Robust Stochastic Bandit Algorithms under Probabilistic Unbounded Adversarial Attack

Ziwei Guan,[1] **Kaiyi Ji,**[1] **Donald J. Bucci Jr.,**[2] **Timothy Y. Hu,**[2]
Joseph Palombo,[2] **Michael Liston,**[2] **Yingbin Liang**[1]
[1]The Ohio State University, ECE Department
2015 Neil Ave, Columbus, OH 43210
[2]Lockheed Martin Advanced Technology Laboratories
Cherry Hill, NJ, 08002, USA

Abstract

The multi-armed bandit formalism has been extensively studied under various attack models, in which an adversary can modify the reward revealed to the player. Previous studies focused on scenarios where the attack value either is bounded at each round or has a vanishing probability of occurrence. These models do not capture powerful adversaries that can catastrophically perturb the revealed reward. This paper investigates the attack model where an adversary attacks with a certain probability at each round, and its attack value can be arbitrary and unbounded if it attacks. Furthermore, the attack value does not necessarily follow a statistical distribution. We propose a novel sample median-based and exploration-aided UCB algorithm (called med-E-UCB) and a median-based ϵ-greedy algorithm (called med-ϵ-greedy). Both of these algorithms are provably robust to the aforementioned attack model. More specifically we show that both algorithms achieve $\mathcal{O}(\log T)$ pseudo-regret (i.e., the optimal regret without attacks). We also provide a high probability guarantee of $\mathcal{O}(\log T)$ regret with respect to random rewards and random occurrence of attacks. These bounds are achieved under arbitrary and unbounded reward perturbation as long as the attack probability does not exceed a certain constant threshold. We provide multiple synthetic simulations of the proposed algorithms to verify these claims and showcase the inability of existing techniques to achieve sublinear regret. We also provide experimental results of the algorithm operating in a cognitive radio setting using multiple software-defined radios.

1 Introduction

Stochastic multi-armed bandit models capture the scenarios where a player devises a strategy in order to access the optimal arm as often as possible. Such models have been used in a broad range of applications including news article recommendation (Li et al. 2010), online advertising (Pandey et al. 2007), medical treatment allocation (Kuleshov and Precup 2014), and adaptive packet routing (Awerbuch and Kleinberg 2004). As security concerns have a critical impact in these applications, stochastic multi-armed bandit models under adversarial attacks have attracted extensive attention. A variety of attack models have been studied under the multi-armed bandit formalism. Below we briefly summarize major models that are relevant to our study.

- The adversarial multi-armed bandit model, in which an adversary is allowed to attack in each round with each attack subject to a bounded value. (Auer et al. 2002) proposed a robust EXP3 algorithm as a defense algorithm and (Audibert and Bubeck 2009; Stoltz 2005; Bubeck and Cesa-Bianchi 2012) further provided tighter bounds. (Jun et al. 2018; Liu and Shroff 2019) showed that a small total attack cost of $\mathcal{O}(\log T)$ makes UCB and ϵ-greedy algorithms fail with regret $\mathcal{O}(T)$.
- The budget-bounded attack model, in which an adversary has a total budget of attack value but can choose to attack only over some time instances. The aim of defense is to achieve a regret that gradually transits between adversarial (the above always attack) and stochastic (never attack) models. (Lykouris, Mirrokni, and Paes Leme 2018) provided a variety of such robust algorithms and (Gupta, Koren, and Talwar 2019) further developed an algorithm that improved the regret in (Lykouris, Mirrokni, and Paes Leme 2018).
- The fractional attack model, in which the total rounds that an adversary attacks is limited either by probability that the adversary can attack or by the ratio of attacked rounds to total rounds. The attack value at each round is also subject to a bounded value. (Kapoor, Patel, and Kar 2019) proposed a robust RUCB-MAB algorithm, which uses sample median to replace sample mean in UCB algorithm. (Seldin and Slivkins 2014) proposed an EXP3-based algorithm which can achieve optimality in both stochastic and adversarial cases.
- The heavy-tail outlier model, in which the observed reward can have heavy-tail values, whose distribution has bounded first moment and unbounded second moment. (Bubeck, Cesa-Bianchi, and Lugosi 2013) proposed a robust Cantoni UCB algorithm that can defend against such heavy-tail outliers.

We observe that all of the above adversarial models assume that the attack value (i.e., the adversarial cost) either is bounded or has vanishing probability of occurrence. In this paper, we study an adversarial attack model where the attack value can be *arbitrary* and *unbounded*. To elaborate

further, *an arbitrary attack value* allows flexible and adaptive attack strategies, which may not follow a probabilistic distribution. *Unboundedness* allows arbitrarily large attack values to occur with constant probability. Under such an attack model, it is impossible to defend if the adversary can attack at each round. Thus, we assume that the adversary attack with a fixed probability ρ at each round (as justified in (Kapoor, Patel, and Kar 2019; Altschuler, Brunel, and Malek 2019)).

Such an attack model turns out to be quite challenging due to arbitrary and unbounded attack values. As we demonstrate in Section 5, the existing popular (robust) algorithms fail to defend, even when attack values are not substantially large all the time. These algorithms include (a) vanilla UCB and ϵ-greedy, which are mean-based and clearly are vulnerable under arbitrarily large attack; (b) EXP3, designed to succeed typically under bounded attack values; (c) Cantoni UCB for heavy-tail bandit (Bubeck, Cesa-Bianchi, and Lugosi 2013), which requires large valued outliers to occur with asymptotically small probability; (d) RUCB-MAB (Kapoor, Patel, and Kar 2019), also designed to succeed typically under bounded attack values.

The contribution of this paper lies in proposing two novel robust median-based bandit algorithms to defend from arbitrary and unbounded attack values, and furthermore developing sharp bounds of their regret performance.

1.1 Our Contributions

We summarize our contributions as follows.

- We propose a novel **median-based exploration-aided UCB algorithm (med-E-UCB)** by incorporating a diminishing number of periodic exploration rounds. In contrast to RUCB-MAB in (Kapoor, Patel, and Kar 2019) (which directly replaces sample mean in vanilla UCB by sample median), our med-E-UCB adds a small amount of exploration, which turns out to be critical for maintaining logarithmic regret. We further propose a **median-based ϵ-greedy algorithm (med-ϵ-greedy)**, for which logarithmic regret is achieved without additional exploration.
- For both med-E-UCB and med-ϵ-greedy, we show that, even under arbitrary and unbounded attacks, they achieve an optimal $\mathcal{O}(\log T)$ pseudo-regret bound for no attack scenarios, as long as the attack probability ρ does not exceed a certain constant threshold. This allows the number of attacks to scale linearly. We also provide a high-probability analysis with respect to both the randomness of reward samples and the randomness of attacks, so that $\mathcal{O}(\log T)$ regret is guaranteed under almost all trajectory.
- We develop a new analysis mechanism to deal with the technical challenge of incorporating the sample median into the analysis of bandit problems. Direct use of the existing concentration bounds on the sample median does not provide a guarantee of $\mathcal{O}(\log T)$ regret for our algorithms. In fact, it turns out to be nontrivial to maintain the concentration of sample median (which requires sufficient exploration of each arm) and at the same time best control the exploration to keep the scaling of the regret at the $\mathcal{O}(\log T)$ level. Such an analysis provides insight for us to design exploration procedure in med-E-UCB.
- We provide the synthetic demonstrations and experimental radio results from a realistic cognitive radio setting that demonstrate the robustness of the proposed algorithms and verify their $\mathcal{O}(\log T)$ regret under large valued attacks. We demonstrate in both sets of experiments that existing algorithms fail to achieve logarithmic regret under the proposed attack model.

All the technical proofs of the theorems in the paper can be found in the full version of this work posted on arXiv.

1.2 Related Works

Stochastic vs adversarial multi-armed bandit. Under an adversarial bandit model where attacks occur at each round, (Jun et al. 2018; Liu and Shroff 2019) showed that UCB and ϵ-greedy fail with regret of $\mathcal{O}(T)$ while sustaining only a small total attack cost of $\mathcal{O}(\log T)$. Then, (Bubeck and Slivkins 2012; Seldin and Slivkins 2014; Auer and Chiang 2016; Seldin and Lugosi 2017; Zimmert and Seldin 2019; Gupta, Koren, and Talwar 2019) designed robust algorithms that achieve $\mathcal{O}(\sqrt{T})$ regret for an adversarial bandit and $\mathcal{O}(\log T)$ for stochastic bandit without attack. Furthermore, (Lykouris, Mirrokni, and Paes Leme 2018; Gupta, Koren, and Talwar 2019) provided robust algorithms and characterized the regret under the model where the fraction of attacks ranging from always-attacking to never-attacking. (Kapoor, Patel, and Kar 2019) proposed a median-based UCB algorithm, and derived the same type of regret for a similar but probabilistic model, where the adversary attacks at each round with a certain probability. Other similar intermediate models were also studied in (Zimmert and Seldin 2019), where either the ratio of attacked rounds to total rounds or the value of attacks are constrained. All these studies assume a bounded range for the attack value at each round, whereas our study allows arbitrary and unbounded attack values.

(Bubeck, Cesa-Bianchi, and Lugosi 2013) proposed Cantoni UCB algorithm for a heavy-tail bandit, and showed that the algorithm can tolerate outlier samples. Though their heavy-tail distributions allow outliers to occur with fairly high probability as compared to sub-Gaussian distributions, our adversarial model is much more catastrophic. It allows the attack distribution to have an unbounded mean, whereas the heavy-tail distribution still requires a finite mean and certain higher order moments. For example, under our attack model, an adversary can attack only the optimal arm with probability ρ by subtracting a sufficiently large constant c, so that the optimal arm no longer has the largest sample mean. Consequently, Cantoni UCB fails with the regret increasing linearly with T. The experiment in Section 5 also shows that Cantoni UCB fails under our attack model.

Median-based robust algorithms. The sample median is well known to be more robust than the sample mean in statistics (Tyler 2008; Zhang, Chi, and Liang 2016). Hence, the sample median has been used in a variety of contexts to design robust algorithms in multi-armed bandit problems (Altschuler, Brunel, and Malek 2019), parameter recovery in phase retrieval (Zhang, Chi, and Liang 2018), and regression (Klivans, Kothari, and Meka 2018). In this paper, the analysis methods are novel and we provide high probability guarantees of $\mathcal{O}(\log T)$ regret (rather than just in average).

We also note that the RUCB-MAB (Kapoor, Patel, and Kar 2019) directly replaces the sample mean by the sample median in UCB, which as shown in Section 5 fails to defend against arbitrary and unbounded attack.

2 Problem Formulation

Consider a K-armed bandit, where each arm i exhibits a stationary reward distribution with a cumulative probability distribution function (CDF) F_i and mean μ_i. Denote the arm with the maximum mean reward as i^*, and assume that it is unique. Let $\mu^* = \mu_{i^*}$ represent the maximum mean reward and $\Delta_i = \mu^* - \mu_i$ for all $i \neq i^*$. Throughout the paper, we assume that μ_i for $i = 1, ..., K$ are fixed constants and do not scale with the number of pulls T.

At each round t, the player can pull any arm, $i \in \{1, ..., K\}$. Then the bandit generates a reward $\tilde{X}_{i,t}$ according to the distribution F_i.

There exists an adversary, who decides to attack with probability $0 < \rho < 1$, independently of the history of arm pulls by the player. If the adversary attacks, it adds an attack value η_t to the reward so that the player observes a perturbed reward $X_{i,t} = \tilde{X}_{i,t} + \eta_t$; If the adversary does not attack, the player observes a clean reward $X_{i,t} = \tilde{X}_{i,t}$. That is,

$$X_{i,t} = \begin{cases} \tilde{X}_{i,t} + \eta_t, & \text{with probability } \rho; \\ \tilde{X}_{i,t}, & \text{with probability } 1 - \rho. \end{cases} \quad (1)$$

We emphasize that in our attack model, with a constant probability $\rho > 0$, the attack value η_t can be arbitrarily large. Furthermore, the realizations of η_t do not follow any statistical model, which is much more catastrophic than the typical heavy-tail distributions(Bubeck, Cesa-Bianchi, and Lugosi 2013). Since attack values can be arbitrary, our attack model allows the adversary to *adaptively* design its attack strategy based on reward history and distributions, as long as it attacks with probability ρ.

For a bandit algorithm, we define the *pseudo-regret* as

$$\bar{R}_T = \mu^* T - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right], \quad (2)$$

where I_t denotes the index of the arm pulled at time t, and the expectation is over both stochastic rewards as well as the random occurrence of attacks. It measures how the reward obtained by the algorithm deviates from that received by the optimal strategy in expectation. Furthermore, in practical adversarial scenarios, it is of great interest to characterize the regret in reward trajectory. Thus, we define the following stronger notion of the *regret*

$$R_T = \mu^* T - \sum_{t=1}^{T} \mu_{I_t}. \quad (3)$$

Here, R_T is a random variable with respect to the random occurrence of attacks and reward values, and in general is a function of attack values.

The goal of this paper is to design algorithms that minimize the pseudo-regret and more importantly minimize the regret with high probability. More importantly, the latter condition guarantees robust operation over almost all reward trajectories.

Notations: For a given cumulative distribution function (CDF) F, we define its generalized p-quantile (where $0 < p < 1$) function as $\theta_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$. For a sample sequence $\{x_i\}_{i=1}^m$, let $\hat{F}$ be its empirical distribution. Then let $\theta_p(\{x_i\}_{i=1}^m)$ be the p-quantile of $\hat{F}$, i.e., $\theta_p(\{x_i\}_{i=1}^m) = \theta_p(\hat{F})$. If $p = 1/2$, we obtain the median of the sequence given by $\mathrm{med}(\{x_i\}_{i=1}^m) := \theta_{1/2}(\{x_i\}_{i=1}^m) = \theta_{1/2}(\hat{F})$.

We use $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with mean μ and variance σ^2, and use $\Phi(\cdot)$ to denote the CDF of the standard Gaussian distribution $\mathcal{N}(0, 1)$.

In this paper, $\lceil \cdot \rceil$ denotes the nearest bigger integer, $\lfloor \cdot \rfloor$ denotes the nearest smaller integer, and $[K]$ represents the set $\{1, 2, ..., K\}$. Furthermore, $y = \mathcal{O}(f(x))$ represents that there exists constants $M > 0, \zeta > 0$ such that $y \leq Mf(x)$ for all $x \geq \zeta$. And $\log(\cdot)$ denotes the natural logarithm with the base e. For a differentiable function f, we write its derivative as f'.

3 Median-based and Exploration-aided UCB

In this section, we first propose a median-based UCB algorithm, and then show that such an algorithm can defend against the attack model described in Section 2.

3.1 Algorithm Overview

We begin by explaining why direct replacement of sample mean by sample median in UCB (Kapoor, Patel, and Kar 2019) cannot defend against large attack values. Consider a catastrophic attack scheme that our attack model allows, where the adversary sets $\eta_t = -\infty$ or a significantly large negative value in the case when the player pulls the optimal arm, and $\eta = 0$ otherwise. Then, the first time that the player pulls the optimal arm, the adversary attacks with a positive probability $\rho > 0$, resulting in the value of $\mathrm{med}_{i^*}(t) + \sqrt{\frac{\omega \log t}{T_j(t)}}$ being $-\infty$, where $\mathrm{med}_{i^*}(t)$ denotes the sample median of the rewards received by arm i^* up to time t. Consequently, the optimal arm will never be pulled in the future, and hence the regret grows linearly with T. The primary reason that median-based vanilla UCB fails in such a case is due to insufficiently enforced exploration for each arm. That is, the sample median can fail with only one catastrophic sample if there are not enough samples. On the other hand, if there are further enforced explorations to pull the optimal arm, the sample median can eventually rule out outlier attack values since such an attack occurs only probabilistically and not all the time. The above understanding motivates us to design an exploration-aided UCB algorithm, and then incorporate the sample median to defend against large attack values. We call this algorithm med-E-UCB and describe it formally in Algorithm 1. This idea is illustrated in Figure 1, where we divide the pulling rounds into blocks. Each block consists of G rounds, where $G \geq Kb \log G$ and $b > 0$ is an arbitrary constant. During the first block (i.e. $k = 0$), the size of pure exploration round is fixed at $b \log G$ pulls per arm. Except for the first block, at the beginning of

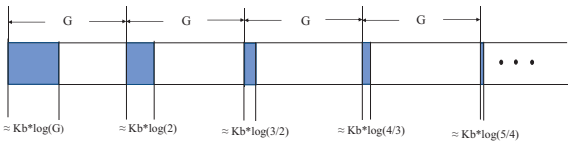


Figure 1: An illustration of med-E-UCB scheme, where the blue blocks denote the pure exploration rounds and the white blocks denote the UCB rounds.

each block, say block k, each arm is approximately explored $b\log\frac{k+1}{k}$ rounds. As a result, each arm is guaranteed to have been pulled $b\log((k+1)G)$ times at block k. So that pure exploration does not significantly affect the regret.

Algorithm 1 med-E-UCB

Input: Number of arms K, group size G, exploration parameters b,ω, and total rounds T.

1: Initialization: for the first $K\lceil b\log G\rceil$ rounds, pull each arm $\lceil b\log G\rceil$ times.
2: **for** $t = K\lceil b\log G\rceil + 1, ..., T$ **do**
3: $k = \lfloor \frac{t}{G} \rfloor$;
4: **if** $kG + 1 \le t \le kG + K(\lceil b\log(k+1)G\rceil - \lceil b\log kG\rceil)$ **then**
5: Pure Exploration:
$I_t = \left\lceil \frac{t-kG}{\lceil b\log(k+1)G\rceil - \lceil b\log kG\rceil} \right\rceil$;
6: **else**
7: UCB round:
$I_t = \underset{j}{\operatorname{argmax}} \left\{ \operatorname{med}_j(t-1) + \sqrt{\frac{\omega\log t}{T_j(t-1)}} \right\}$,
8: **end if**
9: **end for**

3.2 Analysis of Regret

In this subsection, we analyze the regret of med-E-UCB. The distributions associated with the arm are not necessarily Gaussian, and need only satisfy the following assumption.

Assumption 1. *There exists a constant s, such that $\theta_{\frac{1}{2}-s}(F_{i^*}) > \theta_{\frac{1}{2}+s}(F_j)$ for all $j \neq i^*$. Moreover, $F_i(\cdot)$ is differentiable for all $i \in [K]$, and there exist constants $l > 0$ and $\xi > 0$, such that*

$$\inf\{F'_{i^*}(x) :\ \theta_{\frac{1}{2}-s}(F_{i^*}) - \xi < x < \theta_{\frac{1}{2}-s}(F_{i^*})\} \ge l\ ,$$

and for all $j \neq i^*$

$$\inf\{F'_j(x) :\ \theta_{\frac{1}{2}+s}(F_j) < x < \theta_{\frac{1}{2}+s}(F_j) + \xi\} \ge l.$$

The above assumption essentially requires that the median of the optimal arm and the median of the non-optimal arms have gaps such that the optimal arm can stand out statistically. This assumption further requires that the probability density within a ξ-neighborhood of the median to be lower-bounded by a positive l in order to guarantee a good concentration property. Clearly, Gaussian distributions satisfy Assumption 1.

Lemma 1 (Sample median concentration bound)**.** *Let $X_i = \tilde{X}_i + \eta_i, i = 1, ..., n$ be n attacked data samples, where $\tilde{X}_i, i = 1, ..., n$ are original (i.e., un-attacked) data samples i.i.d. drawn from the distribution with CDF F. The η_i's are unbounded attack values. If $\sum_{i=1}^n \mathbb{1}\{\eta_i \neq 0\} \le s\cdot n$ holds for a constant $s \in (0,1)$, then for any $a, b > 0$, we have*

$$\mathbb{P}\left(\operatorname{med}(\{X_i\}_{i=1}^n) - \theta_{\frac{1}{2}-s}(F) \le -a\right) \le \exp\left(-2np_1^2\right),$$

$$\mathbb{P}\left(\operatorname{med}(\{X_i\}_{i=1}^n) - \theta_{\frac{1}{2}+s}(F) \ge b\right) \le \exp\left(-2np_2^2\right),$$

where $p_1 = \frac{1}{2} - s - F(\theta_{\frac{1}{2}-s}(F) - a)$, and $p_2 = F(\theta_{\frac{1}{2}+s}(F) + b) - \frac{1}{2} - s$.

Using Lemma 1, we obtain the following regret bounds for med-E-UCB.

Theorem 1. *Consider the stochastic multi-armed bandit problem as described in* (1)*. Suppose Assumption 1 holds. Further assume that the attack probability $\rho < s$, total number of rounds $T > G$, $b \ge \max\{\frac{\omega}{\xi^2}, \frac{2}{(s-\rho)^2}\}$ and $\omega \ge \frac{2}{l^2}$. Then the pseudo-regret of med-E-UCB satisfies*

$$\begin{aligned}\bar{R}_T \le & \sum_{j=1, j\neq i^*}^{K} \Delta_j b\log(2T) \\ & + \sum_{j=1, j\neq i^*}^{K} \Delta_j \left(\frac{4\omega\log T}{(\theta_{\frac{1}{2}-s}(F_{i^*}) - \theta_{\frac{1}{2}+s}(F_j))^2} \right) \\ & + \sum_{j=1, j\neq i^*}^{K} \Delta_j (2 + \frac{2\pi^2}{3}).\end{aligned}$$

For constant K, Δ_j, b and ω, $\bar{R}_T = \mathcal{O}(\log(T))$.

Theorem 1 implies that med-E-UCB achieves the best possible pseudo-regret bound under the attack-free model (Lai and Robbins 1985). Considering that the number of attacks can scale linearly with the total number of pulling rounds and attack values can be unbounded and arbitrary, Theorem 1 demonstrates that med-E-UCB is robust algorithm against very powerful attacks.

Furthermore, we establish a stronger *high-probability* guarantee for the regret of med-E-UCB with respect to both the randomness of attack occurrence and rewards.

Theorem 2. *Suppose Assumption 1 holds. Assume that the attack probability $\rho < s$, total number of rounds $T > G$, $b \ge \max\{\frac{\omega}{\xi^2}, \frac{2}{(s-\rho)^2}\}$ and $\omega \ge \frac{3.5}{l^2}$. Then, with probability at least $1-\delta$ with respect to the randomness of attack occurrence and rewards, the regret of med-E-UCB satisfies*

$$\begin{aligned}R_T \le & \sum_{j=1, j\neq i^*} \Delta_j b\log(2T) \\ & + \sum_{j=1, j\neq i^*} \Delta_j \left(\frac{4\omega\log T}{(\theta_{\frac{1}{2}-s}(F_{i^*}) - \theta_{\frac{1}{2}+s}(F_j))^2} \right) \\ & + \sum_{j=1, j\neq i^*} \Delta_j \left(e\left(\frac{bK}{2\delta}\right)^{\frac{1}{4}} + \frac{2K}{\delta} + 3 + \frac{\pi^2}{3} \right).\end{aligned}$$

For constant K,Δ_j, b and ω, $R_T = \mathcal{O}(\log T + \frac{1}{\delta})$.

In practice, the high-probability result as Theorem 2 is much more desirable. In such a case, we would like the guarantee of successful defense for almost all realizations of the attack (i.e., with high probability) rather than an on-average performance which does not imply what happens for each attack realization.

Theorems 1 and 2 readily imply the following corollary for Gaussian distributions. To present the result, let the ith arm be associated with $\mathcal{N}(\mu_i, \sigma^2)$. Further let $\Delta_{min} := \min_{i \neq i^*}\{\mu^* - \mu_i\}$, and $l = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\Delta_{min}+4)^2}{32\sigma^2}\right)$, where $\Phi(\cdot)$ denotes the CDF of $\mathcal{N}(0,1)$.

Corollary 1. *Suppose each arm corresponds to a Gaussian distribution. Suppose* $\rho < \Phi(\frac{\Delta_{min}}{4\sigma}) - \frac{1}{2}$, $b \geq \max\{\omega, \frac{2}{(\Phi(\Delta_{min}/(4\sigma))-1/2-\rho)^2}\}$, *and* $\omega \geq \frac{2}{l^2}$. *Then, the pseudo-regret of med-E-UCB satisfies* $\bar{R}_T = \mathcal{O}(\log T)$. *And, with probability at least* $1-\delta$ *with respect to the randomness of both attacks and rewards, the regret of med-E-UCB satisfies* $R_T = \mathcal{O}\left(\log(T) + (\frac{1}{\delta})^{\frac{1}{4}}\right)$.

4 Median-based ϵ-greedy

In this section, we propose a robust ϵ-greedy algorithm based on the sample median, and show that it is robust to defend against the adversarial attack described in Section 2. This algorithm is helpful to compare with med-E-UCB to illustrate that under unbounded attacks med-E-UCB is more exploration-efficient.

4.1 Algorithm Overview

We propose an ϵ-greedy algorithm that incorporates the sample median to defend against adversarial attacks. We call the algorithm med-ϵ-greedy and describe it formerly in Algorithm 2. Compared with the standard ϵ-greedy algorithm, the med-ϵ-greedy algorithm replaces the sample mean by the sample median. In addition, the exploration parameter c needs to be appropriately chosen to provide sufficient exploration to guarantee the concentration of the sample median.

Algorithm 2 med-ϵ-greedy

Input: Number K of arms, total number of T pulling, and exploration parameter c.

1: Initialization: pull each arm $\lceil c \rceil$ times.
2: **for** $t = \lceil c \rceil K + 1, ..., T$ **do**
3: Pull arm

$$I_t = \begin{cases} \underset{j}{\text{argmax}}\{\text{med}_j(t-1)\}, & \text{w.p. } 1 - \frac{cK}{t} \\ \text{Uniformly pick an arm from 1 t o K, w.p. } \frac{cK}{t} \end{cases}$$

4: **end for**

4.2 Analysis of Regret

In this subsection, we analyze both the pseudo-regret and regret of the med-ϵ-greedy algorithm. We first make the following assumption on the reward distributions.

Assumption 2. *There exists a constant* $0 < s < 1$ *and a constant* $x_0 \in \mathbb{R}$, *such that the CDF* $F(\cdot)$ *of the optimal arm satisfies* $F_{i^*}(x_0) < \frac{1}{2} - s$, *and the CDFs of the remaining arms satisfy* $F_j(x_0) > \frac{1}{2} + s$, *for all* $j \neq i^*$.

The above assumption ensures that the sample median of the optimal arm is larger than those of the other arms with a desirable gap. Compared to Assumption 1 for med-E-UCB, Assumption 2 is slightly weaker as it does not need the CDF to be differentiable and its derivative to be bounded below in the neighborhood of $\frac{1}{2} + s$ or $\frac{1}{2} - s$ quantiles. Clearly, a collection of Gaussian distributions with a unique largest mean satisfies Assumption 2.

The following theorem characterizes the pseudo-regret bound for the med-ϵ-greedy algorithm.

Theorem 3. *Consider the stochastic multi-armed bandit problem under adversarial attack as described in* (1). *Let Assumption 2 hold. Suppose* $\rho < s$, *and suppose the exploration parameter* c *satisfies the following condition* $c > \max\{20, \frac{2}{\left(F_j(x_0)-\frac{1}{2}-s\right)^2}, \frac{2}{\left(\frac{1}{2}-s-F_{i^*}(x_0)\right)^2}, \frac{2}{(s-\rho)^2} : j = 1, 2, ...K, j \neq i^*\}$. *Then the pseudo-regret of med-ϵ-greedy satisfies*

$$\bar{R}_T \leq c \sum_{j=1, j\neq i^*}^{K} \Delta_j \log T + 2cKe\mu^* + \sum_{j=1, j\neq i^*}^{K} (2+3c)\Delta_j,$$

For fixed Δ_j, K, *and* c, $\bar{R}_T = \mathcal{O}(\log T)$.

Theorem 3 indicates that even under adversarial attack, med-ϵ-greedy still achieves $\mathcal{O}(\log T)$ regret, which is the same as the optimal pseudo-regret order in attack-free model. In contrast to med-E-UCB, exploration rounds in vanilla ϵ-greedy are already sufficient for the sample median to be effective.

Aside from the pseudo-regret bound, we further provide a high-probability guarantee for the regret below.

Theorem 4. *Given Assumption 2, suppose* $\rho < s$, *and the exploration parameter* c *satisfies the following condition* $c > \max\{40, \frac{4}{\left(F_j(x_0)-\frac{1}{2}-s\right)^2}, \frac{4}{\left(\frac{1}{2}-s-F_{i^*}(x_0)\right)^2}, \frac{1}{(s-\rho)^2} : j = 1, 2, ...K, j \neq i^*\}$. *Then, with probability at least* $1-\delta$ *with respect to the randomness of both attacks and rewards, the regret of med-ϵ-greedy satisfies*

$$R_T \leq \frac{6\lceil c \rceil^2 K^3}{\delta}\mu^* + \sum_{j=1, j\neq i^*}^{K} 2c\Delta_j \log T + \sum_{j=1, j\neq i^*}^{K} 2c\Delta_j.$$

For constant Δ_j, K, *and* c, $R_T = \mathcal{O}(\log T + \frac{1}{\delta})$

Theorems 3 and 4 readily implies the result when all arms correspond to Gaussian distributions, which we state in the following corollary. Similar to Corollary 1 for med-E-UCB, for Gaussian distributions, we have derived the threshold for ρ below which med-ϵ-greedy has the desired regret. To present the result, suppose the ith arm is associated with $\mathcal{N}(\mu_i, \sigma^2)$, and let $\Delta_{min} := \min_{i \neq i^*}\{\mu^* - \mu_i\}$.

Corollary 2. *Suppose each arm corresponds to a Gaussian distribution, and* $\rho < \Phi(\frac{\Delta_{min}}{4\sigma}) - \frac{1}{2}$. *Let*

$c > \max\{10, \frac{1}{\left(\Phi(\frac{\Delta_{min}}{2\sigma})-\Phi(\frac{\Delta_{min}}{4\sigma})\right)^2}, \frac{1}{(\Phi(\frac{\Delta_{min}}{4\sigma})-\frac{1}{2}-\rho)^2}\}$.

Then, the pseudo-regret of med-ϵ-greedy satisfies $\bar{R}_T = \mathcal{O}(\log T)$.

Furthermore, with probability at least $1-\delta$ *with respect the randomness of both attacks and rewards, the regret of med-ϵ-greedy satisfies* $R_T = \mathcal{O}(\log T + \frac{1}{\delta})$.

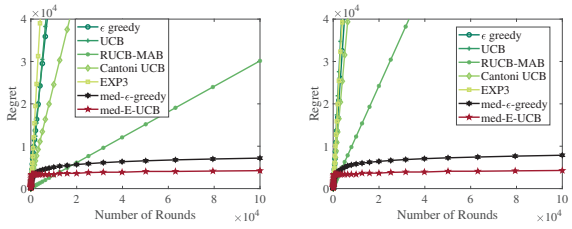
5 Experiments

5.1 Comparison among Algorithms

In this subsection, we provide experiments to demonstrate that existing robust algorithms fail under the attack model considered here, whereas our two algorithms are successful.

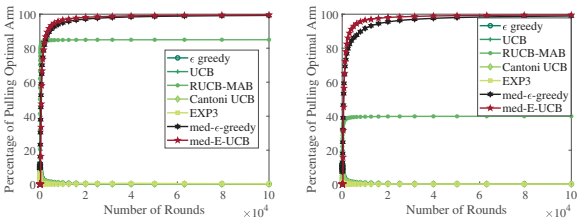
In our experiment, we choose the number of arms to be 10. The reward distribution of the ith arm is $\mathcal{N}(2i, 1)$ for $i \in [K]$. The attack probability is fixed to be ρ ($\rho = 0.125$ and 0.3). The adversary generates an attack value η uniformly at random from the interval $(0, 1800)$ if it attacks, and subtracts the clean reward by η if the optimal arm is pulled and adds to the clean reward otherwise. For each algorithm, each trial contains $T = 10^5$ rounds, and the final results take the average of 20 Monte Carlo trials.

We first compare the performance of our med-E-UCB and med-ϵ-greedy with RUCB-MAB (Kapoor, Patel, and Kar 2019), EXP3 (Auer et al. 2002), and Cantoni UCB (Bubeck, Cesa-Bianchi, and Lugosi 2013). We also include (vanilla) UCB(Auer, Cesa-Bianchi, and Fischer 2002) and (vanilla) ϵ-greedy (Auer, Cesa-Bianchi, and Fischer 2002) in the comparison for completeness. For med-E-UCB, we set $b = 4, \omega = 4$, and $G = 10^3$. For med-ϵ-greedy, we set $c = 10$. Other parameters are set as suggested by the original references. It can be seen from Figure 2 that our med-E-UCB and med-ϵ-greedy algorithms significantly outperform the other algorithms with respect to the average regret. It is also clear that only our med-E-UCB and med-ϵ-greedy algorithms have logarithmically increasing regret, whereas all other algorithms suffer linearly increasing regret. Between our two algorithms, med-E-UCB performs slightly better than med-ϵ-greedy. This implies that the med-E-UCB is more exploration efficient than med-ϵ-greedy. The same observations can also be made in Figure 3, where the performance metric is the percentage of pulling the optimal arm. Only our med-E-UCB and med-ϵ-greedy algorithms asymptotically approaches 100% optimal arm selection rate.



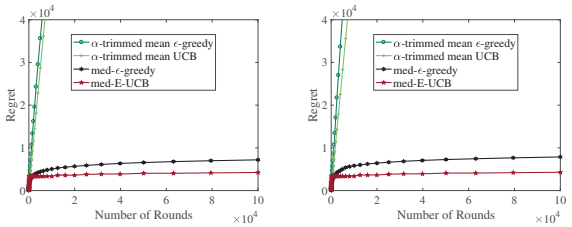
(a) $\rho = 0.125$ (b) $\rho = 0.3$

Figure 2: Comparison of regret among algorithms



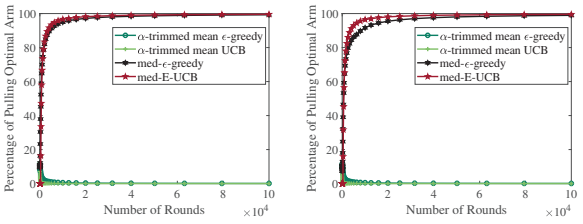
(a) $\rho = 0.125$ (b) $\rho = 0.3$

Figure 3: Comparison of percentage of pulling the optimal arm among algorithms



(a) $\rho = 0.125$ (b) $\rho = 0.3$

Figure 4: Comparison of regret among algorithms



(a) $\rho = 0.125$ (b) $\rho = 0.3$

Figure 5: Comparison of percentage pulling of optimal arm among algorithms

We further note that as shown in Figures 2 and 3, RUCB-MAB in (Kapoor, Patel, and Kar 2019) does not defend against large valued attacks. This fact indicates that direct replacement of sample mean by sample median as in RUCB-MAB does not yield sufficiently robust performance. As a result, the diminishing periodic exploration in our med-E-UCB turns out to play a critical role in its successful defense.

We further compare the performance of our med-E-UCB and med-ϵ-greedy with a so-called α-trimmed scheme (Bednar and Watt 1984), which is a very popular method in signal processing to deal with similar unbounded arbitrary attacks. The idea of such a scheme is to remove the top and bottom α fraction of samples before calculating the sample mean in order to eliminate the influence of outliers. It can be seen from Figures 4 and 5 that med-E-UCB and med-ϵ-greedy significantly outperform the α-trimmed UCB and α-trimmed ϵ-greedy algorithms.

5.2 Experiment over Cognitive Radio Testbed

In this subsection, we present experimental results over a wireless over-the-air radio testbed (Figure 6) to validate the performance of med-E-UCB and med-ϵ-greedy. The testbed models a pair of secondary users in a cognitive radio network opportunistically sharing spectrum resources with a primary user (not shown) while simultaneously defending against a stochastic adversarial jammer. We model the channel selection problem of the secondary users as a multi-armed bandit and use the channel signal to interference and noise ratio (SINR) as a measure of reward. The SINR is approximated using the inverse of error vector magnitude, which has a linear relationship for the range of signal powers we measure in the software defined radio (SDR). The transmitted packet signal power is constant throughout the experiment so that lower noise channels have a greater SINR compared with noisier channels which have a lower SINR. We model our unbounded adversary according to (1). If the adversary attacks, a fixed power noise jamming attack is placed over top the secondary user signal packet, significantly reducing SINR by 40dB (i.e, 4 orders of magnitude). Each radio node uses an Ettus Research Universal Software Radio Peripheral (USRP) B200 Software Defined Radio (SDR).

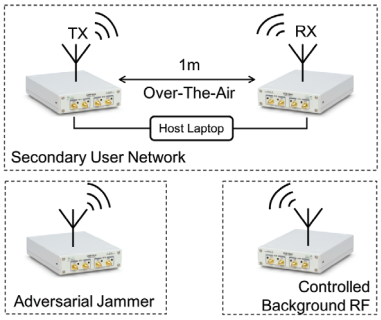


Figure 6: Cognitive radio testbed hardware setup. The receive node (RX) selects a channel according to its assigned policy and communicates the selected channel to the transmit node (TX) via an ACK channel. TX transmits a 500 kHz bandwidth QPSK signal in the 1.2 GHz UHF band. RX receives a reward based on the selected channel conditions.

We center our experiment in the 1.2 GHz UHF frequency band across a 5 MHz RF bandwidth. Channel SINR is controlled using an additional USRP to transmit frequency varying noise to create five contiguous 500 kHz channels centered between 1200 MHz and 1202 MHz (Figure 7). The rewards of the 5 channels without adversarial perturbation are normally distributed with mean SINR of [41, 37, 35, 31, 28] dB and unit variance. Mean SINR of adversarially attacked rounds range between 5 to 10 dB. We use identical parameters as our simulations from Section 5.1, with $c = 10$ for med-ϵ-greedy and $b = 4, \omega = 4$, for med-E-UCB. Similarly, the attack probabilities are fixed to be ρ which equals either 0.125 or 0.3. Due to hardware timing constraints, rounds occur at 2-second intervals. So we reduce the total number of rounds $T = 2000$ for each experiment and set $G = 200$ for med-E-UCB.

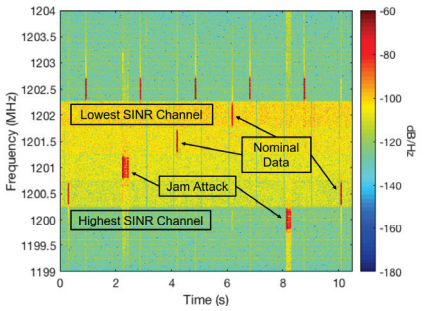
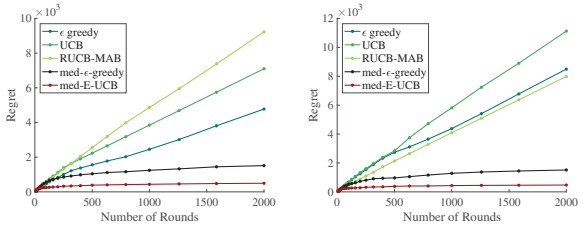
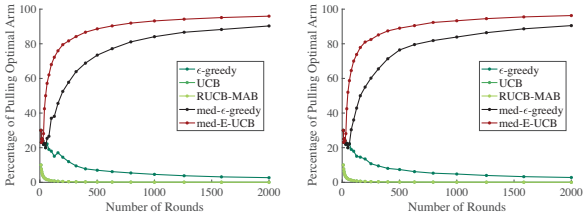


Figure 7: Example spectrogram of received data. Five individual 500 kHz channels are available for selection between center frequencies 1200 MHz and 1202 MHz with increasing amounts of background noise. Data transmission are sent every 2-seconds indicated by thin red data packets. Some proportion of data transmissions are overlaid with thicker red bars, indicating an adversarial attack.



(a) $\rho = 0.125$ (b) $\rho = 0.3$

Figure 8: Comparison of regret among algorithms



(a) $\rho = 0.125$ (b) $\rho = 0.3$

Figure 9: Comparison of percentage of pulling the optimal arm among algorithms

The experiment results are illustrated in Figures 8 and 9, which indicate that med-E-UCB and med-ϵ-greedy achieve logarithmic regret compared with other algorithms including RUCB-MAB as well as mean UCB and mean ϵ-greedy. Similarly, our algorithms both eventually converge to a 100% pull rate of the optimal arm.

6 Conclusion

In this work, we proposed two median-based bandit algorithms, which we show to be robust under probabilistic unbounded valued adversarial attacks. Our median-based

method can be potentially applied to many other models including multi-player bandits (Gai and Krishnamachari 2011) and UCT (Kocsis and Szepesvári 2006).

Acknowledgment

The work of Z. Guan, K. Ji and Y. Liang is also supported in part by U.S. National Science Foundation under the grants CCF-1801846 and ECCS-1818904.

References

Altschuler, J.; Brunel, V.-E.; and Malek, A. 2019. Best arm identification for contaminated bandits. *Proceedings of Machine Learning Research (PMLR)*.

Audibert, J.-Y., and Bubeck, S. 2009. Minimax policies for adversarial and stochastic bandits.

Auer, P., and Chiang, C.-K. 2016. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proc. Conference on Learning Theory (COLT)*, 116–120.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1):48–77.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2-3):235–256.

Awerbuch, B., and Kleinberg, R. D. 2004. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proc. Annual ACM Symposium on Theory of Computing*, 45–53.

Bednar, J., and Watt, T. 1984. Alpha-trimmed means and their relationship to median filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32(1):145–153.

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

Bubeck, S., and Slivkins, A. 2012. The best of both worlds: stochastic and adversarial bandits. In *Proc. Conference on Learning Theory (COLT)*, 42–1.

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

Gai, Y., and Krishnamachari, B. 2011. Decentralized online learning algorithms for opportunistic spectrum access. In *IEEE Global Telecommunications Conference (GLOBECOM)*, 1–6.

Gupta, A.; Koren, T.; and Talwar, K. 2019. Better algorithms for stochastic bandits with adversarial corruptions. *Proc. Conference of Learning Theory (COLT)*.

Jun, K.-S.; Li, L.; Ma, Y.; and Zhu, J. 2018. Adversarial attacks on stochastic bandits. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 3640–3649.

Kapoor, S.; Patel, K. K.; and Kar, P. 2019. Corruption-tolerant bandit learning. *Machine Learning* 108(4):687–715.

Klivans, A.; Kothari, P. K.; and Meka, R. 2018. Efficient algorithms for outlier-robust regression. *Proc. Conference On Learning Theory (COLT)* 1420–1430.

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

Kuleshov, V., and Precup, D. 2014. Algorithms for multi-armed bandit problems. *Journal of Machine Learning Research*.

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

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proc. International Conference on World Wide Web*, 661–670.

Liu, F., and Shroff, N. 2019. Data poisoning attacks on stochastic bandits. In *Proc. International Conference on Machine Learning (ICML)*.

Lykouris, T.; Mirrokni, V.; and Paes Leme, R. 2018. Stochastic bandits robust to adversarial corruptions. In *Proc. Annual ACM SIGACT Symposium on Theory of Computing*, 114–122.

Pandey, S.; Agarwal, D.; Chakrabarti, D.; and Josifovski, V. 2007. Bandits for taxonomies: A model-based approach. In *Proc. SIAM International Conference on Data Mining*, 216–227.

Seldin, Y., and Lugosi, G. 2017. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Proc. Conference on Learning Theory (COLT)*.

Seldin, Y., and Slivkins, A. 2014. One practical algorithm for both stochastic and adversarial bandits. In *Proc. International Conference on Machine Learning (ICML)*, 1287–1295.

Stoltz, G. 2005. *Incomplete information and internal regret in prediction of individual sequences*. Ph.D. Dissertation.

Tyler, D. E. 2008. Robust statistics: Theory and methods. *Journal of the American Statistical Association* 103(482):888–889.

Zhang, H.; Chi, Y.; and Liang, Y. 2016. Provable non-convex phase retrieval with outliers: Median truncated-wirtinger flow. In *Proc. International Conference on Machine Learning (ICML)*, 1022–1031.

Zhang, H.; Chi, Y.; and Liang, Y. 2018. Median-truncated nonconvex approach for phase retrieval with outliers. *IEEE Transactions on Information Theory* 64:7287–7310.

Zimmert, J., and Seldin, Y. 2019. An optimal algorithm for stochastic and adversarial bandits. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.