Near Optimal Adversarial Attack on UCB Bandits

Shiliang Zuo [1]

Abstract

I study a stochastic multi-arm bandit problem where rewards are subject to adversarial corruption. At each round, the learner chooses an arm, and a stochastic reward is generated. The adversary strategically adds corruption to the reward, and the learner is only able to observe the corrupted reward at each round. I propose a novel attack strategy that manipulates a learner employing the upper-confidence-bound (UCB) algorithm into pulling some non-optimal target arm $T - o(T)$ times with a cumulative cost that scales as $\widehat{O}(\sqrt{\log T})$, where T is the number of rounds. I also prove the first lower bound on the cumulative attack cost. The lower bound matches the upper bound up to $O(\log \log T)$ factors, showing the proposed attack strategy to be near optimal.

1. Introduction

Stochastic multi-arm bandit is a framework for sequential decision-making with partial feedback. In its most basic form, a learner interacts with a set of arms giving stochastic rewards, and in each timestep, the learner is able to observe and collect the realized reward of one chosen arm. Past works have extensively studied different algorithms for optimizing the regret in the multi-arm bandit problem (for an overview see (Bubeck et al., 2012)). Some well-known algorithms include the upper-confidence-bound (UCB) algorithm and the ε-greedy algorithm (Auer et al., 2002). Given the fact that these algorithms are widely deployed in practice (e.g. news recommendation (Li et al., 2010), advertisements displayment (Chapelle et al., 2014)), it is important to understand the trustworthiness of these algorithms. Specifically, how do these algorithms respond when faced with adversarial attacks?

Recently, Jun et al. (Jun et al., 2018) initiated studying adversarial attacks on multi-arm bandit algorithms, taking a first step towards understanding the reliability and trustworthiness of these algorithms. In the adversarial attack scenario, an adversary sits between the learner and the environment. At each round t, the learner chooses an arm a_t and a stochastic reward r_t^0 is generated. Before the reward is observed by the learner, the adversary observes the chosen arm a_t and the reward r_t^0, and adds a strategic corruption α_t to the reward r_t^0. Then the learner is only able to observe the corrupted reward $r_t := r_t^0 - \alpha_t$. There is a specific target arm that the adversary wishes to promote, and the goal of the adversary is to manipulate the learner into choosing this target arm $T - o(T)$ times (or equivalently, any non-target arms a sub-linear number of times), while minimizing the cumulative attack cost, defined as the sum of corruptions throughout all rounds: $\sum_{t=1}^{T} |\alpha_t|$.

In this work, I characterize exactly the vulnerability of the UCB algorithm in the adversarial attack scenario. Specifically, I design a novel attack strategy that achieves optimal attack cost and provide matching lower bounds, hence resolving an open problem in (Jun et al., 2018). Assume arm a gives subgaussian rewards with mean μ_a and variance proxy σ^2. Without loss of generality, this work assumes the target arm is K. Let $\Delta_a^+ = \max(0, \mu_a - \mu_K)$. The main result is an attack strategy with cost $\widehat{O}(K\sigma\sqrt{\log T} + \sum_{a \neq K} \Delta_a^+)$ which holds for T uniformly over time, improving the attack cost in (Jun et al., 2018) by a $O(\sqrt{\log T})$ factor. The attack manipulates a learner employing the UCB algorithm into pulling the target arm $T - o(T)$ times and succeeds with high probability. I also conduct numerical experiments to validate the theoretical results. The experiments also show a significant improvement over the attack strategy in (Jun et al., 2018).

Related Work The recent work (Jun et al., 2018) first studied the problem of adversarial attacks on stochastic bandits. They showed an attack strategy against the UCB algorithm with the cumulative attack cost scaling as $O(\log T)$. This work improves the attack cost to $\widehat{O}(\sqrt{\log T})$ and provides matching lower bounds. Liu and Sheroff (Liu & Shroff, 2019) further proposed black-box adversarial attacks against stochastic bandits. Recent works also studied adversarial attacks on adversarial bandits (Ma & Zhou, 2023), gaussian bandits (Han & Scarlett, 2022), and contextual bandits (Ma

[1]Department of Computer Science, University of Illinois Urbana Champaign. Correspondence to: Shiliang Zuo <szuo3@illionis.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

et al., 2018; Garcelon et al., 2020), to name a few. Another line of work takes the viewpoint of the learner and designs algorithms robust to adversarial corruptions (Lykouris et al., 2018; Gupta et al., 2019).

2. Problem Statement

This work considers a stochastic multi-arm bandit problem where rewards are subject to adversarial corruptions. Let T be the time horizon and K the number of arms. The learner chooses arm $a_t \in [K]$ during round t, and a random reward r_t^0 is generated from a subgaussian distribution with variance proxy σ^2. The reward is centered at μ_{a_t}:

$$\mathbb{E}[r_t^0] = \mu_{a_t}.$$

At round t, after the learner chooses an arm a_t and the reward r_t^0 is generated, but before the reward r_t^0 is given to the learner, the adversary adds a strategic corruption α_t to the reward r_t^0. Then the learner only receives the corrupted reward $r_t := r_t^0 - \alpha_t$. Note that the adversary can decide the value of α_t based (a_t, r_t^0) as well as the history H_{t-1}, where the history H_t is defined as

$$H_t = (a_1, r_1^0, \alpha_1, ..., a_t, r_t^0, \alpha_t).$$

The attack framework is summarized in Algorithm 1.

The goal of the adversary is to manipulate the learner into pulling some target arm $T - o(T)$ times, while minimizing cumulative attack cost, defined as $\sum_{t=1}^{T} |\alpha_t|$. Without loss of generality, this work assumes the target arm is K. Let $\tau_a(t) := \{s : a_s = a, 1 \le s \le t\}$ denote the set of timesteps that arm a was chosen up to round t, and let $N_a(t) := |\tau_a(t)|$ denote the number of times arm a has been pulled up until round t. Also let

$$\hat{\mu}_a(t) = \sum_{s \in \tau_a(t)} r_s / N_a(t)$$

denote the post-attack empirical mean for arm a in round t, and let

$$\hat{\mu}_a^0(t) = \sum_{s \in \tau_a(t)} r_s^0 / N_a(t)$$

denote the pre-attack empirical mean for arm a in round t.

Specification of the UCB algorithm This work studies to what extent an adversary can hijack the UCB algorithm's behavior. The UCB algorithm works as follows. In the first K rounds, the learner pulls each arm a once to obtain an initial estimate $\hat{\mu}_a$. Then in later rounds $t > K$, the learner computes the UCB index for arm a as

$$\hat{\mu}_a(t) + 3\sigma\sqrt{\frac{\log t}{N_a(t)}}.$$

The arm with the largest index is then chosen by the learner. The specification of the UCB algorithm follows from (Jun et al., 2018). The UCB algorithm is summarized in Algorithm 2.

Algorithm 1 The general adversarial attack framework

for $t = 1, 2, ..., T$ **do**
  Learner picks arm a_t according to arm selection rule (e.g. UCB)
  Adversary learns a_t and pre-attack reward r_t^0, chooses attack α_t, suffers attack cost $|\alpha_t|$
  Learner receives reward $r_t = r_t^0 - \alpha_t$
end for

Algorithm 2 The UCB algorithm

for $t = 1, 2, ...K$ **do**
  Pull each arm a once and obtain initial estimate $\hat{\mu}_a$
end for
for $t > K$ **do**
  $a_t = \arg\max_a \hat{\mu}_a + 3\sigma\sqrt{\frac{\log t}{N_a(t)}}$
  Choose arm a_t, observe reward and update $\hat{\mu}_a, N_a(t)$
end for

3. Optimal Attack Strategy Against UCB

In this section, I show an optimal attack strategy for the adversary. Recall the goal of the adversary is to manipulate a learner employing the UCB algorithm into choosing some target arm $T - o(T)$ times while keeping the cumulative attack cost low. The main result is an attack strategy that only spends $\widehat{O}(\sqrt{\log T})$ attack cost.

For convenience assume arm K is picked in the first round. The proposed attack strategy works as follows. The adversary only attacks when any non-target arm is pulled, and adds corruption to ensure the difference between the post-attack empirical mean of the pulled arm and the target arm is above a certain gap. Specifically, the attacker ensures that the post-attack empirical means satisfy:

$$\hat{\mu}_{a_t}(t) \le \hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma e^n. \tag{1}$$

The gap $2\beta(N_K(t)) + 3\sigma e^n$ consists of two terms. The first term $\beta(N_K(t))$ is essentially a deviation bound that accounts for the estimation error of the true means (see Lemma 3.1). The second term grows exponentially with the number of times the current arm is pulled; perhaps surprisingly, this term is the key ingredient in ensuring the adversary only needs to spend $\widehat{O}(\sqrt{\log T})$ attack cost (see Lemma 3.2). The attack strategy is summarized in Algorithm 3.

Note that in the actual implementation, the adversary may wish equation Equation (1) to hold with strict inequality.

This can be accomplished by adjusting the attack by an infinitesimal amount. This work will not be concerned with such an issue and simply assume Equation (1) holds with equality when the adversary attacks.

Algorithm 3 Near Optimal Attack on UCB

$\beta(n) = \sqrt{\frac{2\sigma^2}{n} \log \frac{\pi^2 K n^2}{3\delta}}$
for t = 1, 2, ... **do**
  if $a_t \neq K$ **then**
    $n = N_{a_t}(t)$
    Attack with smallest $|\alpha|$, such that $\hat{\mu}_{a_t}(t) \leq \hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma \cdot \exp(n)$
  end if
end for

Recall $\hat{\mu}_a^0(t)$ denotes the pre-attack empirical mean of arm a at round t. Set parameter $\beta(n)$ as:

$$\beta(n) = \sqrt{\frac{2\sigma^2}{n} \log \frac{\pi^2 K n^2}{3\delta}},$$

and define event E as

$$\forall a, t, |\hat{\mu}_a^0(t) - \mu_a| < \beta(N_a(t))$$

which represents the event that pre-attack empirical means are concentrated around the true mean within an error of $\beta(N_a(t))$. By a simple Hoeffding inequality combined with a union bound, one can show event E holds with probability $1 - \delta$.

Lemma 3.1 ((Jun et al., 2018))**.** *Event E happens with probability $1 - \delta$. Further, the sequence $\beta(n)$ is non-increasing in n.*

Using the proposed attack strategy guarantees any non-target arm is pulled $O(\log \log t)$ times for any round t. This lemma lies at the heart of the proposed attack strategy.

Lemma 3.2. *Assume event E holds. At any round t, $N_a(t) \leq \lceil 0.5 \cdot \log \log t \rceil$ for any $a \neq K$.*

Proof. For sake of contradiction suppose some non-target arm a is pulled more than $\lceil 0.5 \cdot \log \log t \rceil$ times. After this arm is pulled for the $\lceil 0.5 \cdot \log \log t \rceil$-th time at round $t_0 < t$, we must have

$$\begin{aligned}\hat{\mu}_a^0(t_0) &< \hat{\mu}_K(t_0) - 2\beta(N_K(t_0)) - 3\sigma \cdot \exp\left(\log \sqrt{\log t}\right) \\ &= \hat{\mu}_K(t_0) - 2\beta(N_K(t_0)) - 3\sigma\sqrt{\log t}. \end{aligned} \tag{2}$$

Now assume arm a has been pulled for the $(\lceil 0.5 \log \log t \rceil + 1)$-th time in round $t_1 \in [t_0 + 1, t]$. Then the UCB index of arm a must be higher than that of arm K in round t_1. However,

$$\begin{aligned}&\hat{\mu}_a^0(t_1 - 1) + 3\sigma\sqrt{\frac{\log t_1}{N_a(t_1 - 1)}} \\ &= \hat{\mu}_a^0(t_0) + 3\sigma\sqrt{\frac{\log t_1}{N_a(t_0)}} \\ &\leq \hat{\mu}_K(t_0) - 2\beta(N_K(t_0)) - 3\sigma\sqrt{\log t} + 3\sigma\sqrt{\frac{\log t_1}{N_a(t_0)}} \\ &\leq \hat{\mu}_K(t_1) - 3\sigma\sqrt{\log t} + 3\sigma\sqrt{\frac{\log t_1}{N_a(t_0)}} \\ &\leq \hat{\mu}_K(t_1).\end{aligned}$$

The second line follows from the fact that arm a has not been chosen since t_0, the third line follows from the design of the attack strategy (specifically Equation (2)), and the fourth line follows from the concentration result given by event E. The UCB index of arm a is lower than that of arm K, hence a contradiction is established, and arm a will not be picked again. □

The main result on the upper bound of the cost of the attack strategy against UCB is given below. Recall $\Delta_a = \mu_a - \mu_K$, $\Delta_a^+ = \max(0, \Delta_a)$.

Theorem 3.3. *With probability $1 - \delta$, for any T, using the proposed attack strategy ensures any non-target arm is pulled $O(\log \log T)$ times and total attack cost is $\widehat{O}(K\sigma\sqrt{\log T} + \sum_{a \in [K]} \Delta_a^+)$.*

Proof. Assume event E holds throughout this proof. Recall $\tau_a(t)$ is the set of timesteps in which arm a was chosen. For any t,

$$\hat{\mu}_a(t) = \frac{\hat{\mu}_a^0(t) N_a(t) - \sum_{s \in \tau_a(t)} \alpha_s}{N_a(t)}.$$

Thus in round t if the adversary attacked arm a, then

$$\begin{aligned}&\frac{\hat{\mu}_a^0(t) N_a(t) - \sum_{s \in \tau_a(t)} \alpha_s}{N_a(t)} \\ &= \hat{\mu}_a(t) \\ &= \hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma e^{N_a(t)}.\end{aligned}$$

Consequently

$$\begin{aligned}&\frac{1}{N_a(t)} \sum_{s \in \tau_a(s)} \alpha_s \\ &= \hat{\mu}_a^0(t) - \hat{\mu}_K(t) + 2\beta(N_K(t)) + 3\sigma e^{N_a(t)} \\ &\leq \Delta_a^+ + \beta(N_a(t)) + 3\beta(N_K(t)) + 3\sigma e^{N_a(t)} \\ &\leq \Delta_a^+ + \beta(N_a(t)) + 3\beta(N_K(t)) + 3\sigma e^{0.5 \log \log t + 1} \\ &\leq \Delta_a^+ + 4\beta(N_a(t)) + 3e \cdot \sigma\sqrt{\log t}.\end{aligned}$$

Here, the third line follows from event E, and the last line comes from the fact that β is nonincreasing and $N_a(t) < N_K(t)$. Thus focusing the attack cost spent on arm a:

$$\sum_{s\in\tau_a(t)} \alpha_s \le N_a(t)(\Delta_a^+ + 4\beta(N_a(t)) + 3e\cdot\sigma\sqrt{\log t}) = \widehat{O}(\sigma\sqrt{\log t} + \Delta_a^+).$$

Summing over all non-target arms, the total attack cost is $\widehat{O}(K\sigma\sqrt{\log t} + \sum_{a\in[K]}\Delta_a^+)$.

Furthur, by Lemma 3.2, any non-target arm is pulled $O(\log\log T)$ times. □

4. Lower Bounds on Attack Cost

In this section, I prove lower bounds on the cumulative attack cost. For a learner employing the UCB algorithm, the lower bounds match the upper bound in the previous section up to $O(\log\log T)$ factors, showing the proposed attack strategy to be near optimal. I also show a lower bound on the attack cost against the ε-greedy algorithm.

Recall that $\tau_a(t)$ represents the set of timesteps that arm a was chosen before round t. Let $\bar{\alpha}_a(t) = \sum_{s\in\tau_a(t)}|\alpha_s|$ denote the cumulative attack cost on arm a. Note that

$$\hat{\mu}_a(t) - \bar{\alpha}_a(t)/N_a(t) \le \hat{\mu}_a^0(t) \le \hat{\mu}_a(t) + \bar{\alpha}_a(t)/N_a(t). \tag{3}$$

4.1. Attack on UCB

For ease of exposition, I will focus on the multi-arm bandit setting with $K = 2$ arms in this section, but the results easily generalize to the case where $K > 2$. The lower bound below shows the previously proposed attack strategy to be optimal up to $\log\log T$ factors.

Theorem 4.1. *Assume the learner is using the UCB algorithm as in Algorithm 2. Consider two arms giving subgaussian rewards, with mean $\mu_1 > \mu_2$ and variance-proxy σ^2. Assume arm 2 is the target arm, and let $\Delta = \mu_1 - \mu_2$. Given $\delta < 0.1$, for any $T > 100$, with probability $1-\delta$, an attack cost of $\Delta + 0.22\sigma\sqrt{\log T}$ is needed to manipulate the learner into pulling the 1st arm no more than $N_1(T)$ times, where $N_1(T) < \min(T/26, \sqrt{3\delta/2}T/\pi)$ times.*

Proof. Throughout this proof assume event E holds. Consider the last round the target arm is pulled. Denote this timestep by $t+1$. Then comparing the UCB index we must have

$$\hat{\mu}_2(t) + 3\sigma\sqrt{\frac{\log t}{N_2(t)}} > \hat{\mu}_1(t) + 3\sigma\sqrt{\frac{\log t}{N_1(t)}}.$$

Therefore by event E and Equation (3)

$$\mu_2(t) + \beta(N_2(t)) + \frac{\bar{\alpha}_{2,t}}{N_2(t)} + 3\sigma\sqrt{\frac{\log t}{N_2(t)}} > \mu_1(t) - \beta(N_1(t)) - \frac{\bar{\alpha}_{1,t}}{N_1(t)} + 3\sigma\sqrt{\frac{\log t}{N_1(t)}}.$$

By the fact that $N_1(t) < N_2(t)/25$:

$$\sqrt{\frac{\log t}{N_2(t)}} < 0.2\sqrt{\frac{\log t}{N_1(t)}},$$

and we can also verify

$$\beta(N_2(t)) < 0.29\beta(N_1(t)).$$

Hence

$$\begin{aligned}
&\frac{\bar{\alpha}_{1,t} + \bar{\alpha}_{2,t}}{N_1(t)} \\
&> \Delta - \beta(N_1(t)) - \beta(N_2(t)) + 3\sigma\sqrt{\frac{\log t}{N_1(t)}} - 3\sigma\sqrt{\frac{\log t}{N_2(t)}} \\
&\ge \Delta - 1.29\beta(N_1(t)) + 2.8\sigma\sqrt{\frac{\log t}{N_1(t)}} \\
&= \Delta - 1.29\sqrt{\frac{2\sigma^2}{N_1(t)}\log\frac{2\pi^2 N_1(t)^2}{3\delta}} + 2.8\sigma\sqrt{\frac{\log t}{N_1(t)}} \\
&\ge \Delta + 0.22\sigma\sqrt{\frac{\log t}{N_1(t)}}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\bar{\alpha}_{1,t} + \bar{\alpha}_{2,t} &\ge N_1(t)\Delta + 0.22\sigma\sqrt{N_1(t)\log t} \\
&\ge \Delta + 0.22\sigma\sqrt{\log t}.
\end{aligned}$$

This finishes the proof. □

4.2. Attack on ε-greedy

For comparison, I establish a lower bound on the cumulative attack cost against ε-greedy. The lower bound shows the attack proposed in (Jun et al., 2018) to be essentially optimal.

The ε-greedy algorithm works as follows. At each round, the learner with probability ε_t does uniform exploration, otherwise, the learner does exploitation and chooses the arm with the largest empirical reward. Assume $\varepsilon_t = {}^{cK}/_{t}$ for some constant c as in (Auer et al., 2002) (thus each arm is chosen for exploration with probability ${}^{c}/_{t}$ each round). The ε-greedy algorithm is summarized in Algorithm 4. In this section, I again focus on the case where there are $K = 2$

Algorithm 4 ε-greedy algorithm

Exploration parameter c
for $t = 1, 2, ..., T$ **do**
With probability cK/t, choose arm u.a.r.
Otherwise choose a_t such that $a_t = \arg\max_{a\in[K]} \hat{\mu}_a(t-1)$
Update $\hat{\mu}_{a_t}$ based on observed reward
end for

arms, though the analysis easily generalizes to the case where $K > 2$.

I first prove a lemma that gives a tight characterization of the number of times each arm is pulled in exploration rounds.

Lemma 4.2. *Fix $\delta \in (0,1)$. Suppose T satisfies $\sum_{t=1}^{T} c/t \geq 16\log(4/\delta)$, then with probability $1-\delta$, the number of times each arm is pulled during exploration rounds is between $0.5c\log T$ and $2c\log T$.*

Proof. Fix arm a. Let X_t be the indicator variable that takes the value 1 if arm a was pulled in round t as exploration. Then

$$\mathbb{E}[X_t] = \frac{c}{t}$$
$$\mathbb{V}[X_t] = \frac{c}{t}(1-\frac{c}{t}).$$

Then by a Freedmans' style inequality (e.g. (Agarwal et al., 2014)), for any $\eta \in (0,1)$, with probability $1-\delta/4$, we have

$$\begin{aligned}\sum_{t=1}^{T}(X_t - \frac{c}{t}) &\leq \eta\sum_{t=1}^{T}\mathbb{V}[X_t] + \frac{\log(4/\delta)}{\eta} \\ &\leq \eta\sum_{t=1}^{T}\mathbb{E}[X_t] + \frac{\log(4/\delta)}{\eta} \\ &= \eta\sum_{t=1}^{T}\frac{c}{t} + \frac{\log(4/\delta)}{\eta}.\end{aligned}$$

Choosing $\eta = \sqrt{\frac{\log(4/\delta)}{\sum_{t=1}^{T} c/t}}$, we obtain

$$\sum_{t=1}^{T} X_t < \sum_{t=1}^{T}\frac{c}{t} + 2\sqrt{\sum_{t=1}^{T}\frac{c}{t}\log(4/\delta)}.$$

A lower bound on $\sum_{t=1}^{T} X_t$ is similar by taking the random variables to be $-X_t$ instead of X_t in Freedmans' inequality, and we can show with probability $1-\delta/4$

$$\sum_{t=1}^{T} X_t > \sum_{t=1}^{T}\frac{c}{t} - 2\sqrt{\sum_{t=1}^{T}\frac{c}{t}\log(4/\delta)}.$$

Thus with probability $1-\delta/2$, for T large enough

$$\frac{1}{2}\sum_{t=1}^{T}\frac{c}{t} < \sum_{t=1}^{T} X_t < 2\sum_{t=1}^{T}\frac{c}{t}.$$

The lemma then follows by taking a union bound over the 2 arms. □

Theorem 4.3. *Assume the learner is using the ε-greedy algorithm as in Algorithm 4 with learning rate $2c/t$ for some fixed constant c. Consider two arms giving subgaussian rewards with mean $\mu_1 > \mu_2$ and variance-proxy σ^2. Assume arm 2 is the target arm and let $\Delta = \mu_1 - \mu_2$. Given $\delta < 0.1$, suppose $T > 100$ satisfies the condition in Lemma 4.2 and that $3\beta(0.5c\log T) < \Delta$. With probability $1-2\delta$, an attack cost of $c\cdot\Delta\log T/6$ is needed to manipulate the learner into pulling the 1st arm no more than $N_1(T)$ times, where $N_1(T) < T/4$.*

Proof. Throughout this proof assume event E and Lemma 4.2 holds. By a union bound, these events hold with probability $1-2\delta$. Consider the last exploitation round before T in which the learner pulled the 2nd arm, and denote the timestep by t. By round T, the number of times the 2nd arm was pulled in exploration rounds is at most $2\log T$. Thus to ensure the 2nd arm is pulled no less than $T - T/4$ rounds, we must have

$$t > T - T/4 - 2\log T > T/2.$$

In this round, the post-attack mean of the 2nd arm must be higher than that of the 1st arm:

$$\hat{\mu}_2(t) > \hat{\mu}_1(t).$$

Therefore by event E and Equation (3):

$$\mu_2(t) + \beta(N_2(t)) + \frac{\bar{\alpha}_{2,t}}{N_2(t)} > \mu_1(t) - \beta(N_1(t)) - \frac{\bar{\alpha}_{1,t}}{N_1(t)}$$

leading to

$$\begin{aligned}\bar{\alpha}_{1,t} + \bar{\alpha}_{2,t} &> N_1(t)(\Delta - 2\beta(N_1(t))) \\ &> c\cdot\Delta\log T/6,\end{aligned}$$

since assuming Lemma 4.2 holds, $N_1(t) > 0.5c\log T$, and by the assumption on T we have $\Delta > 3\beta(0.5c\log T) > 3\beta(N_1(t))$. This finishes the proof. □

5. Experiments

In this section, I describe the results of numerical experiments. In the experiments, the bandit instance has two arms, and the reward distributions are $N(\mu,\sigma^2)$ and $N(0,\sigma^2)$ respectively. The target arm is the second arm. The experiments aim to empirically study how the variance of the

reward σ^2 and the reward gap μ affect the cumulative attack cost. I conduct 9 groups of experiments by varying the parameters of $\sigma \in \{0.1, 1, 2\}$ and $\mu \in \{0.1, 1, 2\}$. In each group, I run 20 trials for the bandit instance with $T = 10^6$. I also run the attack strategy in (Jun et al., 2018) as a baseline for comparison.

The theoretical results in this work indicate non-target arm is pulled at most $0.5 \log \log T$ times; the empirical results validate this, as the non-target arm is only pulled for 1 time, and hence the target arm is pulled almost always. The cumulative attack costs are summarized in Table 1, which fit nicely with the theoretical bound of $\widehat{O}(K\sigma\sqrt{\log T} + \sum_{a \neq K} \Delta_a^+)$ in this work. The results also show a significant improvement over the attack strategy proposed in (Jun et al., 2018).

Table 1. Cumulative attack cost for different choices of (σ, μ)

Approach	σ	$\mu = 0.1$	$\mu = 1$	$\mu = 2$
(Jun et al., 2018)	0.1	23.6	129.4	247.3
This work	0.1	1.3	2.4	3.6
This work	1	14.5	15.9	16.8
This work	2	30.3	30.7	31.0

6. Conclusion

In this work, I studied adversarial attacks that manipulate the behavior of the UCB algorithm for multi-arm bandits. I proposed a novel attack strategy against the UCB algorithm and established the first lower bound on cumulative attack cost. The lower bound matches the attack cost of the proposed attack strategy up to $\log \log T$ factors, showing the attack to be near optimal. The results show that the UCB algorithm is perhaps the most easily exploitable algorithm when compared to other algorithms which are randomized, such as the ε-greedy algorithm.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

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

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

Chapelle, O., Manavoglu, E., and Rosales, R. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1–34, 2014.

Garcelon, E., Roziere, B., Meunier, L., Tarbouriech, J., Teytaud, O., Lazaric, A., and Pirotta, M. Adversarial attacks on linear contextual bandits. *Advances in Neural Information Processing Systems*, 33:14362–14373, 2020.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. *arXiv preprint arXiv:1902.08647*, 2019.

Han, E. and Scarlett, J. Adversarial attacks on gaussian process bandits. In *International Conference on Machine Learning*, pp. 8304–8329. PMLR, 2022.

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

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

Liu, F. and Shroff, N. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pp. 4042–4050. PMLR, 2019.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Ma, Y. and Zhou, Z. Adversarial attacks on adversarial bandits. *arXiv preprint arXiv:2301.12595*, 2023.

Ma, Y., Jun, K.-S., Li, L., and Zhu, X. Data poisoning attacks in contextual bandits. In *Decision and Game Theory for Security: 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29–31, 2018, Proceedings 9*, pp. 186–204. Springer, 2018.