Achieving Nearly-Optimal Regret and Sample Complexity in Dueling Bandits with Applications in Online Recommendations

Lanjihong Ma
National Key Laboratory for Novel Software Technology, School of Artificial Intelligence, Nanjing University
maljh@lamda.nju.edu.cn

Yao-Xiang Ding
State Key Lab of CAD & CG, Zhejiang University
dingyx.gm@gmail.com

Zhen-Yu Zhang
Center for Advanced Intelligence Project, RIKEN
zhen-yu.zhang@riken.jp

Zhi-Hua Zhou*
National Key Laboratory for Novel Software Technology, School of Artificial Intelligence, Nanjing University
zhouzh@lamda.nju.edu.cn

ABSTRACT

We focus on the dueling bandits problem, which has recently drawn significant attention due to its wide-ranging applications in online recommendation systems and the alignment of *large language models* (LLMs), considers an online preference learning scenario where the learner iteratively selects arms based on pairwise comparison feedback to infer user preferences. Two primary objectives are typically considered in dueling bandits: *Regret Minimization* (RM), which aims to improve the overall quality of selected arms over time, and *Best Arm Identification* (BAI), which seeks to efficiently identify the best item with minimal user feedback. For instance, RM is exemplified by the objective of consistently providing high-quality items, while BAI reduces the required human feedback by minimizing the number of necessary comparisons. Conventional research treats RM and BAI as two conflicting objectives, optimizing one at the expense of the other. In this paper, we propose a novel framework that demonstrates the near-consistency of RM and BAI in dueling bandits by reducing the BAI in dueling bandits into a sequential noisy identification problem. Based on our formulation, we propose a black-box reduction technique that transforms any RM algorithm into a BAI algorithm, and prove that such reduction with optimal RM algorithm achieves *optimal sample complexity and nearly-optimal cumulative weak regret simultaneously*. Our proposed algorithm acheives a nearly-optimal BAI sample complexity and attains a cumulative weak regret that is order-wise equivalent to the best-known result simultaneously. Experiments on both synthetic benchmarks and real-world online recommendation tasks validate the effectiveness of the proposed method, providing empirical evidences for our theoretical findings.

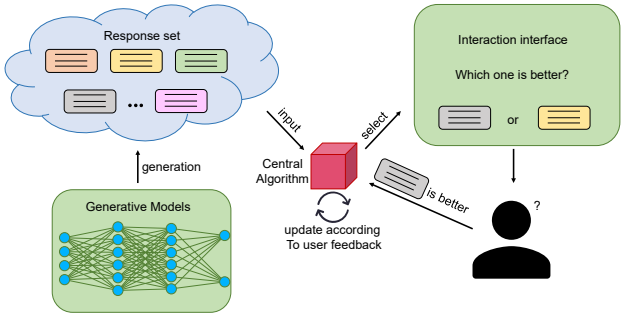

Figure 1: Preferece-based learning from comparison feedback. The generative models first generate several responses, among which the central algorithm selects two responses and return to the user. After the user gives his/her preference, the comparison result is returned to the central algorithm, updating the preferences for better future choices.

CCS CONCEPTS

• **Computing methodologies** → **Online learning settings**; **Learning from implicit feedback**.

KEYWORDS

preference learning, online decision making, regret minimization, best arm identification

ACM Reference Format:
Lanjihong Ma, Yao-Xiang Ding, Zhen-Yu Zhang, and Zhi-Hua Zhou. 2025. Achieving Nearly-Optimal Regret and Sample Complexity in Dueling Bandits with Applications in Online Recommendations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3690624.3709279

1 INTRODUCTION

Preference-based learning is a machine learning approach that focuses on understanding and predicting user preferences through interactions. Unlike conventional methods that rely heavily on fixed datasets, preference-based learning relies on the real-time user feedback to adjust and improve the model's predictions dynamically. This method is particularly relevant in contexts where personalization is key, such as in online recommendation systems [6, 36,

*Zhi-Hua Zhou is the corresponding author.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1245-6/25/08
https://doi.org/10.1145/3690624.3709279

37] and human-machine interactions. Recent advancements have drawn significant interest in preference-based learning, particularly for its critical role in training and fine-tuning Large Language Models (LLMs) [22, 23, 26]. LLMs leverage preference-based learning to adapt to individual user needs by processing feedback and refining their responses, thereby enhancing user experience and engagement. This capability is crucial in applications where user satisfaction and engagement are paramount, as it allows LLMs to provide more relevant and context-aware responses.

According to the types of human feedback, the central algorithmic designs can vary significantly [15]. In this paper, we focus on the scenario where *users provide implicit comparison feedback of the chosen arms*, also known as the dueling bandits problem [3, 33]. We take how dueling bandits works in fine-tuning LLMs as an illustrative example in Figure 1. Specifically, after generating several possible responses to form a response set, the learning system chooses two of the responses and returns them to the user, the user then selects one of them, indicating which of the responses may be preferred by the user. By simplifying the type of user feedback to simple "which one is better?" rather than ratings, demonstrations or prompts, the dueling bandits problem offers straightforward interactions to understand and adapt to the users' preferences.

The central concern within dueling bandits is to sequentially decide which two of the arms are selected to present to users. Mainly two essential but potentially conflicting objectives are considered to guide this selection: *regret minimization (RM)* and *best arm identification (BAI)*. Specifically, in RM, the learner aims to minimize the quality difference between the current selection and the best selection in hindsight, and thus focus more on the overall qualities of the selection sequance; while for BAI, the learner aims to minimize the number of interactions to identify the best response within the response set, and thus whether the current responses are explored is more important than their qualities. Taking the online recommendation tasks [6, 36, 37] for instance: RM ensures that the recommendations provided to users are of high quality, thereby enhancing user satisfaction and engagement over time; while BAI, on the other hand, focuses on efficiently identifying the most preferred items or products within minimal interactions, reducing the amount of feedback needed from users and thus improving the overall user experience by quickly adapting to their preferences. Similarly, in fine-tuning process of LLMs [23], after generating several responses, pairs of responses are returned to the human raters for comparison. The raters' preferences are used to train a reward model, which is further used to evaluate and refine LLMs. Within the process, the objective of RM represents providing high-quality responses during the interaction, corresponding to the quality of the learned reward model; meanwhile, BAI represents minimizing the number of feedback required from human raters' choices, corresponding to the workloads of human raters. In both scenarios, RM and BAI play important roles during these processes.

The pioneer work of Audibert et al. [2] established that regret minimization (RM) and best-arm identification (BAI) are conflicting objectives in conventional multi-armed bandits (MAB) with numerical rewards, where algorithms optimal for RM are sub-optimal for BAI, and vice versa. Given that MAB represents the foundational model in bandit learning, it is widely accepted that RM and BAI may also conflict in other bandit learning scenarios. Most existing research on dueling bandits focuses sorely on either RM or BAI. To the best of our knowledge, our work is the first to systematically investigate the potential compatibility of these two objectives in various bandit scenarios. In this paper, we try to take a step towards bridging this gap by answering the following question:

Can we design a single algorithm for dueling bandits that can achieve near-optimal performance in both objectives of regret minimization (RM) and best arm identification (BAI)?

We provide a *positive* answer to the above question, by proposing a dueling bandits algorithm that is nearly-optimal in both regret for RM and sample complexity for BAI. Specifically, we first reduce the BAI in dueling bandits in the to a sequential noisy identification problem, where we present the intuitions on how dueling bandits problem in the objective of BAI and RM is nearly-consistent. Based on these findings, we then propose a black-box reduction framework that transforms any RM algorithm into a BAI algorithm under fixed confidence, that can achieve *dual optimality*: it safeguards a nearly optimal BAI sample complexity and attains a cumulative weak regret that is order-wise equivalent to the best-known result simultaneously. To the best of our knowledge, this is the first dueling bandit algorithm that achieves such kind of dual optimality via a single algorithm. Finally, we conduct extensive experiments on synthetic benchmarks and real-world tasks to validate the efficacy of the proposed approaches.

2 PROBLEM FORMULATION

In this section, we formalize the dueling bandits problems, including the learning procedure, the assumptions we made, the performance measures for learning objectives, and finally the quantities that reflect learning difficulties.

Consider a scenario with a confidence threshold of $1-\delta$ and a finite set of arms $\mathcal{X} = [N] := \{1, 2, \ldots, N\}$. In each interaction round, the learner must choose two arms from $\mathcal{X}$, according to the selection the human feedback is returned to the learner, indicating a possible preference for one arm over the other. The final target is to identify the best arm, also known as the Condorcet winner[1], with a confidence of at least $1-\delta$. Before entering the specific learning procedure, we first introduce the following assumption and define the Condorcet winner,

Assumption 1 (Total order assumption). *There exists a ground truth order over $\mathcal{X}$, and thus there exists a unique best arm.*

Definition 2.1 (The Condorcet winner). Given a set of arms $\mathcal{X} := [N]$, the Condorcet winner, denoted by a^*, is the arm that has the highest minimum winning probability against the most challenging opponent in the set, formally defined as:

$$a^* \triangleq \arg\max_{a\in\mathcal{X}} \left(\min_{j\in\mathcal{X}, j\neq a} p_{a,j} \right),$$

where $p_{a,j}$ represents the probability of arm a winning against arm j. According to Assumption 1, we have $p_{a^*,j} \geq \frac{1}{2}$ for the Condorcet winner a^* against any other arm j in $\mathcal{X}$.

[1]We use the term Condorcet winner as the best arm in dueling bandits from now on.

The learning procedure. Next, we introduce the learning procedure is details. Suppose a total of T rounds of interactions in a dueling bandit problem over a set of arms $\mathcal{X} := [N]$. At each time step $t = 1, \ldots, T$, the process is as follows:

(1) The learner selects a pair of arms indexed by (i_t, j_t) from $\mathcal{X}$. Upon this selection, the learner incurs a regret $r(i_t, j_t, a^*)$ unknown to the learner. The regret quantifies the loss of not choosing the Condorcet winner a^*.
(2) The learner receives feedback $y_t \sim \text{Bernoulli}(p_{i_t,j_t})$, where $p_{i_t,j_t} \in [0, 1]$ represents the ground truth probability of arm i_t winning against arm j_t. This probability is determined by latent factors unknown to the learner.

We now transite to the objectives in dueling bandits.

Objective 1: Regret Minimization (RM). The cumulative regret, denoted by R_T, is the summation of regret incurred over T time periods, formally defined as $R_T := \sum_{t=1}^{T} r(i_t, j_t, a^*)$. The objective of RM is to minimize R_T through strategic selection of the pair (i_t, j_t) at each time t. There are primarily two types of regret commonly studied: strong and weak regret. Specifically, the *strong regret* at time t is defined as $r_s(i_t, j_t, a^*) \triangleq \mathbf{1}_{\{i_t=a^* \wedge j_t=a^*\}}$ is the selected arm pair and a^* is the Condorcet winner. This metric assesses the average quality of both selected arms. To minimize strong regret, the Condorcet winner must be chosen twice and compared with itself, a requirement often too stringent. In constrast, *weak regret*, despite its name, offers a more pragmatic metric. The weak regret at time t is defined as

$$r_w(i_t, j_t, a^*) \triangleq \mathbf{1}_{\{i_t=a^* \vee j_t=a^*\}}. \tag{1}$$

Unlike strong regret, minimizing weak regret only requires one of the chosen arms to be the Condorcet winner, avoiding selecting the same arm twice. Due to the more practical definition suitable for real-world scenarios, we focus on minimizing weak regret when refer to the objective of RM in this paper.

Objective 2: Best Arm Identification (BAI). The objective of BAI in this paper refers to identifying the Condorcet winner within a fixed confidence level, also known as the δ-PAC BAI. Specifically, the learner is required to identify the Condorcet winner from the arm set with probability at least $1 - \delta$, with the fewest possible interactions. Therefore, the critical metric for BAI is the sample complexity required to meet these criteria.

Finally, we list some critical metrics that quantify the hardness of learning preferences in a dueling bandits problem.

- **Probability Gap:** The probability gap for arms i and j is defined as $\Delta_{i,j} = |p_{i,j} - 0.5|$. This metric indicates the challenge in distinguishing the superior arm in the pair (i, j). The closer $p_{i,j}$ is to 0.5, or equivalently, the closer $\Delta_{i,j}$ is to 0, the more challenging it becomes to differentiate between the two arms through comparisons.
- **Condorcet Minimum Gap:** Similarly, the Condorcet minimum gap, denoted by Δ^*, is then the probability gap between the Condorcet winner a^* and the strongest sub-optimal arm, formally defined as $\Delta^* := \min_{i \neq a^*} \Delta_{a^*,i}$. The Condorcet minimum gap represents how hard it is to distinguish the Condorcet winner from other arms.
- **Global Minimum Gap:** We define the *global minimum gap* as $\Delta = \min_{i \neq j} \Delta_{i,j}$, which represents the smallest probability gap between any two distinct arms within the set. Similar to other gaps, this metric serves as a quantifier for the overall difficulty of the problem instance. By definition, the global minimum gap is strictly less or equal to the Condorcet minimum gap, i.e., $\Delta \leq \Delta^*$.

3 PROPOSED APPROACH

In this section, we first reduce BAI in dueling bandits as a noisy identification process, according to which we explain why BAI and RM in dueling bandits is almost consistent. Upon this finding, we further propose a black-box reduction mechanism that transforms any RM algorithm to a BAI algorithm under the fixed confidence setting, that provably achieves dual optimalities in RM and BAI.

3.1 BAI in dueling bandits: a noisy identification approach

We start by focusing sorely on one of the objectives: Best Arm Identification (BAI) within the context of dueling bandits. Recall that, in each round, after the learners' choice of two arms, feedback $y_t \sim \text{Bernoulli}(p_{i_t,j_t})$ is returned. Given that y_t is a stochastic outcome subject to randomness, a single comparison between arms fails to accurately show their underlying relationship. Hence, multiple rounds of feedback are needed to better understand the true relationship between the arms due to the random nature of each duel's outcome.

To address this challenge, we have developed a two-phase noisy identification procedure. The first phase, denoising, aims to establish reliable pairwise relationships between the chosen arms through repeated comparisons. The second phase, identification, utilizes these established relationships to accurately identify the Condorcet winner, implemented by sequentially eliminating the arms that are weaker in the pair. This process is detailed in Algorithm 1. Specifically, the learner initially selects two arms at random and compares them multiple times. After several comparisons, the average winning rate concentrates to p_{i_t,j_t}. We then construct the confidence region C_{i_t,j_t} as follows:

$$C_{i_t,j_t} := [\hat{\mu}_{i_t,j_t} - c_\delta(i_t, j_t), \hat{\mu}_{i_t,j_t} + c_\delta(i_t, j_t)], \tag{2}$$

where $\hat{\mu}_{i_t,j_t}$ represents the empirical mean, and $c_\delta(i_t, j_t)$ is the confidence radius, formally defined as:

$$\hat{\mu}_{i_t,j_t} = \frac{\mathbf{W}_{i_t,j_t}}{n_{i_t,j_t}}, \quad c_\delta(i_t, j_t) = \sqrt{\frac{\log\left(\frac{8N}{\delta} \cdot n_{i_t,j_t}^2\right)}{2n_{i_t,j_t}}},$$

where $\mathbf{W}_{i_t,j_t}$ maintains the historical winning counts of arm i_t winning arm j_t, and n is the total historical number of comparison between arm i_t and arm j_t, and thus $n_{i_t,j_t} := \mathbf{W}_{i_t,j_t} + \mathbf{W}_{j_t,i_t}$. If $1/2$ does not locate in this confidence region, it represents the learner can correctly identify the better arm in (i_t, j_t) with high probability, and the denoise phase achieves its goal by now. The following lemma states the sample complexity needed in the denoising phase:

Algorithm 1 Simple Pairwise Elimination

Require: Set of arms $\mathcal{X}$, confidence level $1-\delta$.

Randomly select a pair of arms (i_1, j_1) from $\mathcal{X}$.
for $t = 1$ **to** T **do**
  Denoising Phase:
  Compare arms (i_t, j_t) and observe $y_t \sim \text{Bernoulli}(p_{i_t,j_t})$.
  Update the confidence region according to Equation (2), until $1/2$ is excluded from the region.
  Identification Phase:
  Eliminate the weaker arm b from $\mathcal{X}$, i.e., $\mathcal{X} \leftarrow \mathcal{X} \setminus \{b\}$.
  if $|\mathcal{X}| = 1$ **then**
    Break
  end if
end for
return The remaining arm in $\mathcal{X}$ as the Condorcet winner.

Algorithm 2 Pairwise Elimination (PE)

Require: Set of arms $\mathcal{X}$, confidence level $1-\delta$, weak regret minimization algorithm $\mathcal{A}$.

for $t = 1, \ldots, T$ **do**
  Run $\mathcal{A}$ **to select** a pair of arms (i_t, j_t) from $\mathcal{X}$.
  Denoising Phase:
  Compare arms (i_t, j_t) and observe $y_t \sim \text{Bernoulli}(p_{i_t,j_t})$.
  Update the confidence region according to Equation (2), until $1/2$ is excluded from the region.
  Identification Phase:
  Eliminate the weaker arm b from $\mathcal{X}$, i.e., $\mathcal{X} \leftarrow \mathcal{X} \setminus \{b\}$.
  if $|\mathcal{X}| = 1$ **then**
    Break
  end if
end for
return The remaining arm in $\mathcal{X}$ as the Condorcet winner.

Lemma 3.1. *The number of comparisons between pair (i, j) is at most $O\left(\frac{1}{\Delta_{i,j}^2}\log\left(\frac{N}{\Delta_{i,j}\delta}\right)\right)$ with probability at least $1-\frac{\delta}{2N}$, where $\Delta_{i,j}$ is the probability gap between pair (i, j).*

After a limited number of comparisons that denoise the relative preferences between a selected pair, the algorithm progresses to the identification phase. Recall that the objective in BAI is to identify the Condorcet winner rather than obtaining the ranking over all arms, this can be efficiently achieved through simple linear search through the arm set, requiring at most $N-1$ denoising phases.

During the identification phase, the algorithm systematically eliminates the weaker arm b, defined by the condition:

$$\hat{\mu}_{a,b} - c_\delta(a,b) \geq \frac{1}{2}, \quad \text{for } a, b \in \mathcal{X}, a \neq b, \tag{3}$$

where $\hat{\mu}_{a,b}$ and $c_\delta(a,b)$ again are the empirical mean and confidence radius. The algorithm proceeds by sequentially eliminating the weaker arm b from the set of arms, retaining the stronger arm a for the next round of comparisons. In each round, one arm is chosen based on the previous round's winner, while the other is selected randomly. This process continues until only one arm remains, which is then declared the Condorcet winner. Despite the apparent simplicity of this approach, the following theorem establishes that Algorithm 1 can identify the Condorcet winner with high probability and optimal sample complexity.

Theorem 3.2. *Algorithm 1 correctly identifies the Condorcet winner with a sample complexity of at most $O\left(\frac{N}{\Delta^2}\log\left(\frac{N}{\Delta\delta}\right)\right)$, with a probability of at least $1-\frac{\delta}{2}$, where Δ is the global minimum gap.*

Remark 1. *This result not only establishes a high probability of success but also provides an upper bound on the sample complexity required. It is noteworthy that the theoretical lower bound for the sample complexity of any dueling bandits algorithm to identify the Condorcet winner is $\Omega\left(\frac{N}{\Delta^2}\log\left(\frac{1}{\delta}\right)\right)$ [10, 14]. Therefore, Algorithm 1 achieves optimality up to logarithmic factors.*

Though Algorithm 1 gives a promising results in BAI, it is not the case when it comes to Regret Minimization (RM). In the following part, our focus turns to RM and analyze the consistency between BAI and RM in dueling bandits.

3.2 RM is nearly-consistent with BAI in dueling bandits

Our discussion now turns to the objective of Regret Minimization (RM). We recall that the definition of regret for a chosen pair of arms (i_t, j_t) is $\mathbf{1}_{\{i_t=a^* \wedge j_t=a^*\}}$ for strong regret, and $\mathbf{1}_{\{i_t=a^* \vee j_t=a^*\}}$ for weak regret. It is emphasized that strong regret is always incurred whenever two distinct arms are selected, and thus we focus on the more relaxed setting of weak regret in this paper. Notably, the cumulative regret depends on the specific sequence of arm selection, which is not addressed by Algorithm 1 that selects (i_t, j_t) randomly. In this scenario, the cumulative regret is contingent on the timing of the selection of the Condorcet winner a^*. In the optimal case, selecting a^* as either i_1 or j_1 ensures its consistent preference, thereby minimizing the cumulative regret. Conversely, if a^* is chosen as the final arm in the sequence $\mathcal{X}$, regret is incurred in almost every preceding round, leading to a substantial increase in cumulative regret. Consequently, to minimize cumulative regret, the selection sequence should prioritize the early identification and selection of the Condorcet winner.

Additionally, the sample complexity can be further reduced by leveraging problem-specific characteristics. In the denoising phase, repeated comparisons are conducted between pairs of selected arms. According to Lemma 3.1, the sample complexity required during this phase is inversely related to the probability gap $\Delta_{i,j}$ between the selected pair (i, j). Randomly selecting pairs with small $\Delta_{i,j}$ values can result in higher sample complexity due to the increased difficulty in distinguishing between the arms. Therefore, prioritizing the selection of pairs with larger probability gaps is advantageous. In particular, the Condorcet winner, which by definition has the largest probability gap with all other arms, plays a crucial role in optimizing the sample complexity. By strategically focusing on pairs involving the Condorcet winner, the algorithm can achieve more efficient learning and faster convergence.

Table 1: An exemplary instance

$p(i,j)$	$j=1$	$j=2$	$j=3$
$i=1$	\	0.8	0.9
$i=2$	0.2	\	0.6
$i=3$	0.1	0.4	\

To better illustrate this, consider the example in Table 1, which presents three arms with respective winning probabilities $p_{i,j}$. According to the definition, arm 1 is identified as the Condorcet winner. To accurately identify the Condorcet winner from the other arms, two pairs need to be selected for the denoising phase out of the three possible combinations: $(1,2)$, $(1,3)$, and $(2,3)$.

Suppose the first selected pair is either $(1,2)$ or $(1,3)$. In this case, the large probability gaps between the arms lead to fewer comparisons required for the denoising process. Consequently, arm 2 or arm 3 is quickly eliminated by arm 1, thereby incurring low regret. Conversely, if the first pair chosen for comparison is $(2,3)$, the smaller probability gap necessitates more samples for denoising, resulting in higher regret throughout the denoising phase.

To formalize the above findings, we introduce Algorithm 2. Similar to Algorithm 1, the algorithm differs in that the arm-selection sequence is determined by another weak regret minimization algorithm, denoted as $\mathcal{A}$. Hence, Algorithm 2 functions as a reduction framework, that takes an RM algorithm $\mathcal{A}$ as input, and operates as a BAI algorithm. The following theorem states the nearly-consistency of RM and BAI of Algorithm 2, that the sample complexity upper bound can be determined by the cumulative regret,

Theorem 3.3. *Under the reduction framework of Algorithm 2, for any arm-selection sequence determined by the subroutine $\mathcal{A}$, the following properties are guaranteed simultaneously:*

- *With probability at least $1-\frac{\delta}{2}$, the Algorithm identifies the correct Condorcet winner.*
- *With a probability of at least $1-\frac{\delta}{2}$, the required sample complexity is upper bounded by $\left(R_w + O\left(\frac{N}{\Delta^{*2}}\log\left(\frac{N}{\Delta^{*}\delta}\right)\right)\right)$, where R_w represents the cumulative weak regret over the entire time horizon, and Δ^* denotes the minimum gap between the Condorcet winner and any other arm.*

Remark 2. *The first statement ensures a high probability of accurately identifying the Condorcet winner. The second statement demonstrates, independently of the strategy used, that the sample complexity for Best Arm Identification (BAI) is intrinsically linked to the cumulative weak regret. This indicates that effectively optimizing the Regret Minimization (RM) objective can also provide a bound on the sample complexity, ensuring the compatibility of these objectives.*

Remark 3. *Although the sample complexity result includes the additional regret term R_w, which may seem less favorable compared to Theorem 3.2, it is essential to consider the dependency shift from the global minimum gap Δ to the Condorcet minimum gap Δ^*. By definition, it holds that $\Delta \le \Delta^*$ for all instances. Consequently, selecting an appropriate RM algorithm not only ensures low regret but also achieves a potentially lower problem-dependent sample complexity. This shift underscores the importance of the choice of RM algorithm in minimizing both regret and sample complexity.*

In the next part, we take one of the state-of-the-art RM algorithm as an example and fit it into our framework to show how the proposed reduction framework works to achieve dual optimalities in RM and BAI simultaneously.

3.3 An implementation achieving dual optimalities

In this section, we take one of the state-of-the-art RM algorithm as the base algorithm in our reduction framework and illustrate that our proposed framwork achieves nearly-optimal regret and sample complexity simultaneously.

Base RM Algorithm. WS-W, as introduced by Chen and Frazier [5], is one of the state-of-the-art algorithm for minimizing weak regret. The core mechanism of WS-W includes assigning an initial score of zero to each arm. Scores are adjusted based on outcomes: a win increases the score by one, while a loss decreases it by one. The algorithm begins with two randomly selected arms and progresses through a series of epochs, each containing multiple iterations. In each iteration, the winner from the previous iteration (denoted as arm i) competes against a randomly selected arm that has not yet dueled in the current epoch. Each arm, except for the previous winner, duels once per epoch. An iteration within epoch ℓ continues until an arm's score reaches $-\ell$. Once all arms, except for the previous winner, have dueled, the epoch concludes, and the algorithm transitions to the next epoch. The following statement ensures the cumulative weak regret bound for WS-W:

Proposition 3.4 (Theorem 1 of Chen and Frazier [5]). *Under the total order assumption, WS-W's expected cumulative weak regret is upper bounded by $O\left(\frac{N}{\Delta^6}\log(N)\right)$, where N is the number of arms and Δ represents the minimum gap.*

Fitting the Reduction. We incorporate WS-W as the RM subroutine $\mathcal{A}$ within our reduction framework, resulting in the BAI algorithm outlined in Algorithm 3. Unlike the original WS-W, this variant maintains a set of qualified arms Q, restricting arm selection to this set. At each iteration r, two arms (i_r, j_r) are selected from Q according to the WS-W procedure. These arms engage in repeated duels until either one arm's score reaches $S_a = -\ell$ or the confidence interval for the arm pair excludes $1/2$. The loser of the iteration, l_r, is identified as the arm reaching score $-\ell$ or the arm eliminated based on confidence; the other arm is declared the winner, w_r. After each iteration, the losing arm is disqualified. If all remaining arms except the winner are disqualified but not permanently eliminated, they are requalified for competition in the subsequent epoch. Each disqualified arm competes once per epoch, ensuring chances of recovery. This iterative process of comparisons, disqualifications, requalifications, and eliminations continues until a single arm remains, designated as the Condorcet winner.

The following Theorem 3.5 ensures that the proposed reduction from WS-W to WSW-PE preserves the regret guarantees of Proposition 3.4 while achieving near-optimal sample complexity, making WSW-PE both regret-optimal and sample-efficient.

Theorem 3.5. *Let N denote the number of arms, $\delta \in (0,1)$ the confidence level, and R_w the cumulative weak regret over the total time horizon. Algorithm 3, as derived from our reduction, satisfies the following guarantees:*

Algorithm 3 WSW-PE

Input: Set of optional arms $\mathcal{X}$, confidence level $1-\delta$
Initialize: Epoch index $\ell \leftarrow 1$, iteration index $r \leftarrow 1$, qualified arms $Q \leftarrow \mathcal{X}$, pairwise winning counts $\mathbf{W} \leftarrow \mathbf{0}^{N\times N}$, scores $S \leftarrow \mathbf{0}^N$
while $|\mathcal{X}| > 1$ **do**
  if $r > 1$ **then**
    $i_r \leftarrow w_{r-1}$ ▷ Winner of previous iteration
  else
    Randomly initialize i_1
  end if
  Randomly select j_r from $Q \setminus \{i_r\}$
  In iteration r: Repeatedly compare (i_r, j_r), update $\mathbf{W}_{a,b} \leftarrow \mathbf{W}_{a,b} + 1$, $S_a \leftarrow S_a + 1$, $S_b \leftarrow S_b - 1$ **if** arm $a \in \{i_r, j_r\}$ wins, until one of the following conditions is met:
  - Arm a has score $S_a = -\ell$
  - $1/2$ does not lie within the current confidence region as determined by Equation (2)

  Denote w_r and l_r as the winner and loser of the comparison, respectively.
  $Q \leftarrow Q \setminus \{l_r\}$, $r \leftarrow r+1$ ▷ Move to next iteration
  $\mathcal{X} \leftarrow \mathcal{X} \setminus \{l_r\}$ **if** $1/2$ does not lie within the current confidence region ▷ Elimination step
  if $|Q| \le 1$ **then**
    $Q \leftarrow \mathcal{X}$, $\ell \leftarrow \ell + 1$ ▷ Start a new epoch
  end if
end while
Output: $\mathcal{X}$

- ***Sample Efficiency (High-Probability Guarantee):*** *With probability at least $1-\delta$, the algorithm terminates and correctly identifies the Condorcet winner with a sample complexity bounded by*

$$O\left(R_w + \frac{N}{\Delta^{*2}}\log\left(\frac{N}{\Delta^*\delta}\right)\right).$$

- ***Regret Optimality (Expected Guarantee):*** *The expected cumulative weak regret incurred by Algorithm 3 is bounded by*

$$\mathbb{E}[R_w] = O\left(\frac{N}{\Delta^6}\log(N)\right).$$

Remark 4 (Dual optimality). *For some mild instances where the minimum gap Δ is not extremely small, the expected cumulative weak regret is bounded by $O(N\log(N))$, matching the best-known weak regret results [5, 24], and meanwhile indicating an $O(N\log(N))$ sample complexity result that is optimal up to logarithmic factors.*

For worst cases when the minimum gap Δ is small, it may seem the sample complexity result could be dominated by cumulative regret since the Δ^{-6} dependency, we emphasize that these two results are not comparable, as the sample complexity upper bound holds with high probability, while the regret result is given in expectation, the difference hides in the important *variance of the cumulative regret*. Interestingly, the specific proof indicates that the the weak regret guarantee is *independent* of the selection rule of the opponent arm j_r, giving us the chance to apply arbitrary heuristic methods to control variances and meanwhile shares the same weak regret guarantee. We take the following case as an example,

Table 2: Statistics of synthetic and real-world data. Smaller Δ and larger N imply harder instances.

Synthetic Benchmark		Real-World Applications		
# of arms	Δ	Name	# of arms	Δ
5	0.20	arXiv ranking	6	0.040
10	0.10	car preference	10	0.009
20	0.08	sushiA	10	0.205
50	0.05	sushiB50	50	0.027
100	0.03	sushiB	100	0.014
200	0.02	-	-	-

Example 3.6. For arbitrary arm $a \in \{Q \setminus i_r\}$, an empirically efficient strategy is to select $j_r \leftarrow a$ with probability $\frac{\exp(-\eta\ell_a)}{\sum_{b\in\{Q\setminus i_r\}}\exp(-\eta\ell_b)}$, where $\ell_a := \|\mathbf{W}_{:,a}\|_1 - \|\mathbf{W}_{a,:}\|_1$ is the negative winning difference of arm a against all opponents and η is the step size as an input.

Replacing the selection rule of j_r in Algorithm 3 by the example above, we obtain the WSW-PE-EXP algorithm. Due to technical difficulties, the variance guarantee for the proposed WSW-PE-EXP algorithm remains unknown. Instead, in the next section, we show the efficacy of the proposed approaches by empirical evidences.

4 EXPERIMENTS

In this section, we conduct empirical evaluations to evaluate how the proposed reduction performs in *regret* and *sample complexity*.

Global settings. We conduct experiments on both synthetic and real-world applications of the online recommendation tasks. Algorithms generated by our reduction are compared with other state-of-the-art algorithms, whose objectives are either weak regret minimization or BAI. We take the cumulative weak regret over the total time horizon and the total number of comparisons needed to identify the Condorcet winner with given fixed confidence, as the measure for RM or BAI, respectively. All results are averaged on 100 independent trials.

Our Approach. Recall that our proposed reduction framework utilizes weak regret minimization algorithms as sub-routines to derive BAI algorithms. In this part, we employ WS-W [5] and BTW [24] as the foundational sub-routines. The resulting algorithms from our reduction process are designated as WSW-PE and BTW-PE, respectively. We also extend our evaluation on the modified WSW-PE-EXP algorithm, as described in Example 3.6.

Contenders. We compare WSW-PE and BTW-PE with some state-of-the-art (ϵ,δ)-PAC dueling bandits algorithms by setting $\epsilon = \Delta^*$, as the (ϵ,δ)-PAC algorithms become δ-PAC algorithms when $\epsilon \le \Delta^*$. The state-of-the art (ϵ,δ)-PAC algorithms include Knockout [10] and SeqElim [9], both of which are theoretically optimal; MallowsMPI [4] that is almost-optimal under the Mallows assumption [4]. Notice that the above algorithms are not designed for regret minimization, for fair comparisons in turns of regret, we fit them into the famous explore-then-commit (ETC) framework [13], that first run BAI to identify the Condorcet winner, and then repeatedly pull the found arm in the objective of RM. We denote by

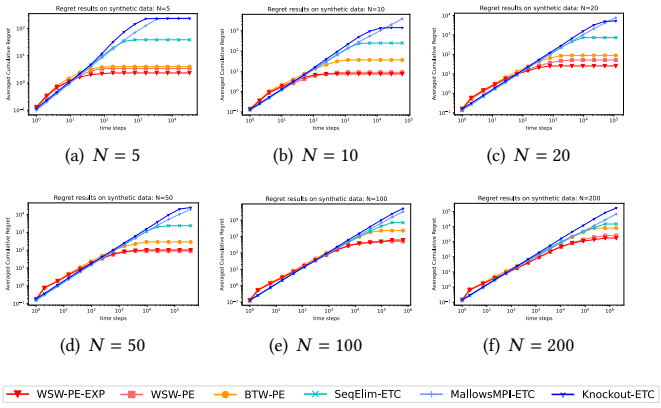

Figure 2: Cumulative weak regret results averaged on 100 independent trials on synthetic benchmarks. The lower the regret, the better. We conclude that our reductions incur relatively smaller regret than the ETC-type of algorithms.

Table 3: Computational efficiency comparisons, the results are the mean computation time in milliseconds (ms) with standard deviation, averaged over 100 trials.

Algorithm	$N = 10$	$N = 20$	$N = 50$	$N = 100$	$N = 200$
PE (Ours) Avg. Time (ms)	20.46 ± 7.53	**50.76 ± 24.51**	**185.26 ± 67.62**	**1247.43 ± 241.08**	**4445.83 ± 522.96**
ETC Avg. Time (ms)	**18.68 ± 22.04**	92.62 ± 71.82	377.56 ± 164.72	2124.50 ± 1528.47	9446.51 ± 7103.22

"algorithm-ETC" as the corresponding fits, for example, Knockout-ETC stands for fitting Knockout into the ETC framwork. We also compare with SeqElim-ETC and MallowsMPI-ETC.

4.1 Synthetic Benchmark

We generate the winning probability by randomly sampled from a uniform distribution for a different number of arms. Specifically, we set the number of arms as $N = 5, 10, 20, 50, 100, 200$. We then construct dueling bandits problems by generating the winning probability $p_{i,j}$ for $i, j \in \mathcal{X}$ according to the pre-specified minimum gap Δ. Table 2 lists some statistics of the experimental instances. Results were averaged over 100 independent trials using a 12th Gen Intel(R) Core(TM) i5-12600KF 3.70 GHz CPU.

Regret results. In Figure 2 we present the cumulative weak regret result averaged over 100 independent trials. The lower the regret, the better. Notice that the regret result only depends on specific sub-routine $\mathcal{A}$, we combine the cumulative regret results of WS-W and BTW with WSW-PE and BTW-PE, respectively. The PE-type of algorithms outperforms the ETC-type of algorithms empirically in most cases.

Sample complexity results. In Figure 3, we present the sample complexity needed to identify the Condorcet winner with probability at least $1 - \delta$. As we can see, the PE-type of algorithms obtain comparable or even better performances to the state-of-the-art contenders in most cases.

Computational complexity results. In Table3, we present the averaged computational efficiency of the PE-type of algorithms and ETC-type of algorithms, the computation time results are displayed in milliseconds (ms) with associated standard deviation. These findings highlight the scalability of the PE algorithm, which consistently outperforms ETC in terms of computational efficiency as the number of arms increases.

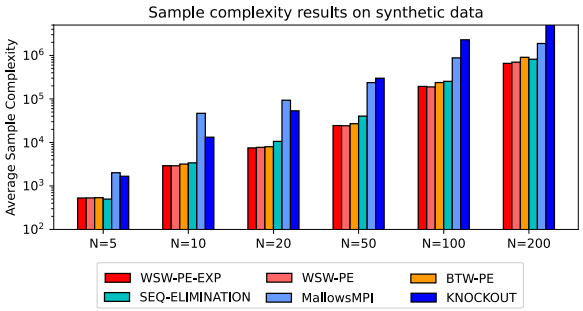

Figure 3: The sample complexity results averaged on 100 trials on synthetic benchmarks. The lower the sample complexity, the better. We can conclude that algorithms generated by our reduction use relatively fewer samples to identify the Condorcet winner with the same fixed confidence, especially when the number of optional arms are large.

4.2 Online Recommendation

The online recommendation task aims sequentially selecting items to present to users, obtaining corresponding preference feedback and learning user preferences to refine future recommendations [31, 32]. This task can be effectively modeled as a dueling bandits problem, where the system presents pairs of items to users, according to which the users return an implicit feedback indicating which item is preferred. RM focuses on consistently recommending items that align with user preferences throughout the learning process, and BAI aims to efficiently identify the user's most preferred item with minimal feedback. Both objectives are important in this scenario.

In this part, we conduct real-world online recommendation tasks on the following three datasets:

- **arXiv ranking** [25]: This dataset consists of six ranking functions for the arXiv.org e-print archive, using implicit retrieval information from user clicks. The goal is to identify the best ranking function based on user interactions. Regret measures the loss in clicks compared to the optimal ranking function, and sample complexity reflects the efficiency of learning the best function. Number of arms: $N = 6$, global minimum gap: $\Delta = 0.040$.
- **Car preference** [8]: This dataset captures user preferences across 10 car types, aiming to identify the most preferred car type based on feedback. Regret quantifies the potential loss in car sales, and sample complexity indicates how quickly sellers can determine the preferred type. Number of arms: $N = 10$, global minimum gap: $\Delta = 0.009$.
- **Sushi** [16]: This dataset provides detailed user preferences for various sushi types, divided into two subsets:
  - **sushiA**: Includes subjective rankings for 10 traditional Japanese sushi types, offering a controlled environment for evaluating regret minimization (RM) and best arm identification (BAI). Number of arms: $N = 10$, global minimum gap: $\Delta = 0.205$.
  - **sushiB**: Represents a more complex scenario with 100 sushi types, analyzed in two configurations:
    - **sushiB50**: A random subset of 50 sushi types, providing moderate complexity. Number of arms: $N = 50$, global minimum gap: $\Delta = 0.027$.

Table 4: Multi-objective scores on real-world applications. The results are presented in mean ± std (rank) format, the smaller the better. The average performance rank over all objectives and all experiments is presented at the last row of the table.

Datasets	Objective	*Our Reduction*			*ETC-Type of Algorithms*		
		WSW-PE-EXP	WSW-PE	BTW-PE	SeqElim-ETC	MallowsMPI-ETC	Knockout-ETC
arXiv ranking	$\alpha = 0$ (BAI)	0.412 ± 0.074 (2)	0.483 ± 0.090 (4)	0.461 ± 0.080 (3)	**0.292 ± 0.105 (1)**	1.000 ± 1.424 (6)	0.720 ± 0.370 (5)
	$\alpha = 1$ (RM)	**0.007 ± 0.010 (1)**	0.113 ± 0.038 (3)	0.071 ± 0.032 (2)	0.145 ± 0.171 (4)	0.879 ± 1.845 (5)	1.000 ± 0.885 (6)
	$\alpha = 0.1$	0.410 ± 0.074 (2)	0.482 ± 0.090 (4)	0.460 ± 0.080 (3)	**0.292 ± 0.105 (1)**	1.000 ± 1.426 (6)	0.722 ± 0.371 (5)
	$\alpha = 0.3$	0.407 ± 0.073 (2)	0.479 ± 0.089 (4)	0.456 ± 0.079 (3)	**0.291 ± 0.106 (1)**	1.000 ± 1.431 (6)	0.725 ± 0.374 (5)
	$\alpha = 0.5$	0.401 ± 0.072 (2)	0.474 ± 0.088 (4)	0.451 ± 0.079 (3)	**0.289 ± 0.107 (1)**	1.000 ± 1.439 (6)	0.731 ± 0.380 (5)
	$\alpha = 0.7$	0.388 ± 0.070 (2)	0.463 ± 0.086 (4)	0.439 ± 0.077 (3)	**0.285 ± 0.109 (1)**	1.000 ± 1.459 (6)	0.744 ± 0.394 (5)
	$\alpha = 0.9$	0.335 ± 0.060 (2)	0.416 ± 0.079 (4)	0.388 ± 0.069 (3)	**0.268 ± 0.120 (1)**	1.000 ± 1.541 (6)	0.800 ± 0.462 (5)
Car preference	$\alpha = 0$ (BAI)	0.494 ± 0.139 (3)	**0.479 ± 0.092 (1)**	0.484 ± 0.106 (2)	0.593 ± 0.265 (4)	1.000 ± 0.161 (6)	0.693 ± 0.294 (5)
	$\alpha = 1$ (RM)	**0.005 ± 0.004 (1)**	0.101 ± 0.023 (4)	0.091 ± 0.022 (3)	0.312 ± 1.836 (5)	1.000 ± 2.982 (6)	0.063 ± 0.061 (2)
	$\alpha = 0.1$	0.494 ± 0.139 (3)	**0.479 ± 0.092 (1)**	0.484 ± 0.106 (2)	0.593 ± 0.265 (4)	1.000 ± 0.161 (6)	0.693 ± 0.294 (5)
	$\alpha = 0.3$	0.494 ± 0.140 (3)	**0.479 ± 0.092 (1)**	0.484 ± 0.106 (2)	0.596 ± 0.266 (4)	1.000 ± 0.161 (6)	0.693 ± 0.294 (5)
	$\alpha = 0.5$	0.495 ± 0.142 (3)	**0.479 ± 0.092 (1)**	0.484 ± 0.106 (2)	0.599 ± 0.268 (4)	1.000 ± 0.161 (6)	0.693 ± 0.294 (5)
	$\alpha = 0.7$	0.498 ± 0.147 (3)	**0.480 ± 0.092 (1)**	0.485 ± 0.106 (2)	0.606 ± 0.274 (4)	1.000 ± 0.161 (6)	0.694 ± 0.294 (5)
	$\alpha = 0.9$	0.509 ± 0.187 (3)	**0.484 ± 0.093 (1)**	0.488 ± 0.106 (2)	0.643 ± 0.326 (4)	1.000 ± 0.161 (6)	0.696 ± 0.293 (5)
sushiA	$\alpha = 0$ (BAI)	**0.015 ± 0.002 (1)**	0.015 ± 0.003 (3)	0.015 ± 0.002 (4)	0.015 ± 0.006 (2)	1.000 ± 1.540 (6)	0.073 ± 0.025 (5)
	$\alpha = 1$ (RM)	**0.001 ± 0.001 (1)**	0.001 ± 0.002 (3)	0.001 ± 0.001 (2)	0.011 ± 0.010 (4)	1.000 ± 1.825 (6)	0.069 ± 0.034 (5)
	$\alpha = 0.1$	**0.015 ± 0.002 (1)**	0.015 ± 0.003 (3)	**0.015 ± 0.002 (1)**	0.015 ± 0.006 (4)	1.000 ± 1.545 (6)	0.073 ± 0.026 (5)
	$\alpha = 0.3$	**0.015 ± 0.002 (1)**	0.015 ± 0.003 (3)	**0.015 ± 0.002 (1)**	0.015 ± 0.006 (4)	1.000 ± 1.557 (6)	0.073 ± 0.026 (5)
	$\alpha = 0.5$	**0.013 ± 0.002 (1)**	0.014 ± 0.003 (3)	0.014 ± 0.002 (2)	0.014 ± 0.006 (4)	1.000 ± 1.577 (6)	0.073 ± 0.027 (5)
	$\alpha = 0.7$	**0.011 ± 0.002 (1)**	0.012 ± 0.002 (2)	0.012 ± 0.002 (2)	0.013 ± 0.007 (4)	1.000 ± 1.613 (6)	0.072 ± 0.028 (5)
	$\alpha = 0.9$	**0.007 ± 0.002 (1)**	**0.007 ± 0.002 (1)**	**0.007 ± 0.002 (1)**	0.012 ± 0.008 (4)	1.000 ± 1.703 (6)	0.071 ± 0.030 (5)
sushiB50	$\alpha = 0$ (BAI)	0.137 ± 0.051 (2)	**0.134 ± 0.045 (1)**	0.149 ± 0.024 (3)	0.167 ± 0.070 (4)	1.000 ± 1.350 (6)	0.975 ± 0.465 (5)
	$\alpha = 1$ (RM)	**0.002 ± 0.007 (1)**	0.002 ± 0.010 (2)	0.007 ± 0.003 (3)	0.078 ± 0.090 (4)	0.500 ± 1.241 (5)	1.000 ± 0.584 (6)
	$\alpha = 0.1$	0.136 ± 0.051 (2)	**0.133 ± 0.044 (1)**	0.148 ± 0.024 (3)	0.167 ± 0.070 (4)	1.000 ± 1.358 (6)	0.983 ± 0.470 (5)
	$\alpha = 0.3$	0.133 ± 0.050 (2)	**0.130 ± 0.043 (1)**	0.145 ± 0.023 (3)	0.166 ± 0.072 (4)	0.997 ± 1.376 (5)	1.000 ± 0.483 (6)
	$\alpha = 0.5$	0.124 ± 0.047 (2)	**0.121 ± 0.041 (1)**	0.135 ± 0.022 (3)	0.161 ± 0.074 (4)	0.964 ± 1.367 (5)	1.000 ± 0.489 (6)
	$\alpha = 0.7$	0.108 ± 0.041 (2)	**0.105 ± 0.036 (1)**	0.118 ± 0.019 (3)	0.149 ± 0.076 (4)	0.901 ± 1.350 (5)	1.000 ± 0.502 (6)
	$\alpha = 0.9$	**0.064 ± 0.019 (1)**	0.064 ± 0.023 (2)	0.073 ± 0.013 (3)	0.120 ± 0.081 (4)	0.740 ± 1.306 (5)	1.000 ± 0.535 (6)
sushiB	$\alpha = 0$ (BAI)	**0.178 ± 0.025 (1)**	0.180 ± 0.024 (2)	0.195 ± 0.028 (3)	0.247 ± 0.074 (4)	0.814 ± 0.784 (5)	1.000 ± 0.433 (6)
	$\alpha = 1$ (RM)	**0.001 ± 0.002 (1)**	**0.001 ± 0.002 (1)**	0.009 ± 0.001 (3)	0.041 ± 0.057 (4)	0.426 ± 1.650 (5)	1.000 ± 0.640 (6)
	$\alpha = 0.1$	**0.175 ± 0.025 (1)**	0.178 ± 0.023 (2)	0.193 ± 0.027 (3)	0.244 ± 0.074 (4)	0.809 ± 0.795 (5)	1.000 ± 0.436 (6)
	$\alpha = 0.3$	**0.169 ± 0.024 (1)**	0.172 ± 0.023 (2)	0.187 ± 0.026 (3)	0.235 ± 0.073 (4)	0.796 ± 0.826 (5)	1.000 ± 0.443 (6)
	$\alpha = 0.5$	**0.159 ± 0.023 (1)**	0.161 ± 0.021 (2)	0.178 ± 0.025 (3)	0.222 ± 0.072 (4)	0.773 ± 0.876 (5)	1.000 ± 0.455 (6)
	$\alpha = 0.7$	**0.139 ± 0.020 (1)**	0.141 ± 0.019 (2)	0.161 ± 0.022 (3)	0.195 ± 0.070 (4)	0.730 ± 0.972 (5)	1.000 ± 0.478 (6)
	$\alpha = 0.9$	**0.087 ± 0.013 (1)**	0.088 ± 0.012 (2)	0.115 ± 0.014 (3)	0.124 ± 0.065 (4)	0.614 ± 1.231 (5)	1.000 ± 0.539 (6)
Avg. Rank		**1.65**	2.20	2.57	3.46	5.63	5.29

- **Full sushiB**: The complete set of 100 sushi types, representing the most challenging scenario. Number of arms: $N = 100$, global minimum gap: $\Delta = 0.014$.

Table 2 summarizes the statistics of these instances.

Performance Metrics. Previous works have primarily focused on either RM or BAI, where the respective performance metrics are cumulative weak regret and sample complexity. In real-world applications, however, it is often desirable to optimize both objectives simultaneously. Therefore, we adopt a multi-objective perspective to evaluate overall performance in this section. Specifically, we introduce the following multi-objective score, $\mathcal{S}(\alpha)$, which balances RM and BAI performance:

$$\mathcal{S}(\alpha) \triangleq \alpha\mathcal{R} + (1-\alpha)C.$$

In this context, α serves as a weight parameter that governs the relative importance of the two objectives, while $\mathcal{R}$ and C denote the normalized cumulative weak regret and sample complexity, respectively. The normalization process is conducted by dividing each result by the worst observed value, ensuring that all normalized scores fall within the interval $[0, 1]$, thereby facilitating their interpretation. To account for a diverse range of practical scenarios, we evaluate the performance for values of α spanning from 0 to 1 in steps of 0.1. Specifically, when $\alpha = 0$, the learning objective is focused on regret minimization (RM), and when $\alpha = 1$, the focus shifts to best arm identification (BAI). A consistently low multi-objective score, independent of α, signifies strong performance across both RM and BAI objectives.

Experimental Results. The multi-objective performance results are summarized in Table 4, presented as mean ± standard deviation (rank). A lower score indicates superior performance, with the best result for each objective within each dataset highlighted in bold. Our reduction-based algorithms outperform existing methods across most datasets and objectives. The average performance

rank across all objectives and experiments is displayed at the bottom of the table. The average rank of our reduction-generated algorithms is lower than that of the state-of-the-art methods, demonstrating the effectiveness of our approach for both objectives. Furthermore, the proposed WSW-PE-EXP algorithm (as detailed in Example 3.6) achieves the best overall performance, highlighting the importance of variance control.

5 RELATED WORK

In this section, we introduce some advances in dueling bandit. As most current works focus on either RM or BAI, we introduce the works separately.

Dueling bandits in the objective of RM. Mainly two kinds of regrets are widely studied in dueling bandits, i.e., strong regret and weak regret. Most works assume specific structural properties and focus on minimizing strong regret. Yue et al. [35] presents an $\Omega(N \log T)$ lower bound of worst-case expected strong regret under the assumption that the Condorcet Winner exists, where N is the number of arms and T is the time horizon. Later, various algorithms are proposed with an $O(N \log T)$ expected strong regret under different additional assumptions [19, 34, 35, 38, 39]. Relative Minimum Empirical Divergence (RMED) proposed by Komiyama et al. [17] is the first algorithm that is optimal without additional assumption. Other exciting works reduce dueling bandits problems to online convex optimization problems [1, 18, 30, 33]. As for weak regret minimization, there is not much attention as any strong regret minimization algorithm can be used for weak regret by definition. It is until recently, when Chen and Frazier [5] states that the $O(N \log T)$ strong regret bound is too loose for weak regret, that weak regret minimization draws attentions. Chen and Frazier [5] proposes WinnerStays-Weak (WS-W) algorithm that enjoys an $O(N^2)$ expected weak regret under the assumption that the Condorcet Winner exists, and the bound can be further improved to $O(N \log N)$ when the arms follow a total order. Subsequently, Peköz et al. [24] proposes Beat The Winner (BTW), which attains the same $O(N^2)$ expected weak regret result with much simpler analysis. Recently, the elegant work of Saad et al. [29] provides a $\Omega(N/\Delta^*)$ lower bound for weak regret minimization in dueling bandits and proposes the WR-TINF algorithm that is provably optimal when the optimality gap is sufficiently large.

Dueling bandits in the objective of BAI. Most BAI dueling bandits algorithms aim to identify the best arm within minimum sample complexity with *fixed confidence*. (ϵ, δ)-PAC is the most general BAI setting that search for an approximately correct arm in a finite number of comparisons with probability at least $1-\delta$. Yue and Joachims [34] propose BTM-PAC with an $O((N/\epsilon^2)\log(N/\epsilon\delta))$ sample complexity bound when the number of arms N is large. Falahatgar et al. [10] improve the bound to $O((N/\epsilon^2)\log(1/\delta))$ that matches the lower bound without requirement on N, and the same bound is achieved with a similar goal [9, 11]. Moreover, some works aim to identify exactly the best arm rather than an approximate one, namely, $\epsilon \leq \Delta^*$ for (ϵ, δ)-PAC algorithms, where Δ^* is the Condorcet minimum gap and will be defined later in the problem statement. Feige et al. [12] first show the sample complexity lower bound of any δ-correct algorithm to identify the exact Condorcet winner is $\Omega((N/\Delta^2)\log(1/\delta))$ when the minimum gap Δ is given in advance. Falahatgar et al. [9] further prove the same lower bound with an unknown Δ. Busa-Fekete et al. [4] propose the MallowsMPI algorithm for the exact BAI problem with an $O\left(N/\Delta^2 \log(N/\delta\Delta)\right)$ sample complexity, under the assumption of Mallows model [20]. Mohajer et al. [21] improve the bound by getting rid of the additional $O(\log(1/\Delta))$ term without assuming Mallows model. Similar results are achieved recently [27, 28].

Simultaneous objective of RM and BAI. Though most existing works of bandit learning focus solely on the objective of either RM or BAI. The simultaneous objective of RM and BAI in bandit learning has drawn increasing attention recently. Garivier et al. [13] considers the varying objective from BAI to RM by the explore-then-commit (ETC) strategy. It reduces the RM problem to the BAI problem by first running a BAI algorithm to identify the best arm with probability at least $1-\delta$, and then repeatedly pulling the found best arm to achieve low regret. The expected regret incurred by this strategy is bounded by $S_{\mathcal{A}} + \delta T$, where $S_{\mathcal{A}}$ is the sample complexity needed for the BAI algorithm $\mathcal{A}$, while extra δT term is the regret incurred if the BAI algorithm finds a sub-optimal arm with probability at most δ. To ensure sub-linear regret, it is necessary to constrain $\delta \leq 1/T$. In this case, according to the sample complexity lower bound mentioned in Remark 1, $S_{\mathcal{A}} \geq N \log T$ for any BAI algorithm $\mathcal{A}$. implying a sub-optimal regret bound of $O(N \log T)$ by ETC with dependency on the total time horizon T. In contrast, our result of $O(N \log N)$ is constant-in-horizon T and therefore offers a better performance guarantee. Degenne et al. [7] proposed a mixed objective approach that interpolates between Regret Minimization (RM) and Best Arm Identification (BAI). Both this and other similar works focus on optimizing a single mixed objective, formulated by combining RM and BAI through weighted parameters. However, this approach inherently involves trade-offs, as the weights dictate the balance between the two objectives, potentially leading to compromises in either RM or BAI performance. In contrast, our research seeks to explore the potential for optimizing both RM and BAI simultaneously without such compromises.

6 CONCLUSION

In this paper, we investigates the dueling bandits problem, with a particular focus on the compatibility of two key objectives: regret minimization (RM) and best arm identification (BAI). These objectives are generally regarded as conflicting in most bandit scenarios. We demonstrate that RM and BAI can be nearly compatible in dueling bandits by reducing the BAI task to a noisy identification process. Building on this insight, we propose a reduction framework that transforms any RM algorithm into a BAI algorithm. Using this framework, we develop an algorithm that achieves dual optimality in RM and BAI: it identifies the Condorcet winner with near-optimal sample complexity and ensures regret optimality by maintaining a cumulative weak regret that is provably constant with respect to the time horizon, matching the best-known results. Finally, we validate the efficacy of our approach through synthetic benchmarks and real-world applications in online recommendation systems, demonstrating its practical relevance. We hope this work will inspire further research on addressing evolving and complex objectives in open and dynamic environments, particularly in domains where balancing competing goals, such as exploration-exploitation trade-offs and adaptive decision-making, is crucial.

ACKNOWLEDGEMENT

We acknowledge the funding provided by the National Natural Science Foundation of China (62206245) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 856–864, 2014.

[2] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 41–53, 2010.

[3] Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22:7:1–7:108, 2021.

[4] Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of mallows. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 1071–1079, 2014.

[5] Bangrui Chen and Peter I Frazier. Dueling bandits with weak regret. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 731–739, 2017.

[6] Shi-Yong Chen, Yang Yu, Qing Da, Jun Tan, Hai-Kuan Huang, and Hai-Hong Tang. Stabilizing reinforcement learning in dynamic environment with application to online recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1187–1196, 2018.

[7] Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to A/B tests. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1988–1996, 2019.

[8] E. V. Bonilla E. Abbasnejad, S. Sanner and P. Poupart. Learning community-based preferences via dirichlet process mixtures of gaussian processes. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2013.

[9] Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 7060–7070, 2017.

[10] Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1088–1096, 2017.

[11] Moein Falahatgar, Ayush Jain, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. The limits of maxing, ranking, and preference learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1427–1436, 2018.

[12] Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

[13] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 784–792, 2016.

[14] Björn Haddenhorst, Viktor Bengs, and Eyke Hüllermeier. Identification of the generalized condorcet winner in multi-dueling bandits. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 25904–25916, 2021.

[15] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 4415–4426, 2020.

[16] Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 583–588, 2003.

[17] Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 1141–1154, 2015.

[18] Wataru Kumagai. Regret analysis for continuous dueling bandit. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 1489–1498, 2017.

[19] Chang Li, Ilya Markov, Maarten de Rijke, and Masrour Zoghi. Mergedts: A method for effective large-scale online ranker evaluation. *ACM Transactions on Information and Systems*, 38(4):40:1–40:28, 2020.

[20] Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

[21] Soheil Mohajer, Changho Suh, and Adel M. Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2488–2497, 2017.

[22] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

[24] Erol Peköz, Sheldon M Ross, and Zhengyu Zhang. Dueling bandit problems. *Probability in the Engineering and Informational Sciences*, pages 1–12, 2020.

[25] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 43–52, 2008.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2024.

[27] Wenbo Ren, Jia Liu, and Ness B. Shroff. On sample complexity upper and lower bounds for exact ranking from noisy comparisons. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 10014–10024, 2019.

[28] Wenbo Ren, Jia Liu, and Ness B. Shroff. The sample complexity of best-k items selection from pairwise comparisons. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 8051–8072, 2020.

[29] El Mehdi Saad, Alexandra Carpentier, Tomáš Kocák, and Nicolas Verzelen. On weak regret analysis for dueling bandits. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, page to appear, 2024.

[30] Aadirupa Saha, Tomer Koren, and Yishay Mansour. Dueling convex optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 9245–9254, 2021.

[31] Xiaoli Tang, Tengyun Wang, Haizhi Yang, and Hengjie Song. AKUPM: attention-enhanced knowledge-aware user preference model for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1891–1899, 2019.

[32] Zimu Wang, Yue He, Jiashuo Liu, Wenchao Zou, Philip S. Yu, and Peng Cui. Invariant preference learning for general debiasing in recommendation. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1969–1978, 2022.

[33] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1201–1208, 2009.

[34] Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 241–248, 2011.

[35] 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.

[36] Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2025–2034, 2016.

[37] Zhou Zhao, Hanqing Lu, Deng Cai, Xiaofei He, and Yueting Zhuang. User preference learning for online social recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2522–2534, 2016.

[38] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31th International conference on machine learning (ICML)*, pages 10–18, 2014.

[39] Masrour Zoghi, Shimon Whiteson, and Maarten de Rijke. Mergerucb: A method for large-scale online ranker evaluation. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 17–26, 2015.

Supplementary Materials

In the appendix, we provide theoretical analysis in Appendix A.

A ANALYZE

In this section, we provide detailed proofs of the theoretical results. Specifically, we first give the proof of Lemma 3.1 and Theorem 3.2, clearing the problem setups. Then we give the core proof of Theorem 3.3, that strategy-independently states the possibility of optimizing the simultaneous objective of RM and BAI in dueling bandits and provide its corresponding sample complexity result that is related to the regret; We finally prove Theorem 3.5 that provide the guarantees of the exemplary case.

A.1 Proof of Lemma 3.1 and Theorem 3.2

Proof. Recall the elimination criterion that any arm i is eliminated by arm j if $\hat{\mu}_{i,j} + c_\delta(n_{i,j}) \leq \frac{1}{2}$, where $\hat{\mu}_{i,j}$ is the empirical winning rate of arm i against arm j and $c_\delta(n_{i,j})$ is the confidence radius defined in (2) after comparing (i, j) for $n_{i,j}$ times. We bound the probability of the event that the denoise criteria is not met after comparing n_δ times as follows,

$$\begin{aligned}
&\mathbf{P}\left[n_{i,j} \geq n_\delta, \left|\frac{1}{2} - \hat{\mu}_{i,j}\right| \leq c_\delta(n_{i,j})\right] \leq \mathbf{P}\left[n_{i,j} \geq n_\delta, \hat{\mu}_{i,j} - c_\delta(n_{i,j}) \leq \frac{1}{2}\right] \leq \mathbf{P}\left[\hat{\mu}_{i,j}(n_\delta) - c_\delta(n_\delta) \leq \frac{1}{2}\right] \\
&= \mathbf{P}\left[\left(\frac{1}{2} + \Delta_{i,j}\right) - \hat{\mu}_{i,j}(n_\delta) \geq \Delta_{i,j} - c_\delta(n_\delta)\right] \leq \mathbf{P}\left[\left|\hat{\mu}_{i,j}(n_\delta) - \mathbb{E}[\hat{\mu}_{i,j}(n_\delta)]\right| \geq \Delta_{i,j} - c_\delta(n_\delta)\right] \leq 2\exp\left(-2n_\delta(\Delta_{i,j} - c_\delta(n_\delta))^2\right),
\end{aligned}$$

where $\hat{\mu}_{i,j}(n_\delta)$ is the empirical average winning rate after exact n_δ comparisons, and the third inequality holds by the difinition of the probability gap, that $\Delta_{i,j} := \left|\frac{1}{2} - \mathbb{E}[\hat{\mu}_{i,j}]\right|$, and the last inequality is by Hoeffding's inequality. When taking $n_\delta = \left\lceil \frac{2}{{\Delta_{i,j}}^2} \log\left(\frac{8Nn_\delta^2}{\delta}\right)\right\rceil$ such that $c_\delta(n_\delta) \leq \Delta_{i,j}/2$, we can bound the result above by

$$\mathbf{P}\left[n_{i,j} \geq n_\delta, \left|\frac{1}{2} - \hat{\mu}_{i,j}\right| \leq c_\delta(n_{i,j})\right] \leq 2\exp\left(-2n_\delta(\Delta_{i,j} - c_\delta(n_\delta))^2\right) \leq 2\exp\left(-2n_\delta c_\delta^2(n_\delta)\right) \leq \frac{\delta}{4Nn_\delta^2} \leq \frac{\delta}{4N} \leq \frac{\delta}{2N}.$$

This implies that whenever $n_{i,j} \geq n_\delta$ for some $n_\delta = \Theta\left(\frac{1}{{\Delta_{i,j}}^2} \log\left(\frac{N}{\Delta_{i,j}\delta}\right)\right)$, it is rare for the event that the denoising phase does not finish as neither of the arm in (i, j) is not eliminated after comparing $n_{i,j}$ times, with probability at most $\delta/(2N)$. This completes the denoising part. Accordingly, by union bound, we add up this procedure over all other arms

$$\mathbf{P}\left[\exists\, j \in \{\mathcal{X} \setminus \{i\}\}, n_{i,j} \geq n_\delta, \left|\frac{1}{2} - \hat{\mu}_{i,j}\right| \leq c_\delta(n_{i,j})\right] \leq (N-1) \cdot \frac{\delta}{2N} \leq \frac{\delta}{2}.$$

Hence, the total number of comparisons is a linear denoising ergodic over all arms, bounded by

$$\sum_{j \in \{\mathcal{X}\setminus\{i\}\}} n_\delta = O\left(\frac{N}{\Delta^2} \log\left(\frac{N}{\Delta\delta}\right)\right),$$

where Δ is the global minimum gap. □

A.2 Proof of Theorem 3.3

We first focus on the error rate of the algorithm.

Lemma A.1. *Under the elimination criterion given by Algorithm 2, the Condorcet winner is eliminated, and thus the wrong arm is returned with probability at most $\delta/2$.*

Proof. Similarly, denote by $n_{a^*,a}$ the number of comparisons between the Condorcet winner a^* and arbitrary other arm a. The algorithm outputs a wrong arm that is not the Condorcet winner only when the Condorcet winner is eliminated by any other arms. We first bound the probability of a^* being eliminated by a as follows,

$$\begin{aligned}
&\mathbf{P}\left[\exists\, n_{a^*,a}, \hat{\mu}_{a^*,a} + c_\delta(n_{a^*,a}) \leq \frac{1}{2}\right] \leq \sum_{n_{a^*,a}=1}^{\infty} \mathbf{P}\left[\hat{\mu}_{a^*,a} - \frac{1}{2} - \Delta^* \leq -c_\delta(n_{a^*,a}) - \Delta^*\right] \leq \sum_{n_{a^*,a}=1}^{\infty} \mathbf{P}\left[\mathbb{E}[\hat{\mu}_{a^*,a}] - \hat{\mu}_{a^*,a} \geq c_\delta(n_{a^*,a}) + \Delta^*\right] \\
&\leq \sum_{n_{a^*,a}=1}^{\infty} \mathbf{P}\left[|\hat{\mu}_{a^*,a} - \mathbb{E}[\hat{\mu}_{a^*,a}]| \geq c_\delta(n_{a^*,a})\right] \leq \sum_{n_{a^*,a}=1}^{\infty} 2\exp\left(-2n_{a^*,a}c_\delta^2(n_{a^*,a})\right) = \sum_{n_{a^*,a}=1}^{\infty} \frac{\delta}{4Nn_{a^*,a}^2} \leq \frac{\pi^2}{6} \cdot \frac{\delta}{4N} \leq \frac{\delta}{2N},
\end{aligned}$$

where the second inequality is due to the fact that $\mathbb{E}[\hat{\mu}_{a^*,a}] = p_{a^*,a} \geq \frac{1}{2} + \Delta^*$ for all $a \in \mathcal{X}$ by the definition of Δ^*, and the third inequality omits Δ^* since $\Delta^* > 0$.

Hence, with probability at most $\delta/(2N)$, there exists an arm $a \neq a^*$ that eliminates the Condorcet winner. Applying union bound over all $a \in \{\mathcal{X} \setminus \{a^*\}\}$ yields

$$\mathbf{P}\left[\exists\, a \in \{\mathcal{X} \setminus \{a^*\}\}, \exists\, n_{a^*,a}, \hat{\mu}_{a^*,a} + c_\delta(n_{a^*,a}) \leq \frac{1}{2}\right] \leq (N-1) \cdot \frac{\delta}{2N} \leq \frac{\delta}{2}.$$

□

Next, we focus on the second statement related to sample complexity. We divide total comparisons into comparisons that the Condorcet winner **duels** and the comparisons that the Condorcet winner **does not duel**. The following lemma bounds the comparisons that the Condorcet winner does not duel,

Lemma A.2. *Denote by R_w the cumulative weak regret incurred by our algorithm, the number of comparisons where the Condorcet winner **does not duel** is at most R_w.*

The following lemma bounds the number of comparisons the Condorcet winner **duels**,

Lemma A.3. *The number of comparisons where the Condorcet winner **duels** is at most $O\left(\frac{N}{\Delta^{*2}} \log\left(\frac{N}{\Delta^* \delta}\right)\right)$ with probability at least $1 - \delta/2$.*

Combining the above lemmas completes the proof.

A.3 Proof of Theorem 3.5

As most parts of Theorem 3.5 are straightforward corollaries of Theorem 3.3. The central proof focuses on the regret part, that we can treat the PE reduction as a sequence of sub-problems, induced from the original dueling bandits problem where none of the arms are eliminated, and thus preserves the regret results in Proposition 3.4. Specifically, as the only difference between the selection sequence of WS-W and WSW-PE comes from the possibly reduced arm set of WSW-PE. We first propose the following two facts: the minimum gap of reduced sub-problem, denoted by Δ_s, is never smaller than that of the original problem; the number of arms of the sub-problem, denoted by N_s, is never larger than that of the original problem. That is, $\Delta_s \geq \Delta, N_s \leq N$. We define some indicator functions as

- Denote by $D(\ell)$ as the indicator of the event that the Condorcet winner is chosen as the main arm i at the first iteration of epoch ℓ.
- Denote by $V(\ell, k)$ as the indicator of the event that $D(\ell) = 1$ and the Condorcet winner is not selected as the main arm i in the the k-th iteration of epoch ℓ, which means the Condorcet winner must be disqualified by other arms during the first iteration to the $k-1$-th iteration of epoch ℓ.
- Denote by $B(\ell, k)$ as the indicator of the event that the main arm i is better than the other arm j in the k-th iteration of epoch ℓ.

Let $\bar{\cdot} = 1 - \cdot$ be the inverse indicator of $\cdot$. With these indicator functions.

Proof. (Theorem 3.5).

Notice that weak regret is incurred during the k-th iteration of epoch ℓ only when $D(\ell) = 0$ or $V(\ell, k') = 1$ for some $k' \leq k$, and each duel incurs at most 1 weak regret by definition. Thus we have

$$\mathbb{E}[R_w] \leq \sum_{\ell=1}^{\infty} \sum_{k=1}^{N_s-1} \mathbb{E}\left[(\bar{D}(\ell) + V(\ell,k))\frac{\tau_{\ell,k}}{2}\right] \leq \text{Term A} + \text{Term B} + \text{Term C} + \text{Term D},$$

where according to [Lemma 1,2,3 of Chen and Frazier [5]], we have

$$\text{Term A} = \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} B(\ell,k)\bar{D}(\ell)\tau_{\ell,k}\right] \leq \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} \frac{1}{2\Delta} \cdot \left(\frac{1-2\Delta}{1+2\Delta}\right)^{\ell-1}\right] \leq O\left(\frac{N}{\Delta^2}\right).$$

$$\text{Term B} = \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} B(\ell,k)V(\ell,k)\tau_{\ell,k}\right] \leq \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} \frac{1}{2\Delta} \cdot \left(\frac{1-2\Delta}{1+2\Delta}\right)^{\ell}\right] \leq O\left(\frac{N}{\Delta^2}\right).$$

$$\text{Term C} = \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} \bar{B}(\ell,k)\bar{D}(\ell)\tau_{\ell,k}\right] \leq \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} \frac{N\ell\left(1+2\Delta^2\right)}{32\Delta^4} \cdot (\log N + 1) \cdot \left(\frac{1-2\Delta}{1+2\Delta}\right)^{\ell-1}\right] \leq O\left(\frac{N\log N}{\Delta^6}\right).$$

$$\text{Term D} = \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} \bar{B}(\ell,k)V(\ell,k)\tau_{\ell,k}\right] \leq \mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^{N-1} \frac{N\ell\left(1+2\Delta^2\right)}{32\Delta^4} \cdot (\log N + 1) \cdot \left(\frac{1-2\Delta}{1+2\Delta}\right)^{\ell}\right] \leq O\left(\frac{N\log N}{\Delta^6}\right).$$

Combining the above results and the statement follows.

□