus the actions are indistinguishable to the learner.

However, we can still provide a bound for larger values of ε provided one is willing to tolerate a linear term in the regret. We outline the argument only for the trimmed mean case since the argument for the shorth mean is very similar. Note that argument for bounding $\mathbb{E}[N_a(T)]$ in Theorem 3 works under the condition

$$\alpha \leq \frac{\Delta_a}{4(\Delta_a + 4\sigma_0\sqrt{6\log(T)})}.$$

Let $\mathcal{S}$ be the set of actions satisfying this condition. The arguments in the proof of Theorem 3 show that

$$\sum_{a>1, a\in\mathcal{S}} \Delta_a \mathbb{E}[N_a(T)] \leq 8\sigma_0\sqrt{KT\log(T)} + \sum_{a>1,a\in\mathcal{S}} 15\Delta_a.$$

Therefore the bound of $\tilde{O}(\sigma_0\sqrt{KT})$ holds only for the regret due to actions $a \in \mathcal{S}$. For any action $a \notin \mathcal{S}$, we have

$$\Delta_a < \frac{16\alpha\sigma_0\sqrt{6\log(T)}}{1-4\alpha}$$

assuming $\alpha < 0.25$. The total regret contribution for $a \notin \mathcal{S}$ is therefore

$$\begin{aligned}\sum_{a>1,a\notin\mathcal{S}} \Delta_a \mathbb{E}[N_a(T)] &\leq \frac{16\alpha\sigma_0\sqrt{6\log(T)}}{1-4\alpha} \sum_{a>1,a\notin\mathcal{S}} \mathbb{E}[N_a(T)] \\ &\leq \frac{16\alpha\sigma_0\sqrt{6\log(T)}}{1-4\alpha} T\end{aligned}$$

So the total regret is $\tilde{O}(\sqrt{KT} + \frac{\alpha}{1-4\alpha}T)$.