Practical Adversarial Attacks on Stochastic Bandits via Fake Data Injection

Qirun Zeng[1] **Eric He**[2] **Richard Hoffmann**[2] **Xuchuang Wang**[3] **Jinhang Zuo**[1*]

[1]City University of Hong Kong, Hong Kong SAR, China
[2]California Institute of Technology, Pasadena, CA, USA
[3]Hong Kong Baptist University, Hong Kong SAR, China

qirun.zeng@my.cityu.edu.hk {ehe, rhoffman}@caltech.edu
xuchuangw@hkbu.edu.hk jinhang.zuo@cityu.edu.hk

Abstract

Adversarial attacks on stochastic bandits have traditionally relied on some unrealistic assumptions, such as per-round reward manipulation and unbounded perturbations, limiting their relevance to real-world systems. We propose a more practical threat model, Fake Data Injection, which reflects realistic adversarial constraints: the attacker can inject only a limited number of bounded fake feedback samples into the learner's history, simulating legitimate interactions. We design effective attack strategies under this model, explicitly addressing both magnitude constraints (on reward values) and temporal constraints (on when and how often data can be injected). Our theoretical analysis shows that these attacks can mislead a class of bandit algorithms into selecting a target arm in nearly all rounds while incurring only sublinear attack cost. Experiments on synthetic and real-world datasets validate the effectiveness of our strategies, revealing vulnerabilities in stochastic bandit algorithms under practical adversarial scenarios.

1 Introduction

Multi-armed bandit (MAB) algorithms are widely used in online decision-making systems for their ability to balance exploration and exploitation using partial feedback. They form the backbone of many interactive applications, including personalized recommendation [Li et al., 2010, Zeng et al., 2026b, Zou et al., 2026], online advertising [Chen et al., 2016], clinical trials [Villar et al., 2015], and adaptive routing [Li et al., 2016]. As these algorithms are increasingly deployed in high-stakes, user-facing systems, growing concerns have emerged regarding their vulnerability to adversarial manipulation. A growing body of work [Jun et al., 2018, Liu and Shroff, 2019, Zuo et al., 2023, Zuo, 2024] has shown that MAB algorithms are vulnerable to adversarial attacks, where an attacker subtly perturbs the observed rewards to mislead the learning process. Remarkably, even with limited intervention, the attacker can steer the learner toward selecting a targeted but suboptimal arm in the vast majority of rounds.

However, prior works on adversarial attacks against stochastic bandits [Jun et al., 2018, Liu and Shroff, 2019, Zuo, 2024] almost exclusively adopt a *feedback-perturbation* threat model, where the attacker observes the learner's chosen arm and the corresponding reward in each round, then arbitrarily modifies that reward before it is revealed to the learner. This assumption effectively grants the attacker *full control of the feedback channel at every timestep*, a level of adversarial power that is unrealistic in real-world systems. For example, in a restaurant recommendation platform, no adversary can continuously overwrite genuine user ratings submitted by thousands of real customers. At most,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

they can register fake accounts and inject fabricated reviews. Likewise, in online advertising or click-through prediction, practical attacks take the form of click fraud or synthetic interactions, which add fake feedback rather than altering genuine user data. In other words, prior models *overestimate* the attacker's per-round control, while simultaneously *underestimating* their ability to inject feedback across arbitrary arms.

These mismatches motivate our shift to a more realistic and constrained threat model, *Fake Data Injection*. Instead of modifying genuine feedback, the attacker influences the learner indirectly by injecting a limited number of fabricated (arm, reward) pairs into its interaction history. These fake samples must conform to valid feedback ranges (e.g., binary clicks or 1–5 star ratings), and their injection is subject to constraints such as system-level detection or resource limits. The learner processes these fake interactions indistinguishably from real ones, updating estimates, counts, and decision logic accordingly.

This new threat model captures practical attack surfaces overlooked by previous works. It removes the unrealistic assumption of per-round reward manipulation, enables feedback injection on arbitrary arms, and respects the bounded nature of real-world feedback. At the same time, it introduces new algorithmic challenges that cannot be addressed by existing techniques. Unlike the standard model where the attacker perturbs rewards in real time, fake data injection raises fundamental questions about *when*, *how strongly*, and *how frequently* to inject samples to effectively influence the learner. The attacker must decide: (i) how many fake samples are required to suppress the selections of a non-target arm, (ii) how to achieve this using only bounded reward values, and (iii) how to distribute injections over time when batch size or injection frequency is constrained. These challenges call for new analytical tools and attack strategies that explicitly account for both *magnitude* constraints (on reward values) and *temporal* constraints (on when and how often data can be injected).

1.1 Our Contributions

We develop a new framework for adversarial attacks on stochastic bandits under realistic fake-data constraints. Our main contributions are summarized as follows:

- We introduce *Fake Data Injection*, a new threat model in which the attacker cannot overwrite genuine feedback and must instead inject fabricated but valid $(\text{arm}, \text{reward})$ records. This model explicitly captures practical constraints on both the magnitude of injected rewards and the timing/frequency of injections, resulting in a more realistic attack setting.
- We develop the *suppression principle* for fake data injection attacks, showing that once a non-target arm becomes sufficiently suppressed, the learner's own exploration mechanism prevents further exploration of that arm. For UCB, we formalize this through an exponential suppression lemma and develop the *Simultaneous Injection (SI)* attack, which achieves successful attacks using only bounded fake injections with sublinear cost.
- We further study temporally constrained attacks where the attacker can inject only limited batches over time. We propose *Periodic Bounded Injection (PBI)*, which maintains suppression through periodic injections and achieves successful attacks with sublinear cost under both temporal and magnitude constraints.
- We extend the suppression analysis beyond UCB through a general *suppression-bounded* framework for a class of bandit algorithms whose exploration of persistently suppressed arms is bounded, including Thompson Sampling (TS) and ϵ-greedy.
- We validate our methods on both synthetic and real-world datasets, demonstrating that even sparse, bounded fake data can significantly bias bandit learners in practice.

Our work bridges theoretical attack models for stochastic bandits and practical data-driven attacks in real systems by formalizing fake data injection under bounded and limited injections. Further analyses for TS and ϵ-greedy are provided in Appendices B and C, respectively.

1.2 Related Work

There is a growing body of work on adversarial attacks against bandit algorithms [Jun et al., 2018, Liu and Shroff, 2019, Garcelon et al., 2020, Ma and Zhou, 2023, Wang et al., 2022]. Jun et al. [2018] showed that with sublinear cost, an attacker can steer UCB and ϵ-greedy toward suboptimal

arms. Building on this, Liu and Shroff [2019] extended the paradigm to unknown algorithms and contextual bandits, while Garcelon et al. [2020] studied perturbations to both contexts and rewards in linear bandits. More recently, Zuo [2024] introduced optimal attack strategies against UCB and TS, establishing matching lower bounds. Most of these works assume a strong *feedback-perturbation* model with direct reward modifications each round [Jun et al., 2018, Liu and Shroff, 2019, Garcelon et al., 2020, Zuo, 2024], or bounded variants that still allow continuous interference [Xu et al., 2021, Wang et al., 2024, Zuo et al., 2023]. Zeng et al. [2026a] shows that attack can succeed even when the target arm's mean reward lies exactly at the lower endpoint of the reward range.

Another line of research has focused on defense algorithms against adversarial attacks and corruptions, including robust MAB [Lykouris et al., 2018, Gupta et al., 2019, Liu and Shroff, 2019, Mukherjee et al., 2021], robust linear bandits [Bogunovic et al., 2021, He et al., 2022], and combinatorial bandits [Xu and Li, 2021]. However, the corruption-based threat models considered in these works are typically weaker than those studied in the attack literature, since the adversary is not allowed to observe the learner's actions before corrupting the feedback.

We differ from both lines by introducing the Fake Data Injection threat model, where adversaries cannot tamper with genuine rewards but may inject fabricated (arm, reward) pairs. This setting reflects realistic constraints in recommendation or advertising platforms, where fake users can be created but authentic feedback remains immutable. Beyond bandits, it parallels data poisoning in broader ML systems (recommenders, ads, A/B testing, RL), and our tools (e.g., the Exponential Suppression Lemma) provide general insights into vulnerabilities of sequential decision-making.

2 Preliminaries

2.1 Stochastic Bandits

We consider the standard stochastic multi-armed bandit setting with the arm set $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$, where each arm $i \in \mathcal{K}$ is associated with an unknown reward distribution with mean μ_i. The reward distributions are assumed to be σ^2-sub-Gaussian, with σ^2 known. In each round $t \in \mathcal{T} \triangleq \{1, 2, \ldots, T\}$, the learner selects an arm a_t and receives a reward r_t drawn from the distribution corresponding to arm a_t.

Upper Confidence Bound (UCB). We consider the UCB algorithm as specified in Jun et al. [2018], Zuo [2024], and the prototype is the (α, ψ) algorithm of Bubeck and Cesa-Bianchi [2012, Section 2.2]. In the first K rounds, the learner pulls each arm once to initialize reward estimates. For subsequent rounds $t > K$, the learner selects the arm with the highest UCB index $a_t = \arg\max_{i \in \mathcal{K}} \Lambda_i(t) \triangleq \hat{\mu}_i(t) + 3\sigma\sqrt{\frac{\log t}{N_i(t)}}$, where $\hat{\mu}_i(t)$ is the empirical mean reward of arm i and $N_i(t)$ is the number of times arm i has been selected up to round t.

2.2 Previous Threat Model and Limitations

Prior work on adversarial robustness in stochastic bandits broadly falls into two categories: corruption-robust approaches (e.g., Kapoor et al. [2019], Gupta et al. [2019]) and adversarial attack models (e.g., Jun et al. [2018], Zuo et al. [2023], Zuo [2024]). The former assumes a weaker, history-only adversary and focuses on designing defenses against such corruption, whereas the latter studies a stronger, adaptive adversary that is standard in adversarial-attack settings. Our paper fits into this adversarial-attack category. Therefore, we begin by reviewing the standard threat model adopted in prior works on adversarial attack models.

In each round t, the learner selects an arm a_t to play, and the environment generates a pre-attack reward r_t^0 drawn from the underlying distribution of arm a_t. The attacker then observes the tuple (a_t, r_t^0) and decides an attack value α_t. The learner receives only the post-attack reward $r_t = r_t^0 - \alpha_t$. Define the cumulative attack cost as $C(T) = \sum_{t=1}^{T} |\alpha_t|$. The attacker's objective is to manipulate the learner into selecting a specific target arm for a linear number of rounds while incurring only sublinear attack cost. Formally, an attack is successful if it induces the learner to select the target arm in $T - o(T)$ rounds while incurring sublinear attack cost $C(T) = o(T)$. Although many attack strategies have been proposed under the standard threat model, it exhibits several critical limitations when applied to practical settings.

First, the model assumes that the attacker can perturb the environment-generated reward in *every* round. This assumption is often unrealistic in real-world applications such as recommender systems. For example, consider an app that recommends restaurants to users and collects their feedback to improve future recommendations. An attacker may wish to bias the system toward recommending a particular restaurant, but it is infeasible to directly modify the feedback of all real users. In practice, a more common attack strategy is to create fake users who submit fabricated feedback. However, even this is constrained by operational or detection limits: adding fake users in every round is highly impractical. Thus, the assumption of per-round attack capability does not reflect realistic adversarial power. *Second*, the model restricts the attacker to modifying only the reward of the chosen arm in each round. In contrast, fake-user-based attacks offer more flexibility. A fake user can submit feedback on any item (i.e., any arm), regardless of what the learner selected in that round. This means fake data injection enables the attacker to fabricate feedback for arbitrary arms, not just the one currently played by the learner. As a result, the standard model underestimates the attacker's flexibility in practice and overestimates their ability to act at every timestep. *Third*, the threat model assumes that both the pre-attack reward and the attack values are unbounded. In many systems, user feedback is naturally bounded (e.g., binary click signals or discrete rating scores). Allowing arbitrarily large attack values could result in out-of-range or clearly invalid feedback, which would either be filtered out by the system or easily flagged as suspicious. Therefore, attacks that rely on large reward perturbations are incompatible with these bounded-feedback environments.

3 New Threat Model: Fake Data Injection

To address practical limitations of the standard adversarial attack model, we introduce a new and more realistic threat model, which we call the Fake Data Injection Threat Model. This model captures how adversaries behave in real-world systems such as recommendation platforms, where direct manipulation of genuine user feedback is infeasible, and attacks are often carried out by injecting fabricated interactions (e.g., fake users with fake feedback).

In the Fake Data Injection model, the attacker does *not* interfere with the feedback received by the learner during normal interactions. Instead, at round t, it may initiate an injection batch $\mathcal{B}_t$, a finite multiset of fake samples (a^F, r^F) satisfying $a^F \in \mathcal{K}$ and $r^F \in [r_{\min}, r_{\max}]$. The learner processes these samples one by one as valid records: each fake sample updates the corresponding empirical mean, increments the arm's sample count, and advances the round counter by one.

Let $N_i^0(t)$ and $N_i^F(t)$ denote the numbers of genuine and fake samples, respectively, of arm i up to round t. The learner's sample count is $N_i(t) = N_i^0(t) + N_i^F(t)$. The total number of fake samples injected up to round t is $N^F(t) \triangleq \sum_{i\in\mathcal{K}} N_i^F(t)$. Enumerate the fake samples injected by horizon T, in order of arrival, as (a_s^F, r_s^F), $s = 1, \ldots, N^F(T)$. We measure attack cost by the number of injected samples $N^F(T)$ and the cumulative reward deviation

$$C^F(T) \triangleq \sum_{s=1}^{N^F(T)} \left| r_s^F - \mu_{a_s^F} \right|. \tag{1}$$

Here, $N^F(T)$ captures the operational cost of injection, while $C^F(T)$ facilitates comparison with prior bandit-attack literature. We consider an attack *successful* if it induces $T - o(T)$ genuine pulls of the target arm while ensuring both $N^F(T) = o(T)$ and $C^F(T) = o(T)$.

This hybrid mechanism distinguishes our model from both prior online perturbation models and offline poisoning settings. In offline data poisoning (e.g., Liu and Shroff [2019]), the attacker modifies a fixed dataset collected from arms actually played, and the learner is then trained once on this corrupted dataset. In contrast, our attacker can inject arbitrary (i, r) pairs at any time, affecting arms that may never have been pulled. The learner immediately incorporates these fake samples and continues to make decisions in an online, round-by-round fashion. Thus, while our model permits batch-style updates, it remains fundamentally online, bridging the gap between classical reward-perturbation attacks and purely offline poisoning. It resolves several key limitations of the previous threat model:

Limited access manipulation. Unlike the standard model which assumes the attacker can modify the reward in every round, our model reflects the more plausible scenario where the attacker can only inject a limited number of fake interactions. For instance, in a restaurant recommendation app, an

attacker cannot tamper with the feedback from real users but can register a finite number of fake accounts to submit biased reviews. It is unrealistic to assume the attacker can do this in every round without detection or resource exhaustion.

Flexible feedback across arms. The standard model restricts the attacker to modifying the reward for the arm chosen by the learner. In contrast, our model allows the attacker to fabricate data for any arm. This mirrors real-world attacks where fake users can submit reviews or feedback on arbitrary items, not just those recommended to them. For example, an attacker aiming to boost a target restaurant can flood the system with positive feedback for that restaurant, regardless of whether it was actually recommended in a specific round.

Bounded and plausible feedback. In our model, the fake rewards must satisfy $r_{\min} \leq r^F \leq r_{\max}$, consistent with many practical systems that collect binary clicks or scaled ratings (e.g., 1 to 5 stars). This avoids the unrealistic assumption of unbounded reward modifications, where a single large perturbation could dominate the learner's behavior. Our bounded injection design ensures the fake data remains indistinguishable from legitimate interactions.

4 Attack Strategies

The central difficulty in fake data injection is that the attacker cannot erase genuine feedback or assign arbitrary corrupted rewards; it can only add valid samples to the learner's history. We address this constraint through a *suppression principle*: inject enough bounded fake feedback so that each non-target arm's empirical estimate falls below a stable lower benchmark for the target arm. Once this separation is certified, the learner's own exploration rule prevents the suppressed arm from being selected often, turning targeted manipulation into a problem of establishing and maintaining non-target suppression. This perspective is the main technical driver of our attacks. For UCB, we show that suppression can persist for an exponentially long interval, which leads to Simultaneous Injection (SI) and Periodic Bounded Injection (PBI) attacks with sublinear cost under different injection constraints. We then abstract the same idea into a suppression-bounded framework that applies beyond UCB. We label the target arm as K and assume it has the lowest expected reward.[1] We first analyze UCB and then extend the suppression-based argument to a broader class of algorithms.

4.1 The Suppression Lemma

Before presenting our concrete attack strategies, we establish the suppression mechanism used throughout the section.

Lemma 4.1 (Exponential suppression under UCB)**.** *For the UCB algorithm specified in Section 2, let* $\delta \in (0, \frac{1}{2})$*, and define* $\hat{\ell}_K(t) \triangleq \hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma\delta_0$ *with* $\beta(N) \triangleq \sqrt{\frac{2\sigma^2}{N} \log \frac{\pi^2 K N^2}{3\delta}}$*. If*

$$\hat{\mu}_i(t) \leq \hat{\ell}_K(t), \tag{2}$$

then arm i *is not selected again until at least round* $\exp\big(N_i(t)\delta_0^2\big)$*, with probability at least* $1 - \delta$*.*

Remark 4.2. Lemma 4.1 establishes a critical property of our attack strategy: once a non-target arm i has been pulled sufficiently and a properly chosen fake data injection is applied, its UCB index becomes *exponentially suppressed*. More precisely, if the following two conditions are satisfied: (1) $N_i(t) \geq \frac{\log T}{\delta_0^2}$ and (2) $\hat{\mu}_i(t) \leq \hat{\ell}_K(t)$, then arm i will not be selected again until after round T.

This suppression effect is crucial: it guarantees that once the attack is applied to arm i, its influence on the learning process becomes negligible for the remaining rounds. The attacker can thus prevent further exploration of non-target arms using only a *simultaneous (or sequential) injection* per arm, ensuring that the learner increasingly concentrates on the target arm. This mechanism forms the backbone of all our attack strategies.

4.2 Warm-up: Simultaneous Injection

We begin our study of fake data injection attacks by considering a setting in which the attacker is nearly free of discovery and can inject fake data points as needed. We formalize this insight

[1] This is the most challenging target for the attacker; the same arguments extend directly to any other target arm.

Algorithm 1: Simultaneous Injection (SI) on UCB

Input: Attack parameter $\delta_0 > 0$

for $t = 1, 2, \ldots$ **do**
  for *each non-target arm* $i < K$ **do**
    if $N_i(t) = \lceil \log T/\delta_0^2 \rceil$ **then**
      Inject $n_i = \left\lceil \frac{\hat{\mu}_i(t) - \hat{\ell}_K(t)}{\hat{\ell}_K(t) - r_{\min}} \cdot \left\lceil \frac{\log T}{\delta_0^2} \right\rceil \right\rceil$ samples $(i, r_{\min})$;
    end
  end
end

through the Simultaneous Injection Algorithm, a simple yet effective attack strategy against the UCB algorithm, as shown in Algorithm 1.

The attack operates as follows. For each non-target arm i, we wait until it has been pulled $N_i(t) = \lceil \log T/\delta_0^2 \rceil$ times. At this point, we inject a batch of fake samples designed to reduce its empirical mean below a high-probability lower bound of the target arm. The injected samples for non-target arm i ensure that after the attack, the empirical mean of arm i satisfies $\hat{\mu}_i(t + n_i) \leq \hat{\ell}_K(t)$, thus making it unlikely to be selected in future rounds. We further require that the lower bound on bounded fake rewards is not too large, formalized by the assumption below.

Assumption 4.3. $r_{\min} \leq \mu_K - 3\beta(1) - 3\sigma\delta_0$.

This assumption ensures that suppression of non-target arms remains feasible even when injected rewards are bounded from below. In particular, it guarantees that the attacker can reduce the empirical mean of any non-target arm below the lower confidence bound of the target arm, despite the limited manipulation strength. In contrast, most prior works (e.g., Jun et al. [2018], Liu and Shroff [2019], Zuo [2024]) implicitly assume unbounded attacks, under which rewards can be decreased arbitrarily and such a margin condition holds automatically. In practice, Assumption 4.3 can be relaxed to the round-dependent condition $r_{\min} < \hat{\mu}_K(t) - 2\beta(N_K(t)) - \eta$, for any fixed constant $\eta > 0$, which only needs to hold at the specific round t when the injection is carried out. For simplicity of exposition, we assume that Assumption 4.3 holds throughout the subsequent analysis. We now state the full guarantee for SI; its proof is given in Appendix A.2.

Theorem 4.4. *Suppose* $T > 2K$, $\delta \in (0, 1/2)$, *and* $r_{\min} < \mu_K - 3\beta(1) - 3\sigma\delta_0$. *Define* $\Delta_i \triangleq \mu_i - \mu_K$. *With probability at least* $1 - \delta$, *SI forces UCB to select the target arm in at least*

$$T - \mathcal{O}\left(\sum_{i=1}^{K-1} \frac{\mu_i + \beta(\log T/\delta_0^2) - r_{\min}}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \right)$$

rounds. For each non-target arm $i < K$, *the number of fake samples satisfies*

$$N_i^F(T) = \mathcal{O}\left(\frac{\Delta_i + 4\beta(1) + 3\sigma\delta_0}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \right).$$

Moreover, the total injection count and reward-deviation cost satisfy

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \quad \textit{and} \quad C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T).$$

Compared with the attack algorithm under the standard threat model in Jun et al. [2018], the SI Algorithm achieves a similar level of target-arm selection with comparable sublinear attack cost. Notably, the parameter δ_0 controls the trade-off between the number of non-target arm pulls and the attack cost: increasing δ_0 reduces the number of non-target pulls but increases the cost per injection. However, the marginal benefit diminishes once $\delta_0 > \sqrt{\log T}$, beyond which the cost grows without improving effectiveness.

Remark 4.5. Using the minimum reward $r_{\min}$ simplifies the analysis and minimizes the number of injected samples, but identical fake rewards are not necessary. The attacker may instead use varying rewards whose batch average satisfies $\bar{r}^F < \hat{\ell}_K(t)$, replacing $r_{\min}$ by $\bar{r}^F$ when computing the injection size. This allows more natural-looking reward mixtures, although a higher injected mean requires more samples to achieve the same suppression.

Algorithm 2: Periodic Bounded Injection on UCB

Input: Attack parameter δ_0, reward lower bound $r_{\min}$, max batch size f

Initialize $n_i = 0$ for every non-target arm i;
for $t = 1, 2, \ldots$ **do**
  for *each non-target arm* $i < K$ **do**
    if $N_i(t) = \lceil \log T/\delta_0^2 \rceil$ **then**
      $n_i \leftarrow \left\lceil \frac{\hat{\mu}_i(t) - \hat{\ell}_K(t)}{\hat{\ell}_K(t) - r_{\min}} \cdot \left\lceil \frac{\log T}{\delta_0^2} \right\rceil \right\rceil$; $\tau_i \leftarrow t$;
      Calculate R_i according to Equation (3);
    end
    if $N_i^F(t) < n_i$ *and* $\tau_i \leq t$ **then**
      Inject $\min\{n_i - N_i^F(t), f\}$ fake samples $(i, r_{\min})$;
      $\tau_i \leftarrow \tau_i + f + R_i$;
    end
  end
end

4.3 Periodic Bounded Injection

The Simultaneous Injection algorithm above assumes that the attacker can inject all required fake samples within a sequence of rounds. However, this assumption may not hold in practice. For example, in a restaurant recommendation system, injecting a large batch of fake (e.g., low-rating) reviews at once may trigger anomaly detection mechanisms, leading the system to filter or ignore the fake data. In contrast, injecting smaller amounts of fake feedback periodically at a controlled rate can be significantly less suspicious and more effective in practice.

To model this scenario, we introduce a more restrictive and realistic setting where:

1. There exists a lower bound $r_{\min}$ that the fake rewards cannot fall below;
2. Each injection batch contains at most f fake samples per arm;
3. Consecutive injections on the same arm i must be separated by a cooldown period of at least R_i rounds, where R_i depends on the maximum batch size f.

To address this setting, we propose the Periodic Bounded Injection (PBI) algorithm, shown in Algorithm 2. Given a maximum batch size f, the algorithm adaptively schedules periodic injections to suppress the empirical mean of non-target arms while respecting both constraints.

The PBI algorithm distributes the injection of fake samples across multiple rounds rather than injecting them all at once. Once a non-target arm i reaches the designated pull threshold ($\lceil \log T/\delta_0^2 \rceil$), the attacker computes both the total number of fake samples n_i required to suppress the empirical mean of arm i, and a waiting interval R_i, which ensures that the fake samples can be injected periodically without allowing arm i to regain a high UCB index. The notation $\tilde{\mu}_i(t+f)$ represents the estimated value of $\hat{\mu}_i(t+f)$ with f fake data injections starting from round t. At each interval of $R_i + f$ rounds, a batch of at most f fake samples is injected until the total n_i is exhausted. This strategy effectively balances stealthiness and attack *efficacy*, making it robust against detection in practical systems with bounded feedback and rate-limited injection constraints.

Lemma 4.6. *Let $t_i(c) = t + (f + R_i)c$ denote the round before the $(c+1)$-th injection. And let*

$$R_i \triangleq \min_{1 \leq c \leq \lceil n_i/f \rceil} \frac{1}{c} \left[\exp\left(\left(\frac{\hat{\mu}_K(t) - 2\beta(N_K(t)) - \tilde{\mu}_i(t_i(c))}{3\sigma} \right)^2 \cdot (N_i(t) + fc) \right) - t \right] - f. \quad (3)$$

The choice of R_i ensures that once a batch of f fake data samples is injected into non-target arm i, the arm will not be selected again for at least the next R_i rounds.

Proof Sketch. The argument follows a one-shot variant of Lemma 4.1. It suffices to consider the first batch of f injected samples, which induces the weakest suppression. This batch creates a sufficient gap between the UCB index of arm i and that of the target arm K. By the definition of R_i, this gap

persists for at least R_i rounds, during which arm i's UCB index remains below that of arm K. Hence arm i is not selected in this interval. □

Based on our experimental observations, setting $c = 1$ generally leads to the smallest practical value of $R_i(c)$. The derivation of R_i and the PBI proofs are given in Appendix A.3.

Theorem 4.7. *Suppose $T > 2K$, $\delta \in (0, 1/2)$, and $r_{\min} < \mu_K - 3\beta(1) - 3\sigma\delta_0$. Define $\Delta_i \triangleq \mu_i - \mu_K$. With probability at least $1 - \delta$, PBI forces UCB to select the target arm in at least*

$$T - \mathcal{O}\left(\sum_{i=1}^{K-1} \frac{\mu_i + \beta(\log T/\delta_0^2) - r_{\min}}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2}\right)$$

rounds. For each non-target arm $i < K$, the number of fake samples satisfies

$$N_i^F(T) = \mathcal{O}\left(\frac{\Delta_i + 4\beta(1) + 3\sigma\delta_0}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2}\right).$$

Moreover, the total injection count and reward-deviation cost satisfy

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \quad \textit{and} \quad C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T).$$

The suppression idea underlying SI can also be implemented by waiting for a non-target arm to be selected and applying a sufficiently large reward perturbation [Jun et al., 2018, Zuo, 2024]. However, this connection does not directly extend to PBI: splitting injections into bounded batches changes arm counts and UCB indices over time, so the learner may revisit a non-target arm before all injections are completed. Moreover, reward perturbation is restricted to the currently selected arm and cannot freely realize PBI's schedule.

4.4 Extension to Suppression-Bounded Algorithms

We now extend the suppression principle beyond UCB to a broader class of bandit algorithms whose exploration of persistently suppressed arms remains bounded. Intuitively, once an arm becomes empirically dominated by another arm, the algorithm can pull that arm only a bounded number of additional times.

Definition 4.8 (Suppression-bounded algorithms)**.** *Fix a horizon T, a confidence level $\delta \in (0, 1)$. A bandit algorithm $\mathcal{A}$ is said to be* suppression-bounded *if, for any arm i, there exists a deterministic bound $B_i^{\mathcal{A}}(t, T, \delta)$ such that the following implication holds with probability at least $1 - \delta$: if*

$$\hat{\mu}_i(s) < \max_{j \in \mathcal{K}} \hat{\mu}_j(s), \qquad \textit{for all } s \in \{t+1, \ldots, T\},$$

then

$$N_i(T) - N_i(t) \le B_i^{\mathcal{A}}(t, T, \delta).$$

We refer to $B_i^{\mathcal{A}}(t, T, \delta)$ as the post-suppression pull bound *for arm i.*

This class covers many standard bandit algorithms, including UCB, Thompson Sampling, and ϵ-greedy. Further analyses for Thompson Sampling and ϵ-greedy appear in Appendices B and C; concrete post-suppression bounds are given in Appendix D.1.

Lemma 4.9 (The Suppression Lemma)**.** *Fix a non-target arm $i < K$ and a round t. Define*

$$U_i(t) \triangleq \hat{\mu}_i^0(t) + 2\beta(N_i^0(t)), \qquad L_K(t) \triangleq \hat{\mu}_K^0(t) - 2\beta(N_K^0(t)).$$

Suppose that before round t, the attacker injects n_i fake samples into arm i with $r_{\min} < L_K(t)$, and injects no further fake samples into arm i afterward. If

$$n_i > \frac{U_i(t) - L_K(t)}{L_K(t) - r_{\min}} \left(N_i^0(t) + B_i(t, T, \delta/K)\right), \tag{4}$$

then, with probability at least $1 - \delta - \delta/K$, arm i is selected at most $B_i(t, T, \delta/K)$ additional times after the injection and before round T. Consequently, after applying the condition to every non-target arm, all non-target arms are suppressed simultaneously with probability at least $1 - 2\delta$.

Theorem 4.10. *Suppose $\mathcal{A}$ is suppression-bounded. Fix $\delta \in (0, 1/2)$ and assume $r_{\min} < \mu_K - 3\beta(1)$. Define $\Delta_i \triangleq \mu_i - \mu_K$. For each non-target arm $i < K$, let t_i be the round at which the attacker first suppresses arm i relative to arm K. Before round t_i, the attacker injects the minimum number of fake samples satisfying Equation (4). Then, with probability at least $1 - 2\delta$, the learner selects the target arm in at least*

$$T - \mathcal{O}\left(\sum_{i=1}^{K-1} \frac{\mu_i + 3\beta(1) - r_{\min}}{\mu_K - 3\beta(1) - r_{\min}} \left(N_i^0(t_i) + B_i\left(t_i, T, \delta/K\right)\right)\right)$$

rounds. For each non-target arm $i < K$, the number of fake samples satisfies

$$N_i^F(T) = \mathcal{O}\left(\frac{\Delta_i + 6\beta(1)}{\mu_K - 3\beta(1) - r_{\min}} \left(N_i^0(t_i) + B_i\left(t_i, T, \delta/K\right)\right)\right).$$

Moreover, the total injection count and reward-deviation cost satisfy

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \quad \textit{and} \quad C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T).$$

Lemma 4.9 and Theorem 4.10 show that any algorithm in which each non-target arm is pulled only $o(T)$ times can be suppressed at an attack cost of only $o(T)$, provided that fake samples are injected at a suitable time. This is consistent with Zuo [2024, Fact 1]. The guarantee is broadly applicable, but it is generally not as sharp as the algorithm-specific analyses for UCB or Thompson Sampling. For example, in the UCB case, the attacker can inject fake samples so that, for each non-target arm i, $\hat{\mu}_i(t) \le \hat{\ell}_K(t) < L_K(t)$, which eliminates post-suppression pulls and effectively yields $B_i(t, T, \delta) = 0$. This highlights the trade-off between generality and sharpness: the abstract theorem applies broadly once a valid post-suppression pull bound is available, whereas algorithm-specific arguments can exploit the precise decision rule to obtain stronger guarantees. Proofs are given in Appendix D.2.

The theorem applies directly to the SI algorithm. For the PBI algorithm, one can adapt the argument by invoking the post-suppression pull bound $B_i(t, T, \delta)$ separately on each suppression period R_i. Aggregating these period-wise bounds yields the corresponding guarantee for PBI.

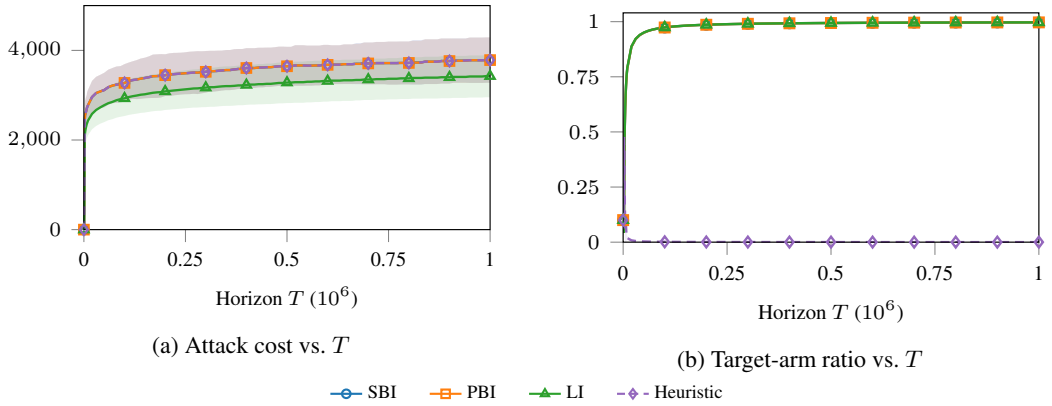
Remark 4.11. Our suppression-bounded framework is related to the mean-based algorithms studied by Xu et al. [2021], but differs in its defining condition. While mean-basedness imposes structural restrictions on the decision policy, suppression-boundedness directly bounds the number of subsequent pulls under persistent empirical domination. Both frameworks cover common algorithms such as UCB, Thompson Sampling, and ϵ-greedy.

Remark 4.12. Our analysis assumes genuine feedback is observed only for the selected arm. However, additional unsolicited clean feedback may dilute injected samples and undermine suppression. If such observations are bounded by $Q_i(T)$, they can be accommodated by replacing $B_i^{\mathcal{A}}(t, T, \delta)$ with $B_i^{\mathcal{A}}(t, T, \delta) + Q_i(T)$ in Equation (4). Thus, sublinear arrivals preserve sublinear attack cost, whereas linear arrivals may require linear cost. This dilution mechanism is similar in spirit to the smoothed myopic response defense of Zuo [2024], although it incurs $\Theta(T)$ regret even without attack.

5 Experiments

We compare SBI (the bounded SI attack), PBI, LI, and a random-timing heuristic. LI uses one fake sample per triggered non-target arm, with its reward chosen to meet the suppression threshold. The heuristic uses SBI's fake rewards and per-arm sample counts but randomizes injection times, giving the same cost. We set $K = 10$, $\delta = 0.05$, $r_{\max} = 1$, and $f = 10$. Methods share genuine reward streams and initialize each arm once; T counts genuine and fake records. We vary T from K to 10^6; both panels use the same 20 runs at each T. PBI withholds uncertified injections, and all trials are included. For UCB, we sample 10 movies from a MovieLens-derived binary rating matrix Harper and Konstan [2015] and generate Bernoulli rewards using their mean ratings, targeting the least popular movie. We set $\sigma = 0.5$ and $\delta_0 = 0.2$. Thompson Sampling settings and results appear in Appendix E. Figure 1 shows attack cost $C^F(T)$ and target-arm selection. At $T = 10^6$, SBI and PBI reach 99.68%, versus 0.02% for the matched-cost heuristic. All SBI/PBI schedules are feasible on this grid, with no non-target pulls after each arm's first injection. Code and raw results are available at GitHub[1].

[1] https://github.com/qirunzeng/bandit-fake-data-injection



(a) Attack cost vs. T

(b) Target-arm ratio vs. T

Figure 1: UCB experiments with 20 repetitions. Shading shows one sample standard deviation.

6 Concluding Remarks

This work introduces *Fake Data Injection*, a practical threat model for adversarial attacks on stochastic bandits. Unlike prior models that allow per-round, unbounded reward perturbations, our framework reflects realistic constraints, including bounded feedback, limited injection capability, and the attacker's inability to modify genuine user data. Under this model, we develop effective attack strategies that manipulate bandit learners using only sublinear-cost injections. Both theory and experiments show that even sparse, bounded fake interactions can significantly bias stochastic bandit algorithms. Our analytical tools, such as the suppression lemma, further reveal a broader vulnerability pattern in sequential decision-making systems. These results expose a gap between classical robustness guarantees and the practical security of online learning under realistic feedback constraints.

Despite these results, several limitations still remain. We study a passive learner that processes all feedback without defense, whereas robust systems may use filtering or auditing. We focus on stochastic bandits; extending the framework to contextual bandits or reinforcement learning remains open. In addition, real-world attackers may face detection risks or adaptive filtering, which our current model does not capture. Our model also abstracts away user identities, collaborative filtering, and richer user behavior. Future work could investigate defenses such as anomaly detection and arm-level auditing, explore the dynamic interaction between attackers and adaptive learners, and extend the analysis to settings where multiple users can inject action–reward pairs.

Acknowledgement

The work of Jinhang Zuo is supported by the Early Career Scheme (No. CityU 21212025) of the University Grants Committee (UGC) and CityU 9610706. The work of Xuchuang Wang is supported in part by the RGC JRFS2526-4S06.

References

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

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 12 2012. ISSN 1935-8237.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 14362–14373. Curran Associates, Inc., 2020.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in Neural Information Processing Systems*, volume 35, pages 34614–34625. Curran Associates, Inc., 2022. doi: 10.52202/068431-2508.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *International conference on machine learning*, pages 1245–1253. PMLR, 2016.

Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050. PMLR, 2019.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pages 114–122, 2018.

Yuzhe Ma and Zhijin Zhou. Adversarial attacks on adversarial bandits. In *The Eleventh International Conference on Learning Representations*, 2023.

Arpan Mukherjee, Ali Tajer, Pin-Yu Chen, and Payel Das. Mean-based best arm identification in stochastic bandits under reward contamination. In *Advances in Neural Information Processing Systems*, volume 34, pages 9651–9662. Curran Associates, Inc., 2021.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Huazheng Wang, Haifeng Xu, and Hongning Wang. When are linear stochastic bandits attackable? In *International Conference on Machine Learning*, pages 23254–23273. PMLR, 2022.

Zichen Wang, Rishab Balasubramanian, Hui Yuan, Mengdi Wang, Huazheng Wang, et al. Adversarial attacks on online learning to rank with stochastic click models. *Transactions on Machine Learning Research*, 2024.

Haike Xu and Jian Li. Simple combinatorial algorithms for combinatorial bandits: Corruptions and approximations. In *Uncertainty in Artificial Intelligence*, pages 1444–1454. PMLR, 2021.

Yinglun Xu, Bhuvesh Kumar, and Jacob D Abernethy. Observation-free attacks on stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 22550–22561. Curran Associates, Inc., 2021.

Qirun Zeng, Manhin Poon, Xiangxiang Dai, Qixin Zhang, and Jinhang Zuo. Beyond reward suppression: Near-optimal offline attacks on warm-start bandits with bounded rewards. *arXiv preprint arXiv:2610.10000*, 2026a.

Qirun Zeng, Xuchuang Wang, Jiayi Shen, Xutong Liu, Fang Kong, and Jinhang Zuo. Best arm identification in generalized linear bandits via hybrid feedback. *arXiv preprint arXiv:2605.05745*, 2026b.

Yurui Zou, Qirun Zeng, and Jinhang Zuo. Conversational generalized linear bandits with hybrid feedback. *arXiv preprint arXiv:2610.35262*, 2026.

Jinhang Zuo, Zhiyao Zhang, Zhiyong Wang, Shuai Li, Mohammad Hajiesmaili, and Adam Wierman. Adversarial attacks on online learning to rank with click feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 41675–41692. Curran Associates, Inc., 2023.

Shiliang Zuo. Near optimal adversarial attacks on stochastic bandits and defenses with smoothed responses. In *International Conference on Artificial Intelligence and Statistics*, pages 2098–2106. PMLR, 2024.

Appendix

Contents

1 Introduction **1**

1.1 Our Contributions . . . 2

1.2 Related Work . . . 2

2 Preliminaries **3**

2.1 Stochastic Bandits . . . 3

2.2 Previous Threat Model and Limitations . . . 3

3 New Threat Model: Fake Data Injection **4**

4 Attack Strategies **5**

4.1 The Suppression Lemma . . . 5

4.2 Warm-up: Simultaneous Injection . . . 5

4.3 Periodic Bounded Injection . . . 7

4.4 Extension to Suppression-Bounded Algorithms . . . 8

5 Experiments **9**

6 Concluding Remarks **10**

A Proofs for UCB **14**

A.1 Proof of the suppression lemma for UCB . . . 14

A.2 Proof of Simultaneous Injection on UCB . . . 14

A.3 Proofs of Periodic Bounded Injection on UCB . . . 15

B Thompson Sampling Algorithm-specific Analysis **17**

B.1 Suppression Lemma for Thompson Sampling . . . 17

B.2 Simultaneous Injection on Thompson Sampling . . . 18

B.3 Periodic Bounded Injection on Thompson Sampling . . . 19

C Epsilon-greedy Algorithm-specific Analysis **22**

D General Suppression Framework Extensions **22**

D.1 Concrete instantiations of the post-suppression bound . . . 22

D.2 Proofs of the General Suppression Framework . . . 23

E Additional Experiments **24**

A Proofs for UCB

For notational simplicity, we treat all numbers here as *integers*; this convention does not affect the results.

Suppose that the reward distributions of arms are σ^2-sub-Gaussian. The following concentration result will be useful throughout our analysis. Recall that $\beta(N) \triangleq \sqrt{\frac{2\sigma^2}{N}\log\frac{\pi^2 KN^2}{3\delta}}$. Let

$$\mathcal{E} \triangleq \left\{\forall i, t, |\hat{\mu}_i^0(t) - \mu_i| < \beta(N_i^0(t))\right\}, \tag{5}$$

where $\hat{\mu}_i^0(t)$ and $N_i^0(t)$ denote the empirical mean and count of genuine samples of arm i up to round t, respectively.

Lemma A.1 (Lemma 1 of Jun et al. [2018]). *For $\delta \in (0,1)$, $\mathbb{P}\{\mathcal{E}\} > 1-\delta$.*

A.1 Proof of the suppression lemma for UCB

Proof of Lemma 4.1. Under event $\mathcal{E}$, we can establish a lower bound on arm K's estimate. For any two rounds $t_2 > t_1$,

$$\hat{\mu}_K(t_2) \geq \mu_K - \beta(N_K(t_2)) \geq \mu_K - \beta(N_K(t_1)) \geq \hat{\mu}_K(t_1) - 2\beta(N_K(t_1)), \tag{6}$$

where the second inequality follows from the monotonicity of $\beta(N)$.

Suppose that at round t_1, the following equation holds,

$$\hat{\mu}_i(t_1) \leq \hat{\mu}_K(t_1) - 2\beta(N_K(t_1)) - 3\sigma\delta_0. \tag{7}$$

which means $\mathrm{UCB}_i(t_1) < \mathrm{UCB}_K(t_1)$, so arm i is not selected at round t_1.

Now consider any subsequent round t_2 with $t_1 \leq t_2 < \exp(m_i\delta_0^2)$, where $m_i \triangleq N_i(t_1)$, and assume that arm i has not been selected in any round between t_1 and t_2. Then $N_i(t_2{+}1) = m_i$ and $\hat{\mu}_i(t_2{+}1) = \hat{\mu}_i(t_1)$, so the UCB index for arm i at round t_2+1 is

$$\begin{aligned}\mathrm{UCB}_i(t_2+1) &= \hat{\mu}_i(t_1) + 3\sigma\sqrt{\tfrac{\log(t_2+1)}{m_i}} \leq \hat{\mu}_K(t_1) - 2\beta(N_K(t_1)) - 3\sigma\delta_0 + 3\sigma\sqrt{\tfrac{\log(t_2+1)}{m_i}} \\ &\leq \hat{\mu}_K(t_1) - 2\beta(N_K(t_1)) \leq \hat{\mu}_K(t_2+1) \leq \mathrm{UCB}_K(t_2+1),\end{aligned}$$

where the third step uses the bound $\sqrt{\frac{\log(t_2+1)}{m_i}} < \delta_0$. This argument shows that the UCB index of arm i remains strictly lower than that of the target arm K for all $t_2 < \exp(m_i\delta_0^2)$. By induction, arm i will not be selected again before round $\exp(m_i\delta_0^2)$. □

A.2 Proof of Simultaneous Injection on UCB

Proof. According to Algorithm 1, after injecting n_i samples with value $r_{\min}$ into arm i at round t, the empirical mean at round $t+n_i$ becomes:

$$\begin{aligned}\hat{\mu}_i(t+n_i) &= \frac{\hat{\mu}_i(t)N_i(t) + n_i \cdot r_{\min}}{N_i(t)+n_i} \leq \frac{\hat{\mu}_i(t)N_i(t) + \frac{\hat{\mu}_i(t)-\hat{\ell}_K(t)}{\hat{\ell}_K(t)-r_{\min}}N_i(t)\cdot r_{\min}}{N_i(t) + \frac{\hat{\mu}_i(t)-\hat{\ell}_K(t)}{\hat{\ell}_K(t)-r_{\min}}N_i(t)} \\ &= \frac{\hat{\mu}_i(t) + \frac{\hat{\mu}_i(t)-\hat{\ell}_K(t)}{\hat{\ell}_K(t)-r_{\min}}r_{\min}}{1 + \frac{\hat{\mu}_i(t)-\hat{\ell}_K(t)}{\hat{\ell}_K(t)-r_{\min}}} = \frac{\hat{\mu}_i(t)\hat{\ell}_K(t) - \hat{\ell}_K(t)r_{\min}}{\hat{\mu}_i(t) - r_{\min}} = \hat{\ell}_K(t),\end{aligned}$$

where the last step ensures that after injection, arm i's empirical mean is suppressed below the target threshold.

On $\mathcal{E}$, we first bound the number of injected samples. At the trigger time t_i, concentration and the positive margin imply

$$\begin{aligned}n_i &= \frac{\hat{\mu}_i(t_i) - \hat{\ell}_K(t_i)}{\hat{\ell}_K(t_i) - r_{\min}}N_i(t_i) \leq \frac{\Delta_i + \beta(N_i(t_i)) + 3\beta(N_K(t_i)) + 3\sigma\delta_0}{\mu_K - 3\beta(N_K(t_i)) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \\ &\leq \frac{\Delta_i + 4\beta(1) + 3\sigma\delta_0}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2}.\end{aligned}$$

Consequently,

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \le \sum_{i=1}^{K-1} n_i \le \sum_{i=1}^{K-1} \frac{\Delta_i + 4\beta(1) + 3\sigma\delta_0}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2},$$

Next, since event $\mathcal{E}$ holds, and by Lemma 4.1, arm i will not be selected after the injection. The total number of history entries for arm i, including fake samples, is therefore bounded by:

$$\begin{aligned} N_i(t) + n_i &\le \frac{\log T}{\delta_0^2} + \frac{\hat{\mu}_i(t) - \hat{\ell}_K(t)}{\hat{\ell}_K(t) - r_{\min}} \cdot \frac{\log T}{\delta_0^2} = \frac{\hat{\mu}_i(t) - r_{\min}}{\hat{\ell}_K(t) - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \\ &\le \frac{\mu_i + \beta(N_i(t)) - r_{\min}}{\hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \le \frac{\mu_i + \beta(\frac{\log T}{\delta_0^2}) - r_{\min}}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2}, \end{aligned}$$

where we used concentration bounds for $\hat{\mu}_i(t)$ and $\hat{\mu}_K(t)$ under event $\mathcal{E}$ and the nonincreasing property of $\beta(\cdot)$. The cumulative reward-deviation cost is:

$$C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T) \le \sum_{i=1}^{K-1} (\mu_i - r_{\min}) \frac{\Delta_i + 4\beta(1) + 3\sigma\delta_0}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \frac{\log T}{\delta_0^2}.$$

Therefore, both the cumulative attack cost and the number of non-target arm pulls are $\mathcal{O}\left(\frac{\log T}{\delta_0^2}\right)$ per arm, and hence sublinear in T, completing the proof. □

A.3 Proofs of Periodic Bounded Injection on UCB

Proof of Lemma 4.6. Suppose event $\mathcal{E}$ holds. We begin by estimating the total number of fake samples needed to demote arm i. This quantity is given by

$$n_i = \frac{\hat{\mu}_i(t) - \hat{\ell}_K(t)}{\hat{\ell}_K(t) - r_{\min}} \cdot \frac{\log T}{\delta_0^2},$$

where $\hat{\ell}_K(t)$ is a conservative lower bound on arm K's empirical mean, and $r_{\min}$ is the value of each injected fake sample. Under the batch size constraint f, the attack spans $\lceil \frac{n_i}{f} \rceil$ periods, with each period injecting f fake samples.

Our goal is to choose an appropriate delay parameter R_i such that after each injection, arm i is not selected again until the next scheduled injection. Specifically, we require that after the c-th batch (for any $c \in \{1, 2, \ldots, \lceil \frac{n_i}{f} \rceil\}$), arm i is not selected for at least R_i rounds.

Let $t_i(c) = t + (f + R_i)c$ denote the round before the $(c+1)$-th injection. We examine the UCB index of arm i at time $t_i(c)$:

$$\text{UCB}_i(t_i(c)) = \hat{\mu}_i(t_i(c)) + 3\sigma\sqrt{\frac{\log t_i(c)}{N_i(t) + fc}} = \frac{N_i(t)\hat{\mu}_i(t) + fcr_{\min}}{N_i(t) + fc} + 3\sigma\sqrt{\frac{\log t_i(c)}{N_i(t) + fc}}. \quad (8)$$

To ensure arm i is not selected before round $t_i(c)$, we want its UCB index to be no larger than that of arm K. A sufficient condition is

$$\text{UCB}_i(t_i(c)) \le \hat{\mu}_K(t) - 2\beta(N_K(t)) \le \text{UCB}_K(t_i(c)), \quad (9)$$

where we use the lower bound $\hat{\mu}_K(t) - 2\beta(N_K(t))$ to conservatively approximate arm K's UCB.

Define $\tilde{\mu}_i(t_i(c))$ as the post-injection empirical mean of arm i:

$$\tilde{\mu}_i(t_i(c)) = \frac{N_i(t)\hat{\mu}_i(t) + fcr_{\min}}{N_i(t) + fc}.$$

Then, the condition in Equation (9) implies that, for any c, the delay parameter $R_i(c)$ must satisfy:

$$R_i(c) \le \frac{\exp\left(\left(\frac{\hat{\mu}_K(t) - 2\beta(N_K(t)) - \tilde{\mu}_i(t_i(c))}{3\sigma}\right)^2 \cdot (N_i(t) + fc)\right) - t}{c} - f. \quad (10)$$

Finally, to ensure that this condition holds for every injection period, we define the overall delay parameter as the minimum over all c:

$$R_i = \min_{1 \le c \le \lceil \frac{n_i}{f} \rceil} \frac{\exp\left(\left(\frac{\hat{\mu}_K(t) - 2\beta(N_K(t)) - \tilde{\mu}_i(t_i(c))}{3\sigma}\right)^2 \cdot (N_i(t) + fc)\right) - t}{c} - f. \tag{11}$$

This choice of R_i ensures that after each batch of f fake samples, arm i will not be pulled again until the next scheduled injection. □

Proof of Theorem 4.7. Consider any non-target arm $i < K$. Under PBI, periodic injections are initiated once $N_i(t) = \frac{\log T}{\delta_0^2}$. The total number of fake samples required, denoted by n_i, is identical to that in SI and is chosen such that the empirical mean of arm i is driven below the lower confidence bound of the target arm K.

Unlike SI, the PBI algorithm distributes these n_i injections over time in batches of size f, separated by delays of length R_i. By Lemma 4.6, after each batch of f injected samples, arm i will not be selected for at least R_i rounds. This establishes a persistent suppression effect: throughout the entire injection process, arm i is not revisited by the learner, and therefore no additional genuine samples of arm i are collected before all n_i fake samples are injected. We now consider two cases:

Case 1: All n_i fake samples are successfully injected before round T. Then the empirical mean of arm i is fully suppressed, and its future selection is prevented. This directly mirrors the SI case analyzed in Theorem 4.4, so both the number of non-target arm pulls and cumulative attack cost are bounded as in that theorem.

Case 2: Only a portion of the n_i fake samples are injected by round T. Since Lemma 4.6 guarantees that arm i is not pulled during this partial injection phase, the number of target-arm pulls remains high. Meanwhile, both the number of injected samples and the cumulative attack cost are no larger than in the SI case.

As a consequence, the two cases both incur no more than

$$n_i = \frac{\hat{\mu}_i(t) - \hat{\ell}_K(t)}{\hat{\ell}_K(t) - r_{\min}} \cdot \frac{\log T}{\delta_0^2}$$

fake samples.

We first bound the number of injected samples. The prescribed n_i obeys the per-arm bound established in the proof of Theorem 4.4, so

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \le \sum_{i=1}^{K-1} n_i \le \sum_{i=1}^{K-1} \frac{\Delta_i + 4\beta(1) + 3\sigma\delta_0}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2}.$$

Next, since no further genuine pulls of arm i occur during injection, its total contribution to the learner's history is bounded by $N_i(t) + n_i$, where

$$\begin{aligned} N_i(t) + n_i &\le \frac{\log T}{\delta_0^2} + \frac{\hat{\mu}_i(t) - \hat{\ell}_K(t)}{\hat{\ell}_K(t) - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \\ &= \frac{\hat{\mu}_i(t) - r_{\min}}{\hat{\ell}_K(t) - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \\ &\le \frac{\mu_i + \beta(N_i(t)) - r_{\min}}{\hat{\mu}_K(t) - 2\beta(N_K(t)) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \\ &\le \frac{\mu_i + \beta\left(\frac{\log T}{\delta_0^2}\right) - r_{\min}}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \cdot \frac{\log T}{\delta_0^2}. \end{aligned}$$

The cumulative reward-deviation cost is:

$$C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T) \le \sum_{i=1}^{K-1} (\mu_i - r_{\min}) \frac{\Delta_i + 4\beta(1) + 3\sigma\delta_0}{\mu_K - 3\beta(1) - 3\sigma\delta_0 - r_{\min}} \frac{\log T}{\delta_0^2}.$$

This proves the claim. □

B Thompson Sampling Algorithm-specific Analysis

Thompson Sampling (TS). We consider the Thompson Sampling algorithm specified in Zuo [2024], and the prototype is the (α, ψ) algorithm of Agrawal and Goyal [2017]. In the first K rounds, the learner pulls each arm once. For rounds $t > K$, a sample ν_i is drawn independently for each arm $i \in \mathcal{K}$ from the distribution $\mathcal{N}\left(\hat{\mu}_i(t), \frac{1}{N_i(t)}\right)$, and the learner selects the arm with the largest sampled value $a_t = \arg\max_{i \in \mathcal{K}} \nu_i$. In the absence of attacks, both UCB and TS are known to achieve sublinear regret by selecting suboptimal arms only $o(T)$ times.

For the Thompson Sampling extensions, we also use an event $\mathcal{F}$ that controls the sampled value ν_i of each arm:

$$\mathcal{F} \triangleq \left\{ \forall i, t, |\nu_i(t) - \hat{\mu}_i(t)| < \frac{\gamma(t)}{\sqrt{N_i(t)}} \right\}, \tag{12}$$

where $\gamma(t) \triangleq \sqrt{2 \log \frac{\pi^2 K t^2}{3\delta}}$.

Lemma B.1. *For $\delta \in (0,1)$, $\mathbb{P}\{\mathcal{F}\} > 1 - \delta$.*

Proof. Since $\nu_i(t) \sim \mathcal{N}\left(\hat{\mu}_i(t), \frac{1}{N_i(t)}\right)$, it follows that $\nu_i(t) - \hat{\mu}_i(t) \sim \mathcal{N}\left(0, \frac{1}{N_i(t)}\right)$. By the Gaussian tail bound, for any $c > 0$,

$$\Pr(|\nu_i(t) - \hat{\mu}_i(t)| > c) < 2 \exp\left(-\frac{c^2}{2/N_i(t)}\right) = 2 \exp\left(-\frac{N_i(t) c^2}{2}\right).$$

Setting $c = \frac{\gamma(t)}{\sqrt{N_i(t)}}$ and $\gamma(t) = \sqrt{2 \log \frac{\pi^2 K t^2}{3\delta}}$, we obtain

$$\Pr\left(|\nu_i(t) - \hat{\mu}_i(t)| > \frac{\gamma(t)}{\sqrt{N_i(t)}} \right) < 2 \exp\left(-\frac{\gamma(t)^2}{2}\right) = \frac{6\delta}{\pi^2 K t^2}.$$

Summing over t yields the claim. □

We collect the Thompson Sampling extensions here. The event $\mathcal{F}$ in Equation (12) controls the posterior samples around their empirical means; combined with the genuine-sample event $\mathcal{E}$, it yields a TS analogue of the UCB suppression lemma. We then state SI-TS and PBI-TS in the same order as the main text develops SI and PBI for UCB.

B.1 Suppression Lemma for Thompson Sampling

Define

$$\Gamma_i(t) \triangleq \sqrt{2 \log \frac{\pi^2 K}{3\delta} + 4 N_i(t) \delta_0^2}, \qquad \kappa_i^{\mathrm{TS}}(t) \triangleq \Gamma_i(t) \left(\frac{1}{\sqrt{N_i(t)}} + \frac{1}{\sqrt{N_K(t)}} \right),$$

and $\hat{\ell}'_K(t) \triangleq \hat{\mu}_K(t) - 2\beta(N_K(t)) - \kappa_i^{\mathrm{TS}}(t)$. The margin $\kappa_i^{\mathrm{TS}}(t)$ is the largest combined posterior-sampling deviation of arms i and K over the suppression window ending at $\exp(N_i(t)\delta_0^2)$.

Lemma B.2 (Exponential suppression under Thompson Sampling)**.** *For each non-target arm $i < K$, if $\hat{\mu}_i(t) \le \hat{\ell}'_K(t)$, then with probability at least $1 - 2\delta$, arm i will not be selected again until at least round $\lfloor \exp(N_i(t)\delta_0^2) \rfloor$.*

Proof. Suppose that at round t_1, the following inequality holds:

$$\hat{\mu}_i(t_1) \le \hat{\mu}_K(t_1) - 2\beta(N_K(t_1)) - \kappa_i^{\mathrm{TS}}(t_1). \tag{13}$$

Let $m_i \triangleq N_i(t_1)$ and consider any round t_2 such that $t_1 < t_2 < \lfloor \exp(m_i \delta_0^2) \rfloor$. Assuming that arm i is not selected from round t_1 to t_2, then $N_i(t_2 + 1) = m_i$ and $\hat{\mu}_i(t_2 + 1) = \hat{\mu}_i(t_1)$. Applying the

concentration bounds from Lemmas A.1 and B.1, the sampled value $\nu_i(t_2+1)$ for arm i satisfies:

$$\nu_i(t_2+1) < \hat{\mu}_i(t_2+1) + \frac{\gamma(t_2+1)}{\sqrt{m_i}} \le \hat{\mu}_K(t_1) - 2\beta(N_K(t_1)) - \kappa_i^{\mathrm{TS}}(t_1) + \frac{\gamma(t_2+1)}{\sqrt{m_i}}$$
$$\le \hat{\mu}_K(t_2+1) - \kappa_i^{\mathrm{TS}}(t_1) + \frac{\gamma(t_2+1)}{\sqrt{m_i}}.$$

Since $t_2+1 \le \exp(m_i\delta_0^2)$, we have $\gamma(t_2+1) \le \Gamma_i(t_1)$. Also, $N_K(t_2+1) \ge N_K(t_1)$. Hence

$$\kappa_i^{\mathrm{TS}}(t_1) \ge \gamma(t_2+1)\left(\frac{1}{\sqrt{m_i}} + \frac{1}{\sqrt{N_K(t_2+1)}}\right).$$

Therefore

$$\nu_i(t_2+1) \le \hat{\mu}_K(t_2+1) - \frac{\gamma(t_2+1)}{\sqrt{N_K(t_2+1)}} \le \nu_K(t_2+1),$$

where the last inequality uses event $\mathcal{F}$ for arm K.

This chain of inequalities implies that the sampled value $\nu_i(t_2+1)$ remains lower than $\nu_K(t_2+1)$ for all $t_2 < \exp(m_i\delta_0^2)$, with probability at least $1-2\delta$. Therefore, arm i will not be selected again until at least round $\exp(m_i\delta_0^2)$. □

B.2 Simultaneous Injection on Thompson Sampling

Algorithm 3: SI on TS

Input: Attack parameter δ_0, target arm K, lower bound $r_{\min}$

for $t = 1, 2, \ldots$ **do**
  for *each non-target arm* $i < K$ **do**
    if $N_i(t) = \lceil \frac{\log T}{\delta_0^2} \rceil$ **then**
      Inject $n_i = \left\lceil \frac{\hat{\mu}_i(t) - \hat{\ell}'_K(t)}{\hat{\ell}'_K(t) - r_{\min}} \cdot \left\lceil \frac{\log T}{\delta_0^2} \right\rceil \right\rceil$ fake samples $(i, r_{\min})$;
    end
  end
end

Theorem B.3. *Suppose $T > 2K$ and $\delta \in (0, 1/2)$. Define $\Delta_i \triangleq \mu_i - \mu_K$. For each non-target arm i, let t_i be the round at which Algorithm 3 triggers the injection and set*

$$\bar{\kappa}_i^{\mathrm{TS}} \triangleq \sqrt{2\log\frac{\pi^2 K}{3\delta} + 4\log T}\left(1 + \sqrt{\frac{\delta_0^2}{\log T}}\right).$$

Assume $\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min} > 0$ for every non-target arm. With probability at least $1-2\delta$, the modified Simultaneous Injection attack forces Thompson Sampling to select the target arm in at least

$$T - \mathcal{O}\left(\sum_{i=1}^{K-1} \frac{\mu_i + \beta(\log T/\delta_0^2) - r_{\min}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2}\right) \tag{14}$$

rounds. For each non-target arm $i < K$, the number of fake samples satisfies

$$N_i^F(T) = \mathcal{O}\left(\frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2}\right).$$

Moreover, the total injection count and reward-deviation cost satisfy

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \quad \textit{and} \quad C^F(T) \le \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T).$$

Proof. According to Algorithm 3, after injecting n_i samples, the empirical mean at round $t+n_i$ becomes:

$$\hat{\mu}_i(t+n_i) = \frac{\hat{\mu}_i(t)N_i(t) + n_i r_{\min}}{N_i(t)+n_i} \le \frac{\hat{\mu}_i(t)N_i(t) + \frac{\hat{\mu}_i(t)-\hat{\ell}'_K(t)}{\hat{\ell}'_K(t)-r_{\min}} N_i(t) r_{\min}}{N_i(t) + \frac{\hat{\mu}_i(t)-\hat{\ell}'_K(t)}{\hat{\ell}'_K(t)-r_{\min}} N_i(t)}$$

$$= \frac{\hat{\mu}_i(t) + \frac{\hat{\mu}_i(t)-\hat{\ell}'_K(t)}{\hat{\ell}'_K(t)-r_{\min}} r_{\min}}{1 + \frac{\hat{\mu}_i(t)-\hat{\ell}'_K(t)}{\hat{\ell}'_K(t)-r_{\min}}} = \frac{\hat{\mu}_i(t)\hat{\ell}'_K(t) - \hat{\ell}'_K(t) r_{\min}}{\hat{\mu}_i(t) - r_{\min}} = \hat{\ell}'_K(t),$$

where the last step ensures that after injection, arm i's empirical mean is suppressed below the target threshold.

On $\mathcal{E} \cap \mathcal{F}$, we first bound the number of injected samples. Applying concentration directly to the ratio defining n_i gives

$$n_i = \frac{\hat{\mu}_i(t_i) - \hat{\ell}'_K(t_i)}{\hat{\ell}'_K(t_i) - r_{\min}} N_i(t_i) \le \frac{\Delta_i + \beta(N_i(t_i)) + 3\beta(N_K(t_i)) + \kappa_i^{\mathrm{TS}}(t_i)}{\mu_K - 3\beta(N_K(t_i)) - \kappa_i^{\mathrm{TS}}(t_i) - r_{\min}} \cdot \frac{\log T}{\delta_0^2}$$

$$\le \frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2}.$$

Consequently,

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \le \sum_{i=1}^{K-1} n_i \le \sum_{i=1}^{K-1} \frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2},$$

which gives the stated injection-count bound.

Next, since events $\mathcal{E}$ and $\mathcal{F}$ hold, and by Lemma B.2, arm i will not be selected after round $t+n_i$. The total number of history entries for arm i, including fake samples, is therefore bounded by:

$$N_i(t) + n_i = \frac{\log T}{\delta_0^2} + \frac{\hat{\mu}_i(t) - \hat{\ell}'_K(t)}{\hat{\ell}'_K(t) - r_{\min}} \frac{\log T}{\delta_0^2} = \frac{\hat{\mu}_i(t) - r_{\min}}{\hat{\ell}'_K(t) - r_{\min}} \frac{\log T}{\delta_0^2}$$

$$\le \frac{\mu_i + \beta(N_i(t)) - r_{\min}}{\hat{\mu}_K(t) - 2\beta(N_K(t)) - \kappa_i^{\mathrm{TS}}(t) - r_{\min}} \frac{\log T}{\delta_0^2}$$

$$\le \frac{\mu_i + \beta(\frac{\log T}{\delta_0^2}) - r_{\min}}{\mu_K - 3\beta(1) - \kappa_i^{\mathrm{TS}}(t) - r_{\min}} \frac{\log T}{\delta_0^2}$$

$$\le \frac{\mu_i + \beta(\frac{\log T}{\delta_0^2}) - r_{\min}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \frac{\log T}{\delta_0^2}.$$

The cumulative reward-deviation cost is:

$$C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T) \le \sum_{i=1}^{K-1} (\mu_i - r_{\min}) \frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \frac{\log T}{\delta_0^2}.$$

This proves the theorem. □

B.3 Periodic Bounded Injection on Thompson Sampling

Let $t_i(c) = t + (f + R_i)c$ denote the round before the $c+1$-th injection and

$$\eta_i(t) \triangleq \frac{\left(\hat{\mu}_K(t) - 2\beta(N_K(t)) - \tilde{\mu}_i(t_i(c))\right)^2}{2\left(1/\sqrt{N_i(t) + fc} + 1/\sqrt{N_K(t)}\right)^2}$$

Algorithm 4: PBI on TS

Input: Attack parameter δ_0, reward bound $[r_{\min}, r_{\max}]$, max batch size f

Initialize $n_i = 0$ for every non-target arm i;
for $t = 1, 2, \ldots$ **do**
  for *each non-target arm* $i < K$ **do**
    if $N_i(t) = \lceil \frac{\log T}{\delta_0^2} \rceil$ **then**
      $n_i \leftarrow \left\lceil \frac{\hat{\mu}_i(t) - \hat{\ell}'_K(t)}{\hat{\ell}'_K(t) - r_{\min}} \cdot \left\lceil \frac{\log T}{\delta_0^2} \right\rceil \right\rceil$; $\tau_i \leftarrow t$;
      $R_i \leftarrow \min_{1 \le c \le \lceil \frac{n_i}{f} \rceil} \left\{ \frac{1}{c} \left(\sqrt{\frac{3\delta}{\pi^2 K} \cdot \exp(\eta_i(t))} - t \right) - f \right\}$;
    end
    if $N_i^F(t) < n_i$ *and* $\tau_i \le t$ **then**
      Inject $\min\{n_i - N_i^F(t), f\}$ fake samples $(i, r_{\min})$;
      $\tau_i \leftarrow \tau_i + f + R_i$;
    end
  end
end

Lemma B.4. *The choice of R_i in Algorithm 4 ensures that once a batch of f fake data samples is injected into non-target arm i, the arm will not be selected again for at least the next R_i rounds.*

Proof. We aim to guarantee that arm i is not selected between successive fake sample injections. Let $\tilde{\nu}_i$ and $\tilde{\nu}_K$ denote the Thompson-sampled values of arm i and arm K, respectively, after the c-th injection. After injecting f fake samples with value $r_{\min}$ for c periods, the empirical mean of arm i becomes

$$\tilde{\mu}_i(t_i(c)) = \frac{N_i(t)\hat{\mu}_i(t) + fcr_{\min}}{N_i(t) + fc}.$$

We want to ensure that arm i is unlikely to be selected before time $t_i(c)$ by ensuring:

$$\begin{aligned} \tilde{\nu}_i(t_i(c)) &\le \tilde{\mu}_i(t_i(c)) + \frac{\gamma(t_i(c))}{\sqrt{N_i(t) + fc}} \\ &\le \hat{\mu}_K(t) - 2\beta(N_K(t)) - \frac{\gamma(t_i(c))}{\sqrt{N_K(t)}} \le \tilde{\nu}_K(t_i(c)), \end{aligned} \tag{15}$$

where $\gamma(t) = \sqrt{2\log\left(\frac{\pi^2 K t^2}{3\delta}\right)}$ bounds the Thompson Sampling deviation under event $\mathcal{F}$.

Rearranging the middle inequality in Equation (15), we require:

$$\gamma(t_i(c)) \left(1/\sqrt{N_i(t) + fc} + 1/\sqrt{N_K(t)}\right) \le \hat{\mu}_K(t) - 2\beta(N_K(t)) - \tilde{\mu}_i(t_i(c)). \tag{16}$$

Solving Equation (16) for $t_i(c)$ gives: $t_i(c) \le \sqrt{\frac{3\delta}{\pi^2 K} \cdot \exp\left(\eta_i(t)\right)}$. Therefore, we define:

$$R_i = \min_{1 \le c \le \lceil \frac{n_i}{f} \rceil} \frac{1}{c} \left[\sqrt{\frac{3\delta}{\pi^2 K} \cdot \exp\left(\eta_i(t)\right)} - t \right] - f. \tag{17}$$

This ensures that after each batch of f fake samples, arm i is suppressed for at least R_i rounds under events $\mathcal{E}$ and $\mathcal{F}$. Hence, the learner will not select arm i between consecutive injections. □

Theorem B.5. *Suppose $T > 2K$ and $\delta \in (0, 1/2)$. Define $\Delta_i \triangleq \mu_i - \mu_K$. For each non-target arm i, let t_i be the round at which Algorithm 4 triggers the injection schedule and set $\bar{\kappa}_i^{\mathrm{TS}} \triangleq \kappa_i^{\mathrm{TS}}(t_i)$. Assume $\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min} > 0$ for every non-target arm. With probability at least $1 - 2\delta$, the modified PBI forces Thompson Sampling to select the target arm in at least*

$$T - \mathcal{O}\left(\sum_{i=1}^{K-1} \frac{\mu_i + \beta\left(\log T/\delta_0^2\right) - r_{\min}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2} \right) \tag{18}$$

rounds. For each non-target arm $i < K$, the number of fake samples satisfies

$$N_i^F(T) = \mathcal{O}\left(\frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2}\right).$$

Moreover, the total injection count and reward-deviation cost satisfy

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \quad \textit{and} \quad C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T).$$

Proof of Theorem B.5. Consider any non-target arm $i < K$, and let $t = t_i$ be the round at which Algorithm 4 triggers the injection schedule for this arm. Under events $\mathcal{E}$ and $\mathcal{F}$, which hold jointly with probability at least $1 - 2\delta$, the choice of R_i in Lemma B.4 guarantees that after each batch of fake samples is injected, arm i is not selected before the next scheduled batch. Hence, during the entire injection process, no additional genuine samples of arm i are collected.

The total number of fake samples prescribed by Algorithm 4 is

$$n_i = \frac{\hat{\mu}_i(t) - \hat{\ell}'_K(t)}{\hat{\ell}'_K(t) - r_{\min}} \cdot \frac{\log T}{\delta_0^2},$$

up to the harmless integer-rounding convention stated at the beginning of the appendix. We now consider two cases.

If all n_i fake samples are injected before round T, then the empirical mean of arm i is reduced below $\hat{\ell}'_K(t)$. By Lemma B.2, arm i is then suppressed for the remainder of the horizon. If only a portion of the n_i fake samples is injected before round T, then Lemma B.4 still ensures that arm i is not selected during this partial-injection phase, and the incurred cost is no larger than the cost of injecting all n_i samples. Thus, in both cases it is enough to upper bound the contribution of $N_i(t) + n_i$ samples in the learner's history.

We first bound the number of injected samples. Applying the same concentration argument as in the proof of Theorem B.3 to the ratio defining n_i yields

$$n_i \le \frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2}.$$

Hence, even if a schedule is incomplete at T,

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \le \sum_{i=1}^{K-1} n_i \le \sum_{i=1}^{K-1} \frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \cdot \frac{\log T}{\delta_0^2},$$

which gives the stated injection-count bound.

Next, since $N_i(t) = \frac{\log T}{\delta_0^2}$, we have

$$\begin{aligned} N_i(t) + n_i &= \frac{\log T}{\delta_0^2} + \frac{\hat{\mu}_i(t) - \hat{\ell}'_K(t)}{\hat{\ell}'_K(t) - r_{\min}} \frac{\log T}{\delta_0^2} = \frac{\hat{\mu}_i(t) - r_{\min}}{\hat{\ell}'_K(t) - r_{\min}} \frac{\log T}{\delta_0^2} \\ &\le \frac{\mu_i + \beta(N_i(t)) - r_{\min}}{\hat{\mu}_K(t) - 2\beta(N_K(t)) - \kappa_i^{\mathrm{TS}}(t) - r_{\min}} \frac{\log T}{\delta_0^2} \le \frac{\mu_i + \beta(\frac{\log T}{\delta_0^2}) - r_{\min}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \frac{\log T}{\delta_0^2}. \end{aligned}$$

Summing this bound over all non-target arms gives the stated lower bound on the number of target-arm selections.

The cumulative reward-deviation cost is:

$$C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T) \le \sum_{i=1}^{K-1} (\mu_i - r_{\min}) \frac{\Delta_i + 4\beta(1) + \bar{\kappa}_i^{\mathrm{TS}}}{\mu_K - 3\beta(1) - \bar{\kappa}_i^{\mathrm{TS}} - r_{\min}} \frac{\log T}{\delta_0^2}.$$

This proves the theorem. □

C Epsilon-greedy Algorithm-specific Analysis

Here we provide the extension of Lemma 4.9 on ϵ-greedy algorithm.

Corollary C.1 (The ϵ-greedy case)**.** *Consider ϵ-greedy with exploration rates $\{\epsilon_s\}_{s=1}^T$. Define*

$$B_i(t,T,\delta) \triangleq \sum_{s=t+1}^{T} \frac{\epsilon_s}{K} + \sqrt{2\left(\sum_{s=t+1}^{T} \frac{\epsilon_s}{K}\right)\log\frac{1}{\delta} + \frac{1}{3}\log\frac{1}{\delta}}.$$

Then ϵ-greedy has the post-suppression pull bound $B_i(t,T,\delta)$, and the condition in Equation (4) becomes

$$n_i > \frac{U_i(t) - L_K(t)}{L_K(t) - r_{\min}}\big(N_i^0(t) + B_i(t,T,\delta/K)\big).$$

Under this condition, with probability at least $1-\delta-\frac{\delta}{K}$, $N_i(T) - N_i(t) \le B_i(t,T,\frac{\delta}{K})$. Applying the condition to all non-target arms suppresses them simultaneously with probability at least $1-2\delta$.

Proof. If $\hat{\mu}_i(s) < \hat{\mu}_K(s)$ throughout rounds $t,\dots,T$, then the arm cannot be selected by the greedy part of ϵ-greedy, so it can only be selected during exploration rounds. Let X_s be the indicator that round s is an exploration round and arm i is chosen. Then $\{X_s\}_{s=t+1}^T$ are independent Bernoulli random variables with means ϵ_s/K. Bernstein's inequality therefore gives

$$\sum_{s=t+1}^{T} X_s \le B_i(t,T,\delta/K)$$

with probability at least $1-\frac{\delta}{K}$, so $B_i(t,T,\frac{\delta}{K})$ is a valid post-suppression pull bound for this algorithm at confidence level δ/K. Applying Lemma 4.9 then yields the stated bound. The simultaneous guarantee follows by a union bound over the concentration event and the Bernstein events for all non-target arms. □

D General Suppression Framework Extensions

D.1 Concrete instantiations of the post-suppression bound

We next instantiate the general suppression framework for UCB and Thompson Sampling. For a simple instantiation, suppose that after the suppression step, throughout the remaining horizon the non-target arm has empirical mean at most μ_i^F and the target arm has empirical mean at least μ_K^F, where $g_i^F \triangleq \mu_K^F - \mu_i^F > 0$. Since our attack injects fake samples only into non-target arms, and these samples can only decrease their empirical means, we may take $\mu_K^F = L_K(t_i)$ and $\mu_i^F = \hat{\mu}_i(t_i) + 2\beta(N_i^0(t_i))$, with $\mu_i^F < \hat{\mu}_K(t_i) - 2\beta(N_K(t_i)) = L_K(t_i)$ after suppression.

For UCB, if arm i has count $N_i(s)$ at a later round $s \le T$, then its index is at most $\mu_i^F + 3\sigma\sqrt{\frac{\log T}{N_i(s)}}$, whereas the target arm's UCB index is at least $\mu_K^F + 3\sigma\sqrt{\frac{\log t_i}{T}}$. Indeed, for every later round $s \ge t_i$, we have $\log s \ge \log t_i$ and $N_K(s) \le T$, so the target-arm exploration bonus is lower bounded by $3\sigma\sqrt{\frac{\log t_i}{T}}$. Hence arm i cannot be selected once

$$N_i(s) \ge \frac{9\sigma^2 \log T}{(g_i^F + 3\sigma\sqrt{\log t_i/T})^2}.$$

Thus a valid post-suppression pull bound for the general framework is

$$B_i^{\mathrm{UCB}}(t,T,\delta/K) = \max\left\{\frac{9\sigma^2\log T}{(\mu_K^F - \mu_i^F + 3\sigma\sqrt{\log t_i/T})^2} - N_i(t), 0\right\}.$$

For Thompson Sampling, by adapting the standard finite-time analysis of Thompson Sampling [Agrawal and Goyal, 2017], one obtains that, with probability at least $1-\frac{\delta}{K}$, $N_i(T) - N_i(t) \le C\frac{\log(KT/\delta)}{(g_i^F)^2}$, for a universal constant $C > 0$. Hence we may take

$$B_i^{\mathrm{TS}}(t,T,\delta/K) = C\frac{\log(KT/\delta)}{(\mu_K^F - \mu_i^F)^2}.$$

This bound follows from the usual TS regret argument: after arm i has been sampled on the order of $\frac{\log(KT/\delta)}{(g_i^F)^2}$ times, its posterior sample exceeds the midpoint $\frac{\mu_i^F+\mu_K^F}{2}$ only with probability $O(\frac{\delta}{KT})$; the remaining pulls are controlled by the standard anti-concentration step for the target arm.

Substituting either B_i^{UCB} or B_i^{TS} into Theorem 4.10 gives a direct but coarser guarantee. The algorithm-specific UCB and Thompson Sampling attacks are sharper because they suppress the arm below a tailored lower confidence or posterior-sampling threshold and then use the exponential suppression lemmas, which eliminate the extra post-suppression term.

D.2 Proofs of the General Suppression Framework

Proof of Lemma 4.9. Let $\mathcal{G}_i(t,T,\frac{\delta}{K})$ be the event in Definition 4.8 for arm i with confidence parameter $\frac{\delta}{K}$. Fix an outcome in $\mathcal{E}\cap\mathcal{G}_i(t,T,\frac{\delta}{K})$, and write $B_i\equiv B_i(t,T,\frac{\delta}{K})$. Immediately after the injection, arm i has $N_i^0(t)+n_i$ samples in the learner's history. We will show that, as long as arm i has been pulled no more than B_i additional times after this injection, its empirical mean stays below the target-arm empirical mean. The post-suppression pull bound in Definition 4.8 then applies.

Let $m\triangleq N_i(s)-N_i(t)$ for some round $s\in\{t,t+1,\dots,T\}$, where $N_i(t)$ denotes the count after the n_i fake samples have been injected. Since no fake samples are injected into arm i after round t, all samples added after round t are genuine. On $\mathcal{E}$, every genuine empirical mean of arm i is upper bounded by $\mu_i+\beta(\cdot)$; in particular, the combined empirical mean of the first $N_i^0(t)+m$ genuine samples of arm i is at most $\mu_i+\beta(N_i^0(t)+m)\le\mu_i+\beta(N_i^0(t))\le\hat{\mu}_i^0(t)+2\beta(N_i^0(t))=U_i(t)$, where we used that $\beta(\cdot)$ is nonincreasing and $\mathcal{E}$ also gives $\mu_i\le\hat{\mu}_i^0(t)+\beta(N_i^0(t))$. Hence

$$\hat{\mu}_i(s)\le\frac{U_i(t)\big(N_i^0(t)+m\big)+n_i r_{\min}}{N_i^0(t)+n_i+m}. \tag{19}$$

Now suppose $m\le B_i$. Since Equation (4) holds and $r_{\min}<L_K(t)$, we have

$$n_i(L_K(t)-r_{\min})>(U_i(t)-L_K(t))(N_i^0(t)+B_i)\ge(U_i(t)-L_K(t))(N_i^0(t)+m).$$

Equivalently, $U_i(t)(N_i^0(t)+m)+n_i r_{\min}<L_K(t)(N_i^0(t)+n_i+m)$. Combining this with Equation (19) yields $\hat{\mu}_i(s)<L_K(t)$. On $\mathcal{E}$, for every $s\ge t$,

$$\hat{\mu}_K(s)\ge\mu_K-\beta(N_K^0(s))\ge\mu_K-\beta(N_K^0(t))\ge\hat{\mu}_K^0(t)-2\beta(N_K^0(t))=L_K(t).$$

Therefore, $\hat{\mu}_i(s)<\hat{\mu}_K(s)$ holds at every subsequent round s. Applying this argument recursively, we conclude that $N_i(T)-N_i(t)\le B_i$. This proves the single-arm claim on $\mathcal{E}\cap\mathcal{G}_i(t,T,\delta/K)$, which has probability at least $1-\delta-\delta/K$. Applying the same argument to all non-target arms and using a union bound over $\mathcal{E}$ and the events $\mathcal{G}_i(t,T,\delta/K)$ gives

$$\mathbb{P}\left(\mathcal{E}\cap\bigcap_{i=1}^{K-1}\mathcal{G}_i(t,T,\delta/K)\right)\ge1-\delta-\frac{K-1}{K}\delta\ge1-2\delta.$$

Thus all non-target arms are suppressed simultaneously with probability at least $1-2\delta$. □

Proof of Theorem 4.10. For each non-target arm i, let t_i be the first round at which the attacker suppresses arm i. Since $r_{\min}<\mu_K-3\beta(1)$, on $\mathcal{E}$ we have $r_{\min}<L_K(t_i)$. Write $B_i^\delta(t_i)=B_i\left(t_i,T,\frac{\delta}{K}\right)$ and $M_i(t_i)=N_i^0(t_i)+B_i^\delta(t_i)$. The attacker injects

$$n_i=\frac{U_i(t_i)-L_K(t_i)}{L_K(t_i)-r_{\min}}M_i(t_i)+1$$

fake samples of arm i, each with reward value $r_{\min}$. The choice of n_i ensures that the sufficient condition in Equation (4) holds with confidence parameter $\frac{\delta}{K}$. We first bound the number of injected samples explicitly. The assumption $r_{\min}<\mu_K-3\beta(1)$ implies $r_{\min}<\mu_K-3\beta(N_K^0(t_i))$ for every trigger time. On $\mathcal{E}$, $U_i(t_i)\le\mu_i+3\beta(N_i^0(t_i))$ and $L_K(t_i)\ge\mu_K-3\beta(N_K^0(t_i))$. Therefore,

$$\frac{U_i(t_i)-r_{\min}}{L_K(t_i)-r_{\min}}\le\frac{\mu_i+3\beta(N_i^0(t_i))-r_{\min}}{\mu_K-3\beta(N_K^0(t_i))-r_{\min}}\le\frac{\mu_i+3\beta(1)-r_{\min}}{\mu_K-3\beta(1)-r_{\min}},$$

and $\frac{U_i(t_i)-L_K(t_i)}{L_K(t_i)-r_{\min}} \le \frac{\Delta_i+6\beta(1)}{\mu_K-3\beta(1)-r_{\min}}$. Therefore $n_i \le \frac{\Delta_i+6\beta(1)}{\mu_K-3\beta(1)-r_{\min}} M_i(t_i)+1$. Consequently,

$$N^F(T) = \sum_{i=1}^{K-1} N_i^F(T) \le \sum_{i=1}^{K-1} n_i \le \sum_{i=1}^{K-1} \left[\frac{\Delta_i + 6\beta(1)}{\mu_K - 3\beta(1) - r_{\min}} M_i(t_i) + 1 \right].$$

Next, by Lemma 4.9, on the event $\mathcal{E} \cap \bigcap_{i=1}^{K-1} \mathcal{G}_i\left(t_i, T, \frac{\delta}{K}\right)$, each non-target arm i is selected at most $B_i^{\delta}(t_i)$ additional times after its injection. Before the injection, arm i contributes $N_i^0(t_i)$ genuine pulls, and the injection itself contributes n_i fake pulls to the learner's history. Thus the total number of non-target-arm pulls in the learner's history is at most $\sum_{i=1}^{K-1}(M_i(t_i)+n_i)$, and

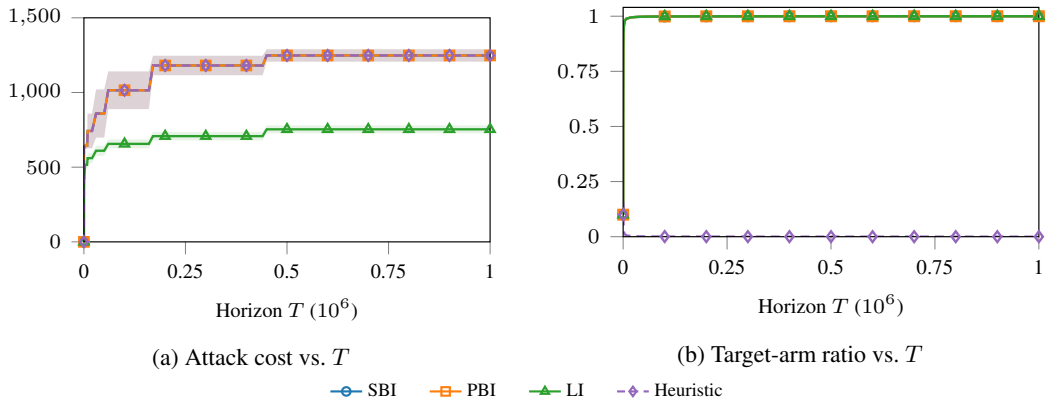
$$M_i(t_i) + n_i \le \frac{U_i(t_i) - r_{\min}}{L_K(t_i) - r_{\min}} M_i(t_i) + 1.$$

This gives the exact-statistic lower bound on the number of target-arm selections. Since $\mathbb{P}(\mathcal{E}) \ge 1-\delta$ and each $\mathcal{G}_i(t_i, T, \frac{\delta}{K})$ fails with probability at most $\frac{\delta}{K}$, a union bound gives success probability at least $1-2\delta$. Substituting the first ratio bound into the exact-statistic pull bound gives the displayed lower bound on target-arm selections. The cumulative reward-deviation cost is:

$$C^F(T) = \sum_{i=1}^{K-1} (\mu_i - r_{\min}) N_i^F(T) \le \sum_{i=1}^{K-1} (\mu_i - r_{\min}) \left[\frac{\Delta_i + 6\beta(1)}{\mu_K - 3\beta(1) - r_{\min}} M_i(t_i) + 1 \right].$$

This yields the displayed cost bound without a separate concentration argument. Because the argument holds for any feasible choice of t_i, the attacker may choose the trigger times that minimize the resulting upper bound. □

E Additional Experiments



(a) Attack cost vs. T

(b) Target-arm ratio vs. T

Figure 2: TS experiments with 20 repetitions. Shading shows one sample standard deviation.

TS follows the shared settings in Section 5, with arm means $(0.9, 0.85, \ldots, 0.45)$ and the smallest-mean arm as target. Genuine rewards are Gaussian with $\sigma = 0.05$; posterior samples have variance $\frac{1}{N_i(t)}$. We use $\delta_0 = 1$. At $T = 10^6$, SBI and PBI reach 99.99% target-arm selection ratio with an average of 17.65 fake samples, while the matched-cost heuristic reaches 0.01%. LI reaches 99.99% with 9 samples.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper introduces the Fake Data Injection threat model to address limitations of prior adversarial bandit frameworks, and develops new attack strategies (SI, PBI) that explicitly account for bounded reward values and injection constraints. Theoretical guarantees and empirical results support the claims that these attacks can successfully mislead bandit algorithms like UCB and Thompson Sampling with sublinear attack cost. The abstract and introduction clearly state these contributions and align well with the content and findings of the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in Section 6.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We propose 2 attack algorithms: 2 tailored attacks for UCB and Thompson Sampling, respectively, and one general attack framework. The main paper presents our principal theoretical results, namely Theorems 4.5 and 4.7 for UCB, and Theorem 4.9 for the general attack. Detailed proofs are deferred to the appendix, where we also provide two additional theorems for Thompson Sampling.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive pseudocode and specify all corresponding hyperparameters to ensure reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have released our code and raw data on Github.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We discussed main experimental settings in Section 5. Full details are included in the Appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Yes, our figures in Section 5 report error bars. The results are consistent with the theory.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Given the lightweight nature of our algorithms, they are executable on most machines without specialized hardware, so detailed CPU or GPU specifications are unnecessary.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is a theoretical and empirical study of adversarial robustness; it does not involve human subjects, private data, or deployment of a harmful system, and we follow the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: As the work is primarily theoretical and methodological, it does not have immediate societal implications. However, future applications of the proposed methods may warrant further consideration of their broader impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release models or datasets with high misuse risk; the experiments use a public benchmark dataset and synthetic simulations.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used the MovieLens dataset, which is publicly available for research purposes. The dataset and its source are properly cited in the paper, and we have adhered to its terms of use.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, models, or software assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The experiments do not involve crowdsourcing or human-subject studies.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects, so IRB approval or equivalent review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only used for writing.