Delayed Adversarial Attacks on Stochastic Multi-Armed Bandits

Pierriccardo Olivieri
Politecnico di Milano
pierriccardo.olivieri@polimi.it

Matteo Castiglioni
Politecnico di Milano
matteo.castiglioni@polimi.it

Nicola Gatti
Politecnico di Milano
nicola.gatti@polimi.it

Abstract

We study adversarial attacks on stochastic bandits when, differently from previous works, the attack of the malicious attacker is delayed and starts after the learning process begins. In particular, we focus on strong attacks and capture the setting where the malicious attacker does not have information about the beginning of the learning process. Interestingly, the lack of such information can dramatically affect the effectiveness of the attack. We analyze this scenario in the case of the UCB algorithm facing an omniscient attacker, providing a more general framework to study adversarial attacks on stochastic bandit algorithms. We characterize the success and profitability of the attack depending on the round in which the attack starts. In particular, we derive an upper and a lower bound on the number of target arm pulls, showing that our bound is tight up to a sublinear factor. Moreover, we provide a condition that identifies when an attack can be successful. Finally, we empirically evaluate the tightness of our theoretical bounds on synthetic instances.

1 Introduction

In recent years, the adoption of machine learning applications accelerates at an unprecedented pace, increasing the impact of this technology in industry and in many aspects of humanity. Evaluating how these systems can be harmed is crucial. Many works address this issue in deep learning and reinforcement learning, trying to understand how a malicious entity could "attack" machine learning systems to alter their behaviour Goodfellow et al. [2014], Sun et al. [2020], Inkawhich et al. [2019]. Other works instead investigate defence strategies, designing robust techniques Chen et al. [2019], Zhang et al. [2020], Pattanaik et al. [2017]. However, the problem of security remains open and is still immature. In optimal decision-making scenarios, the concept of "attack" can be translated as a way to alter the learner's behaviour, fooling the algorithm into selecting specific actions or dramatically reducing the performance. Multi-Armed Bandits (MAB) Auer et al. [2002] are popular and appealing online decision-making algorithms as they provide important theoretical guarantees. Being a balanced framework concerning usability and training, they are used in several real-world scenarios Bouneffouf et al. [2020] such as medical trials [Durand et al., 2018], recommender systems [Zhou et al., 2017], advertising [Castiglioni et al., 2022], and finance [Shen et al., 2015, Huo and Fu, 2017].

In the adversarial attack framework on multi-armed bandits (MAB) there are three entities: a learner, an environment, and an attacker. The learner aims to optimize its policy interacting with the environment. The attacker tries to alter the behaviour of the learner its by corrupting the reward feedback. The attacker aims to fool the learner into selecting a specific target arm which is sub-optimal.

38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

In previous works, the attack starts at the beginning of the learning process and it is considered successful if the learner, upon receiving corrupted observations, selects the target arm $T - o(T)$ times. In this works, the attacker decides the amount of corruption observing the arm played by the learner and the reward generated by the environment. Such framework is denoted as strong attack model and was originally proposed by Jun et al. [2018], becoming a standard in adversarial attacks literature on MAB. The strong attack framework has been proved to be *unrecoverable* for most of the classical bandit algorithms Jun et al. [2018], Liu and Shroff [2019] meaning that a learner under attack will always experience linear regret. However, all previous works designed unrecoverable attacks under the strong assumption that the attack starts synchronously with the start of the bandit learning dynamics. This assumption implicitly requires that the attacker has perfect information about the start of the learning process. This is not always possible in practice. Indeed, in real-world applications, when an attacker decides to attack it does not know the current step t of the learning process. Motivated by this observation, it is natural to investigate the theoretical guarantees of an attack starting at time t^A of the learning process. We call this new framework delayed attack model. To the best of our knowledge, this problem is completely unexplored in bandits with adversarial attacks.

1.1 Original Contributions

We define the delayed attack model, a generalization of the classic adversarial attack framework, where an attack can start at a generic time t^A, providing renewed definition of a successful attack. Intuitively, as t^A increases, it becomes hard to match the definition of success of previous works, i.e., force the learner to selects the target arm $T - o(T)$ times. This requires a more fine-grained analysis of what it is possible to achieve in this setting. In particular, we identify an attack as successful if the learner selects the target arm a linear number of times. Moreover, we define the concept of profitability that is related to how effective is the attack in forcing the learner to select the target arms.

We focus on an UCB learner and an omniscient attacker (i.e., that knows the problem instance) and we derive an upper and lower bound on the number of pulls of the target arm as a function of the beginning of the attack t^A. In particular, we show that under an optimal attack the number of pull of the target arm is approximately

$$(T - t^A) - \frac{\Delta_{o,\tau}}{\epsilon} t^A,$$

where $\Delta_{o,\tau}$ is the difference between the mean reward of the optimal arm and the target arm, and the corruption is $\Delta_{o,\tau} + \epsilon$. Intuitively, as it is customary in the literature, ϵ is a parameter that identify how prone is the attacker to corrupt the rewards more than the minimum requirement $\Delta_{o,\tau}$. Our bounds are tight up to a sublinear factor. This provides a tight characterization of the pulls of the target arm induced by an optimal attack depending on the starting time t^A. Furthermore, we provide a condition to characterize whether an attack can be successful depending on the starting time t^A. In particular, we identify the threshold value

$$\alpha^*(\Delta_{o,\tau}, \epsilon) = \frac{\epsilon}{\epsilon + \Delta_{o,\tau}}$$

that depends on the gap $\Delta_{o,\tau}$, and the parameter ϵ. We show that if the attack starts at time $t^A = \alpha T$ with $\alpha < \alpha^*$, then under our attack the learner plays the target arm a linear number of times. On the other hand, if the attack starts at time $t^A = \alpha T$ with $\alpha > \alpha^*$, then under under any attack the learner plays the target arm a sublinear number of times. Finally, we empirically evaluate our attack via numerical experiments.

1.2 Related Works

The literature on adversarial attack on stochastic bandits can be divided into techniques to craft attacks and design of robust algorithms. Jun et al. [2018] introduce adversarial attacks on stochastic bandits, define the general framework, and propose several attack techniques. In particular, Jun et al. [2018] introduce the oracle attack where the attacker is omniscient, i.e., knows the problem instance, and the first two strong attack algorithms specific for ϵ-greedy and UCB. Liu and Shroff [2019] propose the Adaptive attack by Constant Estimation (ACE), an attack strategy agnostic to the learner's algorithm. Such works operate in the strong attack scenario where the attacker can observe both the played action and the corresponding reward. In contrast to this attack model, in the

weak attack setting, the attacker can only observe the reward vector generated by the environment. In the weak attack setting, Xu et al. [2021] propose an attack technique where the attacker does not need observations and provides a criterion to characterize families of bandits that are naturally vulnerable to adversarial attacks. The weak setting is clearly worse for the attacker and is mainly used in works regarding defence techniques. Concerning robust algorithms against adversarial attacks on stochastic bandits, Lykouris et al. [2018] propose a robust variation of the Active Arm Elimination (AAE) algorithm in the weak attack setting agnostic to the amount of corruption injected. In the same setting, Gupta et al. [2019] propose a corruption agnostic robust algorithm. Rangi et al. [2022] study a defence mechanism against a weak attacker in the stochastic setting when the learner can access limited corruption-free samples. Guan et al. [2020] propose a robust algorithm for a different attack model where the attacker can deal with an unbounded attack with a certain probability. Zhong et al. [2021] proposes Probabilistic Sequential Shrinking (PSS), a robust technique for the best arm identification problem under adversarial corruption. Aside from the stochastic bandit setting, several works analyse the adversarial attack framework also for adversarial bandits [Ma and Zhou, 2023, Yang et al., 2021], for gaussian process bandits [Bogunovic et al., 2020a, Han and Scarlett, 2022], continuous Markov decision processes [Maran et al., 2024], contextual bandits [Garcelon et al., 2020, Bogunovic et al., 2020b, Wang et al., 2022] and combinatorial bandits [Balasubramanian et al., 2024, Dong et al., 2022].

2 Preliminaries

In a MAB Auer et al. [2002] a learner interacts with an environment for T rounds. The learner has K available arms or actions. Each arm $i \in [K]$ [1] is associated with a σ^2-sub-Gaussian reward distribution γ_i with mean μ_i unknown to the learner. At each time $t \in [T]$, the learner selects an arm $i \in [K]$ and observes the corresponding reward $r_i(t) \sim \gamma_{i_t}$ generated by the environment. We denote the optimal arm as $o = \arg\max_{i \in [K]} \mu_i$. Let $\mathbb{I}\{\cdot\}$ be the indicator function. Then, we denote with

$$N_j(t) = \sum_{k=1}^{t} \mathbb{I}\{i_k = j\}$$

the number of times an arm $j \in [K]$ has been pulled until time $t \in [T]$, and with

$$N_j(t_1 \to t_2) = \sum_{k=t_1}^{t_2} \mathbb{I}\{i_k = j\}$$

the number of times an arm $j \in [K]$ has been pulled in the interval $[t_1, t_2]$ with $t_1 < t_2$ and $t_1, t_2 \in [T]$. Moreover, we denote with $\Delta_{i,j} \coloneqq \mu_i - \mu_j$ the gap between the means of two different arms $i, j \in [K]$. Finally, we define as $n_i(t) = \{t' \leq t : i_t = i\}$ the set of rounds where arm i is selected up to round t, and with $\hat{\mu}_i(t) = \sum_{t' \in n_i(t)} r(t')/N_i(t)$ the average reward of arm i up to round t.

The learner's objective is to minimize the regret over the time horizon, where the regret is defined as:

$$R(T) = \mu_o T - \sum_{t=1}^{T} \mu_{i_t}. \tag{1}$$

2.1 Recap on Classical Adversarial Attacks

In the classical adversarial attack framework, an additional entity, called the attacker, sits between the learner and the environment. The attacker, at each round $t \in [T]$, upon observing the arm i played by learner and the reward generated $r_i(t)$ may craft a corruption c_t to alter the reward observed by the learner in the following way:

$$\tilde{r}_i(t) = r_i(t) - c_t. \tag{2}$$

The attacker aims to fool the learner into selecting a target sub-optimal arm τ. We assume that the target arm τ is such that $\mu_\tau < \mu_o$, otherwise, the learner converges to play the target arm even without

[1] In this work, we refer as $[A]$, $A \in \mathbb{N}$, to the set $\{1, \ldots, A\}$.

the attack. [2] The aim of the attacker is to craft the minimal amount of corruption c_t such that the learner, receiving corrupted observations, believes the target arm τ optimal. The attacker is evaluated in terms of successfulness and cost of the attack. An attack is successful if the learner selects the target arm τ for $N_\tau(T) = T - o(T)$ rounds in expectation or high-probablity while the attacker pays a sublinear cost [Jun et al., 2018, Liu and Shroff, 2019]. The cost is defined as the total corruption injected in the time horizon $C(T) = \sum_{t=1}^{T} |c_t|$. In general, there is not a fixed budget for the attack. However, in designing attack techniques we prefer to be *stealth*, i.e., we aim for attacks that minimize the cost c_t inflicted at each round as dealing too much corruption at once can be suspicious in a realistic scenario.

2.2 Oracle Attack

An attack strategy is an online algorithm that, upon observing the arm i played by the learner and the generated reward $r_i(t)$, returns a corruption c_t. The oracle attack model proposed by Jun et al. [2018] is an ideal attack model where the attacker is omniscient, i.e., knows the true means μ_i for all $i \in [K]$. While an omniscient attacker might be unrealistic in practice, this attack model is useful to provide a cleaner attack and analysis. In the following, we introduce the main components of the attack in [Jun et al., 2018]. Suppose the attacker knows the true mean of each arm. When the learner selects an arm i different than the target τ the attacker crafts an attack c_t such that:

$$c_t = \mathbb{I}\{i \neq \tau\} \left[\Delta_{\tau,i} + \epsilon\right]_+ \tag{3}$$

Where, $[k]_+ = \max(0, k)$ and ϵ is an arbitrary constant strictly greater than 0. Several attack strategies provide attacks that do not assume an omniscient attacker, such as the attack to UCB in [Jun et al., 2018] or the more generic Adaptive attack by Constant Estimation (ACE) [Liu and Shroff, 2019].

The standard framework of adversarial attacks on stochastic bandits that we described assumes the attacker injects corruption from $t^A = 1$, i.e., when the bandit algorithm is instantiated and has no prior observations. This attack model have been proven undefendable against an attack. However, such negative result is heavily base on the $t^A = 1$ assumption. In the following section, we present a generalized model in which we relax this assumption analysing attacks that starts at an arbitrary time $t^A \geq 1$.

3 Delayed-Attack Framework

The delayed adversarial attack framework extends the classical attack model proposed by Jun et al. [2018] removing the assumption that the attack starts at $t^A = 1$. In particular, we assume a starting time $t^A \geq 1$ and analisy as the starting time affects the performance of an optimal attack. Clearly, the classical attack framework represents a special case where $t^A = 1$. For each time $t \leq t^A$, the learner acts in a corruption-free scenario, as if the attacker is absent. For $t > t^A$ the attacker starts to inject corruption c_t at each round until the end of the horizon T to fool the learner into selecting the target arm τ instead of the optimal arm. Figure 1 provides a graphical representation of the setting.

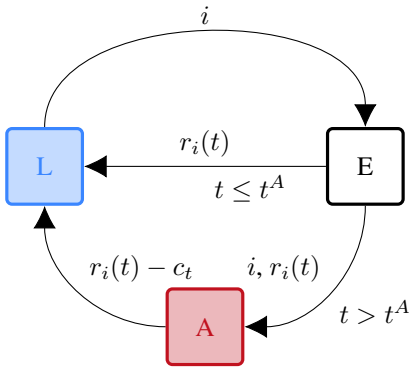


Figure 1: Delayed adversarial attack framework. L represent the learner, E the Environment, and A the attacker.

Formally, we divide the horizon T into two phases. A pre-corruption phase with $T_1 = t^A$ rounds and a post-corruption phase with T_2 rounds such that $T = T_1 + T_2$. Thus in T_1 the learner acts as in a classic bandit problem, i.e., with no corruption. In T_2 the problem shifts toward an adversarial attack model. However, at this point the learner has already good estimates of the environment and this makes the task of performing an attack much more challenging.

[2] We assume that τ is the arm with the lowest average reward i.e., $\mu_\tau = \min_{i \in [K]} \mu_i$. This is w.l.o.g. because all the arms with mean reward lower than the target arm can be eliminated since they are played a sub-linear number of times even without the attack.

3.1 Learner and attacker models

In our theoretical analysis, we focus on the specific case of a learner that employs an UCB algorithm. In particular, we use the same implementation for σ^2-sub-Gaussian rewards proposed by Jun et al. [2018] where the selection rule is:

$$I_t = \begin{cases} t, & \text{if } t \leq K \\ \text{argmax}_i \left\{ \hat{\mu}_i(t-1) + 3\sigma \sqrt{\frac{\log T}{N_i(t-1)}} \right\}, & \text{otherwise.} \end{cases} \tag{4}$$

As the attack model, we consider the oracle attack defined in Eq. (3). To simplify the exposition, we work under the assumption that the attacker knows the true means. This assumption can be easily removed replacing the true mean with the empirical one (see, e.g., [Jun et al., 2018]).

Under these assumptions, we propose a theoretical analysis on (i) the *successfulness* and (ii) *profitability* of the attack. In this scenario, it is necessary to redefine the meaning of successfulness. First of all, it is too strict to assume that the attack is successful only if the target is selected $N_\tau(T) = T - o(T)$ times. Indeed, the target arm is played a sublinear number of times in the rounds preceding the attack. Profitability, instead, is a new more fine-grained metric to measure the effectiveness of an attack.

3.2 Successfulness Analysis

Previous literature defined successfulness as the guarantee that the non-target arms are played a sublinear number of times while the attacker pays a logarithmic cost. As we discussed earlier, this condition is too restrictive for our setting. Hence, we need a weaker notion of successfulness. In particular, with the term *successfulness* we refer to a condition that establishes when the attack force the learner to pull the target arm a linear number of times. Formally:

Definition 3.1 (Successfulness)**.** An Adversarial Attack starting at time t^A to a UCB learner with σ^2-sub-Gaussian rewards acting over a horizon of T rounds is successful if the target arm τ is selected at least:

$$N_\tau(T) = \Omega(T) \tag{5}$$

times in high probability or expectation.

Intuitively, this condition is intimately related to fool the learner into "believe" that the target arm is optimal. Similarly, if the attack is unsuccessful, the UCB learner will not "believe" the target arm is optimal and injecting corruption may be useless. Starting to attack at time $t^A = 1$, the feasibility of the attack is almost always possible since, from the very first rounds, the learner receives corrupted observations. However, in our scenario, every non-corrupted sample fortifies the learner's estimates, increasing the corruption required to convince the learner to bet on the target arm. Depending on the starting time t^A, we can have situations where the attack is impossible, i.e., the learner will never believe the target arm τ optimal.

3.3 Profitability

Successfulness is a binary condition that distinguishes between successful and non-successful attacks depending on the number of times the target arm is selected. However, this condition is verified in very different situations as it only requires a linear number of pulls of the target arm. For this reason, we will also use a more fine-grained metric of successfulness called *profitability*.

The rationale behind this metric is quantifying how many times the attacker makes the learner pulls the target arm τ given that the attack is successful. Profitability responds to the question: *"Given that the attack is successful, how many times the learner will select the target arm?"* The profitability corresponds to the quantification of how many times the target arm has been pulled. Formally, it is defined as:

Definition 3.2 (Profitability)**.** The profitability of a successful attack is defined as the number of times the target arm $N_\tau(T)$ is pulled.

In other words, when an attack is successful, the attacker succeeded in making the learner believing that the target arm τ is optimal, fooling the learner into selecting τ a linear number of times $\Omega(T)$. Instead, profitability is the actual measure of the number of times the target arm has been pulled. These two definitions provide a more fine-grained way to quantify an adversarial attack in the proposed setting.

4 Theoretical Analysis

In this section, we provide theoretical guarantees on the successfulness and profitability of the attack considering a UCB learner and an oracle attacker defined in Section 2. In particular, our core contribution is the derivation of a bound on the number of pulls of the target arm $N_\tau(T)$, given that the attack starts at time $t^A \geq 1$. Such result follows from Lemma 4.2, which provides an upper bound on the number of pulls of the optimal arm, and Lemma 4.3, which instead defines an upper bound on the number of pulls of a sub-optimal arm $i \notin \{o, \tau\}$. In the delayed scenario, we must distinguish between the optimal o and a generic arm i. Since the learner experience $O(\log(T))$ regret and has acted free of corruption for T_1 rounds, it may already have a robust estimate of the optimal arm. Indeed, the optimal arm has been played a linear number of times in T_1 before the attack. This is not the case for sub-optimal arms that have been played a logarithmic number of times and have less consolidated estimates. Before moving into the main theorems, similar to Jun et al. [2018] and Liu and Shroff [2019], we derive a bound for the empirical means for each arm for each round. To this extent, define the event $E = \{|\hat{\mu}_i(t) - \mu_i| \leq \beta(N_i(t)) \quad \forall i, \forall t\}$ where given a probability $\delta > 0$ the function $\beta(N)$ is defined as:

$$\beta(N_i(t)) = \sqrt{\frac{\log(\frac{2KT}{\delta})2\sigma^2}{N_i(t)}}. \tag{6}$$

This is slightly different from the radius proposed by Jun et al. [2018] and Liu and Shroff [2019] since in our work we fix the horizon T. Then, we prove that event E holds with high-probability:

Lemma 4.1. *For any $\delta \in (0, 1), \mathbb{P}(E) > 1 - \delta$.*

Although the proof is standard, it is reported in Appendix A for completeness. Thanks to Lemma 4.1, with probability at least $1 - \delta$, we can bound the mean μ_i of arm i in the interval $[\hat{\mu}_i(t) - \beta(N_i(t)), \hat{\mu}_i(t) + \beta(N_i(t))]$. If the attack starts at time $t^A = 1$, the target arm is immediately recognized as optimal since the corruption will fake the observation from the beginning. In our scenario, the learner constructs a corruption-free estimate of each arm for $t < t^A$ rounds, and upon reaching round t^A the UCB learner will select the optimal arm o approximately $N_o(t^A) \approx t^A$ times, while every other arm i will be selected $N_i(t^A) \approx \log(t^A)$. After t^A the corruption starts. In the following, we provide an upper bound on how many times the learner will select any non-target arm before believing the target arm is optimal. This is equivalent to finding an upper bound on the quantities $N_o(t^A \to t)$ and $N_i(t^A \to t)$. The following inequality determines a sufficient condition to guarantee that an arm i is not pulled at a given time t:

$$\hat{\mu}_i^c(t) + \beta_{UCB}(N_i(t)) \leq \hat{\mu}_\tau(t) + \beta_{UCB}(N_\tau(t)), \tag{7}$$

where $\hat{\mu}_i^c(t)$ represents the partial corrupted estimator for arm i (corrupted after t^A) formally defined as:

$$\hat{\mu}_i^c(t) = \hat{\mu}_i(t) - \frac{\sum_{k=t^A}^{t} c_k}{N_i(t^A \to t)}, \tag{8}$$

with $t^A < t$ and with c_k being the corruption crafted by the attacker during the corruption interval. Now, let

$$\eta \coloneqq \sqrt{\frac{\epsilon\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log T\right)}{\Delta_{o,\tau} T_1}}. \tag{9}$$

For the ease of presentation, we use the term η in Eq. (9) that incorporates the confidence radius β defined in Eq. (6), and the confidence radius of UCB. Analyzing Inequality (7), we obtain the following lemma, which formally states the bound for the optimal arm pulls in the corruption phase:

Lemma 4.2. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select the optimal arm in the corruption phase at most:*

$$N_o(t^A \to T) \leq \frac{(\eta + \Delta_{o,\tau})\, T_1}{\epsilon - \eta}. \tag{10}$$

Similarly, we can bound the number of pulls for a sub-optimal (and not-target) arm i. Let

$$\psi_i \coloneqq \sqrt{\frac{\epsilon\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log T\right)}{\Delta_{i,\tau}\log\left(T_1\right)}}. \tag{11}$$

Intuitively, the term ψ_i incorporates the confidence radius β and the confidence radius of UCB for the specific arm i. Formally, we obtain the following lemma:

Lemma 4.3. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack, with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select a generic non-optimal arm i in the corruption phase at most:*

$$N_i(t^A \to T) \leq \frac{(\psi_i + \Delta_{i,\tau}) \log T_1}{\epsilon - \psi_i}. \tag{12}$$

Then, we can state the core theorem to lower bound the number of target arm pulls in the corruption horizon.

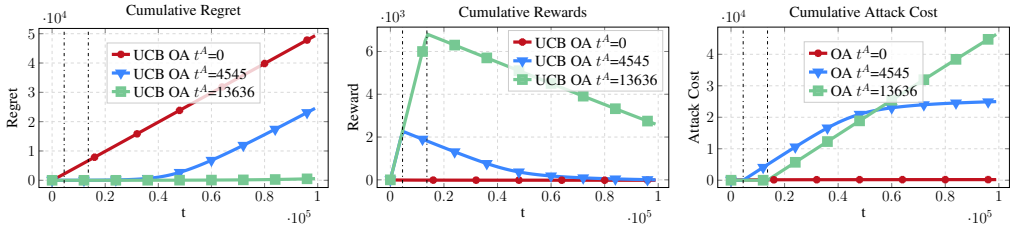
Theorem 4.4. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select the target arm τ in the corruption phase at least:*

$$N_\tau(t^A \to T) \geq T_2 - \frac{\eta + \Delta_{o,\tau}}{\epsilon - \eta} T_1 - \sum_{i \in [K] \setminus \{\tau, o\}} \frac{\psi_i + \Delta_{i,\tau}}{\epsilon - \psi_i} \log T_1.$$

The proof of Theorem 4.4 follows from Lemma 4.2 and Lemma 4.3. The following Corollary 4.5 shows the asymptotic definition of Theorem 4.4:

Corollary 4.5. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select the target arm τ in the corruption phase at least:*

$$N_\tau(t^A \to T) \geq T_2 - \frac{\Delta_{o,\tau}}{\epsilon} T_1 - o(T). \tag{13}$$



(a) (b) (c)

Figure 2: Each figure compares three identical UCB learners, each attacked at a different start time t^A. In particular, we show a comparison of the cumulative regrets in Figure 2a, cumulative rewards in Figure 2b and the attack cost in Figure 2c with a 95% confidence interval over 10 experiments. Each learner is attacked at a different time: from the start ($t^A = 1$) and respectively before and after the successfulness threshold $\alpha^* T$ (the two dotted vertical lines), to show the magnitude of changes in the correspondent metric.

4.1 Successfulness threshold

In this section, we show how the results in the previous section directly imply results on the successfulness of an attack. To this extent, notice that Lemma 4.2 provides the minimum number of rounds $N_o(t^A \to T)$, in the corruption phase, required to change the belief of a UCB learner subject to an oracle attack, for selecting the target arm. In other words, before the learner believes that τ is optimal, there still will be many $N_o(t^A \to T)$ rounds in which the learner will select the optimal arm o. With this in mind, we can have situations where, for some start attack time t^A, even introducing corruption, the learner may never believe the target arm to be optimal, resulting in an unsuccessful attack. As a trivial example, if the attack starts near the horizon's end, the attack cannot select the target arm

a linear number of times. A natural question that arises reasoning about the fact above is whether exists a threshold α^* that identifies the break-even point in the horizon $\alpha^* T$ where any attack starting after $t^A > \alpha^* T$ cannot make the learner pull the target arm τ a linear number of times. Intuitively, the threshold α^*, given $\Delta_{o,\tau}$ and fixed a value for ϵ can discriminate for any starting time t^A if the attack will be successful. Such condition can be derived from Theorem 4.4 and can be expressed in closed form as a parameter $\alpha^*(\Delta_{o,\tau}, \epsilon)$ that depends on the optimal gap $\Delta_{o,\tau}$ and the parameter ϵ. Formally:

Corollary 4.6. *Fixed a constant corruption $\epsilon > 0$. If the attack starts at αT with $\alpha < \alpha^*(\Delta_{o,\tau}, \epsilon)$, a UCB learner will select the target arm τ at least $\Omega(T)$ times with high probability.*

The proof follows from the derivation of the lower bound on $N_\tau(t^A \to T)$. Finally, we prove that our bounds are tight. In particular, with the following theorem, we provide a tight upper bound on the number of target arm pulls.

Theorem 4.7. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption from time αT with $\alpha < \alpha^*$ via oracle attack with $\epsilon > 0$. Then, with high probability, the learner will select the target arm τ in the corruption phase at most:*

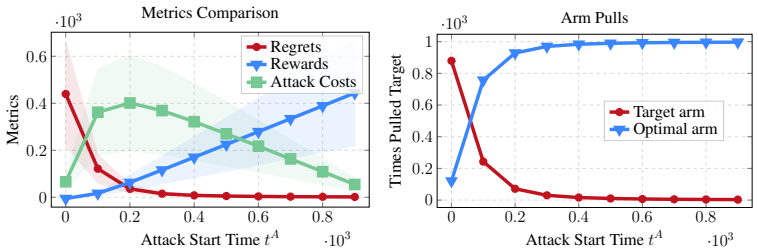
$$N_\tau(t^A \to T) \leq T_2 - \frac{\Delta_{o,\tau}}{\epsilon} T_1 + o(T).$$

Similarly, as a corollary, we can show that if the attack starts from αT with $\alpha > \alpha^*(\Delta_{o,\tau}, \epsilon)$, it is not successful in high probability.

Corollary 4.8. *Fixed a constant corruption $\epsilon > 0$, if the attack starts at αT with $\alpha > \alpha^*(\Delta_{o,\tau}, \epsilon)$, a UCB learner will select the target arm τ at most $o(T)$ times with high probability.*

5 Numerical Experiments

In this section, we provide numerical experiments to support our theoretical claims. In all the experiments, we use the UCB algorithm defined in Equation (4) for the learner and the oracle attack model as attacker defined in Equation (3). We conduct two experiments, one comparing three specific attack times t^A before and after the successfulness threshold defined in Corollary 4.6 and Corollary 4.8. In the second experiment, we show the learner's metrics when the attack can start in every possible round of the horizon T.



(a) (b)

Figure 3: This figure shows the behaviour of a UCB learner, victim of an oracle attack, for each possible attack start time, with a 95% confidence interval on 15 experiments. On the left, Figure 3a shows the value of total *regrets*, total *rewards* and total *cost of the attack* for different times t^A. While on the right, Figure 3b shows the average number of pulls of the target arm τ and the optimal arm o depending on the starting time t^A.

5.1 Comparison between specific starting times

Given the successfulness threshold $\alpha^* = \frac{\epsilon}{\epsilon + \Delta_{o,\tau}}$, we show how the learner behaves when the attack starts in three specific times:

- from time $t^A = 1$, equal to the previous attack framework.
- from time $t^A < \alpha^* T$ where we proved the attack is still successful. We set the attack start time of $t^A = \frac{1}{2}\alpha^* T$.
- from time $t^A > \alpha^* T$ where the attack is not successful. Where we set a value for the start time $t^A = \frac{3}{2}\alpha^* T$

We evaluate a learner with two arms: the optimal arm o with a mean reward set to $\mu_o = \Delta$ with $\Delta > 0$ and a target arm τ with mean reward $\mu_\tau = 0$. We choose $\Delta = 0.5$. Error tolerance δ is set to 0.05, and σ is set to 0.1. The Oracle attacker has the parameter ϵ set to 0.05. As environment parameters, the rewards for each arm i are i.i.d. sampled from a Gaussian distribution $\mathcal{N}(\mu_i, \sigma^2)$. We perform $E = 10$ trials with a horizon $T = 10^5$. For the experiment's reproducibility, we set the random seed to 1234. The results of this experiment are reported in Figure 2. The experiment results show that starting the attack at different times drastically changes the outcome. In particular, focusing on the cumulative regrets of Figure 2a we notice how the attack at time $t^A = 1$ and $t^A < \alpha^* T$, although slightly different, results in a linear regret for the learner, in contrast with the UCB instance attacked at $t^A > \alpha^* T$ which is clearly non-linear. From the perspective of the attacker, starting the attack after the successfulness threshold implies the highest cost in the horizon T, as shown in Figure 2c. For this reason, any attack performed after $\alpha^* T$ is not worth it in terms of cost as the attacker pays the highest price without being capable of fooling the learner into selecting the target arm τ.

5.2 Comparison between each possible attack times

In the second experiment, we run the attack for any possible starting time $t^A \in [T]$. Thus, for any time $t^A \in [T]$ we run an instance of UCB algorithm attacked by an oracle attack starting at t^A. For each t^A we save the sum of the metrics obtained (regrets, rewards and attack cost). Similarly to the previous experiment, we evaluate a learner with two arms $\{o, \tau\}$: the optimal arm mean reward set to $\mu_o = \Delta$ with $\Delta > 0$. Where we choose $\Delta = 0.5$, and the target arm τ has a mean reward of $\mu_\tau = 0$. Error tolerance δ is set to 0.05, and σ is set to 0.1. The Oracle attacker has the parameter ϵ set to 0.05. Again, the rewards for arm i are i.i.d. sampled from a Gaussian distribution $\mathcal{N}(\mu_i, \sigma^2)$. We perform $E = 15$ trials with a horizon $T = 10^3$ for each $t^A \in [1, \dots, 10^3]$. For the experiment's reproducibility, we set the random seed to 1234. Figure 3 shows the results. Fixed a particular t^A, Figure 3a shows the total regret and the total rewards obtained by the learner, as well as the total attack of the attacker. Figure 3b shows how the number of pulls for the optimal o and target arm τ changes depending on the beginning of the corruption. As expected, the target arm pulls degrades gracefully as the beginning of the corruption is delayed.

6 Conclusions

Current state-of-the-art analysis of adversarial attacks in the multi-armed bandits framework assumes the attacker starts to inject corruption at time $t^A = 1$. This assumption is unrealistic in a practical scenario, indeed often leads to unrecoverable results for bandits victims of an attack. We provide a more fine-grained and general framework, namely the delayed adversarial attack model where an attack can start at any time $t^A \geq 1$. We show that results can dramatically change depending on the starting time t^A. In our analysis, we provide a new definition for successfulness that is meaningful even for a delayed starting time of the attack. We provide an upper and lower bound on the number of times that the attacker can fool the learner to select the target arm depending on the starting attack time t^A. Thanks to these results, we define a threshold to discriminate when an attack can be successful depending on t^A and we show the effects via numerical experiments. As a future work, our framework can be further generalized to other algorithms and extended to other classes of bandits. Our paper aims to offer a framework that can represent a standard model to study delayed attacks in the bandit setting. Furthermore, our model of delayed attack might provide new insights on the design of defence strategy. For instance, if the learner experienced a corruption-free phase, employing existing change detection algorithms could be sufficient to detect an attack and stop the algorithm to avoid additional damage.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Rishab Balasubramanian, Jiawei Li, Prasad Tadepalli, Huazheng Wang, Qingyun Wu, and Haoyu Zhao. Adversarial attacks on combinatorial multi-armed bandits, 2024. URL `https://openreview.net/forum?id=vEEWhGjx0M`.

Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-Tolerant Gaussian Process Bandit Optimization, March 2020a.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic Linear Bandits Robust to Adversarial Attacks, October 2020b.

Djallel Bouneffouf, Irina Rish, and Charu C. Aggarwal. Survey on applications of multi-armed and contextual bandits. *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2020.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*, pages 2767–2783. PMLR, 2022.

Tong Chen, Jiqiang Liu, Yingxiao Xiang, Wenjia Niu, Endong Tong, and Zhen Han. Adversarial attack and defense in reinforcement learning-from ai security view. *Cybersecurity*, 2:1–22, 2019.

Jing Dong, Ke Li, Shuai Li, and Baoxiang Wang. Combinatorial bandits under strategic manipulations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 219–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498413. URL `https://doi.org/10.1145/3488560.3498413`.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine Learning in Health Care*, 2018.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial Attacks on Linear Contextual Bandits, October 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

Ziwei Guan, Kaiyi Ji, Donald J. Bucci Jr, Timothy Y. Hu, Joseph Palombo, Michael Liston, and Yingbin Liang. Robust Stochastic Bandit Algorithms under Probabilistic Unbounded Adversarial Attack, February 2020.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better Algorithms for Stochastic Bandits with Adversarial Corruptions, March 2019.

Eric Han and Jonathan Scarlett. Adversarial Attacks on Gaussian Process Bandits, June 2022.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science*, 4, 2017.

Matthew J. Inkawhich, Yiran Chen, and Hai Helen Li. Snooping attacks on deep reinforcement learning. In *Adaptive Agents and Multi-Agent Systems*, 2019.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4042–4050. PMLR, 09–15 Jun 2019.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions, March 2018.

Yuzhe Ma and Zhijin Zhou. Adversarial Attacks on Adversarial Bandits, January 2023.

Davide Maran, Pierriccardo Olivieri, Francesco Emanuele Stradi, Giuseppe Urso, Nicola Gatti, and Marcello Restelli. Online markov decision processes configuration with continuous decision space. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14315–14322, Mar. 2024. doi: 10.1609/aaai.v38i13.29344. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29344`.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish V. Chowdhary. Robust deep reinforcement learning with adversarial attacks. In *Adaptive Agents and Multi-Agent Systems*, 2017.

Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti. Saving stochastic bandits from poisoning attacks via limited data verification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 8054–8061. AAAI Press, 2022.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *International Joint Conference on Artificial Intelligence*, 2015.

Jianwen Sun, Tianwei Zhang, Xiaofei Xie, L. Ma, Yan Zheng, Kangjie Chen, and Yang Liu. Stealthy and efficient adversarial attacks against deep reinforcement learning. *ArXiv*, abs/2005.07099, 2020.

Huazheng Wang, Haifeng Xu, and Hongning Wang. When Are Linear Stochastic Bandits Attackable?, July 2022.

Yinglun Xu, Bhuvesh Kumar, and Jacob D Abernethy. Observation-free attacks on stochastic bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22550–22561. Curran Associates, Inc., 2021.

Lin Yang, Mohammad Hassan Hajiesmaili, Mohammad Sadegh Talebi, John C.S. Lui, and Wing S. Wong. Adversarial bandits with corruptions. 2021.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan D. Liu, Duane S. Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *arXiv: Learning*, 2020.

Zixin Zhong, Wang Chi Cheung, and Vincent Y. F. Tan. Probabilistic Sequential Shrinking: A Best Arm Identification Algorithm for Stochastic Bandits with Corruptions, June 2021.

Qiang Zhou, Xiaofang Zhang, Jin Xu, and Bin Liang. Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*, 2017.

A Omitted Proofs

Lemma 4.1. *For any $\delta \in (0,1), \mathbb{P}(E) > 1-\delta$.*

Proof. Proving that $P(E) \geq 1-\delta$ is equivalent to prove that $P(E^c) \leq \delta$ where E^c is the complementary event. Now let

$$E^c_{i,t} = \left\{ |\hat{\mu}_i(t) - \mu_i| \leq \beta(N_i(t)) \right\},$$

then we have that:

$$\begin{aligned}
P(E^c) &= P\left(\bigcup_{i=1}^{K} \bigcup_{t=1}^{T} E^c_{i,t} \right) \\
&\leq \sum_{i=1}^{K} \sum_{t=1}^{T} P\left(E^c_{i,t} \right) && (14) \\
&\leq \sum_{i=1}^{K} \sum_{t=1}^{T} 2\exp\left\{ -\frac{N_i(t)\beta(N_i(t))^2}{2\sigma^2} \right\} && (15) \\
&\leq \delta, && (16)
\end{aligned}$$

where in Inequality (14) we applied the Union Bound, in Inequality (15) the Hoeffding Bound and Inequality (16) follows by substituting $\beta(N_i(t))$ defined in Equation (6). □

Lemma 4.2. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select the optimal arm in the corruption phase at most:*

$$N_o(t^A \to T) \leq \frac{(\eta + \Delta_{o,\tau})\, T_1}{\epsilon - \eta}. \tag{10}$$

Proof. Consider a UCB learner, experiencing $O(\log(t))$ regret. Consider an omniscient attacker, meaning that at each round, given that the optimal arm has been selected, she corrupts the amount $c_t = \Delta_{o,\tau} + \epsilon$. Let $t > t^A$ any round after the corruption has began. If

$$\hat{\mu}^c_o(t) + \beta_{UCB}\left(N_o(t)\right) \leq \hat{\mu}_\tau(t) + \beta_{UCB}\left(N_\tau(t)\right), \tag{17}$$

where $\mu^c_o(t)$ is a partial corrupted estimator where the corruption only happens in the interval (t^A, t), holds for the optimal arm o, the learner believes that target arm τ is optimal after a corruption phase (we distinguish between optimal arm o and a generic arm i with $i \neq \tau$). Now, the left hand side of Inequality (17) can be upper bounded by:

$$\hat{\mu}^c_o(t) + \beta_{UCB}\left(N_o(t)\right) \leq \hat{\mu}_o - \frac{cN_o(t^A \to t)}{N_o(t)} + \beta_{UCB}\left(N_o(t)\right),$$

where we have extracted the corruption from the partial corrupted estimator $\hat{\mu}^c_o(t)$. The extraction is possible since in the oracle attack, $\forall t \in [T]$ computes a constant, fixed attack $c_t = c = \Delta_{o,\tau} + \epsilon$. Then we can further upper bounding using the fact that event E holds:

$$\begin{aligned}
&\hat{\mu}_o(t) - \frac{cN_o(t^A \to t)}{N_o(t)} + \beta_{UCB}\left(N_o(t)\right) \\
&\qquad \leq \mu_o + \beta\left(N_o(t)\right) - \frac{cN_o(t^A \to t)}{N_o(t)} + \beta_{UCB}\left(N_o(t)\right) \\
&\qquad = \mu_\tau + \Delta_{o,\tau} + \beta\left(N_o(t)\right) - \frac{cN_o(t^A \to t)}{N_o(t)} + \beta_{UCB}\left(N_o(t)\right) \\
&\qquad \leq \hat{\mu}_\tau(t) + \beta_{UCB}(N_\tau(t)) + \Delta_{o,\tau} + \beta\left(N_o(t)\right) - \frac{cN_o(t^A \to t)}{N_o(t)} + \beta_{UCB}\left(N_o(t)\right). \qquad (18)
\end{aligned}$$

Then, if we plug Equation (18) in the Inequality (17) we obtain:

$$\Delta_{o,\tau} + \beta\left(N_o(t)\right) + \beta_{UCB}\left(N_o(t)\right) - \frac{cN_o(t^A \to t)}{N_o(t)} \leq 0. \tag{19}$$

Now notice that $N_o(t)$, with $t > t^A$ can be rewritten as $N_o(T_1) + N_o(t^A \to t)$. Moreover, since $\beta(N)$ is decreasing in the number of arm pulls we can further upper bound Inequality (19) as:

$$\begin{aligned}
&\Delta_{o,\tau} + \beta\left(N_o(t)\right) + \beta_{UCB}\left(N_o(t)\right) - \frac{cN_o(t^A \to t)}{N_o(t)} \\
&\qquad \leq \Delta_{o,\tau} + \beta\left(N_o(t^A \to t)\right) + \beta_{UCB}\left(N_o(t^A \to t)\right) - \frac{cN_o(t^A \to t)}{N_o(t)} \\
&\qquad = \Delta_{o,\tau} + \sqrt{\frac{\log(\frac{2KT}{\delta})2\sigma^2}{N_o(t^A \to t)}} + 3\sigma\sqrt{\frac{\ln(t)}{N_o(t^A \to t)}} - \frac{cN_o(t^A \to t)}{N_o(t)} \\
&\qquad \leq \Delta_{o,\tau} + \sqrt{\frac{\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log t\right)}{N_o(t^A \to t)}} - \frac{cN_o(t^A \to t)}{N_o(t)}.
\end{aligned}$$

Now assume that:

$$N_o(t^A \to t) \geq \frac{\Delta_{o,\tau}}{\epsilon} T_1 \tag{20}$$

Exploiting Equation (20), we have that:

$$\Delta_{o,\tau} + \sqrt{\frac{\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log t\right)}{N_o(t^A \to t)}} - \frac{cN_o(t^A \to t)}{N_o(t)} \leq 0,$$

if

$$\Delta_{o,\tau} + \sqrt{\frac{\epsilon\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log T\right)}{\Delta_{o,\tau} T_1}} - \frac{cN_o(t^A \to t)}{N_o(t)} \leq 0, \tag{21}$$

where we use $t \leq T$. Recalling that

$$\eta \coloneqq \sqrt{\frac{\epsilon\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log T\right)}{\Delta_{o,\tau} T_1}},$$

and plugging it in Inequality (21) we obtain:

$$\begin{aligned}
\Delta_{o,\tau} + \eta - \frac{cN_o(t^A \to t)}{N_o(t)} &= \Delta_{o,\tau} + \eta - \frac{cN_o(t^A \to t)}{N_o(T_1) + N_o(t^A \to t)} \\
&\leq \Delta_{o,\tau} + \eta - \frac{cN_o(t^A \to t)}{T_1 + N_o(t^A \to t)} \\
&= \Delta_{o,\tau} + \eta - \frac{\left(\Delta_{o,\tau} + \epsilon\right) N_o(t^A \to t)}{T_1 + N_o(t^A \to t)},
\end{aligned} \tag{22}$$

where in Inequality (22), we upper bound the number of optimal arm pulls to $N_o(T_1) \approx T_1$. Finally, solving for $N_o(t^A \to t)$ the following inequality:

$$\Delta_{o,\tau} + \eta - \frac{\left(\Delta_{o,\tau} + \epsilon\right) N_o(t^A \to t)}{T_1 + N_o(t^A \to t)} \leq 0,$$

we obtain the following result:

$$N_o(t^A \to t) \geq \frac{(\eta + \Delta_{o,\tau})T_1}{\epsilon - \eta}. \tag{23}$$

□

Lemma 4.3. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack, with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select a generic non-optimal arm i in the corruption phase at most:*

$$N_i(t^A \to T) \leq \frac{\left(\psi_i + \Delta_{i,\tau}\right)\log T_1}{\epsilon - \psi_i}. \tag{12}$$

Proof. This proof follows similar steps of the proof for the Lemma 4.2. However, there are different assumptions. Again we consider a UCB learner, experiencing $O(\log(t))$ regret. Consider an omniscient attacker, meaning that at each round, given that a generic arm i has been pulled, she corrupts the amount $c_t = \Delta_{i,\tau} + \epsilon$. Let $t > t^A$ any round after the corruption has began. If the following Inequality (24) holds for the generic arm i, the learner believes that target arm τ is better than generic arm i after a corruption phase (here we distinguish between optimal arm o and a generic arm i with $i \notin \{o, \tau\}$).

$$\hat{\mu}_i^c(t) + \beta_{UCB}(N_i(t)) \leq \hat{\mu}_\tau(t) + \beta_{UCB}(N_\tau(t)), \tag{24}$$

where $\mu_i^c(t)$ is a partial corrupted estimator where the corruption only happens in the interval (t^A, t). Now, the left hand side of Inequality (24) can be upper bounded by:

$$\hat{\mu}_i^c(t) + \beta_{UCB}(N_i(t)) \leq \hat{\mu}_i - \frac{cN_i(t^A \to t)}{N_i(t)} + \beta_{UCB}(N_i(t)),$$

where we have extracted the corruption from the partial corrupted estimator $\hat{\mu}_i^c(t)$. The extraction is possible since in the oracle attack, $\forall t \in [T]$ computes a constant, fixed attack $c_t = c = \Delta_{i,\tau} + \epsilon$. Then we can further upper bounding using the fact that event E holds:

$$\begin{aligned}
&\hat{\mu}_i(t) - \frac{cN_i(t^A \to t)}{N_i(t)} + \beta_{UCB}(N_i(t)) \\
&\qquad \leq \mu_i + \beta(N_i(t)) - \frac{cN_i(t^A \to t)}{N_i(t)} + \beta_{UCB}(N_i(t)) \\
&\qquad = \mu_\tau + \Delta_{i,\tau} + \beta(N_i(t)) - \frac{cN_i(t^A \to t)}{N_i(t)} + \beta_{UCB}(N_i(t)) \\
&\qquad \leq \hat{\mu}_\tau(t) + \beta_{UCB}(N_\tau(t)) + \Delta_{i,\tau} + \beta(N_i(t)) - \frac{cN_i(t^A \to t)}{N_i(t)} + \beta_{UCB}(N_i(t)).
\end{aligned} \tag{25}$$

Then, if we plug back Equation (25) in the Inequality (24) we obtain:

$$\Delta_{i,\tau} + \beta(N_i(t)) + \beta_{UCB}(N_i(t)) - \frac{cN_i(t^A \to t)}{N_i(t)} \leq 0 \tag{26}$$

Now notice that $N_i(t)$, with $t > t^A$ can be rewritten as $N_i(T_1) + N_i(t^A \to t)$. Moreover, since $\beta(N)$ is decreasing in the number of arm pulls we can further upper bound Inequality (26) as:

$$\begin{aligned}
&\Delta_{i,\tau} + \beta(N_i(t)) + \beta_{UCB}(N_i(t)) - \frac{cN_i(t^A \to t)}{N_i(t)} \\
&\qquad \leq \Delta_{i,\tau} + \beta\left(N_i(t^A \to t)\right) + \beta_{UCB}\left(N_i(t^A \to t)\right) - \frac{cN_i(t^A \to t)}{N_i(t)} \\
&\qquad = \Delta_{i,\tau} + \sqrt{\frac{\log(\frac{2KT}{\delta})2\sigma^2}{N_i(t^A \to t)}} + 3\sigma\sqrt{\frac{\ln(t)}{N_i(t^A \to t)}} - \frac{cN_i(t^A \to t)}{N_i(t)} \\
&\qquad \leq \Delta_{i,\tau} + \sqrt{\frac{\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log t\right)}{N_i(t^A \to t)}} - \frac{cN_i(t^A \to t)}{N_i(t)},
\end{aligned}$$

Now assume that:

$$N_i(t^A \to t) \geq \frac{\Delta_{i,\tau}}{\epsilon}\log(T_1) \tag{27}$$

Exploiting Equation (27), we have that

$$\Delta_{i,\tau} + \sqrt{\frac{\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log t\right)}{N_i(t^A \to t)}} - \frac{cN_i(t^A \to t)}{N_i(t)} \leq 0$$

if

$$\Delta_{i,\tau} + \sqrt{\frac{\epsilon\sigma^2\left(2\log\left(\frac{2KT}{\delta}\right) + 9\log T\right)}{\Delta_{i,\tau}\log(T_1)}} - \frac{cN_i(t^A \to t)}{N_i(t)} \leq 0, \tag{28}$$

where we use $t \leq T$. Recalling that

$$\psi_i \coloneqq \sqrt{\frac{\epsilon\sigma^2 \left(2\log\left(\frac{2KT}{\delta}\right) + 9\log T\right)}{\Delta_{i,\tau}\log\left(T_1\right)}},$$

and plug it in Inequality (28) we obtain

$$\begin{aligned}
\Delta_{i,\tau} + \psi_i - \frac{cN_i(t^A \to t)}{N_i(t)} &= \Delta_{i,\tau} + \psi_i - \frac{cN_i(t^A \to t)}{N_i(T_1) + N_i(t^A \to t)} \\
&\leq \Delta_{i,\tau} + \psi_i - \frac{cN_i(t^A \to t)}{\log T_1 + N_i(t^A \to t)} \\
&= \Delta_{i,\tau} + \psi_i - \frac{\left(\Delta_{i,\tau} + \epsilon\right) N_i(t^A \to t)}{\log T_1 + N_i(t^A \to t)},
\end{aligned} \tag{29}$$

where in Inequality (29), we upper bound the number of a generic arm pulls to $N_i(T_1) \approx \log\left(T_1\right)$. Finally, solving the following Inequality for $N_i(t^A \to t)$

$$\Delta_{i,\tau} + \psi - \frac{\left(\Delta_{i,\tau} + \epsilon\right) N_i(t^A \to t)}{\log T_1 + N_i(t^A \to t)} \leq 0,$$

we obtain the following result

$$N_i(t^A \to t) \geq \frac{\left(\psi_i + \Delta_{i,\tau}\right)}{\epsilon - \psi_i} \log T_1. \tag{30}$$

□

Theorem 4.4. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select the target arm τ in the corruption phase at least:*

$$N_\tau(t^A \to T) \geq T_2 - \frac{\eta + \Delta_{o,\tau}}{\epsilon - \eta} T_1 - \sum_{i \in [K] \setminus \{\tau, o\}} \frac{\psi_i + \Delta_{i,\tau}}{\epsilon - \psi_i} \log T_1.$$

Proof. The proof follows from the application of Lemma 4.2 and Lemma 4.3. The number of pulls of the target arm τ in the corruption phase can be defined as:

$$\begin{aligned}
N_\tau(t^A \to T) &= T_2 - \sum_{i \in [K] \setminus \{\tau\}} N_i(t^A \to T) \\
&= T_2 - N_o(t^A \to T) - \sum_{i \in [K] \setminus \{\tau, o\}} N_i(t^A \to T) \\
&\geq T_2 - \frac{\eta + \Delta_{o,\tau}}{\epsilon - \eta} T_1 - \sum_{i \in [K] \setminus \{\tau, o\}} \frac{\psi_i + \Delta_{i,\tau}}{\epsilon - \psi_i} \log T_1
\end{aligned} \tag{31}$$

□

Corollary 4.5. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption via oracle attack with $\epsilon > 0$ for the remaining T_2 rounds. Then, with high probability, the learner will select the target arm τ in the corruption phase at least:*

$$N_\tau(t^A \to T) \geq T_2 - \frac{\Delta_{o,\tau}}{\epsilon} T_1 - o(T). \tag{13}$$

Proof. Consider the result from Theorem 4.4

$$\begin{aligned}
N_\tau(t^A \to T) &\geq T_2 - \frac{\eta + \Delta_{o,\tau}}{\epsilon - \eta} T_1 - \sum_{i \in [K] \setminus \{\tau, o\}} \frac{\psi_i + \Delta_{i,\tau}}{\epsilon - \psi_i} \log T_1 \\
&\geq T_2 - \frac{\eta + \Delta_{o,\tau}}{\epsilon - \eta} T_1 - o(T).
\end{aligned} \tag{32}$$

Inequality (32) follows from the fact that the term regarding a generic arm $i \in [K] \setminus \{o, \tau\}$, will be selected at most a sub-linear number of times both in pre-corruption and in the corruption phase. This is due to the UCB learner experiencing $O(\log T)$ regret. Furthermore, the term $\frac{\eta+\Delta_{o,\tau}}{\epsilon-\eta}$ in Inequality (32) can be divided in $\frac{\Delta_{o,\tau}}{\epsilon} + \frac{\eta(\epsilon-\Delta_{o,\tau})}{\epsilon(\epsilon+\eta)}$ to obtain:

$$\begin{aligned} T_2 - \frac{\eta + \Delta_{o,\tau}}{\epsilon - \eta} T_1 - o(T) &= T_2 - \left(\frac{\Delta_{o,\tau}}{\epsilon} + \frac{\eta(\epsilon - \Delta_{o,\tau})}{\epsilon(\epsilon + \eta)} \right) T_1 - o(T) \\ &= T_2 - \frac{\Delta_{o,\tau}}{\epsilon} T_1 - \frac{\eta(\epsilon - \Delta_{o,\tau})}{\epsilon(\epsilon + \eta)} T_1 - o(T) \\ &\geq T_2 - \frac{\Delta_{o,\tau}}{\epsilon} T_1 - o(T) \end{aligned} \tag{33}$$

□

Corollary 4.6. *Fixed a constant corruption $\epsilon > 0$. If the attack starts at αT with $\alpha < \alpha^*(\Delta_{o,\tau}, \epsilon)$, a UCB learner will select the target arm τ at least $\Omega(T)$ times with high probability.*

Proof. From results obtained by Theorem 4.4 and Corollary 4.5 we know that:

$$\begin{aligned} N_\tau(t^A \to T) &\geq T_2 - \frac{\Delta_{o,\tau}}{\epsilon} T_1 - o(T) \\ &= (1 - \alpha)T - \frac{\Delta_{o,\tau}}{\epsilon} \alpha T - o(T), \end{aligned} \tag{34}$$

where Equation (34) derives from $T_1 + T_2 = \alpha T + (1 - \alpha)T = T$. Now let $\alpha = \alpha^* - \delta$ where $\alpha^* = \frac{\epsilon}{\epsilon + \Delta_{o,\tau}}$ we can rewrite Equation (34) in:

$$\begin{aligned} (1 - \alpha)T - \frac{\Delta_{o,\tau}}{\epsilon} \alpha T + o(T) &= (1 - \alpha^* + \delta)T - \frac{\Delta_{o,\tau}}{\epsilon} (\alpha^* - \delta) T - o(T) \\ &= \delta T + (1 - \alpha^*)T - \frac{\Delta_{o,\tau}}{\epsilon} (\alpha^* - \delta)T - o(T) \\ &= \delta T + (1 - \alpha^* - \frac{\Delta_{o,\tau}}{\epsilon} \alpha^*)T + \frac{\Delta_{o,\tau}}{\epsilon} \delta T - o(T) \qquad (35) \\ &= \delta T + \frac{\Delta_{o,\tau}}{\epsilon} \delta T - o(T) \\ &= \frac{\epsilon + \Delta_{o,\tau}}{\epsilon} \delta T - o(T) \qquad (36) \end{aligned}$$

The middle term in Equation (35) is exactly 0 thus we obtain Equation (36) as final result. Finally:

$$\begin{aligned} N_\tau(t^A \to T) &\geq \frac{C}{\epsilon} \delta T - o(T) \\ &\geq \Omega(T), \end{aligned}$$

which concludes the proof. □

Theorem 4.7. *Suppose a* UCB *learner acts in a non-corrupted scenario for T_1 rounds. Suppose a strong attacker, knowing the true means, injects corruption from time αT with $\alpha < \alpha^*$ via oracle attack with $\epsilon > 0$. Then, with high probability, the learner will select the target arm τ in the corruption phase at most:*

$$N_\tau(t^A \to T) \leq T_2 - \frac{\Delta_{o,\tau}}{\epsilon} T_1 + o(T).$$

Proof. Consider a bandit instance in which we have only two arms, an optimal arm o and the target arm τ, with true means $\mu_o = 1$ and $\mu_\tau = 1 - \Delta$ respectively. Let $t > t^A$ a generic round t after corruption has began. We want to prove that:

$$N_\tau(t^A \to T) \leq T_2 - \frac{\Delta_{o,\tau} T_1}{\epsilon} + \gamma, \tag{37}$$

where γ is a sub-linear term. To prove Inequality (37), we proceed by contradiction. Consider Inequality (37) false, that is

$$N_\tau(t^A \to T) > T_2 - \frac{\Delta_{o,\tau} T_1}{\epsilon} + \gamma. \tag{38}$$

If Inequality (38) is true, it means that exist a round $t^A < t' < t$ where

$$N_\tau(t^A \to t') \geq T_2 - \frac{\Delta_{o,\tau} T_1}{\epsilon} + \gamma - 1, \tag{39}$$

and the learner selected the arm τ, formally:

$$\mu_o^c + \beta_{UCB}(N_o(t')) \leq \mu_\tau + \beta_{UCB}(N_\tau(t')). \tag{40}$$

Now, to prove that Inequality (38) is a contradiction we need to prove that Inequality (40) is false. Proving Inequality (40) false is equivalent to prove true its contrary, formally:

$$\mu_o^c + \beta_{UCB}(N_o(t')) \geq \mu_\tau + \beta_{UCB}(N_\tau(t')). \tag{41}$$

Finally, to prove Inequality (37) we now reduced to prove Inequality (41) true. Since the instance is defined with only two arms:

$$\begin{aligned} N_o(t^A \to t') &= T_2 - N_\tau(t^A \to t') \\ &\leq \frac{\Delta_{o,\tau} T_1}{\epsilon} - \gamma + 1. \end{aligned}$$

Then, we proceed by lower bounding the left hand side of Inequality (41) obtaining:

$$\begin{aligned} \mu_o^c + \beta_{UCB}(N_o(t)) &\geq \mu_o^c \\ &= \mu_o - \frac{(\epsilon + \Delta_{o,\tau}) N_o(t^A \to t')}{N_o(t')} \\ &= \mu_o - \frac{(\epsilon + \Delta_{o,\tau}) \left(\frac{\Delta_{o,\tau} T_1}{\epsilon} - \gamma + 1 \right)}{T_1 + \frac{\Delta_{o,\tau} T_1}{\epsilon} - \gamma + 1} \\ &\geq \mu_o - \frac{(\epsilon + \Delta_{o,\tau}) \left(\frac{\Delta_{o,\tau} T_1}{\epsilon} + 1 \right)}{T_1 + \frac{\Delta_{o,\tau} T_1}{\epsilon} - \gamma} \end{aligned} \tag{42}$$

Now the right most term in Inequality (42) can be rewritten as:

$$\frac{(\epsilon + \Delta_{o,\tau}) \left(\frac{\Delta_{o,\tau} T_1}{\epsilon} + 1 \right)}{T_1 + \frac{\Delta_{o,\tau} T_1}{\epsilon} - \gamma} = \Delta_{o,\tau} + \frac{\epsilon + \Delta_{o,\tau} + \Delta_{o,\tau}\gamma}{T_1 + \frac{\Delta_{o,\tau} T_1}{\epsilon}},$$

from which we obtain:

$$\mu_o - \Delta_{o,\tau} + \frac{\epsilon + \Delta_{o,\tau} + \Delta_{o,\tau}\gamma}{T_1 + \frac{\Delta_{o,\tau} T_1}{\epsilon}}. \tag{43}$$

Then we upper bounding the right hand side of Inequality (38) obtaining:

$$\begin{aligned} \mu_\tau + \beta_{UCB}(N_\tau(t)) &\leq \mu_\tau + 3\sigma \sqrt{\frac{\log T}{N_\tau(T_1) + N_\tau(t^A \to t')}} \\ &\leq \mu_\tau + 3\sigma \sqrt{\frac{\log T}{N_\tau(t^A \to t')}} \\ &\leq \mu_\tau + 3\sigma \sqrt{\frac{\log T}{T_2 - \frac{\Delta_{o,\tau} T_1}{\epsilon} + \gamma - 1}} \\ &\leq \mu_\tau + 3\sigma \sqrt{\frac{\log T}{T_2 - \frac{\Delta_{o,\tau} T_1}{\epsilon} - 1}}, \end{aligned}$$

Finally, we obtain:

$$\mu_o - \Delta_{o,\tau} + \frac{\epsilon + \Delta_{o,\tau} + \Delta_{o,\tau}\gamma}{T_1 + \frac{\Delta_{o,\tau}T_1}{\epsilon}} \geq \mu_\tau + 3\sigma\sqrt{\frac{\log T}{T_2 - \frac{\Delta_{o,\tau}T_1}{\epsilon} - 1}} \tag{44}$$

Thus, Inequality (41) is true for $\gamma \geq \frac{T_1}{\Delta_{o,\tau}}\left(1 - \frac{\Delta}{\epsilon}\right) 3\sigma\sqrt{\frac{\log T}{T_2 - \frac{\Delta_{o,\tau}T_1}{\epsilon} - 1}} - \frac{\epsilon}{\Delta_{o,\tau}}$ resulting in a contradiction. □

B Experiments

In this section, we provide minor details about the experiments omitted in the main paper.

Experiments details

- Experiment were conducted using `python 3.11.6`
- CPU: `Apple M1`
- RAM: `16 GB`
- Operating System: `macOS 14.2.1`
- System Type: `64 bit`