BEYOND REWARD SUPPRESSION: NEAR-OPTIMAL OFFLINE ATTACKS ON WARM-START BANDITS WITH BOUNDED REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial attacks on bandits aim to mislead a learner toward a target arm while keeping the attack cost small. Existing attacks typically achieve this by suppressing non-target arms. In practice, however, manipulation such as fake reviews often directly promotes the target item. We study this gap through bounded offline attacks on warm-start bandits, where an attacker can inject only valid action–reward pairs into the warm-start history before deployment. We show that target promotion is not merely a heuristic: when the target arm lies near the lower reward boundary, any order-optimal-cost attack against UCB that makes it selected in nearly all online rounds must allocate a nonvanishing fraction of its cost to the target arm. We then design an attack that achieves the optimal sublinear cost and characterize its allocation between target promotion and non-target suppression. We further extend the attack to Thompson Sampling, ϵ-greedy, and a broader class of bandit algorithms. Experiments on real-world and synthetic data validate the effectiveness of our attacks.

1 INTRODUCTION

Multi-armed bandit (MAB) algorithms provide a principled framework for sequential decision-making under uncertainty and have been widely adopted in applications such as recommender systems and online advertising. In many practical applications, however, bandit algorithms are not deployed from scratch. Instead, they are initialized with historical interaction data, giving rise to the warm-start bandit setting (Shivaswamy & Joachims, 2012; Li et al., 2021; Banerjee et al., 2022; Oetomo et al., 2023). While such historical data can improve early decisions and reduce exploration costs, they also create a natural attack surface: fake feedback injected before deployment can bias the learner's initial estimates, and this bias may persist even when all subsequent feedback is genuine.

Adversarial attacks on bandits have been studied extensively, but most existing work (Jun et al., 2018; Liu & Shroff, 2019; Zuo, 2024) focuses on manipulating rewards during online interaction, typically allowing arbitrary modifications to observed rewards and measuring attack cost by the total magnitude of these modifications. Zeng et al. (2025) takes a step further by injecting fake action–reward records to model fake users. Building on these works, we consider a more natural setting in which the learner is initialized with offline logs before online deployment, and the attacker can inject fake action–reward records only during this offline stage. We further measure attack cost by the number of injected samples, which more closely reflects the practical cost of generating fake interactions. This bounded fake-data perspective not only better reflects settings such as review manipulation, but also makes how the attacker allocates its limited effort across arms a central question.

More importantly, existing attacks on stochastic bandits typically steer the learner toward a target arm by suppressing the observed rewards of non-target arms (Jun et al., 2018; Zuo et al., 2023; Zuo, 2024; Zeng et al., 2025). Although such attacks admit strong theoretical guarantees with controlled cost, their mechanism differs from common real-world manipulation practices, where attackers often directly promote a chosen item through favorable feedback. For example, researchers have observed Amazon sellers recruiting reviewers through private Facebook groups and reimbursing purchases in exchange for five-star reviews, sometimes with additional commissions (He et al., 2022b). Direct promotion is also intuitively attractive: a favorable review benefits the target itself, whereas sup-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of bandit attack models and guarantees.

Attack Strategy	Timing	Bounded Rewards	Target Promotion	Allocation Analysis	Opt.
Jun et al. (2018)	online	✗	✗	✗	✗
Liu & Shroff (2019)	offline	✗	✓	✗	✓
Xu et al. (2021)	online	✓	✓	✗	✗
Zuo (2024)	online	✗	✗	✗	✓
Zeng et al. (2025)	online	✓	✗	✗	✗
Hosseini et al. (2026)	offline	✗	✗	✗	✗
Ours (Sec. 4)	offline	✓	✓	✓	✓

Prior work separately studies offline poisoning, bounded attacks, or attack-cost optimality, but does not characterize how attack effort should be allocated across arms. Our work combines bounded offline injection with direct target promotion and proves that cost-optimal attacks require substantial effort on the target arm itself. The online variant of Liu & Shroff (2019) is omitted since it does not satisfy any of the properties in the columns.

pressing a single competitor affects only one of many alternatives. This gap raises a natural question: when is directly promoting the target not only effective, but necessary for a cost-efficient attack?

Motivated by this mismatch, we study bounded offline attacks on warm-start stochastic bandits under two practical constraints. First, every injected reward must lie within the valid feedback range, such as binary clicks or 1–5 star ratings. Second, the attacker can intervene only before deployment, since maintaining access throughout the online learning process may be difficult or impossible. Under these constraints, target promotion and non-target suppression play fundamentally different roles. Injecting favorable feedback on the target improves its comparison against all competitors, but this initial advantage is gradually diluted as genuine online feedback accumulates. In contrast, suppressing a competitor requires injecting unfavorable feedback specifically for that arm, and bounded rewards limit how much each injected sample can change its empirical estimate. The attacker must therefore decide not only how many samples to inject, but also how to allocate these injections across arms. This leads to our central question:

How should an attacker allocate its power between the target arm and its competitors?

We answer this question by jointly optimizing target promotion and non-target suppression. We propose a bounded offline attack against warm-start UCB and establish both an upper bound on its cost and a lower bound for any successful attack. Our lower bound analysis further reveals that the optimal allocation is necessarily target-centric: any attack achieving the leading-order optimal cost that makes the target arm selected in nearly all online rounds must spend a nonvanishing fraction of its cost on promoting the target. We further extend our strategy to the class of suppression-bounded algorithms studied by Zeng et al. (2025), with algorithm-specific analyses for Thompson Sampling and ϵ-greedy in Appendices B and C. These results demonstrate the broader applicability of our strategy beyond UCB.

1.1 Our Contributions

We summarize our contributions as follows:

- We formulate *bounded offline attacks on warm-start stochastic bandits*, measuring attack cost by the number of valid action–reward samples injected before deployment. This model captures settings such as fake-review manipulation, where the cost is associated with generating fake interactions rather than the magnitude of reward modification.
- We establish *upper and lower bounds* on the cost of bounded offline attacks against UCB. In the near-boundary regime, these bounds match to leading order, yielding the optimal sublinear cost scale $\Theta\big(T^{2/3}(K\log T)^{1/3}\big)$ for K arms and T rounds. Notably, we bridge this theory–practice gap by showing that cost-optimal attacks must allocate a nonvanishing fraction of their cost to the target arm.
- We extend our attack strategy beyond UCB to Thompson Sampling, ϵ-greedy, and the broader class of *suppression-bounded* algorithms introduced by Zeng et al. (2025). This

extension shows that the same combination of target promotion and non-target suppression applies to a broad class of bandit algorithms.

- We evaluate our attacks on MovieLens (Harper & Konstan, 2015) against existing baselines. The experiments demonstrate lower attack costs while maintaining effective target-arm manipulation, and empirically validate the predicted target-centric cost allocation.

1.2 Related Work

Research on adversarial manipulation of bandits follows two complementary directions: designing attack algorithms, and developing robust algorithms against adversarial attack.

Adversarial Attacks. Jun et al. (2018) showed that sublinear-cost attacks can steer UCB and ϵ-greedy toward a target arm, while Liu & Shroff (2019) developed an adaptive attack without knowledge of the learner's algorithm. Zuo (2024) established near-optimal attacks against UCB and TS with matching lower bounds. Other work studies restricted intervention: Xu et al. (2021) proposed an observation-free attack during prescribed online phases, and Zeng et al. (2025) introduced bounded fake data injection with constraints on injection timing and frequency. A common strategy in this line of work is to favor the target by suppressing the observed rewards of non-target arms. Attacks have also been studied in contextual, linear, adversarial, and learning-to-rank bandits (Garcelon et al., 2020; Wang et al., 2022; Ma & Zhou, 2023; Zuo et al., 2023). Beyond these works, Liu & Shroff (2019) formulated offline reward-poisoning attacks as convex optimization problems for several bandit algorithms. Hosseini et al. (2026) studied reward-model perturbations before training an offline bandit, without subsequent online interaction.

Robust Bandits. Lykouris et al. (2018) established corruption-dependent regret bounds for stochastic bandits without prior knowledge of the corruption level, and Gupta et al. (2019) improved these guarantees. Kapoor et al. (2019) used robust statistical estimation to tolerate corrupted rewards in multi-armed and linear contextual bandits. For linear bandits, robust phased elimination (Bogunovic et al., 2021) and confidence-weighted regression (He et al., 2022a) control the influence of corrupted observations, with guarantees that account for whether the corruption level is known. Robustness has also been studied in Gaussian process (Bogunovic et al., 2022), Lipschitz (Kang et al., 2023), and combinatorial bandits (Xu & Li, 2021). The 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. Our work follows the attack-design direction and adopts the fake data injection model of Zeng et al. (2025) in an offline-to-online setting.

2 Warm-Start Bandit Setup

We consider a stochastic multi-armed bandit with arm set $\mathcal{K} \triangleq \{1, \dots, K\}$. Each arm $i \in \mathcal{K}$ is associated with an unknown reward distribution with mean μ_i. Rewards are supported on $[r_{\min}, r_{\max}]$ and are σ^2-sub-Gaussian, where σ^2 is known to the learner.

The learner is initialized with an offline log $\mathcal{D}^{\text{off}} \triangleq \big((a_s, r_s)\big)_{s=1}^{T_0}$ containing T_0 action–reward observations, where $a_s \in \mathcal{K}$ denotes the arm selected in observation round s, and $r_s \in [r_{\min}, r_{\max}]$ denotes the corresponding observed reward. For each arm i, let $N_i^{\text{off}} \triangleq \sum_{s=1}^{T_0} \mathbf{1}\{a_s = i\}$ denote its offline count. After initialization, the learner interacts with the bandit environment from round $T_0 + 1$ through round T. Thus, the online rounds are $\mathcal{T}^{\text{on}} \triangleq \{T_0 + 1, \dots, T\}$, with online horizon $H \triangleq |\mathcal{T}^{\text{on}}| = T - T_0$.

At the beginning of each online round $t \in \mathcal{T}^{\text{on}}$, the learner's history consists of the offline log and the preceding online observations:

$$\mathcal{H}_t \triangleq \Big(\mathcal{D}^{\text{off}}, \big((a_s, r_s)\big)_{s=T_0+1}^{t-1}\Big).$$

A general warm-start bandit algorithm is specified by a policy sequence $\Pi \triangleq \{\pi_t\}_{t \in \mathcal{T}^{\text{on}}}$, where $\pi_t(\cdot \mid \mathcal{H}_t) \in \Delta(\mathcal{K})$. In round t, the learner samples $a_t \sim \pi_t(\cdot \mid \mathcal{H}_t)$, observes the corresponding reward r_t, and appends (a_t, r_t) to its history. We write $N_i^{\text{on}}(t) \triangleq \sum_{s=T_0+1}^{t-1} \mathbf{1}\{a_s = i\}$ for the number of online pulls of arm i before round t.

UCB Specialization. We next instantiate this general framework using the UCB algorithm considered by Jun et al. (2018), following the prototype in Bubeck & Cesa-Bianchi (2012, Section 2.2). UCB summarizes the available history $\mathcal{H}_t$ through the number of observations and the empirical mean associated with each arm. For every arm $i \in \mathcal{K}$, define $N_i(t) \triangleq \sum_{s=1}^{t-1} \mathbf{1}\{a_s = i\} = N_i^{\text{off}} + N_i^{\text{on}}(t)$ and $\hat{\mu}_i(t) \triangleq \sum_{s=1}^{t-1} \mathbf{1}\{a_s = i\} r_s / N_i(t)$ whenever $N_i(t) > 0$. The UCB index of arm i is

$$u_i(t) \triangleq \begin{cases} \hat{\mu}_i(t) + 3\sigma\sqrt{\frac{\log t}{N_i(t)}}, & N_i(t) > 0, \\ +\infty, & N_i(t) = 0. \end{cases}$$

The learner selects $a_t \in \arg\max_{i \in \mathcal{K}} u_i(t)$. In the absence of attacks, it achieves sublinear regret.

3 Offline Attack Model

3.1 Attack Setting and Attack Cost

Let $\mathcal{D}^{\text{cl}} \triangleq \big((a_s, r_s)\big)_{s=1}^{T_0^{\text{cl}}}$ denote the clean offline log, where $T_0^{\text{cl}} = |\mathcal{D}^{\text{cl}}|$ is the number of genuine observations. For each arm, its clean-log and subsequent online rewards form a prefix of the same i.i.d. reward stream. Arm choices may depend on past observations; each chosen arm's next reward is then observed and retained, with no filtering based on its value. Before deployment, the attacker observes $\mathcal{D}^{\text{cl}}$ and knows the learner's policy Π, the terminal time T, and the valid reward range $[r_{\min}, r_{\max}]$. The attacker acts only at this stage and cannot modify the learner's actions, observed rewards, or history after online interaction begins.

For each arm i, define its clean offline count and empirical mean as

$$N_i^{\text{cl}} \triangleq \sum_{s=1}^{T_0^{\text{cl}}} \mathbf{1}\{a_s = i\}, \qquad \hat{\mu}_i^{\text{cl}} \triangleq \frac{1}{N_i^{\text{cl}}} \sum_{s=1}^{T_0^{\text{cl}}} \mathbf{1}\{a_s = i\} r_s,$$

where the empirical mean is defined whenever $N_i^{\text{cl}} > 0$. During online interaction, $N_i^{\text{cl}}(t)$ and $\hat{\mu}_i^{\text{cl}}(t)$ denote the count and empirical mean of all genuine observations of arm i available before round t, including the clean offline observations but excluding injected samples. Thus, at the start of online interaction, $N_i^{\text{cl}}(T_0 + 1) = N_i^{\text{cl}}$ and $\hat{\mu}_i^{\text{cl}}(T_0 + 1) = \hat{\mu}_i^{\text{cl}}$.

A candidate injection plan is represented by $(\mathbf{n}, \mathbf{y})$, where $\mathbf{n} \triangleq (n_1, \ldots, n_K) \in \mathbb{Z}_{\geq 0}^K$ specifies the number of fake observations injected for each arm and $\mathbf{y} \triangleq (y_{i,j} : i \in \mathcal{K},\ j = 1, \ldots, n_i)$ specifies the corresponding fake rewards. Every injected reward must lie in $[r_{\min}, r_{\max}]$ and, for discrete feedback, in the legal feedback alphabet.

Let $\mathcal{D}^{\text{F}}(\mathbf{n}, \mathbf{y})$ denote the collection of fake action–reward pairs $\{(i, y_{i,j}) : i \in \mathcal{K},\ j = 1, \ldots, n_i\}$ in the order chosen by the strategy. The learner is initialized using the attacked warm-start log

$$\mathcal{D}^{\text{off}}(\mathbf{n}, \mathbf{y}) \triangleq \mathcal{D}^{\text{cl}} \,\|\, \mathcal{D}^{\text{F}}(\mathbf{n}, \mathbf{y}),$$

where $\|$ denotes sequence concatenation.

Definition 3.1 (Attack Cost)**.** The attack cost is the number of injected samples:

$$\mathcal{C}(\mathbf{n}) \triangleq \sum_{i=1}^{K} n_i. \tag{1}$$

Consequently, $N_i^{\text{off}} = N_i^{\text{cl}} + n_i$ and

$$T_0 = \left|\mathcal{D}^{\text{off}}\right| = T_0^{\text{cl}} + \mathcal{C}(\mathbf{n}), \qquad H = T - T_0.$$

Without loss of generality, we label the *target arm* as K.

Definition 3.2 ((ξ, δ)-Success Criterion)**.** Fix $\delta \in (0, 1)$ and $\xi \colon \mathbb{N} \to \mathbb{R}_{\geq 0}$. An offline attack is (ξ, δ)-successful if

$$\Pr\left(\sum_{i=1}^{K-1} N_i^{\text{on}}(T + 1) \leq \xi\right) \geq 1 - \delta \quad \text{and} \quad \mathcal{C}(\mathbf{n}) = o(T). \tag{2}$$

We summarize the offline injection workflow in Algorithm 1; concrete allocation strategies are developed in the next section.

Algorithm 1: Offline Injection Attack

Input: $\mathcal{D}^{\mathrm{cl}}$, K, Π, T, $[r_{\min}, r_{\max}]$, ξ and δ, Target Arm (K in our analysis)
Output: Attacked warm-start log $\mathcal{D}^{\mathrm{off}}$

Determine an injection plan $(\mathbf{n}, \mathbf{y})$ satisfying Eq. (2);
Construct a fake log $\mathcal{D}^{\mathrm{F}}(\mathbf{n}, \mathbf{y})$ containing $(i, y_{i,j})$ for every $i \in \mathcal{K}$ and $j = 1, \ldots, n_i$;
Set $\mathcal{D}^{\mathrm{off}} \leftarrow \mathcal{D}^{\mathrm{cl}} \,\|\, \mathcal{D}^{\mathrm{F}}(\mathbf{n}, \mathbf{y})$;
return $\mathcal{D}^{\mathrm{off}}$;

4 Attack Strategies

We first instantiate Algorithm 1 for UCB, then extend the strategy to suppression-bounded algorithms from Zeng et al. (2025).

4.1 Attack on UCB

We consider the zero-tolerance case $\xi = 0$, where the attacker aims to make UCB select the target arm in every online round. The *high-level idea* is to (1) enlarge the gap between the target arm and the non-target arms, and (2) make the exploration bonus sufficiently small for each non-target arm, so that $u_K(t) > u_i(t)$ throughout the online phase. To ensure this, we seek to solve

$$\min_{\mathbf{n},\mathbf{y}} \quad \mathcal{C}(\mathbf{n}) \quad \text{subject to} \quad u_K(t) > u_i(t), \qquad \forall i < K,\ t \in \mathcal{T}^{\mathrm{on}}.$$

Let $R \triangleq r_{\max} - r_{\min}$ and assume $N_i^{\mathrm{cl}} \geq 1$ for every arm. Write $b_i(t) = 3\sigma\sqrt{\log t / N_i(t)}$. Our construction keeps the target arm's empirical mean above a threshold z and bounds each non-target arm's UCB index by $z + b_K(t)$. Specifically, we enforce

$$u_i(t) = \hat{\mu}_i(t) + b_i(t) \leq z + b_K(t) < \hat{\mu}_K(t) + b_K(t) = u_K(t), \qquad \forall i < K,\ t \in \mathcal{T}^{\mathrm{on}}.$$

For a confidence parameter $\delta \in (0, 1/2)$, define

$$\beta_\delta(N) \triangleq \sqrt{\frac{2\sigma^2}{N} \log \frac{\pi^2 K N^2}{3\delta}}, \qquad \underline{\mu}_K \triangleq \max\{r_{\min}, \hat{\mu}_K^{\mathrm{cl}} - 2\beta_\delta(N_K^{\mathrm{cl}})\}.$$

Set $\varepsilon \triangleq \underline{\mu}_K - r_{\min}$. We choose

$$z_T^\star = r_{\min} + \left(\frac{18R\sigma^2(K-1)\log T}{T}\right)^{1/3}, \qquad z = r_{\min} + \max\left\{z_T^\star - r_{\min},\, \min\{\varepsilon, R/2\}\right\}.$$

Write $B_i = N_i^{\mathrm{cl}}(\hat{\mu}_i^{\mathrm{cl}} - r_{\min})$ and $d(z) = z - r_{\min} + 3\sigma\sqrt{\log T / T}$. We set

$$n_i = \begin{cases} \left\lceil \dfrac{T(z - \underline{\mu}_K)_+}{r_{\max} - z} \right\rceil + 1, & \text{for } i = K; \\ \left\lceil \left[\left(\dfrac{3\sigma\sqrt{\log T} + \sqrt{9\sigma^2 \log T + 4d(z)B_i}}{2d(z)} \right)^2 - N_i^{\mathrm{cl}} \right]_+ \right\rceil, & \text{for } i < K. \end{cases} \tag{3}$$

Using these counts, construct the fake log in Algorithm 1 as

$$\mathcal{D}^{\mathrm{F}} = \left(\Big\|_{i=1}^{K-1} \big((i, r_{\min})\big)_{j=1}^{n_i} \right) \Big\| \big((K, r_{\max})\big)_{j=1}^{n_K}.$$

We analyze this construction under the following scaling assumptions.

Assumption 4.1. $R = \Theta(1)$, $\sigma = \Theta(1)$, and $N_i^{\mathrm{cl}} = \Theta(1)$ uniformly over $i \in \mathcal{K}$.

In particular, $T_0^{\mathrm{cl}} = \Theta(K) = o(T)$, so every attack satisfying $\mathcal{C}(\mathbf{n}) = o(T)$ has $H = T(1 - o(1))$. Define $S_T \triangleq T^{2/3}(K \log T)^{1/3}$. We now give the theoretical guarantee of our attack strategy.

Theorem 4.2. *Fix $\delta \in (0, 1/2)$. Under Assumption 4.1, for all sufficiently large T, the above attack satisfies*

$$\Pr\left(\sum_{i=1}^{K-1} N_i^{\mathrm{on}}(T+1) = 0\right) \geq 1-\delta \quad \textit{and} \quad \mathcal{C}(\mathbf{n}) = \mathcal{O}\left(\frac{K \log T}{(\varepsilon \vee (S_T/T))^2}\right).$$

Proof Sketch. If $\varepsilon \leq z_T^\star - r_{\min}$, then $z = z_T^\star$; substituting into Eq. (3) gives $n_K = O(S_T)$ and $\sum_{i<K} n_i = O(S_T)$. If $\varepsilon > z_T^\star - r_{\min}$, then $z \leq \underline{\mu}_K$ and $z - r_{\min} = \Theta(\varepsilon)$, giving $n_K = 1$ and $\sum_{i<K} n_i = O(K \log T/\varepsilon^2)$. Combining the two cases yields the stated cost bound. □

Theorem 4.3. *Under Assumption 4.1, fix $\delta \in (0, 1/2)$. Let $\Delta_K \triangleq \mu_K - r_{\min}$. For all sufficiently large T, every $(0, \delta)$-successful bounded offline attack on UCB satisfies*

$$\mathcal{C}(\mathbf{n}) = \Omega\left(\frac{K \log T}{(\Delta_K \vee (S_T/T))^2}\right)$$

with probability at least $1 - \delta - 1/T$.

Proof Sketch. Consider the final round T, when UCB selects the target on the zero-tolerance success event. By then, target injections have been diluted by $\Theta(T)$ genuine observations. Writing $u_K(T) = r_{\min} + \rho$, we have $\rho \leq \Delta_K + O(n_K/T + \sqrt{\log T/T})$ with high probability. Since the target is selected at T, every competitor satisfies $u_i(T) \leq u_K(T)$. Its empirical mean is at least $r_{\min}$, so its exploration bonus cannot exceed ρ, forcing $N_i^{\mathrm{cl}} + n_i \geq 9\sigma^2 \log T/\rho^2$. Combining these constraints on target and non-target injection costs and minimizing over ρ yields the cost lower bound. □

The upper and lower bounds depend on ε and Δ_K, respectively. Their gap-dependent rates match when $\varepsilon = \Omega(\Delta_K)$. When $\varepsilon = \Omega(1)$, our upper and lower bounds match at $\Theta(K \log T)$. This cost order also agrees with the attack costs of Zeng et al. (2025).

Our bounds identify S_T/T as the *boundary scale* for the target arm's gap from the lower reward bound $r_{\min}$. We refer to the regime $\Delta_K = o(S_T/T)$ as the *near-boundary regime*. In this regime, we further sharpen the bounds to match in their leading constants, thereby establishing the asymptotic optimality of our construction. The optimal leading cost is obtained by minimizing the sum of the leading costs of target promotion and non-target suppression. This minimization yields the closed-form value

$$\Lambda_T \triangleq 3\left(\frac{9\sigma^2(K-1)T^2 \log T}{4R^2}\right)^{1/3} = \Theta(S_T).$$

Corollary 4.4. *Under Assumption 4.1, fix $\delta \in (0, 1/2)$ and assume $\Delta_K = o(\Lambda_T/T)$. For all sufficiently large T, our construction has $\mathcal{C}(\mathbf{n}) \leq (1+o(1))\Lambda_T$. And every $(0, \delta)$-successful bounded offline attack on UCB satisfies*

$$\Pr\left[\mathcal{C}(\mathbf{n}) \geq (1 - o(1))\Lambda_T\right] \geq 1 - \delta - \frac{1}{T}.$$

We next characterize the cost allocated to the target arm and show that this allocation is necessary for asymptotic optimality.

Lemma 4.5. *Under Assumption 4.1, fix $\delta \in (0, 1/2)$. For $\varepsilon = o(\Lambda_T/T)$, our construction satisfies $n_K = \left(\frac{2}{3} + o(1)\right)\Lambda_T$. If $\Delta_K = o(\Lambda_T/T)$, then for all sufficiently large T, every $(o(\Lambda_T), \delta)$-successful bounded offline attack on UCB with $\mathcal{C}(\mathbf{n}) \leq (1 + o(1))\Lambda_T$ satisfies*

$$\Pr\left[n_K = \left(\frac{2}{3} + o(1)\right)\Lambda_T\right] \geq 1 - \delta - \frac{1}{T}.$$

Remark 4.6. The $2:1$ cost split balances target promotion against non-target suppression. In the near-boundary regime, raising the target's empirical mean has a leading cost linear in its elevation above the lower reward bound. UCB's $N_i^{-1/2}$ exploration bonuses make each competitor's required injection count, and hence their sum, decrease as the inverse square of that elevation. Minimizing the sum of these leading cost terms yields the $2:1$ split between target and aggregate non-target costs. Appendix E compares this mechanism across algorithms.

For larger tolerances $\xi = O(\Lambda_T)$, Lemma A.2 establishes that near-boundary successful attacks with a deterministic $O(\Lambda_T)$ cost bound must still incur a nonvanishing fraction of their total cost on the target arm. Proofs of Theorems 4.2 and 4.3 appear in Appendices A.1 and A.2, respectively; those of Corollary 4.4 and Lemma 4.5 appear in Appendix A.3.

In the near-boundary regime, our construction is *target-centric*: for any fixed number of competitors, injections on the target account for most of the total cost with high probability, as required for leading-order optimality. This emphasis on favorable feedback for the target qualitatively reflects real-world manipulation practices, such as sellers purchasing positive reviews for their own products (He et al., 2022b).

Comparison to Xu et al. (2021). In the near-boundary regime, the prescribed cost of Xu et al. (2021)'s UCB-based attack scales as $\Theta(K \log T/\Delta_K^2)$. Consequently, the cost is $\Omega(S_T)$ when $\Delta_K = \mathcal{O}(S_T/T)$, ceases to be sublinear when $\Delta_K = \mathcal{O}(\sqrt{K \log T/T})$, and is undefined at $\Delta_K = 0$. In contrast, our attack remains applicable and covers this boundary case.

4.2 Extension to Suppression-Bounded Algorithms

We extend the strategy to algorithms whose additional pulls of persistently dominated arms admit a high-probability bound.

Definition 4.7 (Suppression-Bounded Algorithms from Zeng et al. (2025))**.** Fix a horizon T and 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

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

then

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

The quantity $B_i^{\mathcal{A}}(t, T, \delta)$ is referred to as the *post-suppression pull bound* for arm i. This is a conservative upper bound and $B_i^{\mathcal{A}}(t, T, \delta) \ge B_i^{\mathcal{A}}(t, T', \delta)$ for all $T' < T$.

Theorem 4.8 (Generic Suppression-Bounded Attack)**.** *Suppose $\mathcal{A}$ is suppression-bounded and write $B_i^{\mathcal{A}} \triangleq B_i^{\mathcal{A}}(T_0^{\mathrm{cl}}, T, \delta/K)$. Assume $N_i^{\mathrm{cl}} \ge 1$ for every arm. There exists a bounded offline attack that satisfies:*

$$\Pr\left(\sum_{i=1}^{K-1} N_i^{\mathrm{on}}(T+1) \le \sum_{i=1}^{K-1} B_i^{\mathcal{A}}\right) \ge 1 - 2\delta$$

and

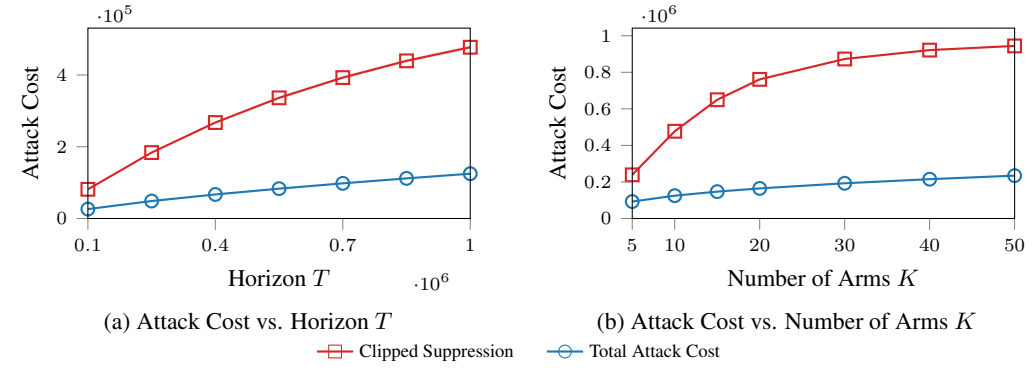
$$\mathcal{C}(\mathbf{n}) \le \mathcal{O}\left(\inf_{\rho \in (0,R)} \left\{ K + T\frac{(\rho - \varepsilon)_+}{R - \rho} + \frac{R - \rho}{\rho} \sum_{i=1}^{K-1} \left(N_i^{\mathrm{cl}} + B_i^{\mathcal{A}}\right) \right\}\right).$$

The construction raises the target's empirical mean above $r_{\min} + \rho$ and keeps each non-target mean at or below this threshold. The target injections counteract dilution by genuine feedback. The non-target injections offset the clean observations and up to $B_i^{\mathcal{A}}$ online observations, even if each has reward $r_{\max}$. Each such observation requires approximately $(R - \rho)/\rho$ compensating injections at $r_{\min}$.

Under Assumption 4.1, if $\sum_{i<K} B_i^{\mathcal{A}} = o(T)$, let $A_T \triangleq \sum_{i<K}(N_i^{\mathrm{cl}} + B_i^{\mathcal{A}}) = o(T)$. Choosing $\rho = R\sqrt{A_T/T}/(1 + \sqrt{A_T/T})$ and using $\varepsilon \ge 0$ gives

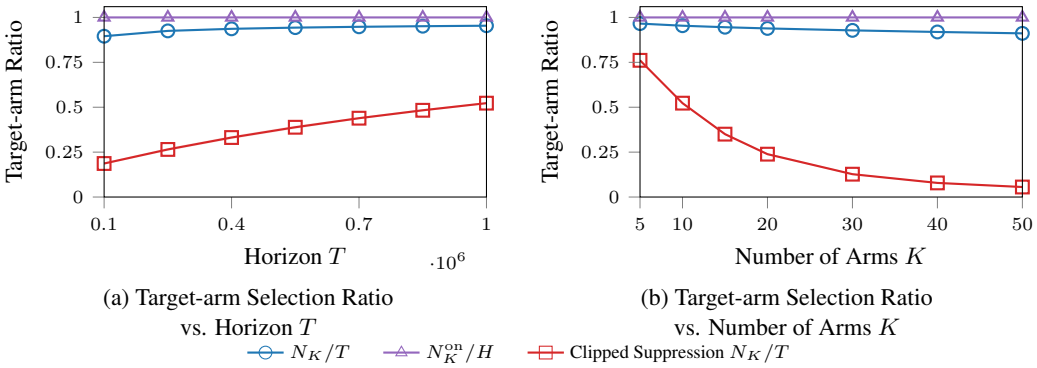
$$\mathcal{C}(\mathbf{n}) = O\left(K + \sqrt{TA_T}\right) = o(T).$$

Consequently, $H = T(1 - o(1))$ and $N_K^{\mathrm{on}}(T+1)/H = 1 - o(1)$ with probability at least $1 - 2\delta$, yielding a successful attack. UCB, TS, and ϵ-greedy all fall into the scope of suppression-bounded algorithms (Zeng et al., 2025).



(a) Attack Cost vs. Horizon T (b) Attack Cost vs. Number of Arms K

Figure 1: Attack Cost for UCB.



(a) Target-arm Selection Ratio vs. Horizon T (b) Target-arm Selection Ratio vs. Number of Arms K

Figure 2: Target-arm Selection Ratio for UCB.

5 Numerical Illustration

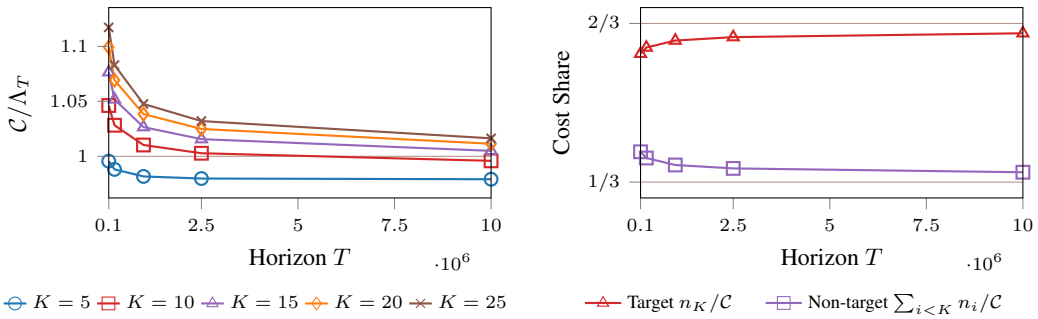
We evaluate our attacks on MovieLens-25M (Harper & Konstan, 2015), mapping ratings of at least four to one and the rest to zero. Each movie defines a Bernoulli arm with its empirical positive-rating rate as the mean; clean and online rewards are sampled independently with replacement. The target, *Glitter* (ID 4775), has the smallest positive mean, $\mu_K = 11/669$; competitors are the most-rated remaining movies. We use $N_i^{\text{cl}} = 5$, $\sigma = 0.5$, and $\delta = 0.05$.

Figure 1 compares our attack strategy with *Clipped Suppression* by varying T at $K = 10$ and K at $T = 10^6$. Our method consistently incurs a substantially lower attack cost. As T increases, the observed cost growth is consistent with the sublinear bound in Theorem 4.2. In contrast, *Clipped Suppression* intervenes whenever a non-target arm is pulled, so its attack cost is $T - N_K$ and grows faster with T. The cost gap also increases as K grows.

Figure 2 further shows that our method achieves lower attack cost while maintaining success. For our method, $N_K^{\text{on}}/H = 1$ in every repeat across all tested values of T and K, meaning that UCB selects the target arm in every online round after deployment. The overall ratio N_K/T, which also includes the warm-start samples, remains close to 1. In contrast, *Clipped Suppression* yields more non-target pulls. Even at $T = 10^6$, the learner selects non-target arms in nearly half of the rounds, and its target-arm selection ratio decreases further as K increases.

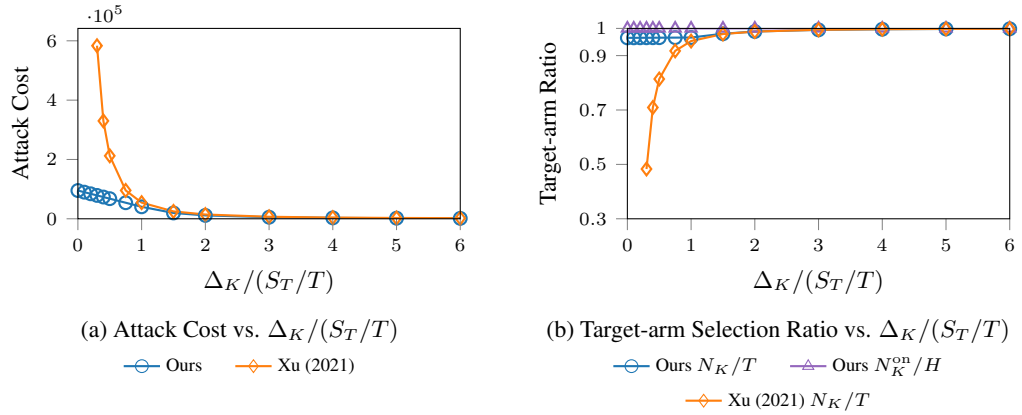
We next illustrate the leading-order optimality and target-centric allocation predicted by our theory. As shown in Figure 3(a), for each fixed K, the normalized attack cost $\mathcal{C}(\mathbf{n})/\Lambda_T$ approaches 1 as T increases, consistent with Corollary 4.4. In Figure 3(b), at $K = 10$, the target-arm cost share $n_K/\mathcal{C}(\mathbf{n})$ approaches $2/3$, consistent with Lemma 4.5.

We then compare our method with the attack of Xu et al. (2021) under the same experimental setting. When $\Delta_K/(S_T/T) < 1$, the target lies within the boundary scale S_T/T, where attacks become challenging. Our attack maintains sublinear cost in this regime and makes UCB select the target arm



(a) Normalized Attack Cost vs. Horizon T (b) Cost Share vs. Horizon T

Figure 3: Near-Boundary Cost and Allocation for UCB.



(a) Attack Cost vs. $\Delta_K/(S_T/T)$ (b) Target-arm Selection Ratio vs. $\Delta_K/(S_T/T)$

Figure 4: Comparison with Xu et al. (2021) for UCB.

in every online round after deployment with high probability. In contrast, the attack cost of Xu et al. (2021) scales as $O(K \log T/\Delta_K^2)$ and grows rapidly as Δ_K approaches zero. Hence, their points are omitted when this cost is undefined at $\Delta_K = 0$ or is at least T, while their target-arm selection ratio remains below 1 in the most challenging near-boundary settings.

These experiments demonstrate the effectiveness of our method, with cost shares consistent with the theoretical analysis. We also provide experiments for TS and ϵ-greedy in Appendix B.2 and Appendix C.2, respectively.

6 Concluding Remarks

In the near-boundary regime, target-centric allocation is necessary for successful UCB attacks with tolerance $\xi = O(S_T)$ and a deterministic $O(S_T)$ cost cap: the target must receive a nonvanishing share of the injections on the high-probability event of our lower bound. Our UCB construction attains this cost scale in the near-boundary regime. We also give constructions for TS and ϵ-greedy in the Appendix, while the suppression-bounded analysis covers a broader class of learners.

However, our analysis assumes a learner that uncritically trusts all offline data. A possible direction for future work is to investigate whether discounting, filtering, or auditing offline observations can mitigate the effects of favorable target injections, and how such defenses trade off the benefits of a clean warm start against robustness to corrupted initialization. Furthermore, our cost and allocation results assume a small clean offline log. More clean observations make empirical means harder to shift but also reduce exploration uncertainty, potentially changing both the optimal attack cost and its allocation. Characterizing how these competing effects depend on the size and distribution of clean observations across arms is an interesting direction for future work.

AI USE STATEMENT

We used generative AI tools to help polish and organize the manuscript and to assist in reviewing the proofs and code. The authors have carefully verified all AI-assisted content and take full responsibility for the final text, theoretical claims, code, and reported results.

ETHICS STATEMENT

This work studies vulnerabilities of bandit learners to support robustness evaluation. The attack methods could also be misused to manipulate recommendations or other sequential decisions. Our experiments are confined to simulations using synthetic rewards and publicly available MovieLens data; no attacks were conducted against live systems or users. The released code is intended for controlled research and authorized evaluation, and does not redistribute the MovieLens dataset.

REPRODUCIBILITY STATEMENT

The model and assumptions are specified in Section 3 and the theoretical results, with proofs in Appendices A–D. Experimental settings are described in Section 5 and Appendices B.2 and C.2. The accompanying source-code supplement includes data-processing and simulation scripts, along with instructions for obtaining MovieLens and reproducing the experiments. Experiment runners record configurations and random seeds in provenance manifests.

REFERENCES

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

Siddhartha Banerjee, Sean R Sinclair, Milind Tambe, Lily Xu, and Christina Lee Yu. Artificial replay: a meta-algorithm for harnessing historical data in bandits. *arXiv preprint arXiv:2210.00025*, 2022.

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

Ilija Bogunovic, Zihan Li, Andreas Krause, and Jonathan Scarlett. A robust phased elimination algorithm for corruption-tolerant gaussian process bandits. *Advances in Neural Information Processing Systems*, 35: 23951–23964, 2022.

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

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

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 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. *Advances in neural information processing systems*, 35:34614–34625, 2022a.

Sherry He, Brett Hollenbeck, Gijs Overgoor, Davide Proserpio, and Ali Tosyali. Detecting fake-review buyers using network structure: Direct evidence from amazon. *Proceedings of the National Academy of Sciences*, 119(47):e2211932119, 2022b. doi: 10.1073/pnas.2211932119.

Seyed Mohammad Hadi Hosseini, Amir Najafi, and Mahdieh Baghshah. Efficient adversarial attacks on high-dimensional offline bandits. In *International Conference on Learning Representations*, volume 2026, pp. 37587–37612, 2026.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. *Advances in neural information processing systems*, 31, 2018.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Robust lipschitz bandits to adversarial corruptions. *Advances in Neural Information Processing Systems*, 36:10897–10908, 2023.

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

Ye Li, Hong Xie, Yishi Lin, and John CS Lui. Unifying offline causal inference and online bandit learning for data driven decision. In *Proceedings of the Web Conference 2021*, pp. 2291–2303, 2021.

Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pp. 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*, pp. 114–122, 2018.

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

Bastian Oetomo, R Malinga Perera, Renata Borovica-Gajic, and Benjamin IP Rubinstein. Cutting to the chase with warm-start contextual bandits. *Knowledge and Information Systems*, 65(9):3533–3565, 2023.

Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Artificial intelligence and statistics*, pp. 1046–1054. PMLR, 2012.

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

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22550–22561. Curran Associates, Inc., 2021.

Qirun Zeng, Eric He, Richard Hoffmann, Xuchuang Wang, and Jinhang Zuo. Practical adversarial attacks on stochastic bandits via fake data injection. *arXiv preprint arXiv:2505.21938*, 2025.

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

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

CONTENTS

1 Introduction **1**

1.1 Our Contributions . . . 2

1.2 Related Work . . . 3

2 Warm-Start Bandit Setup **3**

3 Offline Attack Model **4**

3.1 Attack Setting and Attack Cost . . . 4

4 Attack Strategies **5**

4.1 Attack on UCB . . . 5

4.2 Extension to Suppression-Bounded Algorithms . . . 7

5 Numerical Illustration **8**

6 Concluding Remarks **9**

Appendix **12**

A Proofs for UCB **12**

A.1 Upper Bound Proof . . . 13

A.2 Lower Bound Proof . . . 13

A.3 Optimality Proofs . . . 14

B Analysis for Thompson Sampling **16**

B.1 Attack Strategy . . . 16

B.2 Numerical Illustration . . . 17

C Analysis for ϵ-Greedy **18**

C.1 Attack Strategy . . . 18

C.2 Numerical Illustration . . . 19

D Analysis of Suppression-Bounded Algorithms **20**

E Exploration and Allocation **20**

A PROOFS FOR UCB

For $\delta \in (0,1)$, recall that

$$\beta_\delta(N) = \sqrt{\frac{2\sigma^2}{N} \log \frac{\pi^2 K N^2}{3\delta}}.$$

Define the event

$$\mathcal{E}_\delta \triangleq \left\{ \forall i \in \mathcal{K},\ \forall t \in \mathcal{T}^{\text{on}},\quad \left|\hat{\mu}_i^{\text{cl}}(t) - \mu_i\right| < \beta_\delta(N_i^{\text{cl}}(t)) \right\}.$$

By Jun et al. (2018, Lemma 1), $\Pr(\mathcal{E}_\delta) \geq 1-\delta$. The radius is decreasing on the positive integers when $K \geq 2$ and $\delta \in (0,1)$. On $\mathcal{E}_\delta$, we have $0 \leq \varepsilon \leq \Delta_K$, and

$$\begin{aligned}\hat{\mu}_K^{\mathrm{cl}}(t) &\geq \max\{r_{\min}, \mu_K - \beta_\delta(N_K^{\mathrm{cl}}(t))\} \\ &\geq \max\{r_{\min}, \hat{\mu}_K^{\mathrm{cl}} - 2\beta_\delta(N_K^{\mathrm{cl}})\} = \underline{\mu}_K \qquad (\forall t \in \mathcal{T}^{\mathrm{on}}).\end{aligned}$$

A.1 Upper Bound Proof

Under the event $\mathcal{E}_\delta$ and by Eq. (3), at round $t = T_0 + 1$, we have

$$\hat{\mu}_K(t) \geq \frac{n_K r_{\max} + N_K^{\mathrm{cl}}(t)\underline{\mu}_K}{n_K + N_K^{\mathrm{cl}}(t)} \geq \frac{n_K r_{\max} + T\underline{\mu}_K}{n_K + T} > z.$$

Since $N_K(t) \leq t-1$ and $\log t/(t-1)$ decreases for $t > 1$, we have $u_K(t) = \hat{\mu}_K(t) + 3\sigma\sqrt{\frac{\log t}{N_K(t)}} > z + 3\sigma\sqrt{\frac{\log T}{T}}$. Similarly, for every non-target arm $i < K$,

$$u_i(t) \leq z + 3\sigma\sqrt{\frac{\log T}{T}}.$$

Together, these inequalities imply that

$$u_K(t) > u_i(t), \qquad \forall i < K,$$

and hence UCB selects the target arm at round t. Repeating the same argument inductively from round $T_0 + 1$ through round T shows that UCB selects the target arm in every online round.

Proof of Theorem 4.2. Write $\rho = z - r_{\min}$, $r_T = S_T/T$, and $\rho_T^\star = z_T^\star - r_{\min} = \Theta(r_T)$. We have

$$n_K = \frac{T(\rho-\varepsilon)_+}{R-\rho} + O(1) \leq \frac{T\rho}{R-\rho} + 2.$$

Since $d(r_{\min} + \rho) = \rho + 3\sigma\sqrt{\log T/T} \sim \rho$ and $B_i, N_i^{\mathrm{cl}} = O(1)$ uniformly, Eq. (3) gives

$$n_i = (1+o(1))\frac{9\sigma^2 \log T}{\rho^2} \qquad (i < K).$$

(1) If $\varepsilon \leq \rho_T^\star$, then $\rho = \rho_T^\star$, and the preceding estimates give $n_K = O(Tr_T)$ and $n_i = O(\log T/r_T^2)$ uniformly over $i < K$. Their sum is $O(S_T)$, the asserted scale since $\varepsilon = O(r_T)$.

(2) If $\varepsilon > \rho_T^\star$, then $\rho = \min\{\varepsilon, R/2\} \leq \varepsilon$ and $\rho = \Theta(\varepsilon)$. Thus $n_K = 1$, while Eq. (3) gives

$$n_i = O\left(1 + B_i/\rho + \sigma^2 \log T/\rho^2\right) = O(\log T/\varepsilon^2) \qquad (i < K).$$

Combining these two cases yields the result. □

A.2 Lower Bound Proof

Lemma A.1. *Under Assumption 4.1, there exist constants $c_0, c_1, c_2, c_3 > 0$ such that the following holds for all sufficiently large T. On $\mathcal{E}_{1/T}$, every bounded offline UCB attack realization with at most $\xi \geq 0$ non-target online pulls and $\mathcal{C}(\mathbf{n}) + \xi \leq T/2$ admits a $\rho > 0$ satisfying*

$$\begin{aligned} n_K &\geq c_1 T(\rho - \Delta_K - c_0\sqrt{\log T/T})_+, \\ \sum_{i<K} n_i &\geq \left(c_2 \frac{K \log T}{\rho^2} - \xi - c_3 K\right)_+. \end{aligned}$$

Proof. Since $H - \xi \geq T/2 - T_0^{\mathrm{cl}} > 0$ for large T, the target is selected in at least one online round. Let $t^\star$ be its last such round and set $\rho = u_K(t^\star) - r_{\min} > 0$. We have

$$\begin{aligned} N_K(t^\star) &= N_K^{\mathrm{off}} + N_K^{\mathrm{on}}(T+1) - 1, \\ N_K^{\mathrm{cl}}(t^\star) &= N_K^{\mathrm{cl}} + N_K^{\mathrm{on}}(T+1) - 1. \end{aligned}$$

The bound $N_K^{\mathrm{on}}(T+1) \geq H - \xi$ gives $N_K^{\mathrm{cl}}(t^\star) \geq T/3$ for large T. Since all rounds after $t^\star$ select non-target arms, we also have $t^\star \geq T - \xi \geq T/2$.

On $\mathcal{E}_{1/T}$, the genuine target observations have empirical mean at most $\mu_K + \beta_{1/T}(N_K^{\mathrm{cl}}(t^\star))$, and each injected target reward is at most $r_{\max}$. Meanwhile, every non-target empirical mean is at least $r_{\min}$, and its index is at most the selected target's index $r_{\min} + \rho$. These comparisons give

$$\begin{aligned} \rho &\leq \Delta_K + \frac{(R - \Delta_K)n_K}{N_K(t^\star)} + \beta_{1/T}(N_K^{\mathrm{cl}}(t^\star)) + 3\sigma\sqrt{\frac{\log t^\star}{N_K(t^\star)}}, \\ N_i(t^\star) &\geq \frac{9\sigma^2 \log t^\star}{\rho^2} \qquad (i < K). \end{aligned} \tag{4}$$

The concentration radius and target exploration bonus are both $O(\sqrt{\log T/T})$ for large T. Thus the first inequality gives

$$\rho \leq \Delta_K + \frac{3Rn_K}{T} + c_0\sqrt{\frac{\log T}{T}}.$$

Rearranging and using $R = \Theta(1)$ and $n_K \geq 0$ yields the target cost constraint. Write $M = \sum_{i<K} n_i$. Summing the second inequality and using $N_i(t^\star) = N_i^{\mathrm{cl}} + n_i + N_i^{\mathrm{on}}(t^\star)$ and $\sum_{i<K} N_i^{\mathrm{on}}(t^\star) \leq \xi$ gives

$$M + T_0^{\mathrm{cl}} + \xi \geq \sum_{i<K} N_i(t^\star) \geq \frac{9\sigma^2(K-1)\log t^\star}{\rho^2}.$$

Using $K - 1 \geq K/2$, $t^\star \geq T/2$, $T_0^{\mathrm{cl}} = O(K)$, and $M \geq 0$ yields the non-target cost constraint with uniform constants $c_2, c_3 > 0$. □

Proof of Theorem 4.3. Work on the intersection of the attack's success event with $\mathcal{E}_{1/T}$, which has probability at least $1 - \delta - 1/T$ by a union bound.

Recall $r_T = S_T/T$ and set $a_T = (\Delta_K + c_0\sqrt{\log T/T}) \vee r_T$, with c_0 from Lemma A.1. On paths with $\mathcal{C}(\mathbf{n}) \leq T/2$, add the two constraints in Lemma A.1 at $\xi = 0$. Enlarging the offset to a_T only decreases the linear term, giving

$$\mathcal{C}(\mathbf{n}) + c_3 K \geq \inf_{\rho > 0} \left\{ c_1 T(\rho - a_T)_+ + \frac{c_2 K \log T}{\rho^2} \right\}.$$

(1) If $\rho \leq 2a_T$, the inverse-square term is at least $c_2 K \log T/(4a_T^2)$.

(2) Otherwise, the linear term is at least $c_1 T a_T \geq c_1 K \log T/a_T^2$, since $a_T \geq r_T$ and $T r_T^3 = K \log T$.

Both cases give $\Omega(K \log T/(\Delta_K \vee r_T)^2)$, because $a_T \sim \Delta_K \vee r_T$ by $\sqrt{\log T/T} = o(r_T)$. □

A.3 Optimality Proofs

Proof of Corollary 4.4. Upper bound. Recall $\rho_T^\star = z_T^\star - r_{\min}$. The cost expansions in Appendix A.1 give, uniformly over clean logs and including rounding,

$$\mathcal{C}(\mathbf{n}(z_T^\star)) \leq (1 + o(1)) \left\{ \frac{T\rho_T^\star}{R} + \frac{9\sigma^2(K-1)\log T}{(\rho_T^\star)^2} \right\} = (1 + o(1))\Lambda_T.$$

(1) If $\varepsilon \leq \rho_T^\star$, our construction uses $z = z_T^\star$.

(2) Otherwise, the target count is one at both thresholds, and raising z only reduces non-target counts.

Lower bound. Consider a $(0, \delta)$-successful attack on the intersection of its success event with $\mathcal{E}_{1/T}$, which has probability at least $1 - \delta - 1/T$. It suffices to consider paths with $\mathcal{C}(\mathbf{n}) \leq 2\Lambda_T$, since larger costs already satisfy the conclusion. Write $M = \sum_{i<K} n_i$. Zero tolerance gives $N_K^{\mathrm{on}}(T+1) = H$ and $t^\star = T$. Since $\mathcal{C}(\mathbf{n}) + T_0^{\mathrm{cl}} = o(T)$, the count identities in the proof of Lemma A.1 give

$N_K(T), N_K^{\mathrm{cl}}(T) = T(1-o(1))$. Set $\rho = u_K(T) - r_{\min}$. Since $\Delta_K = o(\Lambda_T/T)$ and $\sqrt{\log T/T} = o(\Lambda_T/T)$, the index constraints in Eq. (4) imply, uniformly over these paths,

$$\rho \le \frac{R}{T}\big(n_K + o(\Lambda_T)\big), \qquad M + T_0^{\mathrm{cl}} \ge \frac{9\sigma^2(K-1)\log T}{\rho^2}.$$

Eliminating ρ gives

$$\big(n_K + o(\Lambda_T)\big)^2 (M + T_0^{\mathrm{cl}}) \ge \frac{9\sigma^2(K-1)T^2\log T}{R^2} = \frac{4\Lambda_T^3}{27}.$$

Since $n_K, M = O(\Lambda_T)$ and $T_0^{\mathrm{cl}} = o(\Lambda_T)$, this yields

$$n_K^2 M \ge (1-o(1))\frac{4\Lambda_T^3}{27}. \tag{5}$$

AM–GM applied to $n_K/2, n_K/2, M$ gives

$$(1-o(1))\frac{4\Lambda_T^3}{27} \le n_K^2 M \le \frac{4(n_K+M)^3}{27} = \frac{4\mathcal{C}(\mathbf{n})^3}{27},$$

so $\mathcal{C}(\mathbf{n}) \ge (1-o(1))\Lambda_T$ on the stated event. □

Proof of Lemma 4.5. $\varepsilon = o(\Lambda_T/T)$ gives $\varepsilon = o(\rho_T^\star)$ and hence $z = z_T^\star$. Eq. (3) then gives

$$n_K = \frac{T(\rho_T^\star - \varepsilon)}{R - \rho_T^\star} + O(1) \sim \frac{T\rho_T^\star}{R} = \frac{2}{3}\Lambda_T.$$

This is deterministic under the stated clean-log condition, which follows from $\Delta_K = o(\Lambda_T/T)$ on $\mathcal{E}_\delta$ because $0 \le \varepsilon \le \Delta_K$.

Necessity. Work on the intersection of the attack's success event with $\mathcal{E}_{1/T}$, which has probability at least $1 - \delta - 1/T$. Write $M = \sum_{i<K} n_i$, let $t^\star$ be the last online round selecting the target, and set $\rho = u_K(t^\star) - r_{\min}$. Since $\xi = o(\Lambda_T)$ and $\mathcal{C}(\mathbf{n}) \le (1+o(1))\Lambda_T$, we still have $t^\star \ge T - \xi$, $\log t^\star \sim \log T$, and $N_K(t^\star), N_K^{\mathrm{cl}}(t^\star) = T(1-o(1))$. The preceding product argument therefore applies with $M + T_0^{\mathrm{cl}} + \xi$ in the non-target constraint. Because $T_0^{\mathrm{cl}} + \xi = o(\Lambda_T)$, eliminating ρ again yields Eq. (5). Together with $n_K^2 M \le 4\mathcal{C}(\mathbf{n})^3/27$ and the cost cap, this implies $\mathcal{C}(\mathbf{n}) \sim \Lambda_T$.

Set $x = n_K/\mathcal{C}(\mathbf{n}) \in [0,1]$. Dividing the product bound by $\mathcal{C}(\mathbf{n})^3$ gives

$$\frac{4}{27} - o(1) \le x^2(1-x) \le \frac{4}{27}.$$

The identity

$$\frac{4}{27} - x^2(1-x) = \left(x - \frac{2}{3}\right)^2\left(x + \frac{1}{3}\right) \ge \frac{1}{3}\left(x - \frac{2}{3}\right)^2$$

forces $x = 2/3 + o(1)$, and hence

$$n_K = x\mathcal{C}(\mathbf{n}) = \left(\frac{2}{3} + o(1)\right)\Lambda_T.$$

□

Lemma A.2 (Necessary Target Cost)**.** *Under Assumption 4.1, fix $\delta \in (0, 1/2)$ and assume $\Delta_K = o(\Lambda_T/T)$ and $\xi = O(\Lambda_T)$. For all sufficiently large T, every (ξ,δ)-successful bounded offline attack on UCB with a deterministic $O(\Lambda_T)$ cost bound satisfies*

$$n_K = \Omega(\Lambda_T), \qquad \frac{n_K}{\mathcal{C}(\mathbf{n})} = \Omega(1)$$

with probability at least $1 - \delta - 1/T$.

Proof. Work on the intersection of success with $\mathcal{E}_{1/T}$, which has probability at least $1 - \delta - 1/T$. Since $\mathcal{C}(\mathbf{n}) + \xi = O(\Lambda_T) = o(T)$, apply Lemma A.1 and use its notation. Since $\sum_{i<K} n_i + \xi + c_3 K = O(\Lambda_T)$, the non-target cost constraint gives $\rho = \Omega(\Lambda_T/T)$. With $\Delta_K, \sqrt{\log T/T} = o(\Lambda_T/T)$, the target cost constraint gives $n_K = \Omega(\Lambda_T)$. The deterministic cost bound then yields $n_K/\mathcal{C}(\mathbf{n}) = \Omega(1)$. □

B Analysis for Thompson Sampling

B.1 Attack Strategy

We consider the Thompson Sampling (TS) algorithm of Zeng et al. (2025), based on the (α, ψ) framework of Agrawal & Goyal (2017). In each online round, it independently samples $\nu_i(t) \sim \mathcal{N}(\hat{\mu}_i(t), 1/N_i(t))$ for every arm and selects

$$a_t = \arg\max_{i \in \mathcal{K}} \nu_i(t).$$

For $\delta \in (0,1)$, define

$$\gamma(t) \triangleq \sqrt{2 \log \frac{\pi^2 K t^2}{3\delta}}, \qquad \mathcal{F} \triangleq \left\{ \forall i \in \mathcal{K},\ \forall t \in \mathcal{T}^{\mathrm{on}},\ |\nu_i(t) - \hat{\mu}_i(t)| < \frac{\gamma(t)}{\sqrt{N_i(t)}} \right\}.$$

By Zeng et al. (2025, Lemma B.1), $\Pr(\mathcal{F}) \geq 1 - \delta$.

Fix $z \in (r_{\min}, r_{\max})$, write $\gamma = \gamma(T)$, and recall $B_i = N_i^{\mathrm{cl}}(\hat{\mu}_i^{\mathrm{cl}} - r_{\min})$. Inject $r_{\max}$ on the target and $r_{\min}$ on each non-target arm, with counts

$$n_i = \left\lceil \left[\left(\frac{\gamma + \sqrt{\gamma^2 + 4(z - r_{\min})B_i}}{2(z - r_{\min})} \right)^2 - N_i^{\mathrm{cl}} \right]_+ \right\rceil, \qquad i < K, \tag{6}$$

and

$$n_K = 1 + \min \left\{ n \in \mathbb{Z}_{\geq 0} :\ (r_{\max} - z)n \geq (z - \underline{\mu}_K)m + \gamma\sqrt{n+m},\ \forall m \in [N_K^{\mathrm{cl}}, T] \right\}. \tag{7}$$

Lemma B.1 (TS Threshold Certificate). *For any $z \in (r_{\min}, r_{\max})$, the above construction makes TS select the target in every online round on $\mathcal{E}_\delta \cap \mathcal{F}$.*

Proof. On $\mathcal{E}_\delta \cap \mathcal{F}$, suppose every preceding online pull selected the target. Applying the count constraints with $m = N_K^{\mathrm{cl}}(t)$ gives

$$\nu_K(t) > \frac{n_K r_{\max} + N_K^{\mathrm{cl}}(t)\underline{\mu}_K}{n_K + N_K^{\mathrm{cl}}(t)} - \frac{\gamma}{\sqrt{n_K + N_K^{\mathrm{cl}}(t)}} \geq z,$$

$$\nu_i(t) < r_{\min} + \frac{B_i}{N_i^{\mathrm{cl}} + n_i} + \frac{\gamma}{\sqrt{N_i^{\mathrm{cl}} + n_i}} \leq z \qquad (i < K).$$

And hence $\nu_K(t) > \nu_i(t)$. Induction from the first online round proves the claim. □

Theorem B.2 (Attack Cost Upper Bound and Allocation for TS). *Fix $\delta \in (0, 1/2)$. Under Assumption 4.1, define*

$$\Lambda_T^{\mathrm{ts}} \triangleq 3 \left(\frac{(K-1)\gamma(T)^2 T^2}{4R^2} \right)^{1/3} = \Theta(S_T).$$

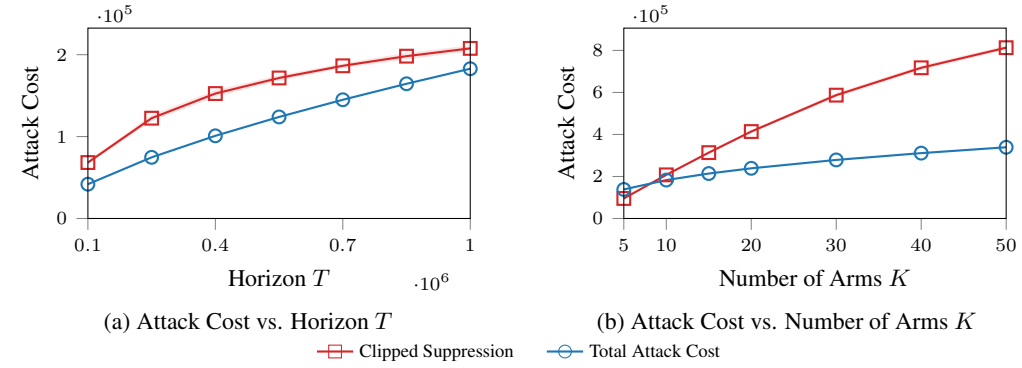
The following hold for all sufficiently large T.

If $\Delta_K = o(S_T/T)$, there exists a threshold for the above allocation satisfying

$$\Pr\left(\sum_{i=1}^{K-1} N_i^{\mathrm{on}}(T+1) = 0 \right) \geq 1 - 2\delta \quad \textit{and} \quad \mathcal{C}(\mathbf{n}) \leq (1 + o(1))\Lambda_T^{\mathrm{ts}}.$$

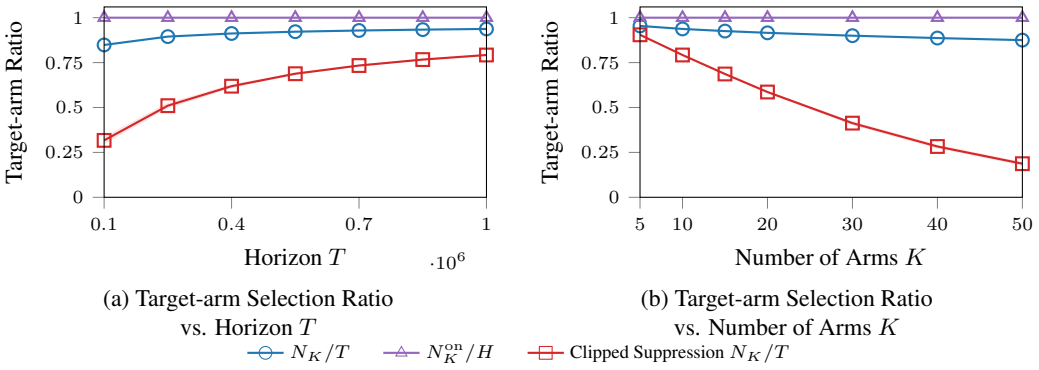
Moreover, under $\mathcal{E}_\delta$,

$$n_K = \left(\frac{2}{3} + o(1) \right) \Lambda_T^{\mathrm{ts}} \quad \textit{and} \quad n_i = (1 + o(1)) \frac{\Lambda_T^{\mathrm{ts}}}{3(K-1)} \textit{ for } i < K.$$



(a) Attack Cost vs. Horizon T (b) Attack Cost vs. Number of Arms K

Figure 5: Attack Cost for TS.



(a) Target-arm Selection Ratio vs. Horizon T (b) Target-arm Selection Ratio vs. Number of Arms K

Figure 6: Target-arm Selection Ratio for TS.

Proof of Theorem B.2. For $\rho = z - r_{\min} = \Theta(S_T/T)$, expanding the count formulas gives, uniformly over clean logs and including rounding,

$$n_K = \frac{T(\rho-\varepsilon)_+}{R} + O(T\rho^2 + \gamma\sqrt{T} + 1), \qquad n_i = (1+o(1))\frac{\gamma^2}{\rho^2} \quad (i < K).$$

The target lower bound follows by setting $m = T$ in its constraint; the upper bound uses $(\rho-\varepsilon)m \le T(\rho-\varepsilon)_+$ and $\sqrt{n+m} \le \sqrt{T} + n/(2\sqrt{T})$. The error is $o(S_T)$. Choose

$$\rho = \left(\frac{2R(K-1)\gamma^2}{T}\right)^{1/3}, \qquad \frac{T\rho}{R} = \frac{2}{3}\Lambda_T^{\text{ts}}, \qquad \frac{(K-1)\gamma^2}{\rho^2} = \frac{1}{3}\Lambda_T^{\text{ts}}.$$

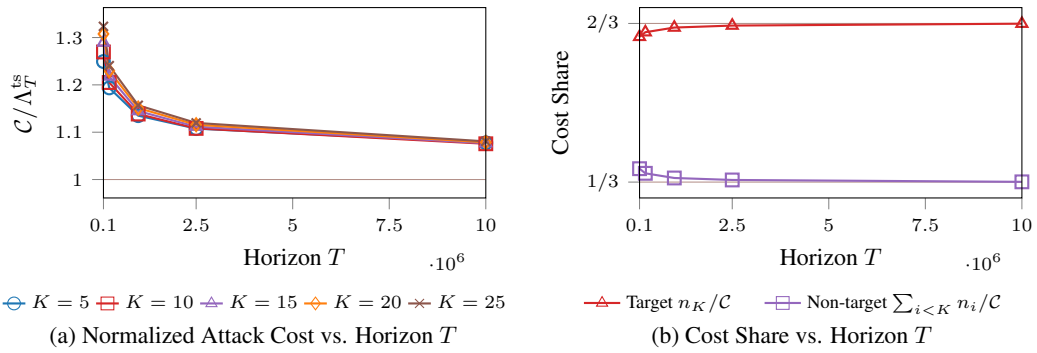
Since $(\rho-\varepsilon)_+ \le \rho$, the cost bound holds deterministically and $T_0^{\text{cl}} + \mathcal{C}(\mathbf{n}) = o(T)$. On $\mathcal{E}_\delta$, $0 \le \varepsilon \le \Delta_K = o(\rho)$, giving the stated allocation. Finally, Lemma B.1 and $\Pr(\mathcal{E}_\delta \cap \mathcal{F}) \ge 1 - 2\delta$ give the success guarantee. □

B.2 Numerical Illustration

We repeat the Clipped Suppression comparisons and near-boundary experiments of Section 5 with TS, and numerically minimize the continuous total count defined by Eq. (6) and Eq. (7) before rounding. All other settings and reporting conventions are unchanged.

Figures 5 and 6 show exclusively target online pulls for our attack in every repeat. Its cost is lower than Clipped Suppression except at the smallest arm count, where the baseline costs less but still pulls non-target arms.

Figure 7 shows normalized cost approaching one and cost shares approaching $2/3$ for the target and $1/3$ for the non-target arms collectively, consistent with the TS construction. Every repeat has zero non-target online pulls.



(a) Normalized Attack Cost vs. Horizon T (b) Cost Share vs. Horizon T

Figure 7: Near-Boundary Cost and Allocation for TS.

C Analysis for ϵ-Greedy

C.1 Attack Strategy

At round s, ϵ-greedy explores with probability ϵ_s by choosing uniformly from all K arms; otherwise it selects an empirical-mean maximizer. The schedule $\{\epsilon_s\}_{s=1}^T \subset [0,1]$ is deterministic, and exploration randomization is independent across rounds.

For simplicity, assume that the entire offline log is constructed by the attacker, so $N_i^{\text{cl}} = 0$ for every arm and $T_0 = \mathcal{C}(\mathbf{n})$.

For $0 \le t < T$ and $\delta \in (0,1)$, define the exploration bound

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

Lemma C.1 (Corollary C.1 of Zeng et al. (2025)). *Fix $T_0 \le t < T$ and $\delta \in (0,1)$. With probability at least $1-\delta$, the following implication holds: if $\hat{\mu}_i(s) < \max_j \hat{\mu}_j(s)$ for all $s \in \{t+1,\dots,T\}$, then*

$$N_i^{\text{on}}(T+1) - N_i^{\text{on}}(t+1) \le B_i^{\text{eg}}(t,T,\delta).$$

Fix $\delta \in (0,1/2)$, write $B = \lceil B_i^{\text{eg}}(0,T,\delta/(K-1))\rceil$ for short, and define

$$r_T^{\text{eg}} = R\sqrt{\frac{(K-1)B}{T}}.$$

Inject reward $r_{\max}$ on the target and $r_{\min}$ on each non-target arm, with

$$n_K = \left\lceil \sqrt{T(K-1)B} \right\rceil + 1, \qquad n_i = \left\lceil \sqrt{\frac{TB}{K-1}} \right\rceil \quad (i < K).$$

Theorem C.2 (ϵ-Greedy Upper Bound). *Fix $\delta \in (0,1/2)$. Suppose $R = \Theta(1)$, $\sum_{s=1}^T \epsilon_s + K\log(K/\delta) = o(T)$. For all sufficiently large T, the above attack satisfies*

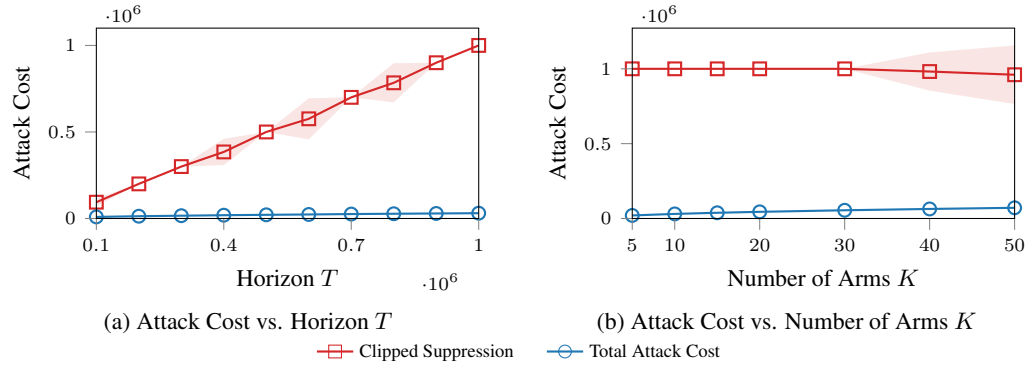
$$\Pr\left(\sum_{i<K} N_i^{\text{on}}(T+1) \le (K-1)B\right) \ge 1-\delta, \qquad \mathcal{C}(\mathbf{n}) = (2+o(1))\sqrt{T(K-1)B}.$$

All greedy online decisions select the target on the same event.

Proof. The definition of B gives $(K-1)B = O(\sum_s \epsilon_s + K\log(K/\delta)) = o(T)$, while rounding yields

$$2\sqrt{T(K-1)B} < \mathcal{C}(\mathbf{n}) \le 2\sqrt{T(K-1)B} + K + 1.$$

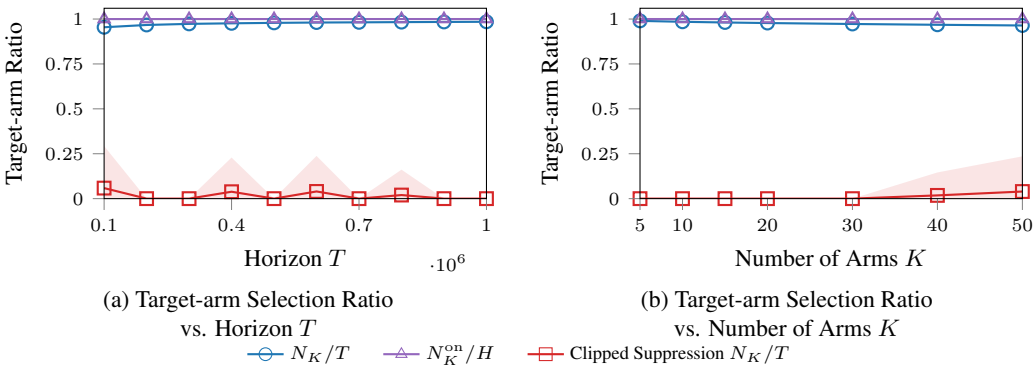
Since $B \ge 1$ and $K = o(T)$, the rounding term is lower order, proving the cost bound and $T_0 = \mathcal{C}(\mathbf{n}) < T$ eventually.



(a) Attack Cost vs. Horizon T

(b) Attack Cost vs. Number of Arms K

Figure 8: Attack Cost for ϵ-greedy.



(a) Target-arm Selection Ratio vs. Horizon T

(b) Target-arm Selection Ratio vs. Number of Arms K

Figure 9: Target-arm Selection Ratio for ϵ-greedy.

The Bernoulli bound underlying Lemma C.1 and a union bound ensure at most B potential exploratory selections per non-target arm over rounds $1, \dots, T$, with probability at least $1 - \delta$. On this event, if preceding greedy decisions selected the target, bounded rewards and $n_K n_i > TB$ give

$$\hat{\mu}_i(s) \le r_{\min} + \frac{RB}{n_i + B} < r_{\min} + \frac{Rn_K}{n_K + T} \le \hat{\mu}_K(s).$$

Induction from the first online round makes every greedy decision select the target, so all non-target pulls are exploratory and total at most $(K-1)B$. □

For $\epsilon_s = \min\{1, cK/s\}$ with fixed $c > 0$ and $K \log T = o(T)$, this gives $r_T^{\text{eg}} = \Theta(\sqrt{K \log T / T})$, cost $O(\sqrt{TK \log T})$, and (ξ, δ)-success with $\xi(T) = O(K \log T)$.

C.2 Numerical Illustration

We use Bernoulli rewards with non-target means equally spaced in $[0.2, 0.8]$ and target mean $r_T^{\text{eg}} / \log T > 0$, with $\epsilon_s = \min\{1, K/s\}$ and $\delta = 0.05$. Online statistics use 50 independent runs per configuration; shaded bands show one sample standard deviation. Greedy ties are broken by arm index.

Clipped Suppression sets every non-target online reward to zero and counts each such round as attack cost. It has no offline log and initializes with one online pull per arm, starting with the target.

Figures 8 and 9 vary T at $K = 10$ and K at $T = 10^6$. Our attack has lower mean costs and higher mean target-arm selection ratios. All greedy decisions selected the target; the remaining non-target pulls were exploratory and stayed below the theoretical bound.

D Analysis of Suppression-Bounded Algorithms

Proof of Theorem 4.8. Fix $\rho \in (0, R)$ and write $B_i^{\mathcal{A}} = B_i^{\mathcal{A}}(T_0^{\mathrm{cl}}, T, \delta/K)$. Choose

$$n_K = \left\lceil T\frac{(\rho-\varepsilon)_+}{R-\rho} \right\rceil + 1, \quad n_i = \left\lceil \frac{R-\rho}{\rho}\left(N_i^{\mathrm{cl}} + B_i^{\mathcal{A}}\right) \right\rceil, \text{ for all } i < K.$$

On $\mathcal{E}_\delta$, every empirical average of the genuine target-arm rewards is at least $\underline{\mu}_K = r_{\min} + \varepsilon$. Thus, for any $t \in \mathcal{T}^{\mathrm{on}}$,

$$\hat{\mu}_K(t) \geq \frac{n_K r_{\max} + N_K^{\mathrm{cl}}(t)\underline{\mu}_K}{n_K + N_K^{\mathrm{cl}}(t)} > r_{\min} + \rho.$$

For any bandit algorithm $\mathcal{A}$, the pull count $N_i(t)$, excluding injected samples, is nondecreasing in t. Thus, for each arm $i < K$, we use the conservative upper bound $B_i^{\mathcal{A}}$ uniformly over $t \in [T_0^{\mathrm{cl}}+1, T]$.

We now use induction over online rounds. Initially, $N_i^{\mathrm{on}}(T_0+1) = 0$. At any later round t, if arm i has been suppressed at every preceding online decision, we have $N_i^{\mathrm{on}}(t) \leq B_i^{\mathcal{A}}$. In either case, $N_i^{\mathrm{cl}}(t) = N_i^{\mathrm{cl}} + N_i^{\mathrm{on}}(t) \leq N_i^{\mathrm{cl}} + B_i^{\mathcal{A}}$, so, since all genuine rewards are at most $r_{\max}$,

$$\hat{\mu}_i(t) \leq \frac{r_{\max}N_i^{\mathrm{cl}}(t) + r_{\min}n_i}{N_i^{\mathrm{cl}}(t) + n_i} \leq r_{\min} + \rho.$$

Hence

$$\hat{\mu}_i(t) \leq r_{\min} + \rho < \hat{\mu}_K(t).$$

Intersecting $\mathcal{E}_\delta$ with the $K-1$ pull-bound events gives probability at least $1-2\delta$ by a union bound. On this event,

$$\sum_{i=1}^{K-1} N_i^{\mathrm{on}}(T+1) \leq \sum_{i=1}^{K-1} B_i^{\mathcal{A}}.$$

Summing the displayed injection counts and taking the infimum yields the claimed cost bound, including an $O(K)$ rounding overhead. □

E Exploration and Allocation

The allocation depends on how injections control exploration. Let $\rho = z - r_{\min}$ and $M = \sum_{i<K} n_i$. In the near-boundary regime, the genuine target rewards contribute little to maintaining the threshold, so the injections must sustain a mean increase of approximately ρ over T observations. This gives the leading target cost $n_K \sim T\rho/R$. The same expression follows from the worst-case reward bound used in our ϵ-greedy construction.

UCB and TS. For UCB, suppressing a competitor requires its exploration bonus to fit below the target's threshold. Since the bonus decays as $N_i^{-1/2}$, a threshold height ρ requires an observation count proportional to ρ^{-2}. Injected samples supply this count before online learning begins. TS has the same dependence because its sampling uncertainty also decays as $N_i^{-1/2}$. Because all competitors share the threshold ρ, summing their leading injection costs changes only the coefficient of ρ^{-2}, so $M = \sum_{i<K} n_i$ has the same dependence. Thus, for both constructions, raising ρ increases target cost linearly but reduces aggregate non-target cost as ρ^{-2}. A small relative increase in ρ adds that fraction of n_K and saves twice that fraction of M. At the minimizing threshold these amounts agree, giving $n_K \sim 2M$: the target accounts for $2/3$ of the total cost.

ϵ-greedy. Here exploration is prescribed by the schedule and cannot be removed by increasing observation counts. Each competitor can receive up to B exploratory rewards. Even if all equal $r_{\max}$, injecting n_i rewards equal to $r_{\min}$ keeps its mean below the threshold whenever

$$\frac{RB}{n_i + B} \leq \rho.$$

For small ρ, this requires approximately RB/ρ injections per competitor. Hence M decreases as ρ^{-1}: a small relative increase in ρ saves only that same fraction of M. Balancing this saving against the increase in target cost gives $n_K \sim M$, so the target accounts for $1/2$ of the total cost.