Dueling Bandit: Adversarial Attack and Robust Defense

Mo Lyu, Chenye Yang, Guanlin Liu and Lifeng Lai, *Senior Member, IEEE*

***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, where T is the time horizon and K is the number of total arms. To counter such attacks, we propose a novel robust defense strategy Attack-Aware RUCB (AA-RUCB) that augments the RUCB algorithm with attack-awareness. Assuming the adversary's budget is upper bounded by A, the proposed algorithm adjusts RUCB's upper confidence bound estimates to account for potential outcome flips. We prove that the defense algorithm preserves the optimal $O(K^2 \ln T)$ regret when $A = 0$ and degrades gracefully to $O(K^2 \ln T + A\sqrt{\ln T})$ as A grows.**

***Index Terms*—dueling bandit, adversarial attack, defense.**

I. Introduction

Multi-armed bandit (MAB) problems introduced by [2] 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 [3]. 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 [4]; advertising platforms employ bandits to optimally allocate limited advertisement slots [5]; dynamic pricing strategies rely on MAB algorithms [6]–[8] to discover the most profitable price points; cognitive radios uses MAB algorithms to identify free spectrum to access [9]–[11] and beam scheduling [12].

While traditional multi-armed bandit methods typically rely on explicit, numerical rewards from a single arm, the dueling bandit framework [13] 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 [14], [15]. In many applications, such as information retrieval [16], [17], product recommendation [18], 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 [19]–[21].

Although classical dueling bandit algorithms, such as the Relative Upper Confidence Bound (RUCB) algorithm [22] 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 [23]–[25]. 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 ex-

M. Lyu is with the Department of Computer Science, C. Yang, G. Liu and L. Lai are with the Department of Electrical and Computer Engineering, University of California, Davis, CA. This work was supported by the National Science Foundation under Grants CCF-2232907. Email:{molyu, cyyyang, glnliu, lflai}@ucdavis.edu. This paper will be presented in part at the 2025 Asilomar Conference on Signals, Systems, and Computers [1].

plored adversarial dueling-bandit scenarios. Gajane et al. [26] achieved a regret bound of $\mathcal{O}(\sqrt{T})$ in purely adversarial settings. Saha et al. [27] 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. [23] further 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 [28]–[34] 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 $T - \log T$ rounds while incurring only $\mathcal{O}(\log T)$ in manipulation costs [35]. This line of work has also been extended to contextual bandits setup [36]–[38]. Despite its practical relevance and formidable nature, the post-action attack model remains relatively underexplored in the dueling bandits literature.

We introduce and study a novel post-action adversarial attack on dueling-bandit learners. In the standard dueling-bandit setting, a learning algorithm (like the RUCB algorithm) repeatedly selects pairs of arms to compare and then observes a stochastic outcome. Prior work on adversarial dueling bandit attacks typically assumes the attacker intervenes in advance or intercepts reward signals in a global way. By contrast, our post-action attacker is much more adaptive: it first observes which pair RUCB has chosen, and only then corrupts the observed outcome of the duel. The proposed attack algorithm intervenes only when the learner compares the adversary's chosen target arm with another arm. In those rounds the attacker flips the feedback, if necessary, so that the target is recorded as the winner more than half the time, while every comparison that does not involve the target is left untouched. This allows the attacker to steer the algorithm's beliefs toward a chosen target arm at minimal cost. We show that this powerful attack can almost entirely hijack RUCB's behavior. Concretely, our analysis proves that an adversary can force RUCB to select a pre-specified "target" arm in $T - \mathcal{O}(K^2 \ln T)$ out of T total rounds, while incurring only $\mathcal{O}(K \ln T)$ total corruption cost. In other words, for essentially the entire time horizon the algorithm is misled into favoring the adversary's desired arm, even though the attacker's budget grows only logarithmically in T. These findings underscore a critical vulnerability of preference-based learners: despite RUCB's optimal logarithmic regret in benign settings, even a low-cost, reactive attack can dramatically subvert its choices. Importantly, this threat model is realistic in many applications. For example, in online ranking or recommendation systems a malicious actor can often observe which items are being evaluated and then launch "click-fraud" or fake-comparison attacks to bias feedback in favor of a particular option.

To defend against such attacks we introduce the Attack Aware RUCB (AA-RUCB), a simple and effective refinement of the original RUCB algorithm. AA-RUCB differs from RUCB by only a single-line modification. In each round it enlarges every upper confidence bound with an additive term that scales with an upper bound of the corruption budget A. This extra margin pads the confidence intervals so the learner can tolerate up to A corrupted outcomes without being misled. Despite the small tweak, AA-RUCB preserves RUCB's statistical guarantees in benign conditions and degrades smoothly when an adversary is present. We show that when $A = 0$, i.e., there is no corruption, AA-RUCB coincides exactly with RUCB and achieves the optimal regret rate $\mathcal{O}(K^2 \ln T)$ for dueling bandits. If an attacker spends budget A, the regret increases only to $\mathcal{O}(K^2 \ln T + A\sqrt{\ln T})$. The computational overhead remains negligible because the modification simply replaces the original confidence radius with a slightly larger one, making the algorithm easy to integrate into existing RUCB implementations. In practice AA-RUCB offers a robust and convenient safeguard: the attacker must spend a much larger budget to influence the learner, yet the algorithm's performance remains unchanged when no attack takes place.

Compared with our earlier conference paper [1], this journal paper makes three significant changes. First, we expand the theoretical treatment by providing complete, self-contained proofs for all key lemmas and theorems, thereby closing several gaps that were left open in the shorter version. Second, we move beyond purely offensive analysis by introducing AA-RUCB and rigorously establishing its robustness under bounded corruption. Third, we broaden the empirical evaluation through an extensive set of simulations that corroborate the tightness of our theoretical guarantees.

The remainder of this paper is organized as follows. Section II formally introduces the dueling bandit problem and our post-action attack model. Section III designs a low-cost attack against the classical RUCB algorithm and derives tight bounds on its flip budget and induced regret. Section IV presents AA-RUCB, a one-line modification of RUCB that provably limits the attacker's impact to an additive $A\sqrt{\ln T}$ term. Complete theoretical guarantees are given in Sections III-C and IV, and Section V corroborates them with synthetic experiments over a wide range of horizons and budgets. We conclude with implications and future directions in Section VI.

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) [39], [40], 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. Accordingly, our objective is to design algorithms that minimize the cumulative convergence regret $R(t)$ over time

B. Post-action Attack

We now introduce post-action 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 $\tilde{Z}^t_{a_c(t),a_d(t)} = 1 - Z^t_{a_c(t),a_d(t)}$. The agent will observe $\tilde{Z}^t_{a_c(t),a_d(t)}$. 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:

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

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:

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

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. Post-Action Attack on RUCB

In this section, we propose a post-action attack targeting the widely used Relative Upper Confidence Bound (RUCB) algorithm [22].

A. RUCB Overview

Before presenting our attack scheme, we first provide an overview of the RUCB algorithm introduced in [22]. To facilitate the presentation, we reproduce the RUCB Algorithm in Algorithm 1 using the same notation as in [22]. 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 [39], [41], [42].

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, thus $N_{ij}(t) = w_{ij}(t) + w_{ji}(t)$. The upper confidence bound [43] 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 $l_{ij}(t) = 1 - u_{ji}(t)$.

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

Algorithm 1 *RUCB*[Algorithm 1 of [22]]

Input: Confidence parameter $\alpha \geq \frac{1}{2}$, horizon $T \in \{1, 2, \dots\} \cup \{\infty\}$
Initialize win matrix $\mathbf{W} = [w_{ij}] \leftarrow \mathbf{0}_{K \times K}$ ▷ w_{ij} counts wins of a_i over a_j.
Initialize set $B \leftarrow \emptyset$
for $t = 1$ **to** T **do**
  $\mathbf{U} \leftarrow \dfrac{\mathbf{W}}{\mathbf{W} + \mathbf{W}^\top} + \sqrt{\dfrac{\alpha \ln t}{\mathbf{W} + \mathbf{W}^\top}}$ ▷ element-wise; define $\frac{x}{0} = 1$.
  $u_{ii} \leftarrow \frac{1}{2}$ **for all** $i = 1, \dots, K$.
  $C \leftarrow \{a_c \mid u_{cj} \geq \frac{1}{2},\ \forall j\}$.
  if $C = \emptyset$ **then**
    Choose a_c uniformly at random from $\{1, \dots, K\}$.
  else
    $B \leftarrow B \cap C$
    if $|C| = 1$ **then**
      $B \leftarrow C$ and set a_c to be the unique element of C.
    else
      Draw a_c from C with distribution

$$p(a_c) = \begin{cases} 0.5, & a_c \in B, \\ \dfrac{1}{2^{|B|}\,|C \setminus B|}, & \text{otherwise.} \end{cases}$$

    end if
  end if
  $a_d \leftarrow \arg\max_j u_{jc}$ ▷ If there is a tie, break ties uniformly; ensure $a_d \neq a_c$.
  Observe comparison (a_c, a_d) and update $\mathbf{W}$:

$$(w_{cd}, w_{dc}) \leftarrow \begin{cases} (w_{cd} + 1,\ w_{dc}), & \text{if } a_c \text{ wins}, \\ (w_{cd},\ w_{dc} + 1), & \text{otherwise.} \end{cases}$$

end for
Return arm a_c that beats the most opponents, i.e. the c maximizing $\#\{j \mid \frac{w_{cj}}{w_{cj}+w_{jc}} > \frac{1}{2}\}$.

the Condorcet winner. The outcomes of comparisons involving pairs where neither element is k remain unchanged.

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

If the learner ever selects the trivial pair (k, k) (i.e., it compares the target arm with itself), the attacker also remains inactive because flipping the outcome of a self-comparison would not alter any preference estimates.

Overall, when (k, i) is chosen, the attacker's strategy ensures that $\tilde{Z}^t_{k,i}$ takes the value 1 (i.e., k wins) with probability $\tilde{p}_{ki}$. 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 $\tilde{\mathbf{P}} = [\tilde{p}_{ij}]$ with $\tilde{p}_{ki} > \frac{1}{2}$ and $\tilde{p}_{ik} = 1 - \tilde{p}_{ki} < \frac{1}{2}$ for all $i \neq k$. For all other pairs (i, j), $\tilde{p}_{ij}$ remains equal to p_{ij}, as the attacker does not flip any outcomes for these pairs.

C. Attack Cost and Regret Analysis

We now analyze the attack cost and attack regret, and show the effectiveness of the proposed attack strategy.

We first introduce some notations. Let $\delta_1, \delta_2 \in (0, 1)$ be fixed confidence parameters. δ_1 controls the length of initial stage

$$\tau(\delta_1) := \left(\tfrac{(4\alpha-1)K^2}{(2\alpha-1)\,\delta_1}\right)^{1/(2\alpha-1)}.$$

We will show that after this initial stage all pairwise estimates lie in their confidence intervals with high probability. For every sub-optimal arm $i \neq 1$ define the true gap $\Delta_i := p_{1i} - \frac{1}{2} > 0$. The level δ_2 is used to bound the comparisons involving the attack target arm a_k. For each non-target arm $a_i \neq a_k$ define

$$N^{\min}_{ik} := \left\lceil \frac{\left(\sqrt{\alpha \ln T} + \sqrt{\frac{1}{2} \ln \frac{1}{\delta_2}}\right)^2}{\left(\tilde{p}_{ki} - \frac{1}{2}\right)^2} \right\rceil,$$

and let $\tilde{t}_i := \min\{t \geq 1 : N_{ik}(t) = N^{\min}_{ik}\}$ be the first round at which the (i, k) comparison count reaches this threshold.

Theorem III.1.

$$\Pr\Big(\forall t > \max_i \tilde{t}_i, \forall i \neq k : a_i \notin C\Big) > 1 - (K-1)\delta_2.$$

Proof. See Appendix A. □

Theorem III.1 establishes that after initial period, 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 III.2. *The total attack regret $R_{\text{attack}}(T)$ of the proposed attack against RUCB is $\mathcal{O}\big(K^2 \ln T\big)$ with probability larger than $1 - \delta_1 - (K-1)\delta_2$.*

Proof. See Appendix B. □

We now highlight the main idea of the proof. We first show that the number of comparisons between the target arm a_k and each sub-optimal arm $a_i \neq a_k$ is bounded by at $\mathcal{O}(\ln T)$; over the $K-1$ target–involving pairs this contributes $\mathcal{O}(K \ln T)$ to the total regret. We then prove that every remaining counter $N_{ij}(T)$ with $i, j \neq k$ is likewise bounded by $\mathcal{O}(\ln T)$; summing over the $\Theta(K^2)$ such pairs yields an additional $\mathcal{O}(K^2 \ln T)$. Hence the overall attack regret satisfies $R_{\text{attack}}(T) = \mathcal{O}(K^2 \ln T)$.

Since the attack regret counts the rounds in which the learner fails to play the target arm, this logarithmic bound

demonstrates that the attacker successfully compels the learner to choose the target arm for almost the entire time horizon.

Theorem III.3. *With probability larger than* $1-\delta_1-(K-1)\delta_2$*, the attack cost* $L_{attack}(T)$ *of the proposed attack against RUCB is*

$$\mathcal{O}\Big((K-1)\cdot\max_i\Big\{\frac{\tilde{p}_{ki}-p_{ki}}{1-p_{ki}}\Big\}\cdot\max_i\left\{\frac{\big(\sqrt{\alpha\ln T}+\sqrt{\frac{1}{2}\ln\frac{1}{\delta_2}}\big)^2}{\big(\tilde{p}_{ki}-\frac{1}{2}\big)^2}\right\}\Big),$$

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

Proof. See Appendix C. □

We now introduce the high level idea of our proof. To establish the $\mathcal{O}(K\ln T)$ bound on the attack cost, we first show 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. We then prove that $N_{ik}(t)$ will not increase further for such an arm, and ensure a_i remains out of C thereafter. By Theorem III.1, 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 III.3.

IV. Defense

As discussed in Section III, RUCB is not robust to adversarial attacks. In this section, we propose a robust variant of the RUCB algorithm, called Attack-Aware RUCB (AA-RUCB), which adjusts the confidence bounds to mitigate post-action attacks. We assume that the total attack budget is upperbounded by A. Knowing such an upper bound is necessary: without any limit on the adversary's power, no learner can separate stochastic noise from arbitrary manipulation, and sub-linear regret becomes impossible. Hence A plays the same role as corruption budgets in other robust bandit and online-learning models, allowing the confidence intervals to be inflated just enough to withstand the worst-case bias while preserving near-optimal performance in benign regimes.

The learning procedure of AA-RUCB closely follows the original RUCB, but with a crucial modification: an additional term is added to each arm's confidence bound to cover the worst-case reward bias an adversary can impose. This modification ensures that, even if rewards are corrupted, the algorithm remains sufficiently optimistic. Consequently, the analysis differs from the clean-feedback case by accounting for the adversarial perturbation, which introduces extra terms in the regret bounds and modifies the concentration arguments. Algorithm 2 summarizes the procedure.

The algorithm inherits the overall structure of the original RUCB: in each round it (i) computes an upper-confidence matrix $\mathbf{U}$ from the cumulative win counts $\mathbf{W}$, (ii) forms a candidate set containing all arms whose upper confidence against every opponent is at least one half, (iii) selects a champion a_c from that set (or uniformly if the set is empty), (iv) pairs it with the opponent a_d that currently maximises u_{dc}, and (v) updates $\mathbf{W}$ with the observed outcome of the duel.

The crucial difference lies in the confidence calculation. Whereas RUCB uses

$$\mathbf{U}_{\text{RUCB}} = \frac{\mathbf{W}}{\mathbf{W}+\mathbf{W}^\intercal} + \sqrt{\frac{\alpha\ln t}{\mathbf{W}+\mathbf{W}^\intercal}},$$

AA-RUCB inflates each element of this matrix with an additional margin that compensates for at most A corrupted outcomes per pair:

$$\mathbf{U} = \frac{\mathbf{W}}{\mathbf{W}+\mathbf{W}^\intercal} + \sqrt{\frac{\alpha\ln t}{\mathbf{W}+\mathbf{W}^\intercal}} + \frac{\alpha A\sqrt{\ln t}}{(\mathbf{W}+\mathbf{W}^\intercal)K^2}.$$

The first term of $\mathbf{U}$ is the empirical win rate for each ordered pair of arms, and the second is the usual Hoeffding-style exploration bonus scaled by $\alpha > \frac{1}{2}$. The third term pads the interval by an amount proportional to the corruption budget A. Because an adversary can flip at most A outcomes for any pair, the empirical win fraction may differ from the truth by at most $A/(\mathbf{W}+\mathbf{W}^\intercal)$. Multiplying by the factor K^{-2} and rescaling with $\sqrt{\ln t}$ yields a margin that safely dominates this bias. Consequently, even after adversarial tampering, every true preference still lies inside the enlarged confidence band with high probability. When $A=0$ the extra term vanishes, so AA-RUCB reduces exactly to RUCB.

The entire confidence matrix is formed entry-by-entry through the scalar definition

$$u_{ij}(t) = \frac{w_{ij}(t)}{N_{ij}(t)} + \sqrt{\frac{\alpha\ln t}{N_{ij}(t)}} + \frac{\alpha A\sqrt{\ln t}}{N_{ij}(t)\,K^2}, \tag{4}$$

where $w_{ij}(t)$ is the number of wins of arm i over arm j up to time t and $N_{ij}(t) = w_{ij}(t) + w_{ji}(t)$ is the total number of comparisons between the two arms. And we set $l_{ij}(t) = 1 - u_{ji}(t)$.

We now state the main theoretical guarantees. The first lemma shows that, with a high probability, the inflated confidence bound indeed covers the true arm means under attack.

Lemma IV.1. *Fix* $\alpha > \frac{1}{2}$, $\delta_1 \in (0,1)$, *and* $K \geq 2$. *Let*

$$\tau_d(\delta_1) := \max\Big\{\exp\big(K^4/\alpha^2\big),\ \big(\tfrac{(4\alpha-1)K^2}{(2\alpha-1)\delta_1}\big)^{\frac{1}{2\alpha-1}}\Big\}. \tag{5}$$

Then

$$\Pr\big\{\,\forall t > \tau_d(\delta_1),\, i,\, j : p_{ij} \in [l_{ij}(t), u_{ij}(t)]\,\big\} \geq 1-\delta_1.$$

Proof. See Appendix D. □

We now discuss the main idea of our proof. Hoeffding's inequality bounds the clean-data deviation by $\sqrt{\alpha\ln t/N_{ij}(t)}$ with high probability. An adversary who flips at most A outcomes per pair can shift the empirical win rate by no more than $A/N_{ij}(t)$. Combining these two contributions yields a radius that safely encloses the true preference, and a union bound over all pairs and times upgrades the per-pair guarantee to the stated high-probability bound.

Algorithm 2 *Attack-Aware RUCB (AA-RUCB)*

Input: Confidence parameter $\alpha \geq \frac{1}{2}$, horizon $T \in \{1, 2, \dots\} \cup \{\infty\}$
Initialize win matrix $\mathbf{W} = [w_{ij}] \leftarrow \mathbf{0}_{K\times K}$ ▷ w_{ij} counts wins of a_i over a_j.
Initialize set $B \leftarrow \emptyset$
for $t = 1$ **to** T **do**
  $\mathbf{U} \leftarrow \dfrac{\mathbf{W}}{\mathbf{W}+\mathbf{W}^\top} + \sqrt{\dfrac{\alpha \ln t}{\mathbf{W}+\mathbf{W}^\top}} + \dfrac{\alpha A\sqrt{\ln t}}{(\mathbf{W}+\mathbf{W}^\top)K^2}$ ▷ element-wise; define $\frac{x}{0} = 1$.
  $u_{ii} \leftarrow \frac{1}{2}$ **for all** $i = 1, \dots, K$.
  $C \leftarrow \{a_c \mid u_{cj} \geq \frac{1}{2},\ \forall j\}$.
  if $C = \emptyset$ **then**
    Choose a_c uniformly at random from $\{1, \dots, K\}$.
  else
    $B \leftarrow B \cap C$
    if $|C| = 1$ **then**
      $B \leftarrow C$ and set a_c to be the unique element of C.
    else
      Draw a_c from C with distribution

$$p(a_c) = \begin{cases} 0.5, & a_c \in B, \\ \dfrac{1}{2^{|B|}\,|C \setminus B|}, & \text{otherwise.} \end{cases}$$

    end if
  end if
  $a_d \leftarrow \arg\max_j u_{jc}$ ▷ If there is a tie, break ties uniformly; ensure $a_d \neq a_c$.
  Observe comparison (a_c, a_d) and update $\mathbf{W}$:

$$(w_{cd}, w_{dc}) \leftarrow \begin{cases} (w_{cd}+1,\ w_{dc}), & \text{if } a_c \text{ wins}, \\ (w_{cd},\ w_{dc}+1), & \text{otherwise.} \end{cases}$$

end for
Return arm a_c that beats the most opponents, i.e. the c maximizing $\#\{j \mid \frac{w_{cj}}{w_{cj}+w_{jc}} > \frac{1}{2}\}$.

Theorem IV.2.

$$\Pr\Big(\forall\, t > \tau_d(\delta_1),\ (i,j) \neq (1,1) : \\ N_{ij}(t) \leq \max\{\tau_d(\delta_1), D_{ij}(\Delta_i, \Delta_j, t)\}\Big) \ \geq\ 1 - \delta_1, \tag{6}$$

where

$$D_{ij}(\Delta_i, \Delta_j, t) \ =\ \Big[N_{\max}(\min\{\Delta_i, \Delta_j\}, t)\Big],$$

and

$$N_{\max}(\Delta, t) \ =\ \left[\frac{2\sqrt{\alpha}\,A/K^2}{\sqrt{1 + \dfrac{2\,A\,\Delta}{K^2\sqrt{\alpha \ln t}}} - 1}\right]^2. \tag{7}$$

Proof. See Appendix E. □

Theorem IV.2 provides a uniform, high-probability upper bound on the number of times any sub-optimal pair of arms is sampled once the burn-in phase up to $\tau_d(\delta_1)$ has elapsed. This bound is crucial for the regret analysis, because every additional comparison of a non-optimal pair contributes two units to the cumulative regret.

Beyond the time index $\tau_d(\delta_1)$ the algorithm always compares the Condorcet winner against itself. Suppose, for contradiction, that some pair (i, j) were drawn more often than the threshold in (7). In that case the inflated confidence interval used by AA-RUCB would already be narrower than the preference gap needed for the algorithm to select that pair, so the comparison could not have been chosen. Since $N_{\max}(\Delta, t)$ is monotone decreasing in the preference gap Δ, the same argument applies to every sub-optimal pair. A union bound over all such pairs then shows that the probability of exceeding the threshold for any pair is at most δ_1, completing the proof.

The following theorem gives the regret bound of the AA-RUCB algorithm. It shows that the regret remains logarithmic in T, with an extra additive term due to the adversary.

Theorem IV.3. *Define* $\Delta^* = \min_i \Delta_i$. *Let* $\tau_d(\delta_1)$ *be the threshold in* (5) *and let* $N_{\max}(\Delta, T)$ *be the quantity in* (7).

Then, with probability at least $1 - \delta_1$, *for all* $T \geq 1$, *the cumulative convergence regret* (1)

$$\begin{aligned} R(T) \ &\leq\ 2\,\tau_d(\delta_1) \ +\ 2\sum_{i>j} N_{\max}\big(\Delta^*,\, T\big) \\ &= O(K^2 \ln T + A\sqrt{\ln T}), \end{aligned} \tag{8}$$

where the sum $\sum_{i>j}$ *ranges over all unordered pairs* $\{i, j\}$ *with* $i, j \neq 1$.

Proof. See Appendix F. □

The main idea of the proof is as follow. Using the high-probability bounds from Lemma IV.1 and Theorem IV.2, the regret is split into two parts. The first $\tau_d(\delta_1)$ rounds can add at most $2\tau_d(\delta_1)$ because any two arms may be pulled. Afterwards, each distinct sub-optimal pair is compared no more than $N_{\max}(\Delta^*, T)$ times, so its contribution is at most $2N_{\max}(\Delta^*, T)$. Summing over all $\binom{K-1}{2}$ such pairs and adding the initial term yields the stated regret bound in $O(K^2 \ln T + A\sqrt{\ln T})$.

V. Experimental data and results

A. Attack

In this section, we provide numerical examples to validate the theoretical results obtained in this paper. 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 build two $K \times K$ preference matrices with the linear–rank rule $P[i, j] = 0.5 + g\,(r_j - r_i)/(2K)$, $P[j, i] = 1 - P[i, j]$ and $P[i, i] = 0.5$. Matrix $\mathbf{P}$ uses the natural order $r_i = i$, making arm 0 the Condorcet winner and arm $K - 1$

(the attacker's target) the weakest; the rank difference between these two arms is $-(K-1)$. Matrix $\tilde{\mathbf{P}}$ is a cyclic shift with $\tilde{r}_{K-1} = 0$ and $\tilde{r}_0 = K-1$, so the same pair has gap $+(K-1)$.

Choosing

$$g_1 = \big(0.5 - p_{ki}^{\min}\big)\frac{2K}{K-1}, \quad g_2 = \big(\tilde{p}_{ki}^{\max} - 0.5\big)\frac{2K}{K-1},$$

gives $P[K-1,0] = p_{ki}^{\min}$ and $\tilde{P}[K-1,0] = \tilde{p}_{ki}^{\max}$. We form $\tilde{\mathbf{P}}$ by copying $\mathbf{P}$ and overwriting only the target arm's row and column with slope g_2; hence the two matrices differ exclusively in entries that involve the target arm.

By default, we set $p_{ki}^{\min} = 0.05$ and $\tilde{p}_{ki}^{\max} = 0.95$. We also choose a time horizon $T = 5e5$ 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 III.3 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 III.2.

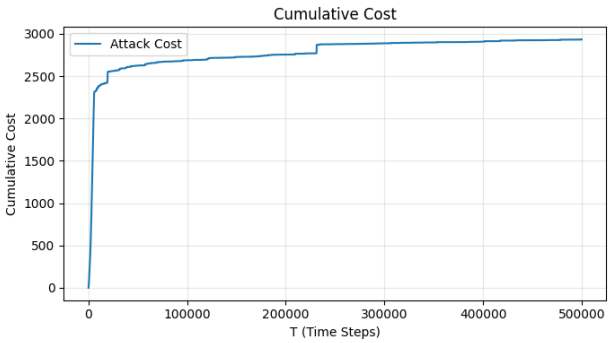

Fig. 1: Cumulative Cost

To probe how different problem characteristics influence both the attacker and the learner, we vary one quantity at a time while holding the others fixed and average each setting over three independent runs.

In Fig. 4 we gradually increase the enforced probability $\tilde{p}_{ki}^{\max}$, thereby enlarging the gap $\tilde{p}_{ki} - \frac{1}{2}$. As predicted by Theorem III.3, the dominant term in the upper bound becomes $1/\big(\tilde{p}_{ki} - \frac{1}{2}\big)^2$. And the attack regret equals the cost plus the number of comparisons that do not involve a_k as the proof in

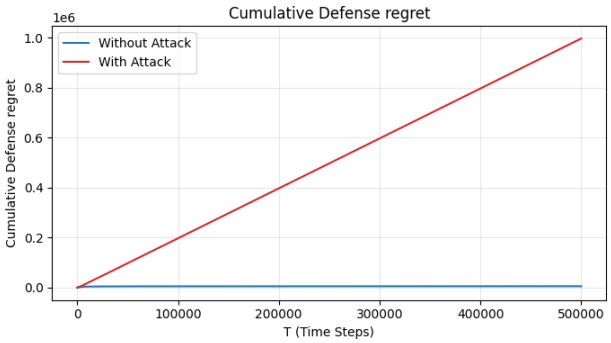

Fig. 2: Cumulative Defense Regret

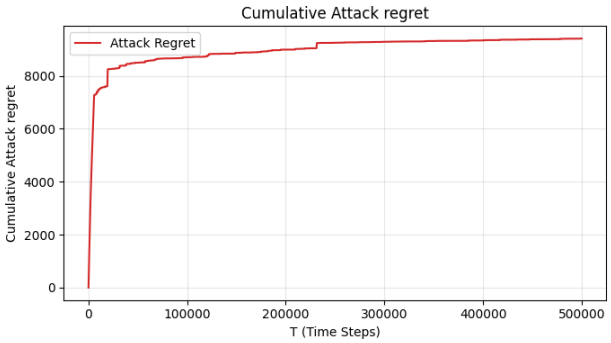

Fig. 3: Cumulative Attack Regret With Attack

Appendix B. Accordingly, both the attack cost and the attack regret fall with the expected quadratic rate.

Figure 5 shows that, as the number of arms K increases, the attack cost exhibits $\mathcal{O}(K \ln T)$ growth, whereas the attack regret scales as $\mathcal{O}(K^2 \ln T)$. These observations are consistent with Theorems III.3 and III.2.

Finally, Figure 6 illustrates that higher values of exploration parameter α induce more exploration and, consequently, larger attack cost and attack regret, in agreement with our theoretical analysis.

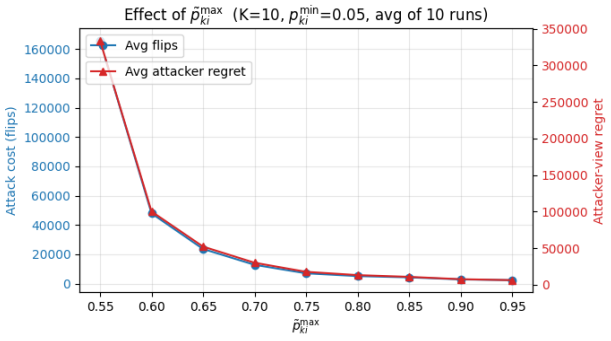

Fig. 4: Average flips and attacker-view regret versus $\tilde{p}_{ki}^{\max}$ ($p_{ki}^{\min} = 0.05$).

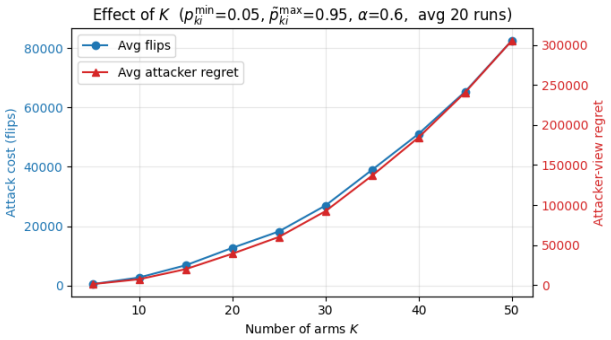

Fig. 5: Average flips and attacker-view regret versus the number of arms K ($p_{ki}^{\min} = 0.05,\ \tilde{p}_{ki}^{\max} = 0.95$).

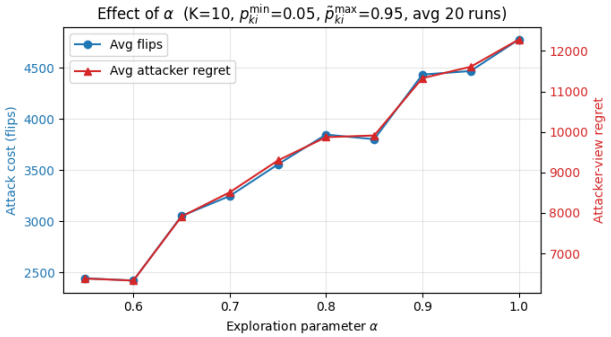

Fig. 6: Average flips and attacker-view regret versus the exploration parameter α ($K = 10,\ p_{ki}^{\min} = 0.05,\ \tilde{p}_{ki}^{\max} = 0.95$).

B. Defense

We complement the theoretical bound of Theorem IV.3 with a set of experiments that illustrate how the proposed defence algorithm behaves under varying time steps, adversarial budgets, and time horizons.

1) Timesteps: We set the horizon $T = 5e5$ and keep the adversary budget at $A = 10\sqrt{T}$. Figure 7 shows the total defense regret as a function of T. The curve grows sub-linearly, corroborating the favourable dependence on T predicted by our analysis.

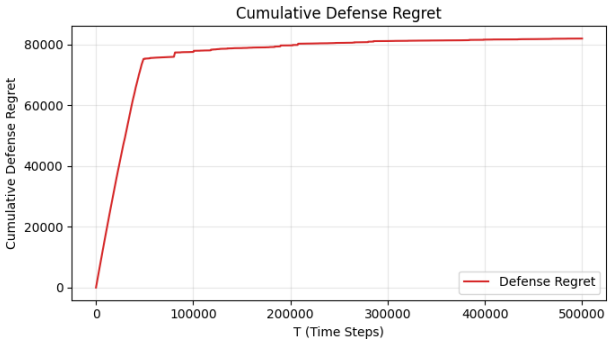

Fig. 7: Total defense regret versus timestep t.

2) Impact of the adversarial budget: Next we fix the horizon at $T = 5e5$ and change the attack budget from $5\sqrt{T}$ to $50\sqrt{T}$. Figure 8 reports the final total defense regret after T rounds. As expected, larger budgets enable the attacker to inflict higher defense regret, yet the curve follows $A \ln T$ growth as in (8) and remains well below the linear worst-case envelope $\mathcal{O}(T)$, indicating the robustness of the defence mechanism.

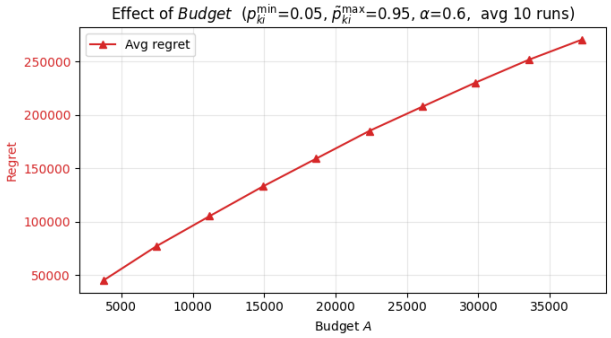

Fig. 8: Final total defense regret as a function of the adversarial budget.

3) Impact of the horizon T: We vary the horizon on the grid $T \in \{1e5, \ldots, 1e6\}$, keep the adversary's budget at $A = 10\sqrt{T}$, and set $(\alpha, \beta) = (0.6, 1/K^2)$. Figure 9 shows the mean total defense regret (sum over the two sub-regrets R_c and R_d) as a function of T. The curve grows sub-linearly, corroborating the favourable dependence on T predicted by our analysis.

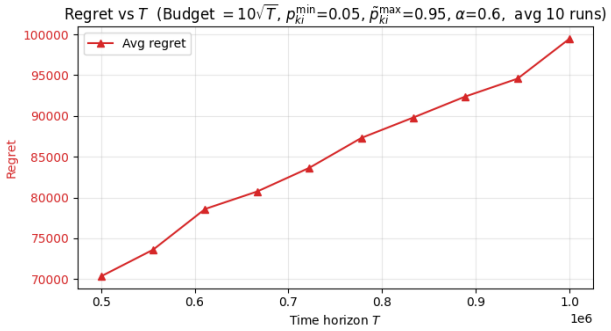

Fig. 9: Final total defense regret as a function of the horizon T.

VI. Conclusion

This work provides the first end-to-end study of both offensive and defensive strategies for dueling bandits under outcome corruption.

On the offensive side, 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.

On the defensive side we have introduced Attack-Aware RUCB (AA-RUCB), which inflates each confidence bound by a corruption-budget term $\frac{\alpha A \sqrt{\ln t}}{K^2 N_{ij}(t)}$. This single-line change retains RUCB's optimal $O(K^2 \ln T)$ regret when $A = 0$ and degrades gracefully to $O(K^2 \ln T + A\sqrt{\ln T})$ in the worst case. We have conducted comprehensive numerical experiments to validate the theoretic results.

REFERENCES

[1] Mo Lyu, Chenye Yang, Guanlin Liu, and Lifeng Lai. Adversarial post-action attacks on dueling bandits. In *Proc. Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2025.

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

[3] 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.

[4] Cem Tekin and Mihaela van der Schaar. Distributed online learning via cooperative contextual bandits. *IEEE Transactions on Signal Processing*, 63(14):3700–3714, 2015.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] Yi Gai and Bhaskar Krishnamachari. Distributed stochastic online learning policies for opportunistic spectrum access. *IEEE Transactions on Signal Processing*, 62(23):6184–6193, 2014.

[11] Saeed Bagheri and Anna Scaglione. The restless multi-armed bandit formulation of the cognitive compressive sensing problem. *IEEE Transactions on Signal Processing*, 63(5):1183–1198, 2015.

[12] V. Krishnamurthy and R.J. Evans. Hidden markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking. *IEEE Transactions on Signal Processing*, 49(12):2893–2908, 2001.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] Chenye Yang, Guanlin Liu, and Lifeng Lai. Stochastic bandits with non-stationary rewards: Reward attack and defense. *IEEE Transactions on Signal Processing*, 72:5007–5020, 2024.

[30] 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.

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

[32] 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.

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

[34] Bingcong Li, Tianyi Chen, and Georgios B. Giannakis. Secure mobile edge computing in iot via collaborative online learning. *IEEE Transactions on Signal Processing*, 67(23):5922–5935, 2019.

[35] Guanlin Liu and Lifeng Lai. Action-manipulation attacks against stochastic bandits: Attacks and defense. *IEEE Transactions on Signal Processing*, 68:5152–5165, 2020.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

APPENDIX

A. Proof of Theorem III.1

Proof. Let $\Gamma(t)$ be the event that $\forall i \neq k$,

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

and $a_i \notin C$ at iteration t.

For any $t > \max_i \tilde{t}_i$ and any $i \neq k$, if $\frac{w_{ik}(t)}{N_{ik}(t)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t)}} < \frac{1}{2}$, $a_i \notin C$, then, with $N_{ik}(t^*) = N_{ik}(t)$:

$$\begin{aligned} u_{ik}(t^*) &= \frac{w_{ik}(t^*)}{N_{ik}(t^*)} + \sqrt{\frac{\alpha \ln t^*}{N_{ik}(t^*)}} \\ &\leq \frac{w_{ik}(t^*)}{N_{ik}(t^*)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t^*)}} \\ &= \frac{w_{ik}(t)}{N_{ik}(t)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t)}} < \frac{1}{2}. \end{aligned} \tag{9}$$

Thus, for $t^* = t+1$, $\forall i \neq k,\ a_i \notin C$.

Since $\Gamma(\max_i \tilde{t}_i + 1)$ is true, from (9), for $t^* = \max_i \tilde{t}_i + 2$, $a_i \notin C$ and $\frac{w_{ik}(t^*)}{N_{ik}(t^*)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t^*)}} < \frac{1}{2}$, which means $\Gamma(t^*)$ is true. Suppose $\Gamma(t_m)$ holds for $t_m > \max_i \tilde{t}_i + 1$, $a_i \notin C$ and $\frac{w_{ik}(t_m)}{N_{ik}(t_m)} + \sqrt{\frac{\alpha \ln T}{N_{ik}(t_m)}} < \frac{1}{2}$. Then, $\Gamma(t_m + 1)$ holds from (9) if let $t = t_m$.

By induction, we can know that if $\Gamma(\max_i \tilde{t}_i + 1)$ holds, then $\Gamma(\max_i \tilde{t}_i + 1 + n)$ holds for any $n > 0$.

If $t = \max_i \tilde{t}_i + 1$, $\Pr\Big(\forall i \neq k,\ a_i \notin C\Big) > 1 - (K-1)\delta_2$ directly from Proposition A.1.

If $t > \max_i \tilde{t}_i + 1$:

$$\begin{aligned} &\Pr\Big(\forall t > \max_i \tilde{t}_i,\ \forall i \neq k,\ a_i \notin C\Big) \\ &= \Pr\Big(\forall t > \max_i \tilde{t}_i + 1,\ \forall i \neq k,\ a_i \notin C \mid \Gamma(\max_i \tilde{t}_i + 1\Big) \\ &\times \Pr\Big(\Gamma(\max_i \tilde{t}_i + 1)\Big) \\ &> 1 \times (1 - (K-1)\delta_2) = 1 - (K-1)\delta_2. \end{aligned}$$

□

B. Proof of Theorem III.2

Proof. Proposition 2 of Zoghi et al. [22] applies to any fixed preference matrix that admits a unique Condorcet winner and to any $\alpha > \frac{1}{2}$. Once the adversary fixes the modified matrix $\tilde{\mathbf{P}} = [\tilde{p}_{ij}]$ with arm a_k satisfying $\tilde{p}_{ki} > \frac{1}{2}$ for all $i \neq k$, the learner's observations are exactly those that would be generated by running Algorithm 1 on $\tilde{\mathbf{P}}$. Hence, from the learner's perspective, all assumptions of Proposition 2 hold with the gaps $\tilde{\Delta}_i = \tilde{p}_{ki} - \frac{1}{2}$ and the Condorcet winner a_k. We may therefore invoke that result to claim that, with probability at least $1 - \delta_1$, for time $T > \tau(\delta)$ any $i \neq k$ and $j \neq k$, $N_{ij}(T) \leq \frac{4\alpha \ln T}{\min(\tilde{\Delta}_i^2, \tilde{\Delta}_j^2)}$. For all combinations of $(i,j), i \neq j, i \neq k, j \neq k$,

$$\sum_{\substack{i \neq k \\ j \neq k}}^{i \neq j} N_{ij}(T) \leq \frac{(K-1)(K-2)}{2} \cdot \frac{4\alpha \ln T}{\min(\tilde{\Delta}_i^2, \tilde{\Delta}_j^2)}. \tag{10}$$

Proposition A.1. *Let a_i be any arm different from the attack target arm a_k,*

$$\Pr\Big(\forall t \geq \max_i \tilde{t}_i,\ \tfrac{w_{ik}(t)}{N_{ik}(t)} + \sqrt{\tfrac{\alpha \ln T}{N_{ik}(t)}} < \tfrac{1}{2}\Big) \ \geq\ 1 - (K-1)\delta_2, \tag{11}$$

so, with the same probability, $\forall i \neq k$, $a_i \notin C$.

Proof. See Appendix G. □

By Proposition A.1, $\Pr\Big(u_{ik}(t) < \frac{1}{2}\Big) \geq 1 - (K-1)\delta_2$ for every $t > \max_i \tilde{t}_i$. Whenever this inequality holds, a_i is excluded from the candidate set C and therefore cannot be selected as the first comparison arm c in round t. If instead $c = k$, the second arm $d = \arg\max_{j \neq k} u_{jk}(t)$ also differs from i because $u_{ik}(t) < u_{kk}(t) = \frac{1}{2} \leq u_{dk}(t)$. Hence the ordered pairs (i,k) and (k,i) are never chosen for comparison in that round, implying $N_{ik}(t+1) = N_{ik}(t)$.

Combining with Theorem III.1, for each arm a_i, where $i \neq k$, there are at most

$$\left\lceil \frac{\left(\sqrt{\alpha \ln T} + \sqrt{\frac{1}{2} \ln \frac{1}{\delta_2}}\right)^2}{\left(\frac{1}{2} - \tilde{p}_{ik}\right)^2} \right\rceil \tag{12}$$

comparisons with a_k.

Putting (10) and (12) with

$$R(t) = \sum_{h=1}^{t} \left(\mathbb{1}_{\{a_c(h) \neq a_k\}} + \mathbb{1}_{\{a_d(h) \neq a_k\}} \right), \tag{13}$$

we derive:

$$\begin{aligned}R_{attack}(T) &= 2 \cdot \sum_{\substack{i\neq k\\ j\neq k}}^{i\neq j} N_{ij}(T) + 1 \cdot \sum_{i\neq k} N_{ki}(T) \\ &\leq 2 \cdot \frac{(K-1)(K-2)}{2} \cdot \frac{4\alpha \ln T}{\min(\tilde{\Delta}_i^2, \tilde{\Delta}_j^2)} \\ &\quad + (K-1) \cdot \max_i \Big\{ \frac{\tilde{p}_{ki} - p_{ki}}{1 - p_{ki}} \Big\} \\ &\quad \cdot \max_i \Big\{ \frac{\left(\sqrt{\alpha \ln T} + \sqrt{\frac{1}{2} \ln \frac{1}{\delta_2}}\right)^2}{\left(\tilde{p}_{ki} - \frac{1}{2}\right)^2} \Big\} \\ &= \mathcal{O}(K^2 \ln T).\end{aligned}$$

And $T \geq R(T) \geq T - R_{attack}(T) = T - \mathcal{O}(K^2 \ln T) = \mathcal{O}(T)$.

□

C. Proof of Theorem III.3

Proof. (12) amounts to a total of

$$(K-1) \cdot \max_i \Big\{ \frac{\left(\sqrt{\alpha \ln T} + \sqrt{\frac{1}{2} \ln \frac{1}{\delta_2}}\right)^2}{\left(\tilde{p}_{ki} - \frac{1}{2}\right)^2} \Big\}$$

comparisons with a_k for all arms a_i, where $i \neq k$ and the $\max$ operation is taken over all i. After that, with $c = k$ and $d = k$, we can infer that the attack has succeeded.

Since the attack cost for each comparison between a_i and a_k, $i \neq k$, can be defined as a random variable that takes on a value of 1 if there is an attack and 0 if there is no attack.

Let PA_{ki} denote the probability that attack reverses the outcome $Z_{k,i}^t$ from 0 to 1, we can calculate the expected attack cost as follows:

$$\begin{aligned}L_{attack}^{ik} &= \begin{cases} 1 & \text{with probability } PA_{ki}, \\ 0 & \text{with probability } 1 - PA_{ki} \end{cases} \\ E(L_{attack}^{ik}) &= 1 \times PA_{ki} + 0 \times (1 - PA_{ki}) \\ &= PA_{ki}.\end{aligned}$$

To express PA_{ki} in terms of the estimated probability $\tilde{p}_{ki}$ and the actual probability p_{ki}, we use the given relationship:

$$PA_{ki} = \frac{\tilde{p}_{ki} - p_{ki}}{1 - p_{ki}}.$$

Substituting this back into the expression for the expected attack cost, we get:

$$E(L_{attack}^{ik}) = \frac{\tilde{p}_{ki} - p_{ki}}{1 - p_{ki}}.$$

It is bounded by $\max_i(\frac{\tilde{p}_{ki} - p_{ki}}{1 - p_{ki}})$ for all arms $i \neq k$. The total attack cost is

$$\begin{aligned}L_{attack}(T) =& (K-1) \cdot \max_i \Big\{ \frac{\tilde{p}_{ki} - p_{ki}}{1 - p_{ki}} \Big\} \\ & \cdot \max_i \Big\{ \frac{\left(\sqrt{\alpha \ln T} + \sqrt{\frac{1}{2} \ln \frac{1}{\delta_2}}\right)^2}{\left(\tilde{p}_{ki} - \frac{1}{2}\right)^2} \Big\}\end{aligned}$$

which is $\mathcal{O}(K \ln T)$. □

D. Proof of Lemma IV.1

Proof. Set $\hat{p}_{ij}(t) = w_{ij}(t)/N_{ij}(t)$. Without attacks, Hoeffding's inequality and a union bound over the K^2 ordered pairs give

$$\begin{aligned}&\Pr\Big(\exists\, t > t_0,\, i, j :\ \big|\hat{p}_{ij}(t) - p_{ij}\big| > \sqrt{\tfrac{\alpha \ln t}{N_{ij}(t)}}\Big) \\ &\qquad \leq \frac{(4\alpha - 1)K^2}{(2\alpha - 1)\, t_0^{2\alpha - 1}}. \quad (14)\end{aligned}$$

Choose

$$t_0 := \Big(\tfrac{(4\alpha-1)K^2}{(2\alpha-1)\delta}\Big)^{\frac{1}{2\alpha-1}}, \quad (15)$$

so that the right-hand side of (14) equals δ_1:

$$\Pr\Big(\exists\, t > t_0,\, i, j :\ \big|\hat{p}_{ij}(t) - p_{ij}\big| > \sqrt{\tfrac{\alpha \ln t}{N_{ij}(t)}}\Big) \leq \delta_1. \quad (16)$$

Let $w_{ij}^\star(t)$ denote the (unobserved) win count had no attacks occurred. Because the adversary flips at most A outcomes,

$$\big|w_{ij}(t) - w_{ij}^\star(t)\big| \leq A \quad \Longrightarrow \quad \big|\hat{p}_{ij}(t) - \hat{p}_{ij}^\star(t)\big| \leq \frac{A}{N_{ij}(t)}.$$

Define

$$t_1 := \exp\big(K^4/\alpha^2\big), \quad (17)$$

so that $A \leq \frac{\alpha\, A \sqrt{\ln t}}{K^2}$ for all $t \geq t_1$. Hence, whenever the no-attack concentration event in (16) holds, we have for every (i, j) and $t \geq t_1$

$$\big|\hat{p}_{ij}(t) - p_{ij}\big| \leq \sqrt{\tfrac{\alpha \ln t}{N_{ij}(t)}} + \tfrac{\alpha\, A \sqrt{\ln t}}{N_{ij}(t)\, K^2} = u_{ij}(t) - \hat{p}_{ij}(t).$$

Since $l_{ij}(t) = 1 - u_{ji}(t)$, the same bound yields $p_{ij} \in [l_{ij}(t), u_{ij}(t)]$.

Let $\tau_d(\delta_1))$ be as in (5); note that $\tau_d(\delta_1) = \max\{t_0, t_1\}$. With (16),

$$\begin{aligned}&\Pr\Big(\exists\, t > \tau_d(\delta_1),\, i, j :\ p_{ij} \notin [l_{ij}(t), u_{ij}(t)]\Big) \\ &\leq \Pr\Big(\exists\, t > t_0,\, i, j :\ \big|\tilde{p}_{ij}(t) - p_{ij}\big| > \sqrt{\tfrac{\alpha \ln t}{N_{ij}(t)}}\Big) \leq \delta_1, \quad (18)\end{aligned}$$

which completes the proof. □

E. Proof of Theorem IV.2

Proof. First, note that the algorithm never actually compares a sub-optimal arm against itself beyond the initial seeding phase. In particular, once $t > \tau_d(\delta_1)$, if the algorithm were to select arm a_i as both the champion and challenger in a round (i.e. $c = i$ and $d = i$), it would contradict the fact that $u_{ii}(t) = \frac{1}{2} < p_{1i} \leq u_{1i}(t)$ (since a_1 is the Condorcet winner). Thus, for $t > \tau_d(\delta_1)$ we have $N_{ii}(t) = 0$ for all $i \neq 1$, and trivial bounds hold for (i, i) pairs. We henceforth consider distinct arms $i \neq j$.

Suppose, for contradiction, that there exist distinct arms $i \neq j$ and a time $t > \tau_d(\delta_1)$ such that

$$N_{ij}(t) \ > \ N_{\max}\big(\min\{\Delta_i, \Delta_j\},\, t\big).$$

Let $s \leq t$ be the *last* round in which a_i and a_j were matched. Without loss of generality, assume the algorithm

chose $c = i$ and $d = j$ at time s (the opposite case is symmetric).

The choice $c = i$ implies

$$u_{ij}(s) \;\geq\; \tfrac{1}{2}, \tag{19}$$

for otherwise arm a_i would not have been a plausible champion against a_j. Likewise, selecting $d = j$ requires $u_{ji}(s) \geq u_{1i}(s)$; because $p_{1i} \leq u_{1i}(s)$ under the high-probability event, this gives

$$l_{ij}(s) = 1 - u_{ji}(s) \;\leq\; 1 - p_{1i} = p_{i1}. \tag{20}$$

Subtracting (20) from (19) yields

$$u_{ij}(s) - l_{ij}(s) \;\geq\; \tfrac{1}{2} - p_{i1} \;=\; \Delta_i. \tag{21}$$

On the other hand, the attack-aware formula for the interval width gives

$$u_{ij}(s) - l_{ij}(s) \;=\; 2\sqrt{\frac{\alpha \ln s}{N_{ij}(s)}}\Big(1 + \sqrt{\tfrac{\alpha}{N_{ij}(s)}}\,\tfrac{A}{K^2}\Big). \tag{22}$$

Rewrite the inequality

$$2\sqrt{\frac{\alpha \ln s}{N_{ij}(s)}}\Big(1 + \sqrt{\frac{\alpha}{N_{ij}(s)}}\,\frac{A}{K^2}\Big) \;\geq\; \Delta_i. \tag{23}$$

Let $x := N_{ij}(s)^{-1/2} > 0$. Inequality (23) becomes the quadratic condition

$$Bc\,x^2 + B\,x - \Delta_i \;\geq\; 0, \qquad B := 2\sqrt{\alpha \ln s}, \quad c := \sqrt{\alpha}\,\frac{A}{K^2}.$$

Since $Bc > 0$, the inequality holds only for $x \leq x_\star$, where

$$x_\star \;=\; \frac{\sqrt{1 + \dfrac{2c\Delta_i}{B}} - 1}{2c}.$$

Re-expressing in terms of $x \leq x_\star$ yields (7) with $\Delta = \Delta_i$: $N_{ij}(s) \leq N_{\max}(\Delta_i, s)$.

Because $N_{ij}(t)$ exceeds the threshold in (7) with $\Delta = \Delta_i$, the right-hand side of (22) is *strictly smaller* than Δ_i.

Consequently

$$u_{ij}(s) - l_{ij}(s) \;<\; \Delta_i. \tag{24}$$

Inequalities (21) and (24) are incompatible, giving the desired contradiction.

The symmetric argument (with i and j interchanged) yields

$$u_{ji}(s) - l_{ji}(s) \;\geq\; \tfrac{1}{2} - p_{j1} \;=\; \Delta_j. \tag{25}$$

Consequently

$$N_{ji}(s) \;\leq\; N_{\max}(\Delta_j, s).$$

The function $N_{\max}(\Delta, t)$ in (7) is decreasing in the gap argument Δ (larger gaps require fewer comparisons) and non-decreasing in the time index t.

Because $\min\{\Delta_i, \Delta_j\} \leq \Delta_i, \Delta_j$ and $s \leq t$, it follows that

$$\begin{aligned} N_{\max}(\Delta_i, s) \;&\leq\; N_{\max}\big(\min\{\Delta_i, \Delta_j\},\, s\big) \\ &\leq\; N_{\max}\big(\min\{\Delta_i, \Delta_j\},\, t\big) \end{aligned}$$

and

$$N_{\max}(\Delta_j, s) \;\leq\; N_{\max}\big(\min\{\Delta_i, \Delta_j\},\, t\big).$$

Hence each pair–specific bound $N_{ij}(s) \leq N_{\max}(\Delta_i, s)$ and $N_{ji}(s) \leq N_{\max}(\Delta_j, s)$ is no larger than the common threshold $N_{\max}\big(\min\{\Delta_i, \Delta_j\},\, t\big)$, justifying the use of this single worst-case bound in the contradiction argument and in the remainder of the proof.

Hence $N_{ij}(t) \;\leq\; N_{\max}\big(\min\{\Delta_i, \Delta_j\},\, t\big)$ must hold for every pair $i \neq j$ when $t > \tau_d(\delta_1)$.

Finally, for $t \leq \tau_d(\delta_1)$ we trivially have $N_{ij}(t) \leq t \leq \tau_d(\delta_1)$. Combining the two time ranges, we conclude that

$$N_{ij}(t) \;\leq\; \max\big\{\tau_d(\delta_1),\, N_{\max}\big(\min\{\Delta_i, \Delta_j\},\, t\big)\big\}$$

for all distinct i, j with probability at least $1 - \delta_1$, completing the proof.

□

F. Proof of Theorem IV.3

Proof. Work on the $1 - \delta_1$ event granted by Lemma IV.1 and Theorem IV.2, so all confidence intervals and comparison bounds hold.

During the first $\tau_d(\delta_1)$ rounds the algorithm may pull any arms. Each round can contribute at most 2 to $R(t)$ (both selected arms could be sub-optimal). Hence the total regret accumulated in rounds $1, \ldots, \tau_d(\delta_1)$ is at most $2\,\tau_d(\delta_1)$.

After round $\tau_d(\delta_1)$, Fix a pair of distinct sub-optimal arms $i, j \neq 1$. Every comparison between a_i and a_j contributes exactly 2 to the regret, and by Theorem IV.2 such a pair is selected at most $N_{\max}(\Delta^*, t)$ times up to round t. Hence the regret due to this pair is bounded by $2\,N_{\max}(\Delta^*, t)$. Summing over all unordered pairs $\{i, j\}$ with $i, j \neq 1$ yields a post-threshold regret bound of $2\sum_{i>j} N_{\max}(\Delta^*, t)$.

Adding the contribution from the initial phase to that from the post-threshold phase gives

$$R(t) \;\leq\; 2\,\tau_d(\delta_1) \;+\; 2\sum_{i>j} N_{\max}\big(\Delta^*,\, t\big),$$

which holds on the $1 - \delta_1$ event, completing the proof.

Putting the explicit formulas for $\tau_d(\delta_1)$ and $N_{\max}(\Delta_{ij}, t)$ back into the regret bound gives the fully expanded inequality

$$\begin{aligned} R(t) \;&\leq 2\,\max\Big\{e^{\,K^4/\alpha^2},\; \Big(\tfrac{(4\alpha-1)K^2}{(2\alpha-1)\,\delta_1}\Big)^{\frac{1}{2\alpha-1}}\Big\} \\ &\quad +\; (K-1)(K-2)\left[\frac{2\sqrt{\alpha}\,A/K^2}{\sqrt{1 + \dfrac{2\,A\,\Delta^*}{K^2\sqrt{\alpha \ln t}}} - 1}\right]^2 \\ &= O(K^2 \ln T + A\sqrt{\ln T}). \end{aligned} \tag{26}$$

□

G. Proof of Proposition A.1

Proof. Abbreviate $N := N_{ik}(t)$ and $a := \sqrt{\alpha \ln T}$.

Because the empirical wins satisfy $w_{ik}(t) \sim \mathrm{Bin}(N, \tilde{p}_{ik})$, the upper-confidence statistic can be written as $u_{ik}(t) = \frac{w_{ik}(t)}{N} + a/\sqrt{N}$. Denote the bonus term by $c := \frac{a}{\sqrt{N}}$. To bound the event $u_{ik}(t) \geq \frac{1}{2}$ observe that $u_{ik}(t) \geq \frac{1}{2} \Leftrightarrow \frac{w_{ik}(t)}{N} - \tilde{p}_{ik} \geq \tilde{\Delta}_i - c$. Hoeffding's inequality therefore gives

$$P\left((u_{ik}(t) \geq \tfrac{1}{2}\right) \leq \exp\!\left[-2N(\tilde{\Delta}_i - c)^2\right].$$

We now show that this exponent does not exceed $-\ln(1/\delta_2)$ once $N \geq N_{ik}^{\min}$. Set $s := \sqrt{N}$; then $c = a/s$ and the inequality $2N(\tilde{\Delta}_i a - c)^2 \geq \ln(1/\delta)$ becomes $2\tilde{\Delta}_i^2 s^2 - 4\tilde{\Delta}_i a\, s + 2a^2 - \ln(1/\delta_2) \geq 0$. Because the quadratic has positive leading coefficient, it is non-negative for every s not less than the larger root $s_+ = (a + \sqrt{\tfrac{1}{2}\ln(1/\delta_2)})/\tilde{\Delta}_i$. Consequently the bound holds for every $N = s^2 \geq s_+^2$, and the ceiling in the definition of $N_{ik}^{\min}$ guarantees precisely this condition. Hence $P\left(u_{ik}(t) \geq \tfrac{1}{2}\right) \leq \delta_2$ and (11) follows.

Finally we justify that $\tilde{\Delta}_i - c > 0$ under the same requirement. Since $x \mapsto a/\sqrt{x}$ is decreasing, $c \leq a/\sqrt{N_{ik}^{\min}}$. But $N_{ik}^{\min} \geq (a + \sqrt{\tfrac{1}{2}\ln(1/\delta_2)})^2/\tilde{\Delta}_i^2$, so $c \leq a\tilde{\Delta}_i/(a + \sqrt{\tfrac{1}{2}\ln(1/\delta_2)}) < \tilde{\Delta}_i$.

Now consider all $K-1$ non-target arms and let $E_i := \{\forall t \geq \tilde{t}_i,\ u_{ik}(t) < \tfrac{1}{2}\}$. Applying the union bound yields

$$\Pr\Big(\bigcap_{i \neq k} E_i\Big) \;\geq\; 1 - \sum_{i \neq k} \Pr(E_i^{\mathrm{c}}) \;\geq\; 1 - (K-1)\delta_2.$$

Hence, with probability at least $1 - (K-1)\delta_2$, there exists the deterministic time $\tilde{t}_{\max} := \max_{i \neq k} \tilde{t}_i$ after which $u_{ik}(t) < \frac{1}{2}$ for every $i \neq k$ and every $t \geq \tilde{t}_{\max}$. Because Line 7 of RUCB includes an arm in the candidate set C only if its upper bound is at least $\frac{1}{2}$, no non-target arm can belong to C after $\tilde{t}_{\max}$. □