Adversarial Multi-dueling Bandits

Pratik Gajane [1]

Abstract

We introduce the problem of regret minimization in adversarial multi-dueling bandits. While adversarial preferences have been studied in dueling bandits, they have not been explored in multi-dueling bandits. In this setting, the learner is required to select $m \geq 2$ arms at each round and observes as feedback the identity of the most preferred arm which is based on an arbitrary preference matrix chosen obliviously. We introduce a novel algorithm, MiDEX (Multi Dueling EXP3), to learn from such preference feedback that is assumed to be generated from a pairwise-subset choice model. We prove that the expected cumulative T-round regret of MiDEX compared to a Borda-winner from a set of K arms is upper bounded by $O((K \log K)^{1/3} T^{2/3})$. Moreover, we prove a lower bound of $\Omega(K^{1/3} T^{2/3})$ for the expected regret in this setting which demonstrates that our proposed algorithm is near-optimal.

1. Introduction

Multi-armed bandits (MAB) is a sequential decision making framework that involves selecting from multiple options (symbolized as *arms*) with unknown outcomes to optimize performance over time. This framework can be useful in impactful applications like e-healthcare, clinical trials, recommendation systems, and online advertising.

In a classical MAB problem, the learner selects an arm in each round and observes absolute feedback i.e., a numerical value as feedback for the selected arm. However, in some tasks, especially those requiring human feedback, it is often more practical to elicit preference feedback than absolute feedback. Motivated by such scenarios, there has been a growing body of work on *dueling bandits* in which the learner selects a pair of arms to be compared in each round, and receives preference feedback about the selected pair. Recently, a few works have extended this setup to *multi-dueling bandits* in which the learner selects a subset of $m \geq 2$ arms in each round, and receives preference feedback about the selected arms (Brost et al., 2016; Saha & Gopalan, 2018; 2019; Ren et al., 2019; Agarwal et al., 2020; Sui et al., 2017; Haddenhorst et al., 2021; Du et al., 2020).

Preferences, either over a pair of arms or for $m \geq 2$ arms, can be expressed as stochastic *stationary* preferences or *adversarial* preferences. Stochastic stationary preferences represent scenarios where preferences are assumed to be generated through stochastic models that do not change over time. Such preferences might be unable to capture real-world applications where preferences might vary significantly and unpredictably over time. These preferences would find more faithful representation within a robust worst-case (*adversarial*) model, which avoids the stringent stochastic assumption and allows for an arbitrary sequence of preferences over time. For dueling bandits, several algorithms have been proposed for stochastic stationary preferences (Yue & Joachims, 2009; 2011; Yue et al., 2012; Urvoy et al., 2013; Zoghi et al., 2014; Komiyama et al., 2015) and for adversarial preferences (Gajane et al., 2015; Saha et al., 2021). However, to the best of our knowledge, all the previous work on multi-dueling bandits assumes stochastic stationary preferences, and adversarial preferences have not been studied in this context.

Our Contributions

- We introduce and formalize the problem of regret minimization in adversarial multi-dueling bandits, where the learner is required to select $m \geq 2$ arms at each round and observes as feedback the identity of the most preferred arm. In this general adversarial model, the sequence of preference matrices is allowed to be entirely arbitrary and they are chosen obliviously by the environment.
- We propose a novel algorithm called, MiDEX, considering a *pairwise-subset choice model* for feedback (exact definitions will follow in Section 3).
- We analyze the expected cumulative regret of MiDEX compared to a *Borda*-winner (which, unlike the alternative of *Condorcet*-winner, always exists and

[1]Eindhoven University of Technology, Netherlands. Correspondence to: Pratik Gajane <pratik.gajane@gmail.com>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria.

may suit the adversarial model better). Our analysis demonstrates that the expected cumulative regret of MiDEX is upper bounded by $O((K \log K)^{1/3} T^{2/3})$.

- Furthermore, we establish a lower bound of $\Omega(K^{1/3} T^{2/3})$ for the expected cumulative regret, indicating the near-optimality of our proposed algorithm.

2. Related Work

In the multi-dueling bandits problem considered in Brost et al. (2016); Schuth et al. (2016); Sui et al. (2017); Du et al. (2020), the learner is assumed to receive some subset of the possible $\binom{m}{2}$ pairwise comparisons amongst the selected m arms. In contrast, Saha & Gopalan (2018); Agarwal et al. (2020) assume a more limited form of feedback, referred to as *winner feedback*, where the learner receives only the identity of the arm that is most preferred among the selected arms. In this article, we consider winner feedback.

In multi-dueling bandits (and dueling bandits), several notions of an optimal arm have been considered in the literature. Many works on multi-dueling bandits use the notion of *Condorcet winner*: an arm being preferred when compared to any other arm. For instance, Saha & Gopalan (2018); Brost et al. (2016); Du et al. (2020) consider regret minimization in multi-dueling bandits for stochastic preferences with Condorcet winner. Agarwal et al. (2020) extend this notion to a generalized Condorcet winner: an arm that has the greatest probability of being the winner in each subset containing it and propose an algorithm for regret minimization. Haddenhorst et al. (2021) also use the notion of a generalized Condorcet winner and propose an algorithm for best arm identification with bounds on its sample complexity. Saha & Gopalan (2019) study the problem of identifying a near-best arm with high confidence where the Condorcet winner is considered to be the best arm. All of these works in the framework of multi-dueling bandits assume that the underlying preferences are of a stationary stochastic nature.

As highlighted by Jamieson et al. (2015), using the notion of a Condorcet winner may pose several drawbacks. Chief among these is the potential non-existence of a Condorcet winner, as illustrated by the absence of one in widely used datasets like MSLR-WEB10k (Qin et al., 2010). Moreover, in the context of adversarial preferences addressed in this study, assuming the presence of a Condorcet winner would imply preferences where a certain fixed arm is consistently preferred to all the other arms at all rounds. Such a constraint might render the framework of adversarial (multi-)dueling bandits that presupposes the existence of a Condorcet winner unsuitable for many real-world applications with non-stationary preferences.

Alternatively, the notion of a *Borda winner* has been used in adversarial dueling bandits (Saha et al., 2021). A Borda winner is an arm with the highest *Borda score* where the Borda score of an arm is the probability that it is preferred over another arm chosen uniformly at random. Firstly, the advantage of using the notion of Borda winner is that it always exists, unlike a Condorcet winner. Secondly, as argued in Jamieson et al. (2015), in certain cases a Borda winner represents a better reflection of preferences than a Condorcet winner when they are distinct, and the former is more robust to estimation errors in preferences. Consequently, in this article, we use the notion of a Borda winner.

Other notions of an optimal arm have also been considered for dueling bandits with stochastic preferences: *Copeland winner* (Zoghi et al., 2015; Komiyama et al., 2016; Wu & Liu, 2016) and *von Neumann* winner (Balsubramani et al., 2016; Dudík et al., 2015).

Another tangentially related problem is the one considered in Ren et al. (2019) where there exists a unique unknown ranking $r_1 \succ r_2, \ldots \succ r_K$ such that $i \succ j$ indicates that i is more preferred than j; the learner receives winner feedback for the selected $m \geq 2$ arms; and the learner's goal is to recover this true ranking.

Multi-armed bandits with preference feedback can also be formulated as partial monitoring games which is a rich framework for sequential decision making under uncertainty (Gajane & Urvoy, 2015; Kirschner et al., 2023).

3. Problem Setting

We consider an online decision making task over a finite set of arms $[K] := \{1, 2, \ldots, K\}$ which spans T rounds [1]. At each round $t = 1, 2, \ldots, T$,

- the learner selects, possibly at random, a multiset of arms $\mathcal{A}_t$ such that $|\mathcal{A}_t| = m$ where $2 \leq m \leq K$; and
- the learner observes a 'winner': an arm that is preferred over all the other arms in $\mathcal{A}_t$ at time t.

The selection of a winner from a multiset of arms is governed by the underlying *subset choice model*. Given a multiset of arms, a subset choice model determines the probability of one of the arms being preferred over the rest in the multiset. In this article, we consider the *pairwise-subset choice model*, introduced by Saha & Gopalan (2018). There also exist other subset choice models in the related literature such as a popular class of models called *Random Utility Models* (Soufiani et al., 2012).

[1] Throughout the article, we use the shorthand of $[V]$ to represent $\{1, 2, 3 \ldots, V\}$ for any positive integer V.

3.1. Pairwise-subset Choice Model

We assume that the environment obliviously fixes a sequence of T preference matrices $P_1, P_2, \ldots, P_T$ where each $P_t(i,j)$ is the probability that arm i is preferred when compared to arm j at round $t \in [T]$. Each $P_t \in [0,1]^{K\times K}$ satisfies $P_t(i,j) = 1 - P_t(j,i)$ and $P_t(i,i) = 1/2$ for all $i, j \in [K]$. These preference matrices are not revealed to the learner.

Given a multiset of arms $\mathcal{A} = \{\mathcal{A}(1), \mathcal{A}(2), \ldots, \mathcal{A}(m)\}$ and a corresponding preference matrix P, the probability of any index $i \in [m]$ being selected as the winner is defined as

$$W(i \mid \mathcal{A}, P) \coloneqq \sum_{j=1, j\neq i}^{m} \frac{2P\left(\mathcal{A}(i), \mathcal{A}(j)\right)}{m(m-1)}.$$

As noted in Saha & Gopalan (2018), the above forms a valid probability distribution over the indices $i \in [m]$, and when $m = 2$ (which corresponds to the dueling bandits case), it simplifies to the probability of an arm winning the pairwise duel.

3.2. Performance Measure: Regret

The performance of the learner's arm selection strategy is measured against the performance of an optimal arm in hindsight. As noted earlier in Section 2, we use the notion of a Borda winner which is defined using the Borda score defined below.

Definition 1 (Borda Score)**.** The Borda score of an arm $i \in [K]$ according to a preference matrix P_t is defied as

$$b_t(i) \coloneqq \frac{1}{K-1} \sum_{j\in[K]\setminus\{i\}} P_t(i,j).$$

Accordingly, the optimal arm i^* is defined as the arm with the highest cumulative Borda score up to horizon T i.e.,

$$i^* \coloneqq \underset{i\in[K]}{\operatorname{argmax}} \sum_{t=1}^{T} b_t(i).$$

Definition 2 (Regret)**.** Let $\mathcal{A}_t$ be the subset of arms selected by an algorithm at $t = 1, \ldots, T$ such that $|\mathcal{A}_t| = m$. Then regret of the algorithm at the end of horizon T is defined as

$$R_T \coloneqq \sum_{t=1}^{T} \left[b_t(i^*) - \frac{1}{m} \sum_{i\in\mathcal{A}_t} b_t(i) \right].$$

In our proposed algorithm, we make use of the *Shifted Borda Score* (Saha et al., 2021).

Definition 3 (Shifted Borda Score)**.** The shifted Borda score of an arm $i \in [K]$ according to a preference matrix P_t is defined as

$$s_t(i) \coloneqq \frac{1}{K} \sum_{j\in[K]} P_t(i,j).$$

Definition 4 (Shifted Borda Regret)**.** Let $\mathcal{A}_t$ be the subset of arms selected by an algorithm at $t = 1, \ldots, T$ such that $|\mathcal{A}_t| = m$. Then shifted Borda regret of the algorithm at the end of horizon T is defined as

$$R_T^s \coloneqq \sum_{t=1}^{T} \left[s_t(i^*) - \frac{1}{m} \sum_{i\in\mathcal{A}_t} s_t(i) \right].$$

The following proposition lets us interpret the shifted Borda score of an arm in terms of its Borda score.

Proposition 1. *The shifted Borda score $s_t(i)$ of any arm $i \in [K]$ is related to its Borda score $b_t(i)$ by the equation*

$$s_t(i) = \frac{K-1}{K} b_t(i) + \frac{1}{2K}.$$

Proof.

$$\begin{aligned} s_t(i) &= \frac{1}{K} \sum_{j\in[K]} P_t(i,j) \\ &= \frac{1}{K} \sum_{j\in[K]\setminus\{i\}} P_t(i,j) + \frac{1}{K} P_t(i,i) \\ &= \frac{K-1}{K} b_t(i) + \frac{1}{2K}, \end{aligned}$$

where the last equality follows from Definition 1 and the fact that $P_t(i,i) = \frac{1}{2}$ for any $i \in [K]$. □

Using the above, we can state the following for optimal arm i^* and regret R_T.

Proposition 2. $i^* \coloneqq \operatorname{argmax}_{i\in[K]} \sum_{t=1}^{T} b_t(i) = \operatorname{argmax}_{i\in[K]} \sum_{t=1}^{T} s_t(i)$.

Proposition 3. $R_T = \frac{K}{K-1} R_T^s$.

4. Our Algorithm and Performance Guarantee

In this Section, we provide our proposed algorithm MiDEX (Multi Dueling EXP3). It falls under the class of *Exponential Weight* algorithms — a well-known class of algorithms for MAB problems that can be traced back to Auer et al. (2002).

In MiDEX, firstly at each round $t = 1, 2, \ldots, T$, two arms x_t and y_t are sampled from q_t. Then each of x_t and y_t is replicated about $\frac{m}{2}$ times to constitute the multiset of arms $\mathcal{A}_t$ to be selected at time t. After receiving the winner index from $\mathcal{A}_t$ according to the pairwise-subset choice model as defined in Eq. (1), Procedure 2 transforms the received feedback which is then used to compute an estimate of the shifted Borda score $\hat{s}_t(i)$ for each arm i. These estimates are, in turn, used to compute q_{t+1}. A parameter $\gamma \in (0,1]$ is incorporated to ensure that for all $t \in [T]$ and $i \in [K]$, $q_t(i) \geq \gamma/K$ which translates to the selection probability of any arm always being above zero.

Algorithm 1 MiDEX (Multi Dueling EXP3)

1: **Input:** Set of arms $[K]$, horizon T, number of arms to be selected at each round m, exploration parameter $\gamma \in (0,1]$ and learning rate $\eta > 0$.
2: **Initialize:** Initial arm-selection probability distribution $q_1(i) = 1/K, \forall i \in [K]$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Sample $x_t, y_t \sim q_t$ i.i.d. with replacement.
5: Construct $\mathcal{A}_t = \{x_t, x_t, \ldots, x_t, y_t, y_t, \ldots, y_t\}$ by replicating x_t for $\lceil \frac{m}{2} \rceil$ times and y_t for $\lfloor \frac{m}{2} \rfloor$ times with probability $\frac{1}{2}$, or x_t for $\lfloor \frac{m}{2} \rfloor$ times and y_t for $\lceil \frac{m}{2} \rceil$ time with probability $\frac{1}{2}$.
6: Receive winning index

$$\text{INDEX}_t \sim W_t(i|\mathcal{A}_t) \tag{1}$$

where $W_t(i|\mathcal{A}_t) = \sum_{j=1, j\neq i}^{m} \frac{2P_t(\mathcal{A}_t(i), \mathcal{A}_t(j))}{m(m-1)}$ for any $i \in [m]$ and $\mathcal{A}_t(i)$ is the i^{th} item in $\mathcal{A}_t$.
7: **if** $\mathcal{A}_t(\text{INDEX}_t) = x_t$ **then**
$o_t = x_t$,
else
$o_t = y_t$.
end if
8: Estimates scores, for all $i \in [K]$:

$$\hat{s}_t(i) := \frac{\mathbb{1}(i = x_t)}{K\, q_t(i)} \sum_{j \in [K]} \frac{\mathbb{1}(j = y_t)\, g(m, o_t, x_t)}{q_t(j)}, \tag{2}$$

where $g(m, o_t, x_t)$ is computed as shown in Procedure 2.
9: Update, for all $i \in [K]$:

$$\tilde{q}_{t+1}(i) := \frac{\exp\left(\eta \sum_{\tau=1}^{t} \hat{s}_\tau(i)\right)}{\sum_{j=1}^{K} \exp\left(\eta \sum_{\tau=1}^{t} \hat{s}_\tau(j)\right)};$$

$$q_{t+1}(i) := (1-\gamma)\, \tilde{q}_{t+1}(i) + \frac{\gamma}{K}. \tag{3}$$

10: **end for**

Theorem 1. *Let $\gamma = \sqrt{\frac{3\eta K}{2}}$ and $\eta = \left(\frac{2\log K}{T\sqrt{K}\, m'}\right)^{2/3}$ where $m' = \left(\sqrt{\frac{3}{2}} + \sqrt{\frac{2}{3}}\, \frac{(3m+1)^2}{4(m+1)^2}\right)$. For any T, $K \geq 2$ and $m \geq 2$, the expected regret of MiDEX satisfies*

$$\mathbb{E}[R_T] \leq 3.78\, (m')^{2/3}\, (K \log K)^{1/3}\, T^{2/3}.$$

The regret upper bound can be further simplified to

$$\mathbb{E}[R_T] \leq 8.13\, (K \log K)^{1/3}\, T^{2/3},$$

for any $m \geq 2$.

Procedure 2 $g(m, o_t, x_t)$

if m is even **then**
return $\frac{\mathbb{1}(o_t = x_t) - \frac{(m-2)}{4(m-1)}}{\frac{m}{2(m-1)}}$,
else
return $\frac{\mathbb{1}(o_t = x_t) - \frac{(m-1)}{4m}}{\frac{m+1}{2m}}$.
endif

5. Mathematical Analysis

The proof of Theorem 1 builds upon the following lemmas. The most important lemmas are Lemma 1 and Lemma 2. Lemma 1 proves how the transformed feedback can be interpreted as the probability of x_t winning the duel against y_t. Lemma 2 proves that the score $\hat{s}_t(i)$ being computed in Eq. (2) is an unbiased estimate of the true shifted Borda score $s_t(i)$. Proofs for the following lemmas can be found in the Appendix.

Lemma 1. $\mathbb{E}[g(m, o_t, x_t)] = P_t(x_t, y_t)$.

Lemma 1 is proved using Procedure 2, the construction of $\mathcal{A}_t$ and the definition of $W_t(i \mid \mathcal{A}_t)$.

Lemma 2. *For all $t \in [T]$ and $i \in [K]$, it holds that $\mathbb{E}[\hat{s}_t(i)] = s_t(i)$.*

Lemma 2 is proved using Lemma 1 and the fact that x_t and y_t are sampled i.i.d. from q_t with replacement.

Next, in Lemma 3, we bound the magnitude of the transformed feedback $g(m, o_t, x_t)$.

Lemma 3. *For all $t \in [T]$ and $m \geq 2$, $g(m, o_t, x_t) \leq \frac{3m+1}{2m+2}$.*

Lemma 3 follows from expanding the construction of $g(m, o_t, x_t)$ given in Procedure 2.

In Lemma 4, we bound the magnitude of the shifted Borda score estimates.

Lemma 4. *Let $\gamma \geq \sqrt{3\eta K/2}$. Then, for any $t \in [T]$, $i \in [K]$ and $\eta > 0$, it holds that $\eta \hat{s}_t(i) \in [0, 1]$.*

Lemma 4 is proved using Lemma 3, the definition of q_t given in Eq. (3) and the definition of $\hat{s}_t$ given in Eq. (2).

Let $\mathcal{H}_{t-1} := (q_1, P_1, x_1, y_1, o_1, \ldots, q_t, P_t)$ denote the history up to round t.

Lemma 5. *For all $t \in [T]$, it holds that $\mathbb{E}_{\mathcal{H}_t}\left[q_t^\intercal\, \hat{s}_t\right] = \mathbb{E}_{\mathcal{H}_{t-1}}[\mathbb{E}_{i \sim q_t}\left[s_t(i) \mid \mathcal{H}_{t-1}\right]]$.*

Lemma 5 follows from the proof of Lemma 2.

Lemma 6. *At any time $t \in [T]$, it holds that $\mathbb{E}\left[\sum_{i=1}^{K} q_t(i)\, \hat{s}_t(i)^2\right] \leq \frac{(3m+1)^2}{4(m+1)^2}\, \frac{K}{\gamma}$.*

Lemma 6 is proved using Lemma 3, the definition of $\hat{s}_t$ given in Eq. (2), and the fact that $\forall i' \in [K]$ and $\forall t \in [T]$, $q_t(i') \geq \gamma/K$ according to Eq. (3), the initialization of q_t, and $\gamma \in (0, 1]$.

Lemma 7. *For any $i \in [K]$, $j \in [m]$ and $t \in [T]$,*

$$\mathbb{P}\Big(\mathcal{A}_t(j) = i\Big) = q_t(i).$$

Lemma 7 follows from the construction of $\mathcal{A}_t$ and the fact that x_t and y_t are sampled i.i.d. from q_t with replacement.

5.1. Proof of Theorem 1

Proof. We start by expanding the expression for the expectation of shifted Borda regret R_T^s.

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_T}[R_T^s] &= \mathbb{E}_{\mathcal{H}_T}\left[\sum_{t=1}^{T}\left[s_t(i^*) - \frac{1}{m}\sum_{j \in \mathcal{A}_t} s_t(j)\right]\right] \\
&= \sum_{t=1}^{T} s_t(i^*) - \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\left[\frac{1}{m}\sum_{j \in \mathcal{A}_t} s_t(j)\right] \\
&= \sum_{t=1}^{T} s_t(i^*) - \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{i \sim q_t}\left[s_t(i) \mid \mathcal{H}_{t-1}\right]\right]
\end{aligned} \tag{4}$$

In the above, the second equality holds because the preference matrices P_t are chosen obliviously, and hence s_t and the identity of i^* remain independent of the randomness of the algorithm. Moreover, the last equality uses that all the m arms in $\mathcal{A}_t$ are $\sim q_t$ (Lemma 7).

For any $\gamma \geq \sqrt{3\eta K/2}$ and $\eta > 0$, we have $\eta \hat{s}_t(i) \in [0, 1]$ using Lemma 4. Using the regret guarantee of standard Exponential Weight algorithm (Auer et al., 2002) over the completely observed fixed sequence of reward vectors $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T$, for $i^* := \operatorname{argmax}_{i \in [K]} \sum_{t=1}^{T} b_t(i) = \operatorname{argmax}_{i \in [K]} \sum_{t=1}^{T} s_t(i)$, one can state that

$$\sum_{t=1}^{T} \hat{s}_t(i^*) - \sum_{t=1}^{T} \tilde{q}_t^{\intercal} \hat{s}_t \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T}\sum_{i=1}^{K} \tilde{q}_t(i)\hat{s}_t(i)^2.$$

Using $\tilde{q}_t = \frac{q_t - \frac{\gamma}{K}}{1-\gamma}$ and $\gamma \in (0, 1)$, with the above inequality, we have that

$$\begin{aligned}
&(1-\gamma)\sum_{t=1}^{T} \hat{s}_t(i^*) - \sum_{t=1}^{T} q_t^{\intercal} \hat{s}_t \\
&\quad \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T}\sum_{i=1}^{K} q_t(i)\hat{s}_t(i)^2 \\
\implies &(1-\gamma)\sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_T}\left[\hat{s}_t(i^*)\right] - \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_T}\left[q_t^{\intercal} \hat{s}_t\right] \\
&\quad \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_T}\left[\sum_{i=1}^{K}\left[q_t(i)\hat{s}_t(i)^2\right]\right] \\
\overset{(a)}{\implies} &(1-\gamma)\sum_{t=1}^{T} s_t(i^*) - \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{i \sim q_t}\left[s_t(i) \mid \mathcal{H}_{t-1}\right]\right] \\
&\quad \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \frac{(3m+1)^2}{4(m+1)^2}\frac{K}{\gamma} \\
\implies &\sum_{t=1}^{T} s_t(i^*) - \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{i \sim q_t}\left[s_t(i) \mid \mathcal{H}_{t-1}\right]\right] \\
&\quad \leq \gamma \sum_{t=1}^{T} s_t(i^*) + \frac{\log K}{\eta} + \frac{(3m+1)^2}{4(m+1)^2}\frac{\eta K T}{\gamma} \\
\overset{(b)}{\implies} &\mathbb{E}_{\mathcal{H}_T}[R_T^s] \\
&\quad \leq \gamma T + \frac{\log K}{\eta} + \frac{(3m+1)^2}{4(m+1)^2}\frac{\eta K T}{\gamma} \\
\overset{(c)}{\implies} &\mathbb{E}_{\mathcal{H}_T}[R_T^s] \\
&\quad \leq \sqrt{\frac{3\eta K}{2}}\,T + \frac{\log K}{\eta} + \frac{(3m+1)^2}{4(m+1)^2}\sqrt{\frac{2\eta K}{3}}\,T \\
\overset{(d)}{\implies} &\mathbb{E}_{\mathcal{H}_T}[R_T^s] \\
&\quad \leq 1.89\,(m')^{2/3}\,(K\log K)^{1/3}\,T^{2/3},
\end{aligned}$$

where $m' = \left(\sqrt{\frac{3}{2}} + \sqrt{\frac{2}{3}}\frac{(3m+1)^2}{4(m+1)^2}\right)$. In the above, (a) follows from Lemma 2, Lemma 5 and Lemma 6; (b) follows from Eq. (4) and using $s_t(i^*) \leq 1$; (c) follows from setting $\gamma = \sqrt{3\eta K/2}$; and (d) follows from optimizing over η which gives $\eta = \left(\frac{2\log K}{T\sqrt{K}\,m'}\right)^{2/3}$.

The theorem follows by using $R_T = \frac{K}{K-1}R_T^s$ for any $K \geq 2$ and $T > 0$. □

5.2. Varying m_t

Note that MiDEX is also applicable when the number of arms to be selected is time-dependent. In this setting, at each round t, the learner receives an integer $m_t|_{2 \leq m_t < K}$ which indicates the number of arms to be selected at time t. MiDEX can be employed here with m being replaced with m_t and the corresponding regret bound would feature $m'' = \max_{m \in \{m_1, m_2, \ldots, m_T\}}\left(\sqrt{\frac{3}{2}} + \sqrt{\frac{2}{3}}\frac{(3m+1)^2}{4(m+1)^2}\right)$ instead of m'. The proof structure remains the same with the upper bound in Lemma 3 being updated to $\max_{m \in \{m_1, m_2, \ldots, m_T\}} \frac{(3m+1)}{2(m+1)}$. The subsequent proofs and computations build upon this updated bound to arrive at the regret upper bound featuring m''.

6. Lower Bound

To prove the lower bound for adversarial multi-dueling bandits, we use a reduction from adversarial dueling bandits to adversarial multi-dueling bandits given in Algorithm 3. That is we show how an algorithm $\mathfrak{A}_{MB}$ designed for adversarial multi-dueling bandits can be used to solve an instance of adversarial dueling bandits DB.

Algorithm 3 $\mathfrak{A}_{DB}$: Reduction from adversarial dueling bandits to adversarial multi-dueling bandits

for t=1,2,... **do**
- $\mathcal{A}_t = \{\mathcal{A}_t(1), \mathcal{A}_t(2), \dots, \mathcal{A}_t(m)\} \leftarrow$ multiset of arms played by $\mathfrak{A}_{MD}$ at round t.
- Sample i_t, j_t from $[m]$ uniformly at random without replacement.
- Play $\Big(\mathcal{A}_t(i_t), \mathcal{A}_t(j_t)\Big)$ where $\mathcal{A}_t(i)$ is the i^{th} item in $\mathcal{A}_t$.
- Receive $w_t \sim \text{BERNOULLI}\Big(P_t\Big(\mathcal{A}_t(i_t), \mathcal{A}_t(j_t)\Big)\Big)$.
- Return $\text{INDEX}_t = i_t w_t + j_t(1 - w_t) \in \{i_t, j_t\}$ as the winning index to $\mathfrak{A}_{MB}$.

end for

Note that even though this reduction is the same as the reduction suggested by Saha & Gopalan (2018) for stochastic multi-dueling bandits, our novel contribution is the lemma below which shows that Algorithm 3 preserves the expected regret for any arbitrary sequence of preference matrices.

Lemma 8. *Using $\mathfrak{A}_{DB}$ given in Algorithm 3,*

$$\mathbb{E}[R_T(\mathfrak{A}_{DB})] = R_T(\mathfrak{A}_{MB}),$$

for any arbitrary sequence of preference matrices $P_1, P_2, \dots, P_T$.

The complete proof can be found in the Appendix. Here we provide a brief outline of the proof.

Proof Outline.

Let $r_t(\mathfrak{A}_{DB}) := b_t(i^*) - \frac{b_t\big(\mathcal{A}_t(i_t)\big) + b_t\big(\mathcal{A}_t(j_t)\big)}{2}$ be the instantaneous regret of $\mathfrak{A}_{DB}$ at round t. Correspondingly, let $r_t(\mathfrak{A}_{MB}) := b_t(i^*) - \frac{1}{m}\left[\sum_{i=1}^{m} b_t\Big(\mathcal{A}_t(i)\Big)\right]$ be the instantaneous regret of $\mathfrak{A}_{MB}$ at round t. Firstly, we show that $\mathbb{E}_{i_t, j_t \overset{Unif}{\sim} [m], i_t \neq j_t}[r_t(\mathfrak{A}_{DB})] = r_t(\mathfrak{A}_{MB})$. Then,

$$\mathbb{E}[R_T(\mathfrak{A}_{DB})] = \sum_{t=1}^{T} \mathbb{E}[r_t(\mathfrak{A}_{DB})] = \sum_{t=1}^{T} r_t(\mathfrak{A}_{MB}) = R_T(\mathfrak{A}_{MB}).$$

□

Using the above reduction and Lemma 8, along with the lower bound proved for adversarial dueling bandits (Saha et al., 2021)[Theorem 16], we can state the following lower bound for the expected regret of adversarial multi-dueling bandits measured against a Borda winner.

Theorem 2. *For any learning algorithm $\mathfrak{A}$, there exists an instance of adversarial multi-dueling bandits with $T \geq K$, $K \geq 4$ and a sequence of preferences $P_1, P_2, \dots, P_T$, such that the expected regret of $\mathfrak{A}$ for that instance is at least $\Omega(K^{1/3}T^{2/3})$.*

7. Concluding Remarks

In conclusion, we have introduced and formalized the problem of regret minimization in adversarial multi-dueling bandits, extending previous research on multi-armed bandits with preference feedback. Our work addresses a gap in the literature by considering scenarios where the learner selects multiple arms at each round and observes the identity of the most preferred arm, based on arbitrary preference matrices. Central to our contribution is the development of a novel algorithm, MiDEX, tailored to learn from preference feedback following a pairwise-subset choice model. Through rigorous analysis, we have demonstrated that MiDEX achieves near-optimal performance in terms of its expected cumulative regret measured against a Borda winner. Specifically, our upper bound on the expected cumulative regret of MiDEX is of the order $O((K \log K)^{1/3}T^{2/3})$. We also prove a matching lower bound of $\Omega(K^{1/3}T^{2/3})$, thereby demonstrating the near-optimality of our proposed algorithm up to a logarithmic factor. Future research directions include conducting high-probability regret analysis and exploring the dynamic regret objective with respect to a time-varying benchmark. Another valuable direction would be to investigate alternative notions for optimal arm and subset choice models. It would also be advantageous to develop a meta-algorithm for multi-dueling bandits which can make use of the corresponding algorithm for dueling bandits as a black-box leading us to incorporate the advancements in dueling bandits into multi-dueling bandits as done for other problems (e.g., Gajane et al. (2023)).

Acknowledgements

This work is supported by the Dutch Research Council (NWO) in the framework of EDIC project (grant number 628.011.021).

Impact Statement

This article presents work whose goal is to advance the field of reinforcement learning theory. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A., Johnson, N., and Agarwal, S. Choice bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18399–18410. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d5fcc35c94879a4afad61cacca56192c-Paper.pdf.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Balsubramani, A., Karnin, Z., Schapire, R. E., and Zoghi, M. Instance-dependent regret bounds for dueling bandits. In Feldman, V., Rakhlin, A., and Shamir, O. (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 336–360, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/balsubramani16.html.

Brost, B., Seldin, Y., Cox, I. J., and Lioma, C. Multi-dueling bandits and their application to online ranker evaluation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pp. 2161–2166, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983659. URL https://doi.org/10.1145/2983323.2983659.

Du, Y., Wang, S., and Huang, L. Dueling bandits: From two-dueling to multi-dueling. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pp. 348–356, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Proceedings of The 28th Conference on Learning Theory*, pp. 563–587, 2015.

Gajane, P. and Urvoy, T. Utility-based dueling bandits as a partial monitoring game. In *the 12th European Workshop on Reinforcement Learning*, 2015.

Gajane, P., Urvoy, T., and Clérot, F. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 218–227, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/gajane15.html.

Gajane, P., Auer, P., and Ortner, R. Autonomous exploration for navigating in mdps using blackbox rl algorithms. In Elkind, E. (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 3714–3722. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/413. URL https://doi.org/10.24963/ijcai.2023/413. Main Track.

Haddenhorst, B., Bengs, V., and Hüllermeier, E. Identification of the generalized condorcet winner in multi-dueling bandits. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=omDF-uQ_OZ.

Jamieson, K., Katariya, S., Deshpande, A., and Nowak, R. Sparse Dueling Bandits. In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 416–424, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/jamieson15.html.

Kirschner, J., Lattimore, T., and Krause, A. Linear partial monitoring for sequential decision making: Algorithms, regret bounds and applications. *Journal of Machine Learning Research*, 24(346):1–45, 2023. URL http://jmlr.org/papers/v24/22-1248.html.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. Regret lower bound and optimal algorithm in dueling bandit problem. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1141–1154, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Komiyama15.html.

Komiyama, J., Honda, J., and Nakagawa, H. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1235–1244, New York, New York, USA, 20–22 Jun 2016. PMLR.

URL https://proceedings.mlr.press/v48/komiyama16.html.

Qin, T., Liu, T.-Y., Xu, J., and Li, H. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4): 346–374, Aug 2010. ISSN 1573-7659. doi: 10.1007/s10791-009-9123-y. URL https://doi.org/10.1007/s10791-009-9123-y.

Ren, W., Liu, J., and Shroff, N. B. *On sample complexity upper and lower bounds for exact ranking from noisy comparisons*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Saha, A. and Gopalan, A. Battle of bandits. In Globerson, A. and Silva, R. (eds.), *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pp. 805–814. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/290.pdf.

Saha, A. and Gopalan, A. Pac battling bandits in the plackett-luce model. In Garivier, A. and Kale, S. (eds.), *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pp. 700–737. PMLR, 22–24 Mar 2019. URL https://proceedings.mlr.press/v98/saha19a.html.

Saha, A., Koren, T., and Mansour, Y. Adversarial dueling bandits. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9235–9244. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/saha21a.html.

Schuth, A., Oosterhuis, H., Whiteson, S., and de Rijke, M. Multileave gradient descent for fast online learning to rank. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pp. 457–466, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450337168. doi: 10.1145/2835776.2835804. URL https://doi.org/10.1145/2835776.2835804.

Soufiani, H. A., Parkes, D. C., and Xia, L. Random utility theory for social choice. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 126–134, Red Hook, NY, USA, 2012. Curran Associates Inc.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multi-dueling bandits with dependent arms. *CoRR*, abs/1705.00253, 2017. URL http://arxiv.org/abs/1705.00253.

Urvoy, T., Clerot, F., Féraud, R., and Naamane, S. Generic exploration and K-armed voting bandits. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 91–99, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/urvoy13.html.

Wu, H. and Liu, X. Double thompson sampling for dueling bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 649–657, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Yue, Y. and Joachims, T. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1201–1208, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553527. URL https://doi.org/10.1145/1553374.1553527.

Yue, Y. and Joachims, T. Beat the mean bandit. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 241–248, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, sep 2012. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.028. URL https://doi.org/10.1016/j.jcss.2011.12.028.

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. Relative upper confidence bound for the k-armed dueling bandit problem. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 10–18, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/zoghi14.html.

Zoghi, M., Karnin, Z. S., Whiteson, S., and de Rijke, M. Copeland dueling bandits. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/9872ed9fc22fc182d371c3e9ed316094-Paper.pdf.

A. Proof of Lemma 1

Lemma 1. $\mathbb{E}[g(m, o_t, x_t)] = P_t(x_t, y_t)$.

Proof. **Case 1: m is even.**

$$\mathbb{E}[g(m, o_t, x_t)] = \frac{\mathbb{E}[\mathbb{1}(o_t = x_t)] - \frac{(m-2)}{4(m-1)}}{\frac{m}{2(m-1)}} \tag{5}$$

Using the construction of $\mathcal{A}_t$, one can write

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(o_t = x_t)] &= \sum_{i=1}^{m/2} W_t(i \mid \mathcal{A}_t) \\
&= \frac{m}{2}\left(2\frac{\left(\frac{m}{2}-1\right)P_t(x_t, x_t) + \frac{m}{2}P_t(x_t, y_t)}{m(m-1)}\right) \\
&= \frac{\left(\left(\frac{m}{2}-1\right)\frac{1}{2}\right)}{m-1} + \frac{m}{2(m-1)}P_t(x_t, y_t) \\
&= \frac{m}{2(m-1)}P_t(x_t, y_t) + \frac{m-2}{4(m-1)}.
\end{aligned} \tag{6}$$

In the above, the second equality follows from the definition of $W_t(i \mid \mathcal{A}_t)$ and (Saha & Gopalan, 2018)[Lemma 1]. Substituting Eq. (6) in Eq. (5), we get

$$\mathbb{E}[g(m, o_t, x_t)] = P_t(x_t, y_t).$$

Case 2: m is odd.
We proceed on similar lines to Case 1.

$$\mathbb{E}[g(m, o_t, x_t)] = \frac{\mathbb{E}\left[\mathbb{1}(o_t = x_t)\right] - \frac{(m-1)}{4m}}{\frac{m+1}{2m}} \tag{7}$$

Using the construction of $\mathcal{A}_t$, one can write

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(o_t = x_t)] &= \frac{1}{2}\sum_{i=1}^{(m-1)/2} W_t(i \mid \mathcal{A}_t) + \frac{1}{2}\sum_{i=1}^{(m+1)/2} W_t(i \mid \mathcal{A}_t) \\
&= \frac{1}{2}\left(\frac{m-1}{2}\right)\left(2\frac{\left(\frac{m-1}{2}-1\right)P_t(x_t, x_t) + \frac{m+1}{2}P_t(x_t, y_t)}{m(m-1)}\right) \\
&\quad + \frac{1}{2}\left(\frac{m+1}{2}\right)\left(2\frac{\left(\frac{m+1}{2}-1\right)P_t(x_t, x_t) + \frac{m-1}{2}P_t(x_t, y_t)}{m(m-1)}\right) \\
&= \left(\frac{1}{2}\right)\left(\frac{\left(\frac{m-1}{2}-1\right)\frac{1}{2} + \frac{m+1}{2}P_t(x_t, y_t)}{m}\right) \\
&\quad + \left(\frac{m+1}{2}\right)\left(\frac{\left(\frac{m+1}{2}-1\right)\frac{1}{2} + \frac{m-1}{2}P_t(x_t, y_t)}{m(m-1)}\right) \\
&= \left(\frac{1}{2}\right)\left(\frac{\left(\frac{m-3}{4}\right) + \frac{m+1}{2}P_t(x_t, y_t)}{m}\right) + \left(\frac{m+1}{2}\right)\left(\frac{\left(\frac{m-1}{4}\right) + \frac{m-1}{2}P_t(x_t, y_t)}{m(m-1)}\right) \\
&= \frac{m-3}{8m} + \left(\frac{m+1}{2}\right)\frac{(m-1)}{4m(m-1)} + \left(\frac{m+1}{4m}\right)P_t(x_t, y_t) + \left(\frac{m+1}{2}\right)\frac{m-1}{2m(m-1)}P_t(x_t, y_t) \\
&= \frac{m-3}{8m} + \frac{m+1}{8m} + \left(\frac{m+1}{4m}\right)P_t(x_t, y_t) + \left(\frac{m+1}{4m}\right)P_t(x_t, y_t)
\end{aligned}$$

$$= \left(\frac{m+1}{2m}\right) P_t(x_t, y_t) + \frac{m-1}{4m}. \tag{8}$$

In the above, the second equality follows from the definition of $W_t(i \mid \mathcal{A}_t)$ and (Saha & Gopalan, 2018)[Lemma 1]. Substituting Eq. (8) in Eq. (7), we get

$$\mathbb{E}[g(m, o_t, x_t)] = P_t(x_t, y_t).$$

□

B. Proof of Lemma 2

Lemma 2. *For all $t \in [T]$ and $i \in [K]$, it holds that $\mathbb{E}[\hat{s}_t(i)] = s_t(i)$.*

Proof.

$$\begin{aligned}
\mathbb{E}[\hat{s}_t(i)] &= \mathbb{E}_{\mathcal{H}_t}\left[\frac{\mathbb{1}(i = x_t)}{K\, q_t(i)} \sum_{j\in[K]} \frac{\mathbb{1}(j = y_t)\, g(m, o_t, x_t)}{q_t(j)}\right] \\
&= \frac{1}{K}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{(x_t, y_t, o_t)}\left[\frac{\mathbb{1}(i = x_t)}{q_t(i)} \sum_{j\in[K]} \frac{\mathbb{1}(j = y_t)\, g(m, o_t, x_t)}{q_t(j)} \,\middle|\, \mathcal{H}_{t-1}\right]\right]\right) \\
&= \frac{1}{K}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{x_t}\left[\frac{\mathbb{1}(i = x_t)}{q_t(i)} \sum_{j\in[K]} \mathbb{E}_{y_t}\left[\frac{\mathbb{1}(j = y_t)\, \mathbb{E}_{o_t}\left[g(m, o_t, x_t) \mid o_t\right]}{q_t(j)} \,\middle|\, x_t\right] \,\middle|\, \mathcal{H}_{t-1}\right]\right]\right) \\
&= \frac{1}{K}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{x_t}\left[\frac{\mathbb{1}(i = x_t)}{q_t(i)} \sum_{j\in[K]} \mathbb{E}_{y_t}\left[\frac{\mathbb{1}(j = y_t)\, P_t(x_t, y_t)}{q_t(j)} \,\middle|\, x_t\right] \,\middle|\, \mathcal{H}_{t-1}\right]\right]\right) \\
&= \frac{1}{K}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{x_t}\left[\frac{\mathbb{1}(i = x_t)}{q_t(i)} \sum_{j\in[K]} \sum_{j'\in[K]} \left[\frac{\mathbb{1}(j = j')\, P_t(x_t, j')\, q_t(j')}{q_t(j)}\right] \,\middle|\, \mathcal{H}_{t-1}\right]\right]\right) \\
&= \frac{1}{K}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\mathbb{E}_{x_t}\left[\frac{\mathbb{1}(i = x_t)}{q_t(i)} \sum_{j\in[K]} P_t(x_t, j) \,\middle|\, \mathcal{H}_{t-1}\right]\right]\right) \\
&= \frac{1}{K}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{i'\in[K]}\left[\frac{\mathbb{1}(i = i') q_t(i')}{q_t(i)} \sum_{j\in[K]} P_t(i', j)\right]\right]\right) \\
&= \frac{1}{K}\sum_{j\in[K]} P_t(i, j) \\
&= s_t(i).
\end{aligned}$$

In the above, the fourth equality is due to Lemma 1. Moreover, the fifth equality and the seventh equality use that $x_t, y_t \sim q_t$ iid with replacement. □

C. Proof of Lemma 3

Lemma 3. *For all $t \in [T]$ and $m \geq 2$, $g(m, o_t, x_t) \leq \frac{3m+1}{2m+2}$.*

Proof. **Case 1: m is even.**

$$g(m, o_t, x_t) \leq \frac{1 - \frac{(m-2)}{4(m-1)}}{\frac{m}{2(m-1)}}$$

$$= \frac{\left(\frac{4(m-1)-(m-2)}{4(m-1)}\right)}{\frac{m}{2(m-1)}}$$
$$= \left(\frac{(3m-2)}{4(m-1)}\right)\left(\frac{2(m-1)}{m}\right)$$
$$= \frac{(3m-2)}{2m}$$

Case 2: m is odd.

$$g(m, o_t, x_t) \leq \frac{1-\frac{(m-1)}{4m}}{\frac{m+1}{2m}}$$
$$= \left(\frac{4m-m+1}{4m}\right)\left(\frac{2m}{m+1}\right)$$
$$= \frac{3m+1}{2(m+1)}$$

For $m \geq 2$,

$$\frac{(3m-2)}{2m} < \frac{3m+1}{2(m+1)}.$$

□

D. Proof of Lemma 4

Lemma 4. *Let $\gamma \geq \sqrt{3\eta K/2}$. Then, for any $t \in [T]$, $i \in [K]$ and $\eta > 0$, it holds that $\eta \hat{s}_t(i) \in [0, 1]$.*

Proof. From the definition of q_t given in (3), it can be seen that, for all $t \in [T]$ and $i \in [K]$,

$$q_t(i) \geq \frac{\gamma}{K}.$$

Using the above along with the definition of $\hat{s}_t$ given in Eq. (2) and Lemma 3, it can be seen that, for all $t \in [T]$ and $i \in [K]$,

$$\hat{s}_t(i) \leq \frac{(3/2)}{(K)\,(\gamma/K)^2} = \frac{3K}{2\gamma^2}.$$

Then, using $\gamma \geq \sqrt{3\eta K/2}$ and the above inequality,

$$\eta \hat{s}_t(i) \leq \frac{3\eta K}{2\gamma^2} = 1.$$

Furthermore, $0 \leq \eta \hat{s}_t(i)$ also holds using the definition of $\hat{s}_t(i)$ given in Eq. (2), Lemma 3 and $\eta > 0$.

□

E. Proof of Lemma 5

Lemma 5. *For all $t \in [T]$, it holds that $\mathbb{E}_{\mathcal{H}_t}\left[q_t^\top\, \hat{s}_t\right] = \mathbb{E}_{\mathcal{H}_{t-1}}[\mathbb{E}_{i\sim q_t}\left[s_t(i) \mid \mathcal{H}_{t-1}\right]]$.*

Proof.

$$\mathbb{E}_{\mathcal{H}_t}\left[q_t^\top\, \hat{s}_t\right] = \mathbb{E}_{\mathcal{H}_t}\left[\sum_{i=1}^{K} q_t(i)\, \hat{s}_t(i)\right] = \mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{i=1}^{K} q_t(i)\, \mathbb{E}_{x_t,y_t,o_t}\left[\hat{s}_t(i) \mid \mathcal{H}_{t-1}\right]\right]$$
$$= \mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{i=1}^{K} q_t(i)\, s_t(i)\right] = \mathbb{E}_{\mathcal{H}_{t-1}}[\mathbb{E}_{i\sim q_t}\left[s_t(i) \mid \mathcal{H}_{t-1}\right]].$$

In the above, the third equality follows from the proof of Lemma 2. □

F. Proof of Lemma 6

Lemma 6. *At any time $t \in [T]$, it holds that* $\mathbb{E}\left[\sum_{i=1}^{K} q_t(i)\ \hat{s}_t(i)^2\right] \leq \frac{(3m+1)^2}{4(m+1)^2}\ \frac{K}{\gamma}$.

Proof.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} q_t(i)\ \hat{s}_t(i)^2\right] \\
&= \mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{i=1}^{K} q_t(i)\ \mathbb{E}_{(x_t,y_t,a_t)}\left[\frac{\mathbb{1}(i=x_t)}{K\ q_t(i)} \sum_{j\in[K]} \frac{\mathbb{1}(j=y_t)\, g(m,o_t,x_t)}{q_t(j)}\ \Bigg|\ \mathcal{H}_{t-1}\right]^2\right] \\
&= \frac{1}{K^2}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{i=1}^{K} \frac{q_t(i)}{q_t(i)^2}\mathbb{E}_{(x_t,y_t)}\left[\sum_{j\in[K]} \frac{\mathbb{1}(i=x_t)\,\mathbb{1}(j=y_t)\,\mathbb{E}_{o_t}[g^2(m,o_t,x_t)\mid x_t,y_t]}{q_t(j)^2}\ \Bigg|\ \mathcal{H}_{t-1}\right]\right]\right) \\
&\leq \frac{(3m+1)^2}{4(m+1)^2K^2}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{i=1}^{K} \frac{1}{q_t(i)}\left[\sum_{j\in[K]} \frac{\mathbb{E}_{x_t}[\mathbb{1}(i=x_t)]\,\mathbb{E}_{y_t}[\mathbb{1}(j=y_t)]}{q_t(j)^2}\ \Bigg|\ \mathcal{H}_{t-1}\right]\right]\right) \qquad \text{(using Lemma 3)} \\
&= \frac{(3m+1)^2}{4(m+1)^2K^2}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{i=1}^{K} \frac{1}{q_t(i)}\left[\sum_{j\in[K]} \frac{q_t(i)\ q_t(j)}{q_t(j)^2}\right]\right]\right) \\
&= \frac{(3m+1)^2}{4(m+1)^2K^2}\left(\mathbb{E}_{\mathcal{H}_{t-1}}\left[K \sum_{j\in[K]} \frac{1}{q_t(j)}\right]\right) \\
&\leq \frac{(3m+1)^2}{4(m+1)^2K}\left(\sum_{j\in[K]} \frac{1}{\gamma/K}\right) \qquad (\because \forall i' \in [K] \text{ and } \forall t \in [T], q_t(i') \geq \gamma/K \text{ using Eq. (3)}) \\
&= \frac{(3m+1)^2}{4(m+1)^2K}\left(\frac{K}{\gamma/K}\right) \\
&= \frac{(3m+1)^2}{4(m+1)^2}\ \frac{K}{\gamma}.
\end{aligned}
$$

□

G. Proof of Lemma 7

Lemma 7. *For any $i \in [K]$, $j \in [m]$ and $t \in [T]$,*

$$\mathbb{P}\Big(\mathcal{A}_t(j) = i\Big) = q_t(i).$$

Proof.

$$
\begin{aligned}
\mathbb{P}\Big(\mathcal{A}_t(j) = i\Big) &= \mathbb{P}(\mathcal{A}_t(j) = x_t)\,\mathbb{P}(x_t = i) + \mathbb{P}(\mathcal{A}_t(j) = y_t)\,\mathbb{P}(y_t = i) \\
&= \mathbb{P}(\mathcal{A}_t(j) = x_t)\, q_t(i) + \mathbb{P}(\mathcal{A}_t(j) = y_t)\, q_t(i) \\
&= q_t(i).
\end{aligned}
$$

□

H. Proof of Lemma 8

Lemma 8. *Using $\mathfrak{A}_{DB}$ given in Algorithm 3,*

$$\mathbb{E}[R_T(\mathfrak{A}_{DB})] = R_T(\mathfrak{A}_{MB}),$$

for any arbitrary sequence of preference matrices $P_1, P_2, \ldots, P_T$.

Proof. Let $r_t(\mathfrak{A}_{DB})$ be the instantaneous regret of $\mathfrak{A}_{DB}$ at round t. Correspondingly, let $r_t(\mathfrak{A}_{MB})$ be the instantaneous regret of $\mathfrak{A}_{MB}$ at round t.

$$
\begin{aligned}
&\mathbb{E}_{i_t, j_t \overset{Unif}{\sim} [m], i_t \neq j_t}[r_t(\mathfrak{A}_{DB})] \\
&= b_t(i^*) - \frac{1}{2}\left[\mathbb{E}_{i_t, j_t \overset{Unif}{\sim} [m], i_t \neq j_t}\left[b_t\Big(\mathcal{A}_t(i_t)\Big) + b_t\Big(\mathcal{A}_t(j_t)\Big)\right]\right] \\
&= b_t(i^*) - \frac{1}{2}\left[\sum_{i=1}^{m}\left(\frac{b_t\Big(\mathcal{A}_t(i)\Big)}{m} + \frac{\sum_{j=1, j\neq i}^{m} b_t\Big(\mathcal{A}_t(j)\Big)}{m(m-1)}\right)\right] \\
&= b_t(i^*) - \frac{1}{2m}\left[\sum_{i=1}^{m}\left(b_t\Big(\mathcal{A}_t(i)\Big) + \frac{\sum_{j=1}^{m} b_t\Big(\mathcal{A}_t(j)\Big) - b_t\Big(\mathcal{A}_t(i)\Big)}{(m-1)}\right)\right] \\
&= b_t(i^*) - \frac{1}{2m}\left[\sum_{i=1}^{m}\left(b_t\Big(\mathcal{A}_t(i)\Big) - \frac{b_t\Big(\mathcal{A}_t(i)\Big)}{(m-1)}\right) + \frac{m\sum_{j=1}^{m} b_t\Big(\mathcal{A}_t(j)\Big)}{(m-1)}\right] \\
&= b_t(i^*) - \frac{1}{2m}\left[\frac{(m-2)}{(m-1)}\sum_{i=1}^{m} b_t\Big(\mathcal{A}_t(i)\Big) + \frac{m\sum_{i=1}^{m} b_t\Big(\mathcal{A}_t(i)\Big)}{(m-1)}\right] \\
&= b_t(i^*) - \frac{1}{2m}\left[\frac{\sum_{i=1}^{m} b_t\Big(\mathcal{A}_t(i)\Big)}{(m-1)}(m-2+m)\right] \\
&= b_t(i^*) - \frac{1}{2m}\left[\frac{\sum_{i=1}^{m} b_t\Big(\mathcal{A}_t(i)\Big)}{(m-1)}(2m-2)\right] \\
&= b_t(i^*) - \frac{1}{m}\left[\sum_{i=1}^{m} b_t\Big(\mathcal{A}_t(i)\Big)\right] \\
&= r_t(\mathfrak{A}_{MB}).
\end{aligned}
$$

In the above, the second equality uses that i_t, j_t are sampled uniformly at random from $[m]$ without replacement.

Then,

$$
\mathbb{E}[R_T(\mathfrak{A}_{DB})] = \sum_{t=1}^{T}\mathbb{E}[r_t(\mathfrak{A}_{DB})] = \sum_{t=1}^{T} r_t(\mathfrak{A}_{MB}) = R_T(\mathfrak{A}_{MB}).
$$

□