Adversarial Post-Action Attacks on Dueling Bandits

Mo Lyu
Department of Computer Science
University of California, Davis
Email: molyu@ucdavis.edu

Chenye Yang, Guanlin Liu and Lifeng Lai
Department of Electrical and Computer Engineering
University of California, Davis
Email: {cyyyang, glnliu, lflai}@ucdavis.edu

2025 59th Asilomar Conference on Signals, Systems, and Computers | 979-8-3315-8745-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IEEECONF67917.2025.11443548

***Abstract*—Dueling bandit algorithms excel in learning from pairwise comparisons, offering robust performance guarantees in benign environments. However, recent evidence suggests that even state-of-the-art methods can be highly susceptible to adversarial manipulation. In this work, we introduce and analyze a post-action attack model on the Relative Upper Confidence Bound (RUCB) algorithm, a widely used dueling bandit algorithm. Unlike pre-action attack considered in the existing work where the attacker can observe all comparisons beforehand, our post-action adversary intercepts only the feedback from the specific arm pair chosen by the learner at each round. Despite this limited access, we show that such targeted interference can coerce the learner into favoring a predetermined target arm for almost the entire time horizon. Specifically, the attacker incurs a total cost of only $\mathcal{O}(K \ln T)$ while ensuring that the learner pulls the target arm in $T - \mathcal{O}(K^2 \ln T)$ comparisons. These findings underscore the vulnerability of dueling bandit algorithms to post-action adversarial interference and highlight the need for more robust dueling bandits strategies.**

I. INTRODUCTION

Multi-armed bandit (MAB) problems introduced by [1] form a foundational framework in online learning and decision making, capturing the essential trade-off between exploring new actions to gather information and exploiting the best known option for immediate reward [2]. In a typical MAB setting, a learner faces a set of "arms" (actions) to choose from repeatedly, each yielding stochastic rewards according to an unknown probability distribution. MAB methods have proven indispensable in a wide variety of applications: for example, recommendation systems leverage bandits to personalize content by adapting to user feedback in real time; advertising platforms employ bandits to optimally allocate limited advertisement slots [3]; dynamic pricing strategies rely on MAB algorithms [4]–[6] to discover the most profitable price points; cognitive radios uses MAB algorithms to identify free spectrum to access [7].

While traditional multi-armed bandit methods typically rely on explicit, numerical rewards from a single arm, the dueling bandit framework [8] instead provides pairwise feedback: given two arms, the learner observes which of the two "wins" in a head-to-head comparison. This subtle yet significant shift introduces additional challenges, as global preference rankings must be inferred from potentially noisy pairwise outcomes, which need not satisfy transitivity or consistency across all arms. Despite these complexities, the dueling bandit problem is of critical importance in real-world scenarios where direct rewards are hard to measure but relative preferences are more natural [9], [10]. In many applications, such as information retrieval [11], [12], product recommendation [13], and preference elicitation—users or systems can more easily compare two items than provide a robust numerical score for each. By leveraging these pairwise comparisons, dueling bandit algorithms capture nuanced user preferences and avoid pitfalls associated with subjective or hard-to-calibrate reward scales. Consequently, dueling bandits are increasingly recognized as a powerful framework for scenarios in which relative feedback is more natural and more accurately reflects the underlying value of different choices [14]–[16].

Although classical dueling bandit algorithms, such as the Relative Upper Confidence Bound (RUCB) algorithm [17] offer strong theoretical guarantees in benign settings, they can be surprisingly vulnerable to adversarial feedback manipulations. In many real-world scenarios, attackers can intercept or distort the pairwise outcomes being observed, thereby misleading the learner's estimates of arm preferences. Moreover, an attacker often needs to manipulate only a small fraction of pairwise outcomes to achieve significant disruption. By shaping the outcomes of crucial comparisons—either before the learner begins interacting with the environment (pre-action) or after each decision (post-action)—the attacker can engineer persistent misestimates, ultimately steering the learner toward suboptimal arms.

The pre-action attack has been investigated in the dueling bandits setup [18]–[20]. Under this model, the adversary corrupts or manipulates the environment by flipping some pairwise comparison outcomes before the learner takes any action. The adversary, having full knowledge of all pairwise outcomes or preference probabilities in advance, flips or distorts these outcomes before the learner takes any action. By the time the learner begins to interact with the environment and observe feedback, the feedback has already been corrupted, thereby altering the ground truth of the environment. As a result, queries by the learner are based on this tampered environment. While much of the early work on pre-action attacks was devoted to the stochastic case, several researchers have explored adversarial dueling-bandit scenarios. Gajane et al. [21] achieved a regret bound of $\mathcal{O}(\sqrt{T})$ in purely adversarial settings. Saha et al. [22] proposed algorithms with regret bounds of $\tilde{\mathcal{O}}(K^{1/3}T^{2/3})$, along with tighter results in a fixed-gap adversarial setup. Agarwal et al. [18] further

This work was supported by the National Science Foundation under Grant CCF-2232907.

analyzed regret as a sum of terms dependent on the adversary's corruption and the baseline stochastic bound. Despite these valuable insights, they typically assume the adversary has full access to all outcomes in advance—an assumption that may not hold in many practical scenarios. Moreover, such attacks may lack efficiency since they often target pairs other than the ones selected by the learner, resulting in unnecessary modifications that do not directly influence the learner's decisions.

In this paper, we propose a novel post-action attack on dueling bandits. In the proposed post-action attack model, we restrict the adversary to only observe and corrupt the feedback from the specific arm pair selected by the learner at each round. This restriction may appear to weaken the attacker, yet recent analyses of reward-manipulation or action-manipulation attacks in stochastic bandits [23]–[29] show that a post-action style of interference can be extremely potent. For example, an attacker can steer the learner into pulling a targeted arm for nearly T rounds while incurring only $\mathcal{O}(\log T)$ in manipulation costs [30]. Because the manipulations only occur when the "desired" arm pair is chosen, the attacker can focus its efforts more efficiently, thereby amplifying its impact without needing complete environment access. Some notable works provide guarantees of $\mathcal{O}(\log T + C)$ [31] or $\tilde{\mathcal{O}}(d\sqrt{T}+dC)$ [32] for contextual extensions. Despite its practical relevance and formidable nature, the post-action attack model remains relatively underexplored in the dueling bandits literature. In this paper, we show that a post-action attacker can force a well-known dueling bandit algorithm (RUCB) to select a specific target arm for almost the entire time horizon, while incurring only a minimal corruption cost. Concretely, we prove that the attacker can achieve $T-\mathcal{O}(K^2 \ln T)$ "successful manipulations" (i.e., rounds in which the target arm is chosen) at a total attack cost of $\mathcal{O}(K \ln T)$. These findings underscore the vulnerability of dueling bandit algorithms under post-action threats and motivate the development of adversarially resilient techniques that can withstand this potent attack model.

The remainder of this paper is organized as follows. Section II formally introduces the dueling bandit problem and our post-action attack model. Next, Section III presents the main ideas behind our proposed attack strategy and explains how it exploits the RUCB algorithm. In Section IV, we provide detailed theoretical analysis, bounding both the adversary's attack cost and the resulting attack regret. We then validate our findings empirically in Section V, demonstrating how the attack effectively misleads RUCB on synthetic datasets. Finally, Section VI summarizes our results and discusses potential avenues for future work, particularly in developing robust defenses against adversarial interference.

II. PROBLEM FORMULATION

A. Dueling bandit problem

Let there be K arms with an unknown $K \times K$ pairwise preference matrix $\mathbf{P} = [p_{ij}]$, where each entry $p_{ij} \in [0, 1]$ represents the probability that arm a_i is preferred over arm a_j in a head-to-head comparison. The matrix satisfies the property $p_{ji} = 1 - p_{ij}$, ensuring that comparisons are probabilistically consistent. Each p_{ij} is an unknown constant.

Following standard dueling bandit assumptions, we assume the existence of a Condorcet winner (the unique best arm) [33], [34], which, without loss of generality, is designated as arm 1. This implies that arm 1 is preferred over every other arm, satisfying $p_{1i} > \frac{1}{2}$ for all $i > 1$. At each time step t, the learner selects a pair of arms $(a_c(t), a_d(t)) \in [K] \times [K]$ and observes the outcome of their comparison.

Let $Z^t_{i,j} = \mathbb{1}_{i \succ j}$ denote the outcome of a pairwise comparison between arms i and j. It equals 1 if arm i wins over arm j, and 0 otherwise with the probability:

$$P(Z^t_{i,j} = 1) = p_{ij}, \quad P(Z^t_{i,j} = 0) = 1 - p_{ij} = p_{ji}.$$

In dueling bandits, a well known performance metric is the regret of the algorithm, which captures how effectively the algorithm converges to the Condorcet winner arm 1. Formally, at time t, we define the cumulative convergence regret:

$$R(t) = \sum_{h=1}^{t} \Big(\mathbb{1}_{\{a_c(h) \neq 1\}} + \mathbb{1}_{\{a_d(h) \neq 1\}} \Big), \tag{1}$$

where $a_c(h)$ and $a_d(h)$ denote the two arms compared at round h. Note that any pull of an arm other than arm 1 contributes to the cumulative regret.

B. Post-action Attack

We now introduce our attack model. The attacker has a target arm. Without loss of generality, we set arm k as the attack target. The attacker's goal is to coerce the learner to frequently select or compare against arm k, effectively misleading the algorithm into favoring arm k over other arms.

The attacker aims to achieve this goal by selectively flipping some comparison results provided by the nature. In particular, at each time step t, the agent will select a pair $(a_c(t), a_d(t))$. The nature will then provide a comparison result $Z^t_{a_c(t),a_d(t)}$ based on dueling bandit model discussed above. The attacker will observe the outcome $Z^t_{a_c(t),a_d(t)}$ and decide whether it would like to attack at this time based on $Z^t_{a_c(t),a_d(t)}$ and all past observations. If the attacker decides to attack at time t, the attacker changes $Z^t_{a_c(t),a_d(t)}$ to $\hat{Z}^t_{a_c(t),a_d(t)} = 1 - Z^t_{a_c(t),a_d(t)}$. The agent will observe $\hat{Z}^t_{a_c(t),a_d(t)}$. If attacker does not attack at time t, the agent will observe $Z^t_{a_c(t),a_d(t)}$.

The performance of an attack schemes are quantified by two metrics: attack cost and attack regret. The cumulative post-action attack cost at time t is defined as:

$$L_{attack}(t) = \sum_{h=1}^{t} \left| Z^h_{a_c(h),a_d(h)} - \hat{Z}^h_{a_c(h),a_d(h)} \right|. \tag{2}$$

The attack regret quantifies the extent to which the adversary successfully deviates the algorithm from the Condorcet winner to the target arm k. Formally, the attack regret at time t is defined as:

$$R_{attack}(t) = \sum_{h=1}^{t} \Big(\mathbb{1}_{\{a_c(h) \neq k\}} + \mathbb{1}_{\{a_d(h) \neq k\}} \Big), \tag{3}$$

where $a_c(h)$ and $a_d(h)$ denote the two arms compared at round h. The attacker aims to minimize the attack regret while incurring the least possible attack cost.

III. PROPOSED METHOD

In this paper, we consider a post-attack strategy on a widely used dueling bandit algorithm: Relative Upper Confidence Bound (RUCB) algorithm, introduced by [17].

A. RUCB Algorithm of [17]

Before presenting our attack scheme, we first provide an overview of the RUCB algorithm introduced in [17]. To facilitate the presentation, we reproduce the RUCB Algorithm in Algorithm 1 using the same notation as in [17]. RUCB leverages an extension of the classic UCB principle to estimate pairwise preferences, selecting the arm most likely to outperform others while simultaneously updating its confidence bounds with the winner as a benchmark. Theoretically, RUCB achieves a finite-time high-probability regret bound of $\mathcal{O}(K \log T)$, which is the best known rate under minimal assumptions [33], [35], [36].

As shown in Algorithm 1, RUCB updates pairwise win counts in $\mathbf{W}$, then calculates upper confidence bounds $\mathbf{U}$ (line 5). Based on these bounds, it constructs a *candidate set* C (line 9). In particular, C contains every arm that currently appears at least as strong as all other arms, with its upper confidence bound being at least 0.5 against every competitor. Next, the algorithm either randomly picks an arm or samples one from C according to the specified rule (lines 10–18): if there is a single candidate, it is chosen; if multiple remain, the selection among them follows a specific sampling distribution. Finally, RUCB selects its comparison pair by choosing the candidate arm a_c and pairing it with a_d, where a_d maximizes $u_{j,c}$ (line 19). The observed outcome from comparing $\{a_c, a_d\}$ then updates $\mathbf{W}$ (line 20).

In this paper, we use the following notations throughout the analysis. The parameter α is an input to Algorithm 1 and is used to control the confidence intervals. For any pair of arms a_i and a_j, $N_{ij}(t)$ denotes the total number of comparisons between these arms up to time t, while $w_{ij}(t)$ represents the total number of wins of a_i over a_j. The upper confidence bound [37] for p_{ij}, the probability of a_i being preferred over a_j, is given by $u_{ij}(t) = \frac{w_{ij}(t)}{N_{ij}(t)} + \sqrt{\frac{\alpha \ln t}{N_{ij}(t)}}$. The corresponding lower confidence bound is denoted as $l_{ij}(t) = 1 - u_{ji}(t)$. And δ is the probability of failure, which determines the high-confidence guarantees of the algorithm. Lastly, define

$$\tau(\delta) := \left(\frac{(4\alpha - 1)K^2}{(2\alpha - 1)\delta} \right)^{\frac{1}{2\alpha - 1}}.$$

B. Proposed attack strategy

The proposed attack strategy aims to ensure that a_k wins the pairwise comparisons against any arm i where $i \neq k$ with probability larger than $1/2$, thereby establishing a_k as the Condorcet winner. The outcomes of comparisons involving pairs where neither element is k remain unchanged.

Algorithm 1 *RUCB*[Algorithm 1 of [17]]

1: **Parameters:** $\alpha \geq \frac{1}{2}$, $T \in \{1, 2, \ldots\} \cup \{\infty\}$
2: **Initialization:** $\mathbf{W} = [w_{ij}] \leftarrow \mathbf{0}_{K \times K}$ ▷ w_{ij} is the number of times arm i beats arm j
3: $B \leftarrow \emptyset$
4: **for** $t = 1$ **to** T **do**
5: $\quad \mathbf{U} = [u_{ij}] \leftarrow \dfrac{\mathbf{W}}{\mathbf{W} + \mathbf{W}^T} + \sqrt{\dfrac{\alpha \ln t}{\mathbf{W} + \mathbf{W}^T}}$ ▷ Element-wise operations, $\frac{x}{0} := 1$ for any x
6: $\quad$ **for** $i = 1$ **to** K **do**
7: $\quad\quad u_{ii} \leftarrow \frac{1}{2}$ ▷ Set diagonal elements to $\frac{1}{2}$
8: $\quad$ **end for**
9: $\quad C \leftarrow \{\, a_c \mid \forall j \colon u_{cj} \geq \frac{1}{2} \}$
10: $\quad$ **if** $C = \emptyset$ **then**
11: $\quad\quad$ Pick c randomly from $\{1, \ldots, K\}$
12: $\quad$ **end if**
13: $\quad B \leftarrow B \cap C$
14: $\quad$ **if** $|C| = 1$ **then**
15: $\quad\quad B \leftarrow C$
16: $\quad\quad c \leftarrow$ the unique element in C
17: $\quad$ **else if** $|C| > 1$ **then**
18: $\quad\quad$ Sample a_c from C using:
19: $\quad\quad p(a_c) = \begin{cases} 0.5, & \text{if } a_c \in B, \\ \dfrac{1}{2^{|B|} \cdot |C \setminus B|}, & \text{otherwise.} \end{cases}$
20: $\quad$ **end if**
21: $\quad a_d \leftarrow \arg\max_j u_{j,c}$ ▷ Ties broken randomly, and $a_d \neq a_c$ if there is a tie
22: $\quad$ **Compare** arms a_c and a_d and increment w_{cd} or w_{dc} accordingly
23: **end for**
24: **Return:** Arm a_c that "beats" the most arms, i.e.,

$$c = \arg\max_i \left| \left\{ j \;\middle|\; \tfrac{w_{ij}}{w_{ij} + w_{ji}} > \tfrac{1}{2} \right\} \right|.$$

Specifically, if the learner chooses a pair at time t does not involve k, the attacker does nothing and sets $\hat{Z}^t_{i,j} = Z^t_{i,j}$. However, if (k, i) is selected (i.e., either $a_c(t) = k$, $a_d(t) = i$ or vice versa), the adversary can simply draw $\hat{Z}^t_{k,i}$ directly from a Bernoulli distribution with parameter $\hat{p}_{ki}$, where $\hat{p}_{ki}$ is a modified preference value chosen by the attacker, satisfying $\hat{p}_{ki} > \frac{1}{2}$, independently of the original sample $Z^t_{k,i}$. If the original outcome $Z^t_{k,i}$ already equals $\hat{Z}^t_{k,i}$, the attacker leaves it unchanged (i.e., no flip occurs); otherwise the attacker flips the outcome.

Overall, when (k, i) is chosen, the attacker's strategy ensures that $\hat{Z}^t_{k,i}$ takes the value 1 (i.e., k wins) with probability $\hat{p}_{ki}$. Note that, the proposed attack strategy does not need to know the true underlying values of p_{ij}.

Under this attack strategy, effectively, the observations observed by the agent is generated by a modified preference matrix $\hat{\mathbf{P}} = [\hat{p}_{ij}]$ with $\hat{p}_{ki} > \frac{1}{2}$ and $\hat{p}_{ik} = 1 - \hat{p}_{ki} < \frac{1}{2}$ for all $i \neq k$. For all other pairs (i, j), $\hat{p}_{ij}$ remains equal to p_{ij}, as the attacker does not flip any outcomes for these pairs.

IV. THEORETICAL ANALYSIS

A. Attack cost

Lemma IV.1. *Fix any confidence level $\delta \in (0,1)$ and any parameter $\alpha > \frac{1}{2}$. There exists a finite round index $\tau(\delta)$ such that*

$$P\left(a_k \in C \text{ for all } t > \tau(\delta)\right) \;\geq\; 1-\delta.$$

In other words, with probability at least $1-\delta$ the target arm a_k remains in the candidate set C from round $\tau(\delta)+1$ onward.

Lemma IV.2. *For any $\alpha > \frac{1}{2}$ and $\delta \in (0,1)$, we have*

$$P\left(\forall\, t > \tau(\delta),\ i, j,\ \hat{p}_{ij} \in [l_{ij}(t), u_{ij}(t)]\right) > 1-\delta, \quad (4)$$

where $l_{ij}(t) = 1 - u_{ji}(t)$ [17].

Lemmas IV.1 and IV.2 provide the foundational probabilistic tools that ensure, with high probability, each estimated pairwise preference $\hat{p}_{ij}$ remains within a confidence interval $[l_{ij}(t), u_{ij}(t)]$ once t exceeds a threshold $\tau(\delta)$ and the target arm a_k consistently remains in the candidate solution set C.

Let sequence $\{t_{ik}^m\}_{m=1}^{N_{ik}}$ denote the time sequence when a_i and a_k are picked.

By the definition of $\mathbf{W}$, $w_{ik}(t) = \sum_{m=1}^{N_{ik}(t)} \hat{Z}_{i,k}^{t_{ik}^m}$.

Proposition IV.3. *Fix any confidence level $\delta \in (0,1)$. Let a_i be any arm different from the attack target arm a_k. Define*

$$N_{ik}^{\min} \;:=\; \left\lceil \frac{\left(\sqrt{\alpha \ln T} + \sqrt{\frac{1}{2}\ln\frac{1}{\delta}}\right)^2}{\left(\frac{1}{2} - \hat{p}_{ik}\right)^2} \right\rceil,$$

and let $\hat{t}_i := \min\{t \geq 1 : N_{ik}(t) = N_{ik}^{\min}\}$ be the first round at which the number of (i,k) comparisons reaches this value. For every round $t \geq \hat{t}_i$,

$$P\left(\frac{w_{ik}(t)}{N_{ik}(t)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t)}} < \frac{1}{2}\right) \;\geq\; 1-\delta,$$

so, with the same probability, $a_i \notin C$.

Proposition IV.3 specifies how and when a suboptimal arm a_i fails to maintain a winning rate above $\frac{1}{2}$ against the attacker's target arm a_k. Once the number of comparisons $N_{ik}(t)$ between a_i and a_k becomes sufficiently large, the empirical preference for a_i will, with high probability, fall below $\frac{1}{2}$. As a result, a_i will be excluded from the candidate set C, ensuring that the learner no longer considers it as a viable option

Corollary IV.4. *For any $t > \hat{t}_i$, if*

$$\frac{w_{ik}(t)}{N_{ik}(t)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t)}} < \frac{1}{2},$$

$a_i \notin C$, then for $t^* = t+1$, we have $N_{ik}(t^*) = N_{ik}(t)$.

Corollary IV.5. *For any $t > \hat{t}_i$, if $\frac{w_{ik}(t)}{N_{ik}(t)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t)}} < \frac{1}{2}$, $a_i \notin C$, then for $t^* = t+1$, $a_i \notin C$ and*

$$\frac{w_{ik}(t^*)}{N_{ik}(t^*)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t^*)}} < \frac{1}{2}.$$

Corollary IV.6. *Let $\Gamma(t)$ be the event that*

$$\frac{w_{ik}(t)}{N_{ik}(t)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t)}} < \frac{1}{2}$$

and $a_i \notin C$ at iteration t. Then,

$$P\left(\forall t > \hat{t}_i + 1, \Gamma(t) \mid \Gamma(\hat{t}_i + 1)\right) = 1.$$

Corollaries IV.4–IV.6 illustrate a key implication: once an arm a_i has been conclusively "outperformed" by a_k, it exits the candidate set C and is no longer compared against a_k. In essence, the adversary's strategic reversals ensure that every other arm eventually shows empirical evidence of being worse than a_k.

Theorem IV.7.

$$P\left(\forall t > \hat{t}_i, a_i \notin C\right) > 1-\delta.$$

Theorem IV.7 establishes that, with high probability, each suboptimal arm $a_i \neq k$ remains excluded from the candidate set, ensuring a_k is the only viable choice in future rounds.

Theorem IV.8. *With probability larger than $1-\delta$, the attack cost $L_{attack}(T)$ of the proposed attack against RUCB is*

$$\mathcal{O}\Big((K-1)\cdot \max_i\Big\{\frac{\hat{p}_{ki}-p_{ki}}{1-p_{ki}}\Big\} \cdot \max_i\left\{\frac{\left(\sqrt{\alpha\ln T}+\sqrt{\frac{1}{2}\ln\frac{1}{\delta}}\right)^2}{\left(\frac{1}{2}-\hat{p}_{ik}\right)^2}\right\}\Big),$$

which scales as $\mathcal{O}(K \ln T)$.

To establish the $\mathcal{O}(K \ln T)$ bound on the attack cost, we first rely on Proposition IV.3, which specifies that when the number of comparisons between a non-target arm $a_i \neq k$ and the target arm a_k exceeds a certain threshold, the winning rate of a_i drops below $1/2$, leading to a_i's removal from the candidate set C. Corollary IV.4 then confirms that $N_{ik}(t)$ will not increase further for such an arm, and Corollaries IV.5 and IV.6 ensure a_i remains out of C thereafter. By Theorem IV.7, every non-target arm eventually stays out of C. Summing over all $i \neq k$ yields the overall $\mathcal{O}(K \ln T)$ cost in Theorem IV.8.

B. Attack regret

Theorem IV.9.

$$\exists t > \tau(\delta),\ i \neq k,\ j \neq k:$$
$$P\left(N_{ij}(t) > \frac{4\alpha \ln t}{\min\left({\Delta_i}^2, {\Delta_j}^2\right)}\right) < \delta, \quad (5)$$

where $\Delta_i = \hat{p}_{ki} - \frac{1}{2}$.

Theorem IV.9 ensures that, after a sufficiently large time, the number of comparisons $N_{ij}(t)$ between any two non-target arms a_i and a_j cannot exceed $\frac{4\,\alpha\,\ln t}{\min(\Delta_i^2, \Delta_j^2)}$ with high probability. This ensures the learner's focus shifts away from non-target pairs, helping control the overall regret.

Theorem IV.10. *The total attack regret $R_{attack}(T)$ of the proposed attack against RUCB is $\mathcal{O}(K^2 \ln T)$.*

The proof proceeds by contradiction. First, it assumes that the number of comparisons $N_{ij}(t)$ between two non-target arms a_i and a_j grows larger than the threshold $\frac{4\alpha \ln t}{\min(\Delta_i^2, \Delta_j^2)}$. Then it shows that this assumption forces the pairwise confidence intervals $\big(u_{ij}(t),\, l_{ij}(t)\big)$ to contract too much, implying a contradiction with the conditions under which the algorithm selects a_i and a_j.

V. EXPERIMENTAL DATA AND RESULT

We conduct experiments using synthetic preference matrices generated by a simple rank-based procedure. Specifically, we set $K = 10$, designate arm 0 as the Condorcet winner, and fix arm 9 as the adversary's attack target.

We construct two $K \times K$ preference matrices with diagonal entries set to 0.5. Matrix $\mathbf{P}$ assigns ranks in ascending order, making arm 0 the top choice and arm 9 the weakest. Matrix $\hat{\mathbf{P}}$ shifts these ranks, placing arm 9 at the top. To fill each off-diagonal entry, we add a fraction of the rank difference between the two arms to a baseline 0.5, giving higher win probabilities to arms with better (lower) ranks. The reverse probability for each pair is set so that the two values sum to 1. This process captures two distinct ranking scenarios while maintaining the same set of arms.

We also choose a time horizon $T = 10^6$ and a parameter $\alpha = 0.6$ to guide the update rules in our algorithms.

Our experimental results align well with the theoretical guarantees. Figure 1 illustrates how the cumulative attack cost grows on the order of $K \ln T$, exactly as Theorem IV.8 predicts.

Turning to the learner's performance, Figure 2 shows the defense regret, defined as the frequency with which arms other than the Condorcet winner are chosen. In the no-attack scenario, RUCB promptly identifies the best arm, and the defense regret remains nearly zero. However, once adversarial manipulation begins, the regret curve slopes upward (with an approximate slope of 2), signifying that the learner is frequently misled away from the Condorcet winner.

The attack regret, illustrated in Figure 3, represents the number of times arms other than the target arm are selected during the adversarial attack. The figure provides a focused view of the attack regret under adversarial conditions, showcasing its growth at a rate of $\mathcal{O}(K^2 \ln T)$ over the entire time horizon, as formally established in Theorem IV.10.

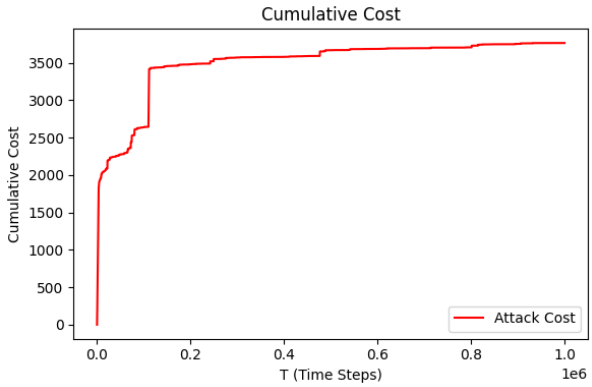

Fig. 1. Cumulative Cost

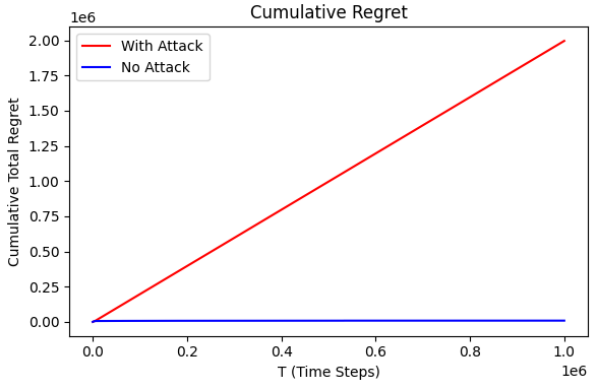

Fig. 2. Cumulative Defense Regret

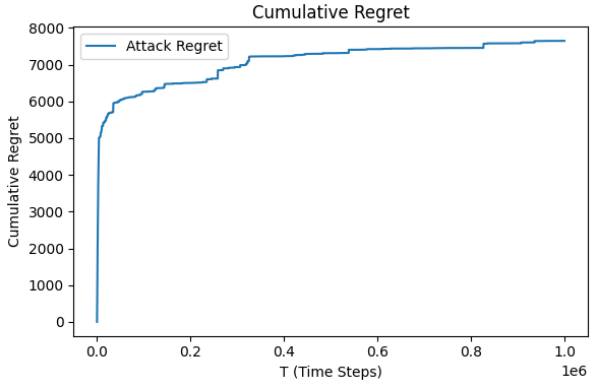

Fig. 3. Cumulative Attack Regret With Attack

VI. CONCLUSION

In this paper, we have examined a potent post-action attack model for the dueling bandit problem, focusing specifically on the RUCB algorithm. Our theoretical analysis shows that, by manipulating only the feedback from the chosen arm pair, an adversary can force the learner to select a predetermined target arm almost exclusively. A learner that continues to run RUCB unchanged under this attack incurs *linear* regret, $\mathcal{O}(T)$, while the adversary needs only a total intervention budget of $\mathcal{O}(K \ln T)$ and suffers at most $\mathcal{O}(K^2 \ln T)$ attack regret. Empirical results reinforce these findings, demonstrating how post-action interference severely compromises RUCB's performance, despite its strong guarantees in benign scenarios. These insights highlight the urgency of developing adversarially resilient dueling bandit algorithms, spurring future research on robust defenses that safeguard online learning against targeted feedback corruption.

REFERENCES

[1] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952.

[2] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, Lille, France, July 2015.

[3] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, page 661–670, Raleigh, North Carolina, USA, April 2010.

[4] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Journal of Machine Learning Research (JMLR)*, 47(2-3):235–256, 2002.

[5] Touqir Sajed and Or Sheffet. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pages 5579–5588, Long Beach, California, USA, June 2019.

[6] Zohar S Karnin. Verification based solution for structured mab problems. In *Neural Information Processing Systems*, Barcelona, Spain, December 2016.

[7] Lifeng Lai, Hesham El Gamal, Hai Jiang, and H. Vincent Poor. Cognitive medium access: Exploration, exploitation, and competition. *IEEE Transactions on Mobile Computing*, 10(2):239–253, 2011.

[8] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[9] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7–es, 2007.

[10] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[11] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, 30(1):1–41, 2012.

[12] Xinyi Yan, Chengxi Luo, Charles L. A. Clarke, Nick Craswell, Ellen M. Voorhees, and Pablo Castells. Human preferences as dueling bandits. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 567–577, Madrid, Spain, July 2022.

[13] Marco De Gemmis, Leo Iaquinta, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. Learning preference models in recommender systems. In *Preference Learning*, pages 387–407. Springer, 2010.

[14] Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on Learning Theory*, pages 1141–1154, Paris, France, July 2015.

[15] Wei Chen, Yihan Du, Longbo Huang, and Haoyu Zhao. Combinatorial pure exploration for dueling bandit. In *International Conference on Machine Learning*, pages 1531–1541, Online, July 2020.

[16] Aadirupa Saha and Pierre Gaillard. Dueling bandits with adversarial sleeping. In *Neural Information Processing Systems*, pages 27761–27771, Online, December 2021.

[17] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *International Conference on Machine Learning*, pages 10–18, Bejing, China, June 2014.

[18] Arpit Agarwal, Shivani Agarwal, and Prathamesh Patil. Stochastic dueling bandits with adversarial corruption. In *International Conference on Algorithmic Learning Theory*, pages 217–248, Online, March 2021.

[19] Aadirupa Saha and Shubham Gupta. Optimal and efficient dynamic regret algorithms for non-stationary dueling bandits. In *International Conference on Machine Learning*, pages 19027–19049, Baltimore, Maryland, USA, July 2022.

[20] Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864, Bejing, China, June 2014.

[21] Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *International Conference on Machine Learning*, page 218–227, Lille, France, July 2015.

[22] Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, pages 9235–9244, Online, July 2021.

[23] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. In *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, December 2018.

[24] Chenye Yang, Guanlin Liu, and Lifeng Lai. Stochastic bandits with non-stationary rewards: Reward attack and defense. *IEEE Transactions on Signal Processing*, 2024.

[25] Huazheng Wang, Haifeng Xu, and Hongning Wang. When are linear stochastic bandits attackable? In *International Conference on Machine Learning*, pages 23254–23273, Baltimore, Maryland, USA, July 2022.

[26] Guanlin Liu and Lifeng Lai. Action-manipulation attacks on stochastic bandits. In *IEEE International Conference on Acoustics, Speech and Signal*, pages 3112–3116, Online, May 2020.

[27] Yuzhe Ma and Zhijin Zhou. Adversarial attacks on adversarial bandits. In *International Conference on Learning Representations*, Kigali, Rwanda, May 2023.

[28] Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Thirty-Second Conference on Learning Theory*, pages 1562–1578, Phoenix, Arizona, USA, June 2019.

[29] Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, volume 97, pages 4042–4050, June 2019.

[30] 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.

[31] Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, pages 19011–19026, Baltimore, Maryland, USA, July 2022.

[32] Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback, 2024.

[33] Tanguy Urvoy, Fabrice Clerot, Raphael Féraud, and Sami Naamane. Generic exploration and K-armed voting bandits. In *International Conference on Machine Learning*, pages 91–99, Atlanta, Georgia, USA, June 2013.

[34] El Mehdi Saad, Alexandra Carpentier, Tomáš Kocák, and Nicolas Verzelen. On weak regret analysis for dueling bandits. In *Neural Information Processing Systems*, Vancouver, Canada, December 2024.

[35] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *International Conference on Machine Learning*, page 1201–1208, Montreal, Quebec, Canada, June 2009.

[36] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pages 23–37, Porto, Portugal, October 2009.

[37] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188, Espoo, Finland, October 2011.