Near Optimal Adversarial Attacks on Stochastic Bandits and Defenses with Smoothed Responses

Shiliang Zuo
University of Illinois Urbana-Champaign

Abstract

I study adversarial attacks against stochastic bandit algorithms. 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. Two sets of results are presented in this paper. The first set studies the optimal attack strategies for the adversary. The adversary has a target arm he wishes to promote, and his goal is to manipulate the learner into choosing this target arm $T - o(T)$ times. I design attack strategies against UCB and Thompson Sampling that only spends $\widehat{O}(\sqrt{\log T})$ cost. Matching lower bounds are presented, and the vulnerability of UCB, Thompson sampling and ε-greedy are exactly characterized. The second set studies how the learner can defend against the adversary. Inspired by literature on smoothed analysis and behavioral economics, I present two simple algorithms that achieve a competitive ratio arbitrarily close to 1.

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. This framework models many real-world applications, including news recommendation [Li et al., 2010], advertisements displayment [Chapelle et al., 2014], and medical experiments [Kuleshov and Precup, 2014]. Past works have extensively studied algorithms for optimizing the regret in the multi-arm bandit problem (for an overview see [Bubeck et al., 2012]), and some well-known algorithms include the upper-confidence-bound (UCB) algorithm [Auer et al., 2002], the Thompson sampling algorithm [Agrawal and Goyal, 2012], and the ε-greedy algorithm [Auer et al., 2002]. When deploying multi-arm bandit algorithms in practice, it is crucial to understand the robustness of these algorithms.

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 robustness of these algorithms. In the adversarial attack scenario, an adversary sits between the learner and the environment. The adversary can add corruption to the reward, and he has a specific target arm he wishes to promote. The adversary's goal is to manipulate the learner into choosing this target arm almost always by only strategically adding small corruptions to the rewards. [Jun et al., 2018] showed that when the learner is employing the UCB or ε-greedy algorithm, the adversary can hijack the algorithm into choosing the target arm by only spending $O(\log T)$ corruption budget, and they left open the question on whether $O(\log T)$ is indeed the optimal attack cost. This work fully characterizes the optimal attack strategy against UCB, Thompson sampling, and ε-greedy. I show optimal attack strategies with $\hat{O}(\sqrt{\log T})$ attack cost against UCB and Thompson sampling, and provide matching lower bounds. I also give a $\Omega(\log T)$ lower bound on the attack cost against ε-greedy. Hence, the vulnerability of UCB, Thompson sampling, and ε-greedy are exactly characterized.

Given the fact that these well-known stochastic bandit algorithms are vulnerable, how can the learner defend against such an adversary? [Lykouris et al., 2018] initiated the design of bandit algorithms robust to adversarial corruptions under the *weak* adversary model. While the strong adversary (the

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

model in [Jun et al., 2018]) decides on the corruption after observing the learner's action and the realized reward, the weak adversary decides on the corruption before observing the learner's chosen action. Their work designed bandit algorithms with regret bounds that degrade gracefully as the corruption budget C increases. The follow-up work [Gupta et al., 2019, Zimmert and Seldin, 2021] further improved their regret bounds. However, when the learner faces a strong adversary, [Liu and Shroff, 2019], [He et al., 2022] showed that any low-regret bandit algorithms can be hijacked (i.e. suffer linear regret) by only spending a sublinear attack cost. While it is apparent that the measure of regret is no longer a suitable benchmark when facing a *strong* adversary in the stochastic bandit setting, is there some other benchmark in which robustness can be measured? This work proposes using the competitive ratio as the benchmark and gives two simple algorithms inspired by behavioral economics that achieve a competitive ratio arbitrarily close to 1.

1.1 Contributions

This work studies adversarial attacks in stochastic bandits under the *strong* adversary model and studies both attack and defense strategies. The first set of results studies optimal attack strategies. It is shown that, when the learner employs the UCB and Thompson sampling algorithm, the adversary only needs to spend $O(\sqrt{\log T})$ attack cost to hijack the learner into choosing the target arm. This work further provides matching lower bounds and thus characterizes exactly the vulnerability of these algorithms under the adversarial attack scenario. This solves several open problems in [Jun et al., 2018].

The second set of results studies defense strategies for the learner. While it is known no algorithm can achieve a sublinear regret in the presence of a *strong* adversary ([Liu and Shroff, 2019]), this work studies the competitive ratio as a benchmark. I design two algorithms for the learner that achieve a competitive ratio arbitrarily close to 1. The algorithms draw connections from smoothed analysis and behavioral economics. I believe the use of competitive ratio under corruptions and the new connections merits further research.

In the stochastic multi-arm bandit setting, there are K arms, and arm $a \in [K]$ gives subgaussian rewards with mean μ_a and variance proxy σ^2. In the study of optimal attack strategies, the target arm the adversary wishes to promote is arm K, and denote $\Delta_a^+ = \max(0, \mu_a - \mu_K)$. The outline and main results of each section can be summarized as follows.

- Section 2 begins with the problem statement.
- Section 3 shows my design of an attack strategy against UCB with cost $\widehat{O}(K\sigma\sqrt{\log T} + \sum_{a\neq K} \Delta_a^+)$. This holds for T uniformly over time, improving the attack cost in [Jun et al., 2018] by a $O(\sqrt{\log T})$ factor.
- Section 4 shows my design of an attack strategy against Thompson sampling with cost $\widehat{O}(K\sqrt{\log T}+\sum_{a\neq K} \Delta_a^+ + K(\sigma+1)\sqrt{\log K})$. This holds for T uniformly over time. To the best of the author's knowledge, no prior work explicitly studies adversarial attacks on Thompson sampling.
- Section 5 proves lower bounds on the attack cost against UCB and Thompson sampling with order $\Omega(\sqrt{\log T})$. This shows the proposed attack strategy against UCB and Thompson sampling to be nearly optimal. A lower bound $\Omega(\log T)$ on the attack cost against ε-greedy is also given.
- Section 6 presents the study of defense strategies and gives algorithms that achieve a competitive ratio arbitrarily close to 1, as long as the total corruption budget $C = o(T)$.
- Section 7 describes numerical experiments. In the study of attack strategies, the experiments show significant improvements over [Jun et al., 2018].

1.2 Related Works

The problem of adversarial attack on stochastic bandits was initiated by [Jun et al., 2018], in which they proposed attack strategies against UCB and ε-greedy. The work by Liu and Shroff [Liu and Shroff, 2019] studied black-box attacks against stochastic bandit algorithms. Recent works also studied adversarial attacks in other problems, including adversarial bandits ([Ma and Zhou, 2023]), contextual bandits ([Ma et al., 2018, Garcelon et al., 2020]) gaussian process bandits ([Han and Scarlett, 2022]) and reinforcement learning ([Zhang et al., 2020]) etc.

Another line of work studies the design of robust algorithms in the presence of adversarial corruption, sometimes under different adversary models. In the stochastic bandit setting, [Lykouris et al., 2018, Gupta et al., 2019] design learning algorithms for the learner robust to adversarial corruptions under the *weak* adversary model, and their result was subsequently improved by [Zimmert and Seldin, 2021]. The *weak* adversary needs to decide on the corruption before observing the action of the learner, while the *strong* adversary (the model in [Jun et al., 2018]) can observe the action of the learner before deciding on the corruption. The study of corruption-

robust algorithms has also been studied in other problems, e.g. linear bandits ([He et al., 2022]), contextual search ([Leme et al., 2022, Zuo, 2023]), and reinforcement learning ([Wei et al., 2022]), to name a few.

Smoothed analysis first appeared as a way to analyze an algorithm's performance beyond worst case, and was first introduced by [Spielman and Teng, 2004]. In their paper, they showed that while the simplex algorithm exhibits exponential time complexity in the worst case, adding a Gaussian noise to the input guarantees a polynomial time complexity. The idea of smoothed analysis has also been explored in machine learning, online learning, as well as game-theoretic contexts ([Sivakumar et al., 2020, Haghtalab et al., 2020, Kannan et al., 2018]).

Behavioral economics is mostly concerned with the study of bounded rationality (for a textbook treatment see e.g.[Camerer, 2011]). The quantal response was proposed in [McKelvey and Palfrey, 1995] as a solution concept when agents have bounded rationality. The quantal response enjoys many nice statistical properties; for example, it naturally arises from the logit model ([McFadden, 1976, Luce, 2012]) and is equivalent to selecting the maximum after a perturbation with Gumbel distribution ([Jang et al., 2016]). The quantal response also appears in the machine learning literature under different contexts, e.g. the softmax activation function usually used in training machine learning models ([Dunne and Campbell, 1997]) and the multiplicative weight update algorithms in online learning ([Arora et al., 2012]). Connections between behavioral economics and online learning has also been explored in ([Wu et al., 2022]).

2 PRELIMINARIES

Algorithm 1 The general adversarial attack framework

for $t = 1, 2, ...$ **do**
  Learner picks arm a_t according to arm selection rule (e.g. UCB, Thompson sampling)
  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

This work studies 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}:

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

The work studies the *strong* adversary, who can observe the learner's chosen arm before deciding on the attack. 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, where the history H_t is defined as

$$H_t = (a_1, r_1^0, \alpha_1, ..., a_{t-1}, r_{t-1}^0, \alpha_{t-1}).$$

The attack framework is summarized in algorithm 1.

In the rest of this work, $\tau_a(t) := \{s : a_s = a, 1 \le s < t\}$ denotes the set of timesteps that arm a was chosen up to round t, and $N_a(t) := |\tau_a(t)|$ denotes the number of times arm a has been pulled up until round t. Also let $\hat{\mu}_a(t)$ denote the post-attack empirical mean for arm a in round t:

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

and let $\hat{\mu}_a^0(t)$ denote the pre-attack empirical mean for arm a in round t:

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

This work will study both attack strategies for the adversary and defense strategies for the learner. In the study of attack strategies, 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. In the study of defense strategies, the goal of the learner is to optimize the cumulative reward, while being agnostic to the attack strategy of the adversary.

2.1 A Concentration Result

The following concentration result will be useful throughout the analysis. 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 $\beta(N_a(t))$. The following has been shown in [Jun et al., 2018], which follows from a Hoeffding inequality combined with a union bound.

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

3 ATTACK STRATEGY AGAINST UCB

In this section, I first give the specification of the UCB algorithm (algorithm 2), then propose a near-optimal attack strategy against it (algorithm 3).

Algorithm 2 UCB, adapted from [Bubeck et al., 2012]

For each arm a, learner maintains empirical mean $\hat{\mu}_a(t)$ and the number of times a has been pulled $N_a(t)$
for $t = 1, 2, \ldots, K$ **do**
  Pull each arm a once
  Update $\hat{\mu}_a(t+1), N_a(t+1)$
end for
for $t > K$ **do**
  $a_t = \arg\max_a \hat{\mu}_a(t) + 3\sigma\sqrt{\frac{\log t}{N_a(t)}}$
  Choose arm a_t and observe reward
  Update $\hat{\mu}_a(t+1), N_a(t+1)$
end for

Algorithm 3 Optimal Attack on UCB

$\beta(n) = \sqrt{\frac{2\sigma^2}{n}\log\frac{\pi^2 K n^2}{3\delta}}$
for $t = 1, 2, \ldots$ **do**
  if $a_t \neq K$ **then**
    Adversary observe reward r_t^0
    Compute and inject the smallest corruption α_t with $\alpha_t \geq 0$, such that:

$$\hat{\mu}_{a_t}(t) \leq \hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma \cdot \exp(N_{a_t}(t))$$

where $\hat{\mu}_{a_t}(t)$ is the post-attack empirical mean:

$$\hat{\mu}_{a_t}(t) = (\hat{\mu}_{a_t}(t-1) \cdot N_a(t) + r_t^0 - \alpha_t)/(N_a(t)+1)$$

  end if
end for

3.1 UCB

The UCB algorithm is summarized in algorithm 2 and the specification follows from [Jun et al., 2018]. 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.

3.2 Adversarial Attack Strategy

I now show an optimal attack strategy for the adversary against the UCB algorithm that only spends $\widehat{O}(\sqrt{\log T})$ attack cost. The attack strategy is summarized in algorithm 3. Recall the goal of the adversary is to manipulate a learner employing the UCB algorithm into choosing the target arm (arm K) at least $T - o(T)$ times while keeping the cumulative attack cost low.

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:

$$\hat{\mu}_{a_t}(t) \leq \hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma\exp(N_a(t)). \quad (1)$$

The key insight is that in order to minimize attack cost, the number of non-target arm pulls should be kept as low as possible. In fact, using the proposed attack strategy guarantees any non-target arm is pulled only $O(\log\log t)$ times for any round t.

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 1. *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 Sketch. Each time a non-target arm is pulled, the adversary injects corruption so that the gap in eq. (1) holds. The gap $2\beta(N_K(t)) + 3\sigma\exp(N_a(t))$ consists of two terms. The first term $\beta(N_K(t))$ is a deviation bound that accounts for the estimation error of the true means (see lemma 1). The second term grows exponentially with the number of times the current arm is pulled and guarantees that any non-target arm is only pulled for $0.5\log\log T$ times for any round T. This in turn implies for any round t, the adversary needs only spend $\widehat{O}(\exp(0.5\log\log T)) = \widehat{O}(\sqrt{\log T})$ attack cost to ensure the gap holds. □

Remark 1. *Note that in the actual implementation, the adversary may wish equation eq.* (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 assumes eq.* (1) *holds with equality when the adversary attacks.*

4 ATTACK STRATEGY AGAINST THOMPSON SAMPLING

In this section, I first give the description of the Thompson sampling algorithm (algorithm 4), then propose a near-optimal attack strategy against it (algorithm 5).

Algorithm 4 Thompson sampling, adapted from [Agrawal and Goyal, 2017]

For each arm a, learner maintains empirical mean $\hat{\mu}_a(t)$ and the number of times a has been pulled $N_a(t)$
for $t = 1, 2, \ldots, K$ **do**
  Pull each arm a once
  Update $\hat{\mu}_a(t+1), N_a(t+1)$
end for
for $t > K$ **do**
  Sample $\nu_a = \mathcal{N}(\hat{\mu}_a(t), \frac{1}{N_a(t)})$
  Choose $a_t = \arg\max_a \nu_a$ and observe reward
  Update $\hat{\mu}_a(t+1), N_a(t+1)$
end for

Algorithm 5 Optimal Attack on Thompson Sampling

$\beta(n) = \sqrt{\frac{2\sigma^2}{n} \log \frac{\pi^2 K n^2}{3\delta}}$
for $t = 1, 2, \ldots$ **do**
  if $a_t \neq K$ **then**
    Compute and inject the smallest corruption α_t with $\alpha_t \geq 0$, such that:

$$\hat{\mu}_{a_t}(t) \leq \hat{\mu}_K - 2\beta(N_K(t)) - 4\exp(N_{a_t}(t)) - \sqrt{8 \log \frac{\pi^2 K}{3\delta}}$$

  where $\hat{\mu}_{a_t}(t)$ is the post-attack empirical mean:

$$\hat{\mu}_{a_t}(t) = (\hat{\mu}_{a_t}(t-1) \cdot N_a(t) + r_t^0 - \alpha_t)/(N_a(t)+1)$$

  end if
end for

4.1 Thompson Sampling

The Thompson Sampling is summarized in algorithm 4 and the specification is adapted from [Agrawal and Goyal, 2017]. In the first K rounds, the learner pulls each arm a once. Then in later rounds, for each arm a, a random variable ν_a is generated from the distribution $\mathcal{N}(\hat{\mu}_a(t), \frac{1}{N_a(t)})$. The arm with the largest ν_a is then chosen.

4.2 Adversarial Attack Strategy

I now show an optimal attack strategy for the adversary against the Thompson sampling algorithm, which only spends $\widehat{O}(\sqrt{\log T})$ attack cost. The attack strategy is summarized in algorithm 5 and shares some similar insights as the attack strategy against the UCB algorithm. Specifically, any non-target arm pulls are also upper bounded by $O(\log \log t)$. This is achieved by injecting corruption whenever non-target arms are pulled to ensure the post-attack empirical means satisfy:

$$\hat{\mu}_{a_t}(t) \leq \hat{\mu}_K(t) - 2\beta(N_K(t)) - 4\exp(N_{a_t}(t)) - \sqrt{8 \log \frac{\pi^2 K}{3\delta}}.$$

Similar to the attack on UCB, the gap contains a term $4\exp(N_{a_t}(t))$ that grows exponentially in the number of times that the non-target arm is pulled. Hence, this exponential term exhibits some generality and may act as a general attack principle for the adversary.

Theorem 2. *With probability $1 - 2\delta$, for any T, using the proposed attack strategy ensures any non-target arm is pulled $O(\log \log T)$ times and the total attack cost is $\widehat{O}(K\sqrt{\log T} + \sum_{a \neq [K]} \Delta_a^+ + K(\sigma+1)\sqrt{\log \frac{\pi^2 K}{3\delta}})$.*

Remark 2. *Note that in the attack cost $\widehat{O}(K\sigma\sqrt{\log T} + \sum_{a \neq K} \Delta_a^+)$ on UCB bandits, the $\sqrt{\log T}$ factor is multiplied by σ, whereas for Thompson sampling the $\sqrt{\log T}$ factor is not. Hence, the attack cost for UCB is expected to grow faster than the attack cost for Thompson sampling as σ gets large, and vice versa. This point is in fact illustrated in the numerical experiments in the following sections.*

5 LOWER BOUNDS ON ATTACK COST

In this section, I prove lower bounds on the cumulative attack cost. For a learner employing the UCB or Thompson sampling algorithm, the lower bounds of $\Omega(\sqrt{\log T})$ 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 of $\Omega(\log T)$ on the attack cost against the ε-greedy algorithm. I shall focus on the setting where $K = 2$, but the results also generalize to the case where $K > 2$. In this section, the bandit environment consists of two arms giving Gaussian rewards $\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)$. The mean $\mu_1 > \mu_2$ and the 2nd

arm is the target arm. Let $\Delta = \mu_1 - \mu_2$ and let B be a sufficiently large constant.

The below two theorems characterize the lower bound on attack cost against UCB and Thompson sampling.

Theorem 3 (UCB Attack Cost Lower Bound)**.** *Assume the learner is using the UCB algorithm as in algorithm 2. For any $T > B$, if the adversary spend an attack cost less than $\Delta + 0.22\sigma\sqrt{\log T - 1}$, then with probability 0.9, the learner pulls the first arm more than $T/26$ times.*

Proof Sketch. Consider the last time the target arm has been pulled, denote this round by t_0. At round t_0, the UCB index for the non-target arm (the optimal arm) must be lower than that of the target arm. Since $N_2(t_0) = \Theta(T)$:

$$\hat{\mu}_1(t_0) + \sqrt{\frac{\log T}{N_1(t_0)}} \lesssim \hat{\mu}_2(t_0).$$

To drag down the UCB index of the non-target arm to achieve this, an attack cost of

$$\left(\Delta + \sqrt{\frac{\log T}{N_1(t_0)}}\right) N_1(t_0) = \Omega(\sqrt{\log T})$$

is needed. □

Theorem 4 (Thompson Sampling Attack Cost Lower Bound)**.** *Assume the learner is using the Thompson Sampling algorithm as in algorithm 4. For any $T > B$, if the adversary spent an attack cost no more than $\Delta + 0.1\sqrt{\log T}$, then with probability 0.81, the learner pulls the 1st arm more than $T/10$ times.*

Remark 3. *Note the lower bounds hold for when the adversary knows the time horizon T, whereas the upper bounds in the previous section hold for T uniformly over time. Therefore, it would seem the $O(\log\log T)$ factor is the price the adversary has to pay when moving from the fixed T setting to the uniform T setting.*

For comparison, I establish a lower bound on the cumulative attack cost against ε-greedy. Interestingly, the ε-greedy exhibit a $\Omega(\log T)$ lower bound, in contrast with the $\Omega(\sqrt{\log T})$ lower bounds for UCB and Thompson sampling. This lower bound shows the attack strategy 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 exploration parameter c as in [Auer et al., 2002] (each arm is chosen for exploration with probability ${}^{c}/_{t}$ each round).

Theorem 5 (ε-greedy Attack Cost Lower Bound)**.** *Assume the learner is using the ε-greedy algorithm with a learning rate $2c/t$ for some fixed constant c. For any $T > B$, if the adversary spent an attack cost no more than $c \cdot \Delta \log T/6$, then with probability 0.8, the learner pulls the 1st arm more than $T/4$ times.*

6 DEFENSES WITH SMOOTHED RESPONSES

This section studies defense strategies for the learner, and uses the competitive ratio as the benchmark instead of regret. This is because a sublinear regret is not possible in the presence of a strong adversary. Specifically, it is known for any low-regret bandit algorithm, a strong adversary can spend $o(T)$ attack cost and make the learner suffer linear regret ([He et al., 2022]).

This section will prove bounds on the collected rewards of the form

$$\texttt{REW} \geq (1-\varepsilon)\texttt{OPT} - o(T),$$

where `REW` is the total reward collected, as measured by the true empirical means, and `OPT` is the expected reward of the optimal policy, i.e.:

$$\texttt{REW} = \sum_{t=1}^{T} \mu_{a_t}, \quad \texttt{OPT} = T\mu^*.$$

If an algorithm satisfies the above lower bound on `REW` (with probability $1-\delta$), then the algorithm is said to achieve $(1-\varepsilon)$ competitive ratio (with probability $1-\delta$). In this section, pre-attack rewards and post-attack rewards are assumed to be bounded in $[0, 1]$. This assumption is made since the competitive ratio is used as a benchmark; however, note even in this bounded rewards setting, [Rangi et al., 2022] showed that no-regret bandit algorithms are prone to adversarial attacks.

I first begin with a more detailed discussion on the use of competitive ratio as benchmark. Then, I show how concepts from smoothed analysis and behavioral economics can be used to design algorithms that achieve a competitive ratio arbitrarily close to 1 in the presence of a strong adversary.

6.1 Discussion on the Use of Competitive Ratio

The strongest form of guarantee one can hope for is a sublinear regret that scales with C. This is possible when the learner faces a *weak* adversary. However, obtaining a sublinear regret is impossible when the adversary is *strong*.

Fact 1. *(From [He et al., 2022]) If some algorithm* ALG *achieves regret* REG(T) *when* $C = 0$*, then there exists a strong adversary with budget* $C = \Theta(\texttt{REG}(T))$ *that can make the learner suffer linear regret.*

From the above fact, if the algorithm were not to suffer linear regret when $C = 0$, then there exists a scenario in which the adversary uses a sublinear attack budget and makes the algorithm suffer linear regret. In either case, there is a scenario where the learner must suffer linear regret. Now, since sublinear regret is not possible, the next natural possible benchmark is the competitive ratio. To understand how good this benchmark is, consider the following proposition.

Proposition 1. *Fix any algorithm* ALG*. There exists some constant* $\varepsilon > 0$ *and an attack strategy with a sublinear budget such that the learner achieves no better than* $1 - \varepsilon$ *competitive ratio, specifically,*

$$\liminf_{T\to\infty} (\texttt{REW}(T)/\texttt{OPT}(T)) \leq 1 - \varepsilon.$$

This can be seen as a consequence of Fact 1 above. For, considering first when $C = 0$, if the above is not satisfied for any $\varepsilon > 0$ (otherwise we already have our ε), then the conclusion must be ALG achieves sublinear regret when $C = 0$. Then applying the above Fact 1 gives a suitable attack strategy with a suitable ε. Though using the competitive ratio as benchmark gives us non-trivial robustness guarantees, it is unclear whether there exists a better benchmark under which robustness can be measured.

6.2 Smoothed Myopic Response

Motivated by smoothed analysis, this work proposes the smoothed myopic response as a defense strategy. At each round, let the empirically best arm be a_t^*. The response is then a ρ-smoothed version of the myopic response. In other words, every arm $a \neq a_t^*$ is pulled with probability ρ, and the empirical best arm is pulled with probability $1-(K-1)\rho$. When the learner puts a small constant as exploration probability on each arm, he will eventually discover the best arm as long as the total corruption budget of the adversary is sublinear in T.

Theorem 6. *Fix any* $\varepsilon > 0$*, for a sufficiently large* T*, with probability* $1 - 1/T$*, algorithm 6 achieves the following:*

$$\texttt{REW} \geq (1 - \varepsilon)\texttt{OPT} - o(T).$$

Proof Sketch. Let a be any suboptimal arm. At any round, arm a has a probability of at least ρ of being chosen. After $T = \Omega(\frac{C}{\rho\Delta_a})$ rounds, arm a has been chosen at least $\Omega(\frac{C}{\Delta_a})$ times and the effect of the corruption diminishes. Specifically, the gap between pre-attack mean and post-attack mean drops below $O(\Delta_a)$. Consequently, after enough rounds (which is sublinear in T), the suboptimality is identified and the learner will choose arm a with probability ρ. Setting ρ to be a sufficiently small constant achieves a competitive ratio arbitrarily close to 1. □

Algorithm 6 Smoothed Response

Input: ε, target competitive ratio is $(1 - \varepsilon)$
$\rho := \varepsilon/K$
for $t = 1, 2, \ldots, T$ **do**
  Let $a_t^* = \arg\max \hat{\mu}_a(t)$
  Let

$$p_{t,a} = \begin{cases} \rho & \text{if } a \neq a_t^* \\ 1 - (K-1)\rho & \text{if } a = a_t^* \end{cases}$$

  Choose arm sampled from p_t
end for

Algorithm 7 Quantal Response

Input: ε: target competitive ratio is $(1 - \varepsilon)$;
$\lambda := 2 \ln \frac{K}{\varepsilon}$
for $t = 1, 2, \ldots, T$ **do**
  $\hat{\mu}^*(t) = \arg\max_a \hat{\mu}_a(t)$
  $\psi_a(t) = \hat{\mu}_a(t)/\hat{\mu}^*(t)$
  Let $p_{t,a} = \exp(\lambda\psi_a(t))/\sum_b \exp(\lambda\psi_b(t))$
  Choose arm sampled from p_t
end for

6.3 Quantal Response

The second response model is motivated by literature on bounded rationality, specifically the quantal response model. At each round, the learner computes the empirical mean and the ratio to the empirically best arm $\psi_a(t) = \hat{\mu}_a(t)/\hat{\mu}^*(t)$. The learner then assigns a probability to pull each arm proportional to $\exp(\lambda\psi_a(t))$, where λ is a parameter chosen by the learner. Thus, the probability can be interpreted as performing a softmax on the empirical means. In addition, the parameter λ controls the 'sharpness' of the smoothed distribution, the larger the λ, the less exploration the learner takes. To illustrate this point, if λ is taken to be $+\infty$, the learner always acts myopically, and if $\lambda = 0$, the learner always does uniform exploration.

Theorem 7. *Fix any constant* $\varepsilon > 0$*, for a sufficiently large* T*, with probability* $1-1/T$*, algorithm 7 achieves:*

$$\texttt{REW} \geq (1 - \varepsilon)\texttt{OPT} - o(T).$$

Table 1: Comparisons of cumulative attack cost. The first entry 'Baseline' runs attack strategy in [Jun et al., 2018] against UCB. Second and third entry runs the proposed attack strategy against UCB and Thompson Sampling (TS). Results on proposed attack strategy all have standard deviation within 1.0 after 10 random trials.

Setting	σ	$\mu = 0.1$	$\mu = 1$	$\mu = 2$
Baseline	0.1	23.6	129.4	247.3
	1	114.4	241.7	360.3
	2	239.4	367.6	475.5
UCB	0.1	1.3	2.4	3.6
	1	14.5	15.9	16.8
	2	30.3	30.7	31.0
TS	0.1	13.0	13.9	15.0
	1	19.0	19.7	20.6
	2	23.8	25.0	26.7

7 EXPERIMENTS

This section describes the numerical simulations on the proposed optimal attack strategies and defense strategies [1].

7.1 Experiments on Attack Strategies

In this subsection, I simulate the proposed attack strategies on UCB and Thompson sampling and describe the results of numerical experiments. In the experiments, the bandit instance has two arms, and the reward distributions are $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(0, \sigma^2)$ 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. For both UCB and Thompson sampling, 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 against UCB in [Jun et al., 2018] as a baseline for comparison.

In the experiments for both UCB and Thompson sampling, any non-target arm is pulled no more than 2 times in any trial, while the target arm is pulled almost every round. This validates the theoretical results, which indicate that any non-target arm gets pulled no more than $0.5 \log \log T$ times.

The cumulative attack costs for different choices of (μ, σ) are summarized in Table 1. For UCB, the empirical results fit nicely with the theoretical bound of $\widehat{O}(\sigma\sqrt{\log T} + \mu)$ in this work (specializing the upper bound on attack cost on UCB to the 2 arm setting, similar for the following bounds). The results also show a significant improvement over the attack strategy proposed in [Jun et al., 2018], which had a theoretical bound of $O(\mu \log T + \sigma \log T)$. For Thompson sampling, the empirical results fit nicely with the theoretical bound of $\widehat{O}(\sqrt{\log T} + \mu + \sigma)$. Also note that in the attack cost on UCB bandits, the $\sqrt{\log T}$ factor is multiplied by σ but not for Thompson sampling. Thus, the attack cost for UCB is expected to grow faster than the attack cost for Thompson sampling as σ gets large, and vice versa, as shown is Table 1.

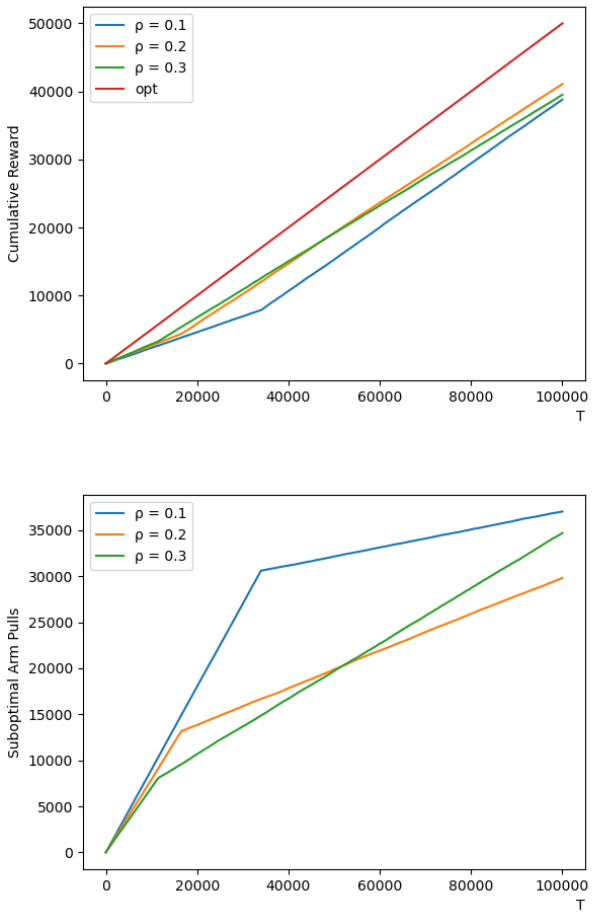

Figure 1: Top subfigure shows cumulative reward under smoothed myopic response, bottom subfigure shows the number of suboptimal arm pulls. The learner does more exploration as ρ increases and will be quicker to identify the optimal arm, but the excess exploration hurts performance in the long run. All values have coefficient of deviation within 0.02 after 10 random trials.

The experiments validate the theoretical results and empirically demonstrate that adding very small cor-

[1] Code available at `https://github.com/ShiliangZuo/BanditAttack.git`

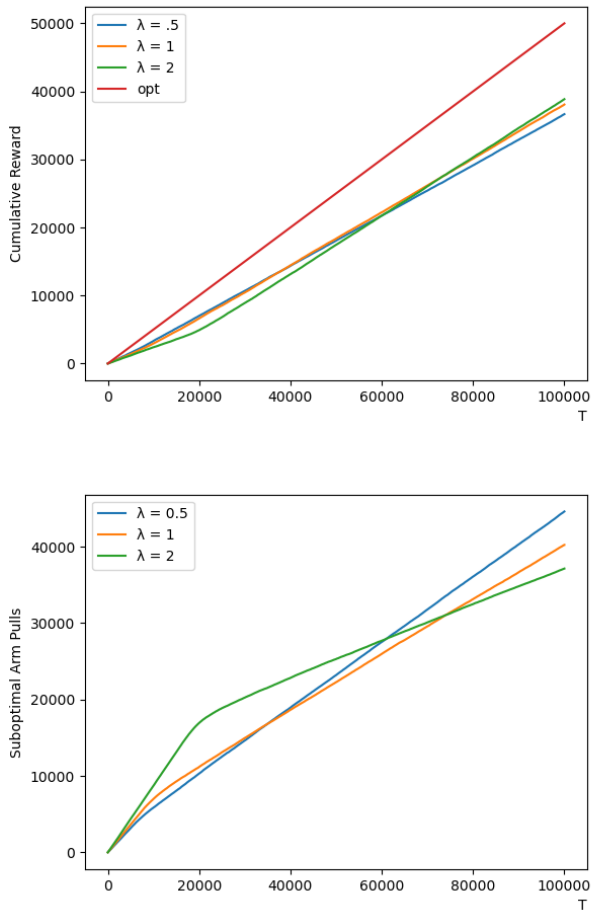

Figure 2: Top subfigure shows cumulative reward under quantal response, bottom subfigure shows the number of suboptimal arm pulls. The learner does more exploration as λ decreases and will be quicker to detect the optimal arm, but the excess exploration hurts performance in the long run. All values have coefficient of deviation within 0.02 after 10 random trials.

ruptions allows the adversary to steer the learner away from the actual optimal arm and manipulates the learner into choosing the adversary's target arm.

7.2 Experiments on Defense Strategies

This subsection describes the second set of experiments on the two proposed defense strategies. There are 2 arms giving Bernoulli rewards. The first arm is the optimal arm and gives rewards with mean $\mu_1 = 0.5$, the second arm is the suboptimal arm and gives rewards with mean $\mu_2 = 0.2$. The time horizon $T = 10^5$ and the corruption budget of the adversary is $C = 10^3$. The adversary adopts the following attack strategy: whenever the learner chooses the first arm (i.e. the optimal arm) and the realized reward is 1, the adversary changes this reward to 0. Figure 1 and figure 2 show the cumulative reward as a function of T, using the smoothed myopic response and the quantal response respectively. In general, when the learner does more exploration (larger ρ in smoothed myopic response and smaller λ in quantal response), he will be quicker to identify the optimal arm. But this excess exploration will hurt performance in the long run.

8 CONCLUSION

In this work, I study adversarial attacks that manipulate the behavior of stochastic bandit algorithms by corrupting the reward the learner observes. Both attack and defense strategies are studied. From the adversary's perspective, I give nearly optimal attack strategies. Tight characterizations are given on the attack cost needed to manipulate the UCB, Thompson sampling, and ε-greedy algorithm into pulling some target arm of the adversary's choosing. For UCB and Thompson sampling, I propose optimal attack strategies with attack cost $\widehat{O}(\sqrt{\log T})$ and establish matching lower bounds $\Omega(\sqrt{\log T})$ on the cumulative attack cost. For the ε-greedy algorithm, I give a lower bound of $\Omega(\log T)$ on the attack cost. I also stdudy defense strategies for the learner. Motivated by literature from smoothed analysis and behavioral economics, I give two simple algorithms that achieve a competitive ratio arbitrarily close to 1.

References

[Abramowitz et al., 1988] Abramowitz, M., Stegun, I. A., and Romer, R. H. (1988). Handbook of mathematical functions with formulas, graphs, and mathematical tables.

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

[Agrawal and Goyal, 2012] Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings.

[Agrawal and Goyal, 2017] Agrawal, S. and Goyal, N. (2017). Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24.

[Arora et al., 2012] Arora, S., Hazan, E., and Kale, S. (2012). The multiplicative weights update method:

a meta-algorithm and applications. *Theory of computing*, 8(1):121–164.

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

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

[Camerer, 2011] Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton university press.

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

[Dunne and Campbell, 1997] Dunne, R. A. and Campbell, N. A. (1997). On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In *Proc. 8th Aust. Conf. on the Neural Networks, Melbourne*, volume 181, page 185. Citeseer.

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

[Gupta et al., 2019] Gupta, A., Koren, T., and Talwar, K. (2019). Better algorithms for stochastic bandits with adversarial corruptions. *arXiv preprint arXiv:1902.08647*.

[Haghtalab et al., 2020] Haghtalab, N., Roughgarden, T., and Shetty, A. (2020). Smoothed analysis of online and differentially private learning. *Advances in Neural Information Processing Systems*, 33:9203–9215.

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

[He et al., 2022] He, J., Zhou, D., Zhang, T., and Gu, Q. (2022). Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in Neural Information Processing Systems*, 35:34614–34625.

[Jang et al., 2016] Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

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

[Kannan et al., 2018] Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., and Wu, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31.

[Kuleshov and Precup, 2014] Kuleshov, V. and Precup, D. (2014). Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*.

[Leme et al., 2022] Leme, R. P., Podimata, C., and Schneider, J. (2022). Corruption-robust contextual search through density updates. In *Conference on Learning Theory*, pages 3504–3505. PMLR.

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

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

[Luce, 2012] Luce, R. D. (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

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

[Ma et al., 2018] Ma, Y., Jun, K.-S., Li, L., and Zhu, X. (2018). 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*, pages 186–204. Springer.

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

[McFadden, 1976] McFadden, D. L. (1976). Quantal choice analaysis: A survey. *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 363–390.

[McKelvey and Palfrey, 1995] McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38.

[Rangi et al., 2022] Rangi, A., Tran-Thanh, L., Xu, H., and Franceschetti, M. (2022). Saving stochastic bandits from poisoning attacks via limited data verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8054–8061.

[Sivakumar et al., 2020] Sivakumar, V., Wu, S., and Banerjee, A. (2020). Structured linear contextual bandits: A sharp and geometric smoothed analysis. In *International Conference on Machine Learning*, pages 9026–9035. PMLR.

[Spielman and Teng, 2004] Spielman, D. A. and Teng, S.-H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463.

[Wei et al., 2022] Wei, C.-Y., Dann, C., and Zimmert, J. (2022). A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR.

[Wu et al., 2022] Wu, J., Shen, W., Fang, F., and Xu, H. (2022). Inverse game theory for stackelberg games: the blessing of bounded rationality. *Advances in Neural Information Processing Systems*, 35:32186–32198.

[Zhang et al., 2020] Zhang, X., Ma, Y., Singla, A., and Zhu, X. (2020). Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*, pages 11225–11234. PMLR.

[Zimmert and Seldin, 2021] Zimmert, J. and Seldin, Y. (2021). Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358.

[Zuo, 2023] Zuo, S. (2023). Corruption-robust lipschitz contextual search. *arXiv preprint arXiv:2307.13903*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

9 PROOFS: ATTACK STRATEGY AGAINST UCB

This section details the proof of Theorem 1. I begin with a lemma characterizing the number of pulls for any non-target arm. Recall event E is the event that pre-attack empirical means concentrate around the true mean (Lemma 1).

Lemma 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

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

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,

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

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 eq. (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. □

Proof. (of Theorem 1) Assume event E holds throughout this proof. By lemma 2, any non-target arm is pulled $O(\log\log T)$ times. Recall $\tau_a(t)$ is the set of timesteps in which arm a was chosen.

For any t,

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

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

$$\hat{\mu}_a(t) = \hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma e^{N_a(t)}.$$

Consequently by the above two equations:

$$\frac{1}{N_a(t)} \sum_{s \in \tau_a(s)} \alpha_s = \hat{\mu}_a^0(t) - \hat{\mu}_K(t) + 2\beta(N_K(t)) + 3\sigma e^{N_a(t)}$$

$$\begin{aligned}
&\le \Delta_a^+ + \beta(N_a(t)) + 3\beta(N_K(t)) + 3\sigma e^{N_a(t)} \\
&\le \Delta_a^+ + \beta(N_a(t)) + 3\beta(N_K(t)) + 3\sigma e^{0.5\log\log t+1} \\
&\le \Delta_a^+ + 4\beta(N_a(t)) + 3e\cdot\sigma\sqrt{\log t}.
\end{aligned}$$

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:

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

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

10 PROOFS: ATTACK STRATEGY AGAINST THOMPSON SAMPLING

This section details the proof of Theorem 2. I first give a concentration result that will be useful in this section. Let $\zeta(t) := \sqrt{2\log\frac{\pi^2Kt^2}{3\delta}}$. Denote the event in the following lemma by E_1.

Lemma 3. *With probability $1-\delta$, for any round t, we have $|\nu_a(t) - \hat{\mu}_a(t)| < \zeta(t)/N_a(t)$.*

Proof. Fix round t and an arm a. Then by a standard Gaussian tail bound:

$$\begin{aligned}
\Pr[|\nu_a(t) - \hat{\mu}_a(t)| > \zeta(t)/N_a(t)] &< 2\exp\big(-\zeta(t)^2/2\big) \\
&= \frac{6\delta}{\pi^2}\cdot\frac{1}{Kt^2}
\end{aligned}$$

The lemma then follows from a union bound over t and arms a. □

Lemma 4. *Assume event E and E_1 hold. At any round t, arm a is pulled for $\lceil 0.5\cdot\log\log t\rceil$ times for any non-target arm $a \neq K$.*

Proof. Assume for the sake of contradiction arm a is pulled more than $\lceil 0.5\log\log t\rceil$ times. Suppose at round t_0 arm a is pulled for the $\lceil 0.5\log\log t\rceil$-th time. After round t_0, we must have:

$$\begin{aligned}
\hat{\mu}_a(t_0) &\le \hat{\mu}_K(t_0) - 2\beta(N_K(t_0)) - 4\exp(N_a(t)) - \sqrt{8\log\frac{\pi^2K}{3\delta}} \\
&\le \hat{\mu}_K(t_0) - 2\beta(N_K(t_0)) - 4\sqrt{\log t} - \sqrt{8\log\frac{\pi^2K}{3\delta}}
\end{aligned} \tag{3}$$

Assume arm a has been chosen again at round $t_1 > t_0$. Then we must have

$$\nu_a(t_1) > \nu_K(t_1).$$

However,

$$\begin{aligned}
\nu_a(t_1) &< \hat{\mu}_a(t_0) + \zeta(t_1) \\
&< \hat{\mu}_K(t_0) - 2\beta(N_K(t_0)) - 4\sqrt{\log t} - \sqrt{8\log\frac{\pi^2K}{3\delta}} + \zeta(t_1) \\
&< \hat{\mu}_K(t_1) - 4\sqrt{\log t} - \sqrt{8\log\frac{\pi^2K}{3\delta}} + \zeta(t_1) \\
&< \hat{\mu}_K(t_1) - \zeta(t_1) \\
&< \nu_K(t_1)
\end{aligned}$$

Here, the first line follows from event E_1, the second line follows from the design of the attack strategy (eq. (3)), the third line follows from event E, the fourth line follows from the definition of ζ, and the final line follows from event E_1 again. Hence, a contradiction is established, and arm a will not be picked again before round t. □

Proof. (of Theorem 2) In a similar fashion as the proof in theorem 1, the attack cost on arm a can be bounded by:

$$\begin{aligned}\frac{1}{N_a(t)}\sum_{s\in\tau_a(s)}\alpha_s &\le \hat{\mu}_a^0(t) - \hat{\mu}_K(t) + 2\beta(N_K(t)) + 2.9\exp(N_a(t)) + 2\sqrt{\log\frac{\pi^2 K}{3\delta}} \\ &\le \Delta_a^+ + 4\beta(N_a(t)) + 2.9\sqrt{\log T} + 2\sqrt{\log\frac{\pi^2 K}{3\delta}}.\end{aligned}$$

Hence the attack cost on arm a can be bounded as:

$$\sum_{s\in\tau_a(s)}\alpha_s = \widehat{O}(\Delta_a^+ + \sqrt{\log T} + (\sigma+1)\sqrt{\log\frac{\pi^2 K}{3\delta}}).$$

Summing over all non-target arms completes the proof. □

11 PROOFS: LOWER BOUNDS

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

$$\hat{\mu}_a(t) - \frac{C_a(t)}{N_a(t)} \le \hat{\mu}_a^0(t) \le \hat{\mu}_a(t) + \frac{C_a(t)}{N_a(t)}. \tag{4}$$

11.1 UCB Lower Bound

Proof. (of Theorem 3) Throughout this proof assume event E holds with $\delta = 0.1$. Suppose the non-target arm has been pulled no more than $T/26$ times. We show the attack cost is at least $\Delta + 0.22\sigma\sqrt{\log T - 1}$. Consider the last round the target arm is pulled. Denote this timestep by t, then $t > T/2$. 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 eq. (4)

$$\begin{aligned}&\mu_2(t) + \beta(N_2(t)) + \frac{C_2(t)}{N_2(t)} + 3\sigma\sqrt{\frac{\log t}{N_2(t)}} \\ >\ &\mu_1(t) - \beta(N_1(t)) - \frac{C_1(t)}{N_1(t)} + 3\sigma\sqrt{\frac{\log t}{N_1(t)}}.\end{aligned}$$

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

$$\begin{aligned}&\frac{C_1(t) + C_2(t)}{N_1(t)} \\ &> \Delta - \beta(N_1(t)) - \beta(N_2(t)) + 3\sigma\sqrt{\frac{\log t}{N_1(t)}} - 3\sigma\sqrt{\frac{\log t}{N_2(t)}} \\ &\ge \Delta - 1.29\beta(N_1(t)) + 2.8\sigma\sqrt{\frac{\log t}{N_1(t)}}\end{aligned}$$

$$= \Delta - 1.29\sqrt{\frac{2\sigma^2}{N_1(t)}\log\frac{2\pi^2 N_1(t)^2}{3\delta}} + 2.8\sigma\sqrt{\frac{\log t}{N_1(t)}}$$
$$\geq \Delta + 0.22\sigma\sqrt{\frac{\log t}{N_1(t)}}.$$

Finally,

$$C_1(t) + C_2(t) \geq N_1(t)\Delta + 0.22\sigma\sqrt{N_1(t)\log t}$$
$$\geq \Delta + 0.22\sigma\sqrt{\log t}.$$

This finishes the proof. □

11.2 Thompson Sampling Lower Bound

In the following let $\Pr\big[\mathcal{N}(\mu,\sigma^2) > t\big]$ denote the complementary CDF (tail probability) of a Gaussian random variable with specified mean and variance.

Lemma 5 ([Abramowitz et al., 1988])**.** *The tail for a Gaussian distribution can be lower bounded by:*

$$\Pr[\mathcal{N}(-\mu,1) > 0] > \sqrt{\frac{2}{\pi}}\frac{\exp\big(-\mu^2/2\big)}{\mu + \sqrt{\mu^2+2}}.$$

Lemma 6. *Let C be the total attack cost. For a round t, if $N_1(t) < N_2(t)$, then the probability that the non-target arm gets pulled is at least* $\min\Big(\Pr\Big[\mathcal{N}(\frac{\Delta - C}{N_1(t)}, \frac{1}{N_1(t)}) > 0\Big], 1/2\Big)$.

Proof. Fix a round t. If $\hat{\mu}_1(t) > \hat{\mu}_2(t)$, then $\Pr[a_t = 1] > 1/2$. Now assume $\hat{\mu}_1(t) < \hat{\mu}_2(t)$.

$$\begin{aligned}
\Pr[a_t = 1] &= \Pr[\nu(t) > 0] \\
&= \Pr\left[\mathcal{N}(\hat{\mu}_1(t) - \hat{\mu}_2(t), \frac{1}{N_1(t)} + \frac{1}{N_2(t)}) > 0\right] \\
&\geq \Pr\left[\mathcal{N}(\hat{\mu}_1^0(t) - \hat{\mu}_2^0(t) - \frac{C}{N_1(t)}, \frac{1}{N_1(t)} + \frac{1}{N_2(t)}) > 0\right] \\
&\geq \Pr\left[\mathcal{N}(\mu_1(t) - \mu_2(t) - \frac{C}{N_1(t)}, \frac{1}{N_1(t)}) > 0\right] \\
&= \Pr\left[\mathcal{N}(\Delta - \frac{C}{N_1(t)}, \frac{1}{N_1(t)}) > 0\right] \\
&\geq \Pr\left[\mathcal{N}(\frac{\Delta - C}{N_1(t)}, \frac{1}{N_1(t)}) > 0\right]
\end{aligned}$$

Here, the third line is because of eq. (4), the fourth line is because the tail probability cannot increase if we decrease the variance of the Gaussian. □

Lemma 7. *Assume the adversary spends an attack cost no more than $C := \Delta + 0.1\sqrt{\log T}$. Then the non-target arm has been chosen $T^{0.8}$ times during period $[T/5, T/2]$.*

Proof. Consider rounds during the period $[T/5, T/2]$. If at any point t, $N_1(t) > N_2(t)$, then the non-target arm has already been chosen $T/10$ times. Hence we can assume $N_1(t) < N_2(t)$.

$$\Pr[a_t = 1] \geq \Pr\left[\mathcal{N}(\frac{\Delta - C}{N_1(t)}, \frac{1}{N_1(t)}) > 0\right]$$
$$\geq \Pr\Big[\mathcal{N}(-0.1\sqrt{\log T}, 1) > 0\Big]$$

$$\geq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{0.2\sqrt{\log T}} \cdot \exp(-0.005 \log T)$$
$$> T^{-0.1}$$

The expected number of pulls during this period is at least $0.3T^{0.9}$. For a sufficiently large T, by a simple Hoeffding inequality, with probability 0.9, the learner has chosen the non-target arm for at least $T^{0.8}$ rounds during this period. □

Proof. (of Theorem 4) Assume the adversary spends an attack cost no more than $C := \Delta + 0.1\sqrt{\log T}$. We will show with probability at least 0.8, the learner chose the non-target arm more than $T/10$ times. By the previous lemma, the learner choose the non-target arm at least $T^{0.8}$ times during the period $[T/5, T/2]$.

Now, consider rounds during the period $[T/2, T]$. The following holds:

$$\begin{aligned}
\Pr[a_t = 1] &= \Pr\left[\mathcal{N}(\frac{\Delta - C}{N_1(t)}, \frac{1}{N_1(t)}) > 0\right] \\
&\geq \Pr\left[\mathcal{N}(\frac{\Delta - C}{T^{0.8}}, \frac{1}{T}) > 0\right] \\
&\geq \Pr\left[\mathcal{N}(\frac{-0.1\sqrt{\log T}}{T^{0.8}}, \frac{1}{T}) > 0\right] \\
&\geq \Pr\left[\mathcal{N}(\frac{-0.1}{T^{0.5}}, \frac{1}{T}) > 0\right] \\
&\geq \Pr[\mathcal{N}(-0.1, 1) > 0] \\
&> 0.4.
\end{aligned}$$

The expected number of pulls during this period is at least $T/5$. For a sufficiently large T, by Hoeffding inequality, with probability 0.9, the learner chooses the non-target arm at least $T/10$ times.

Hence, with a probability of at least 0.8, the learner has chosen the non-target arm at least $T/10$ times. □

11.3 ε-greedy Lower Bound

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

Lemma 8. *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

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

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

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. □

Proof. (of Theorem 5) Throughout this proof assume event E and lemma 8 holds with $\delta = 0.1$. By a union bound, these events hold simultaneously with probability at least 0.8. Assume the target arm has been pulled at least $3T/4$ rounds. We will show the adversary spend an attack cost at least $\Omega(\Delta \log T)$. 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 $2c \log T$. Thus to ensure the 2nd arm is pulled no less than $T - T/4$ rounds, we must have

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

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 eq. (4):

$$\mu_2(t) + \beta(N_2(t)) + \frac{C_2(t)}{N_2(t)} > \mu_1(t) - \beta(N_1(t)) - \frac{C_1(t)}{N_1(t)}$$

leading to

$$\begin{aligned} C_1(t) + C_2(t) &> N_1(t)(\Delta - 2\beta(N_1(t))) \\ &> c \cdot \Delta \log T/6, \end{aligned}$$

since assuming lemma 8 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. □

12 PROOFS: COMPETITIVE RATIO WITH SMOOTHED RESPONSES

Lemma 9. *Assume for each arm a, the probability of pulling a in each round t can be lower bounded by ρ. Then with probability $1 - \delta$, for every $t > \frac{10}{\rho} \cdot \log \frac{K}{\delta}$, the following holds:*

$$N_a(t) \geq \frac{\rho t}{2}.$$

Moreover, with probability $1 - 2\delta$, for any $t > t_0 := \max(\frac{10}{\rho} \cdot \log \frac{K}{\delta}, \frac{16C}{\rho \Delta_a}, \frac{20 \log(T/\delta)}{\rho \Delta_a^2})$, the learner has:

$$\hat{\mu}_a(t) < \hat{\mu}_{a^*} - \Delta_a/2.$$

Proof. Fix arm a and round t. With probability $1-\delta$, by Chernoff bound:

$$\Pr\left[N_a(t) < \frac{\rho t}{2}\right] < \exp(-0.25\rho t/2)$$

Moreover, it is easy to verify that for any $t > \frac{10}{\rho} \cdot \log \frac{K}{\delta}$:

$$\exp(-0.25\rho t/2) \le \frac{\delta}{Kt^2}.$$

A union bound then completes the first part of the proof.

Now, fix some suboptimal arm a and let $t > t_0$. Then $N_a(t) > \max(\frac{8C}{\Delta_a}, \frac{10\log T}{\Delta_a^2})$. Consequently by eq. (4),

$$\hat{\mu}_a(t) \le \hat{\mu}_a^0(t) + \frac{C}{N_a(t)} \le \hat{\mu}_a^0(t) + \frac{\Delta_a}{8},$$
$$\hat{\mu}_{a^*}(t) \ge \hat{\mu}_{a^*}^0(t) - \frac{C}{N_a(t)} \ge \hat{\mu}_{a^*}^0(t) - \frac{\Delta_a}{8}.$$

And by event E,

$$\hat{\mu}_a^0(t) \le \hat{\mu}_a + \sqrt{\frac{\log(T/\delta)}{N_a(t)}} \le \mu_a + \frac{\Delta_a}{8},$$
$$\hat{\mu}_{a^*}^0(t) \le \hat{\mu}_{a^*} + \sqrt{\frac{\log(T/\delta)}{N_a(t)}} \ge \mu_{a^*} - \frac{\Delta_a}{8}.$$

Rearranging the above inequalities completes the proof. □

12.1 Proof for Smoothed Myopic Response

Proof. (of Theorem 6) Choose $\delta = 1/T$ in Lemma 9. After round t_0, any suboptimal arm has a probability ρ of being pulled. Hence the total pulls of some suboptimal arm a can be bounded by:

$$O(\frac{C}{\Delta_a} + \frac{\log T}{\Delta_a^2}) + 2\rho T.$$

The total regret contributed by suboptimal arm a can be bounded by

$$O(C + \frac{\log T}{\Delta_a}) + 2\rho T\Delta_a \le o(T) + 2\rho T\mu^*.$$

The optimal reward is $T\mu^*$. The expected collected reward is:

$$\begin{aligned}\texttt{REW} &\ge T\mu^* - \sum_a 2\rho T\mu^* - o(T)\\ &\ge T\mu^* - 2\rho KT\mu^* - o(T)\\ &\ge (1-2\rho K)T\mu^* - o(T).\end{aligned}$$

Therefore setting $\rho = \varepsilon/2K$ achieves the final regret. □

12.2 Proof for Quantal Response

In the following let $\psi_a(t) = \frac{\hat{\mu}_a(t)}{\hat{\mu}^*(t)}$ as in the algorithm, then $\psi_a(t) \in [0,1]$. Also let $\rho(\lambda) = \frac{1}{K\exp(\lambda)}$.

Lemma 10. *For any round t and any arm a, $p_{t,a} > \rho(\lambda)$.*

Proof.

$$\begin{aligned} p_{t,a} &= \frac{\exp(\lambda\psi_a(t))}{\sum_b \exp(\lambda\psi_b(t))} \\ &= \frac{\exp(\lambda\psi_a(t))}{\exp(\lambda\psi_a(t)) + \sum_{b\neq a} \exp(\lambda\psi_b(t))} \\ &\geq \frac{1}{1 + (K-1)\exp(\lambda)} \\ &\geq \frac{1}{K\exp(\lambda)}. \end{aligned}$$

□

Proof. (of Theorem 7)

Choose $\delta = 1/T$ in Lemma 9. After $t > t_0$, $\hat{\mu}_a(t) < \hat{\mu}_{a^*}(t) - \Delta_a/2$. Thus

$$\psi_a(t) \leq \frac{\hat{\mu}_{a^*}(t) - \Delta_a/2}{\hat{\mu}_{a^*}(t)} \leq 1 - \frac{\Delta_a/2}{\mu^* + \Delta_a/2} \leq 1 - \frac{1}{4}\frac{\Delta_a}{\mu^*}.$$

The probability that arm a gets pulled after t_0 is then:

$$\begin{aligned} p_{t,a} &\leq \frac{\exp(\lambda\psi_a(t))}{\exp(\lambda\psi_a(t)) + \exp(\lambda\psi_{a^*}(t)))} \\ &\leq \frac{1}{1 + \exp(\lambda - \lambda\psi_a(t))} \\ &\leq \frac{1}{2\exp(\lambda - \lambda\psi_a(t))} \\ &\leq \frac{1}{2\exp\left(\frac{\lambda\Delta_a}{2\mu^*}\right)} \\ &\leq \frac{\mu^*}{\lambda\Delta_a}. \end{aligned}$$

The total regret contributed by arm a can be bounded by:

$$O(\frac{C}{\rho(\lambda)} + \frac{\log T}{\rho(\lambda)\Delta_a}) + \frac{2\mu^*}{\lambda\Delta_a} \cdot T \cdot \Delta_a \leq o(T) + \frac{2T\mu^*}{\lambda}.$$

Hence the total collected reward can be bounded by:

$$\begin{aligned} \texttt{REW} &\geq T\mu^* - \frac{2K}{\lambda}T\mu^* - o(T) \\ &\geq (1 - \frac{2K}{\lambda})T\mu^* - o(T). \end{aligned}$$

Choosing $\lambda = \frac{2K}{\varepsilon}$ finishes proof. □