Adversarial Dueling Bandits

Aadirupa Saha 1  Tomer Koren 2  Yishay Mansour 2

Abstract
We introduce the problem of regret minimization in Adversarial Dueling Bandits. As in classic Dueling Bandits, the learner has to repeatedly choose a pair of items and observe only a relative binary ‘win-loss’ feedback for this pair, but here this feedback is generated from an arbitrary preference matrix, possibly chosen adversarially. Our main result is an algorithm whose $T$-round regret compared to the Borda-winner from a set of $K$ items is $\tilde{O}(K^{1/3}T^{2/3})$, as well as a matching lower bound. We also prove a similar high probability regret bound. We further consider a simpler fixed-gap adversarial setup, which bridges between two extreme preference feedback models for dueling bandits: stationary preferences and an arbitrary sequence of preferences. For the fixed-gap adversarial setup we give an $\tilde{O}(K/\Delta^2 \log T)$ regret algorithm, where $\Delta$ is the gap in Borda scores between the best item and all other items, and show a lower bound of $\Omega(K/\Delta^2)$ indicating that our dependence on the main problem parameters $K$ and $\Delta$ is tight (up to logarithmic factors). Finally, we corroborate the theoretical results with empirical evaluations.

1. Introduction
Dueling Bandits is an online decision making framework similar to the well known (stochastic) multi-armed bandit (MAB) problem (Auer et al., 2002a; Slivkins, 2019), that has gained widespread attention in the machine learning community over the past decade (Yue et al., 2012; Zoghi et al., 2014b; 2015a). In Dueling Bandits, a learner repeatedly selects a pair of items to be compared to each other in a “duel,” and consequently observe a binary stochastic preference feedback, which can be interpreted as the winning item in this duel. The goal of the learner is to minimize the regret with respect to the best item in hindsight, according to a certain score function.

Numerous real-world applications are naturally modelled as dueling bandit problems, including movie recommendations, tournament ranking, search engine optimization, retail management, etc. (see also Busa-Fekete & Hüllemeier, 2014; Yue & Joachims, 2009). Indeed, in many of these scenarios, users with whom the algorithm interacts with find it more natural to provide binary feedback by comparing two alternatives rather than giving an absolute score for a single alternative. Over the years, several algorithms have been proposed for addressing dueling bandit problems (Ailon et al., 2014; Zoghi et al., 2014a; Komiyama et al., 2015; Zoghi et al., 2014b) and there has been some work on extending the pairwise preference to more general subset-wise preferences (Sui et al., 2017; Brost et al., 2016; Saha & Gopal, 2018; 2019; Ren et al., 2018).

While almost all of the existing literature on dueling bandits focus on stochastic stationary preferences, in reality preferences might vary significantly and unpredictably over time. For example, in video recommendation systems, user preferences may evolve according to daily and hourly viewing trends; in web-search optimization, relevance of various websites may vary rather unpredictably. In other words, many of the real-world applications of dueling bandits actually deviate from the stochastic feedback model, and would more faithfully be modelled in a robust worse-case (adversarial) model that alleviates the strong stochastic assumption and allows for an arbitrary sequence of preferences over time. For similar reasons, the MAB problem, and more generally, online learning, are frequently studied in a non-stochastic adversarial setup (Lattimore & Szepesvári, 2018; Bubeck & Cesa-Bianchi, 2012; Cesa-Bianchi & Lugosi, 2006; Seldin & Slivkins, 2014; Seldin & Lugosi, 2017; Neu, 2015; Bubeck & Slivkins, 2012).

Surprisingly, however, a non-stochastic version of dueling bandits has not been well studied (with the only exception being Gajane et al., 2015, discussed below). The first challenge in eschewing stationarity in dueling bandits lies in the performance benchmark compared to which regret is defined. Indeed, most works on stochastic dueling bandits rely on the existence of a Condorcet winner: an item being preferred (and often by a gap) when compared with
any other item. In an adversarial environment, however, assuming a Condorcet winner makes little sense as it would constrain the adversary to consistently prefer a certain item at all rounds, ultimately defeating the purpose of a non-stationary model in the first place. Another main challenge is the inherent disconnect between the feedback observed by the learner and her payoff at any given round; while this disparity already exists in stochastic models of dueling bandits, in an adversarial setup it becomes more tricky to attribute preferential information to the instantaneous quality of items.

### 1.1. Our contributions

In this paper, we introduce and study an adversarial version of dueling bandits. To mitigate the issues associated with Condorcet winner assumptions, and following recent literature on dueling bandits (e.g., Jamieson et al., 2015; Ramamohan et al., 2016; Falahatgar et al., 2017), we focus on the so-called Borda score criterion. The Borda score of an item is the probability that it is preferred over another item chosen uniformly at random. A Borda winner (i.e., an item with the highest Borda score) always exists for any preference matrix, and more generally, this notion naturally extends to any arbitrary sequence of preference matrices. However, the second challenge from above remains: the Borda score of an item is not directly related in nature to the preferential feedback observed for this item on rounds where it is chosen for a duel.

The main contributions of this paper can be summarized as follows:

- We introduce and formalize an adversarial model for K-armed dueling bandits with standard binary “win-loss” preferential feedback (and where regret is measured with respect to Borda scores). To the best of our knowledge, we are first to study such a setup.

- In the general adversarial model, where the sequence of preference matrices is allowed to be entirely arbitrary, we present an algorithm with expected regret bounded by \( \tilde{O}(K^{1/3}T^{2/3}) \).\(^1\) We further demonstrate how to modify our algorithm so as to guarantee a similar bound with high probability. We also give a lower bound of \( \Omega(K^{1/3}T^{2/3}) \), showing our algorithm is nearly optimal.

- We consider a more specialized fixed-gap adversarial model, that bridges between the two extreme preference feedback models for dueling bandits: the well-studied stationary stochastic preferences, and fully adversarial preferences. Here, we assume that there is a fixed item whose average Borda score at any point in time exceeds that of any other item by at least \( \Delta > 0 \), where \( \Delta \) is a gap parameter unknown to the learner. (Other than constraining this fixed gap, the preference assignment may change adversarially.) We present an algorithm that achieves regret \( \tilde{O}(K/\Delta^2) \), and show that it is near-optimal by proving a regret lower bound of \( \Omega(K/\Delta^2) \).

- Finally, we corroborate our theoretical findings with an empirical evaluation.

Our results thus reveal an inherent gap in the achievable regret between dueling bandits and standard multi-armed bandits: in the adversarial model, the optimal regret in dueling bandits grows like \( \Theta(T^{2/3}) \) whereas in standard bandits \( \Theta(\sqrt{T}) \)-type bounds are possible; likewise, in the fixed-gap model the optimal regret for dueling bandits is \( \Theta(K/\Delta^2) \), versus the well-known \( \Theta(K/\Delta) \) regret performance for standard fixed-gap (stochastic) bandits.

The reason for this substantial gap, as we explain in more detail in our discussion of lower bounds, is the following. For gaining information about the identity of the best item in terms of Borda scores, the learner might be forced to choose items the scores of which are already (or even initially) known to be suboptimal, and for which she would unavoidably suffer constant regret. Indeed, the Borda score of an item inherently depends on its relative performance compared to all other items, and it may be that the identity of the Borda winner is determined solely by its comparison to poorly-performing items.

### 1.2. Related work

Dueling bandits were investigated extensively in the stochastic setting. The most frequently used performance objective in this literature is the regret compared to the Condorcet Winner (Yue et al., 2012; Zoghi et al., 2014a; 2015b; Komiyama et al., 2015; Yue & Joachims, 2011). However, there are quite a few well-established shortcomings of this objective; most importantly, the Condorcet winner often fails to exist even for a fixed preference matrix. (See Jamieson et al., 2015 for more detailed discussion.) In absence of Condorcet winners, there are other preference notions studied in the literature, most notably the Borda Winner (Busa-Fekete & Hüllermeier, 2014; Jamieson et al., 2015; Ramamohan et al., 2016; Falahatgar et al., 2017), Copeland Winner (Zoghi et al., 2015a; Komiyama et al., 2016; Wu & Liu, 2016),\(^2\) and Von-Neumann Winner (Dudík et al., 2015; Balsubramani et al., 2016). In this work, we focus on the Borda Winner, which appears to be the most common alternative.

The only previous treatment of dueling bandits in an ad-

\(^1\)Throughout, the notation \( \tilde{O}(\cdot) \) hides logarithmic factors.

\(^2\)It is worth noting that for the Copeland winner to be at all learnable, a gap assumption is required.
versarial setting is Gajane et al. (2015), which considers utility-based preferences and thereby imposes a complete ordering of the items in each time step rather than a general preference matrix. Further, their feedback model includes not only the winning item but also a transfer function which is the difference in utilities between the compared items, thus being more similar to standard MAB and largely departs from the original motivation of dueling bandit. For the identity transfer function, they show in their adversarial utility-based dueling bandit model a tight regret bound of $\Theta(\sqrt{KT})$. In contrast, we show for the adversarial dueling bandit model a tight regret bound of $\tilde{\Theta}(K^{1/3}T^{2/3})$. This shows that when one does not have a direct access to a transfer function and is faced with arbitrary preferences, the regret scales substantially different, i.e., $\Theta(T^{2/3})$ versus $\tilde{\Theta}(T^{1/2})$.

Jamieson et al. (2015) show an instance dependent $\tilde{\Omega}(K/\Delta^2)$ sample complexity lower bound for the Borda-winner identification problem in stochastic dueling bandits. In contrast, our lower bound which is similar in magnitude, applies to the regret which is always smaller (and often strictly smaller) than the sample complexity.

2. Problem Setup

We consider an online decision task over a finite set of items $[K] := \{1, 2, \ldots, K\}$ which spans over $T$ decision rounds. Initially, and obliviously, the environment fixes a sequence of $T$ preference matrices $P_1, \ldots, P_T$, where each $P_t \in [0, 1]^{K \times K}$ satisfies $P_t(i, j) = 1 - P_t(j, i)$, and $P_t(i, i) = \frac{1}{2}$ for all $i, j \in [K]$. The value of $P_t(i, j)$ is interpreted as the probability that item $i$ wins when matched against item $j$ at time $t$. Then, at each round $t$ the learner selects, possibly at random, two items $x_t, y_t \in [K]$ and a feedback $o_t \sim Ber(P_t(x_t, y_t))$ for the selected pair is revealed, where $Ber(p)$ denotes a Bernoulli random variable with parameter $p$. Here, feedback of $o_t = 1$ implies that item $x_t$ wins the duel, while $o_t = 0$ corresponds to $y_t$ being the winner.

The Borda score of item $i \in [K]$ with respect to the preference matrix $P_t$ at time $t$ is defined as

$$b_t(i) := \frac{1}{K - 1} \sum_{j \neq i} P_t(i, j),$$

and

$$i^* := \arg \max_{i \in [K]} \sum_{t=1}^T b_t(i),$$

i.e., $i^*$ is the item with the highest cumulative Borda score at time $T$. The learner’s $T$-round regret $R_T$ is then defined as follows:

$$R_T := \sum_{t=1}^T r_t,$$

where

$$r_t := b_t(i^*) - \frac{1}{2}(b_t(x_t) + b_t(y_t)).$$

We will consider two settings of preference assignments. In the general adversarial setting, $P_1, \ldots, P_T$ is an arbitrary sequence of preference matrices. In the fixed-gap setting, preferences are set so that there is an item $i^* \in [K]$ for which, at all rounds $t \in [T]$, we have $b_t(i^*) \geq b_t(j) + \Delta$ for any other $j \neq i^*$, where $b_t(j) := \frac{1}{T} \sum_{\tau=1}^T b_\tau(j)$ is the average Borda score of item $j \in [K]$ up to time $t$.

3. General Adversarial Dueling Bandits

We first consider the general adversarial setup for an arbitrary sequence of preference matrices. We give an algorithm, called Dueling-EXP3 (D-EXP3), which has an expected regret of $O((K \log K)^{1/3}T^{2/3})$. We also show how a simple modification of the D-EXP3 algorithm guarantees regret $O(K^{1/3}T^{2/3} \sqrt{\log(K/\delta)})$ with probability at least $1 - \delta$.

3.1. The Dueling-EXP3 Algorithm

Our algorithm, detailed in Algorithm 1, is motivated from the classical EXP3 algorithm for adversarial MAB (Auer et al., 2002a), and relies on constructing unbiased estimates for scores of individual items at all rounds. However, in the dueling setup one has to establish such estimates using only binary preference feedback corresponding to a choice of a pair of items. Technically, the algorithm will estimate a shifted version of the Borda score, defined as follows.

Definition 1. The shifted Borda score of item $i \in [K]$ at time $t \in [T]$ is $s_t(i) := \frac{1}{K} \sum_{j \in [K]} P_t(i, j)$. The shifted regret is then defined as $R^*_T := \sum_{t=1}^T [s_t(i^*) - \frac{1}{2}(s_t(x_t) + s_t(y_t))].$

Since all scored are “shifted” by the same value, this will not have any impact and the differences between Borda scores will be maintained (albeit multiplied by $\frac{K}{K-1}$). In particular, the best item is unchanged, i.e., $i^* = \arg \max_{i \in [K]} \sum_{t=1}^T b_t(i) = \arg \max_{i \in [K]} \sum_{t=1}^T s_t(i)$, and for any $K \geq 2$ and $T > 0$ we have $R_T = \frac{K}{K-1} R^*_T$.

At every round $t$, D-EXP3 maintains a weight distribution $q_t \in \Delta[K]$ ($\Delta[K]$ is the $K$-simplex), and compute a score estimate $\hat{s}_t(i)$ for each item $i$, being an unbiased estimate of $s_t(i)$ (Lemma 4). Thus, the cumulative estimated score $\sum_{t=1}^T \hat{s}_t(i)$ can be seen as the estimated cumulative reward of item $i$ at round $t$, and hence $q_{t+1}$ is simply updated running an exponential weight update on these estimated cumulative scores along with a $\gamma$-uniform exploration.

We now state the expected regret guarantee we establish for
Algorithm 1 Dueling-EXP3 (D-EXP3)

1: **Input:** Item set indexed by $[K]$, learning rate $\eta > 0$, parameters $\gamma \in (0, 1)$
2: **Initialize:** Initial probability distribution $q_1(i) = 1/K, \forall i \in [K]$
3: for $t = 1, \ldots, T$ do
4: Sample $x_t, y_t \sim q_t$ i.i.d. (with replacement)
5: Receive preference $o_t(x_t, y_t) \sim \text{Ber}(P_t(x_t, y_t))$
6: Estimate scores, for all $i \in [K]$: 
   $$\tilde{s}_t(i) = \frac{1(x_t = i)}{Kq_t(i)} \sum_{j \in [K]} 1(y_t = j) q_t(x_t, y_t)$$
7: Update, for all $i \in [K]$: 
   $$\tilde{q}_{t+1}(i) = \frac{\exp(\eta \sum_{t=1}^t \tilde{s}_r(i))}{\sum_{j=1}^K \exp(\eta \sum_{t=1}^t \tilde{s}_r(j))}$$
   $$q_{t+1}(i) = (1 - \gamma)\tilde{q}_{t+1}(i) + \frac{\gamma}{K}$$
8: **end for**

Algorithm 1.

**Theorem 2.** Let $\eta = ((\log K)/(T\sqrt{K}))^{2/3}$ and $\gamma = \sqrt{\eta K}$. For any $T$, the expected regret of Algorithm 1 satisfies $E[R_T] \leq 6(K \log K)^{1/3}T^{2/3}$.

The proof of the expected regret bound crucially relies on the the following key lemmas regarding the estimates for the shifted Borda scores. We bound their magnitude, show that they are unbiased estimates, bound their instantaneous regret, and bound their second moment.

We first bound the magnitude of the estimates $\tilde{s}_t(i)$, using the fact that $q_t(j) \geq \gamma/K$.

**Lemma 3.** For all $t \in [T]$ and $i \in [K]$ it holds that $\tilde{s}_t(i) \leq K/\gamma^2$.

Next, we show that $\tilde{s}_t(i)$ is an unbiased estimate of the shifted Borda score $s_t(i)$.

**Lemma 4.** For all $t \in [T]$ and $i \in [K]$ it holds that $E[\tilde{s}_t(i)] = s_t(i)$.

Let $\mathcal{H}_{t-1} := (q_1, P_1, (x_1, y_1), o_1, \ldots, q_t, P_t)$ denotes the history up to time $t$. We compute the expected instantaneous regret at time $t$ as a function of the true shifted Borda scores at time $t$.

**Lemma 5.** For all $t \in [T]$ it holds that $E_{\mathcal{H}_t}[\tilde{q}_t \tilde{s}_t] = E_{\mathcal{H}_{t-1}}[E_{x \sim q_t}[s_t(x) | \mathcal{H}_{t-1}]]$.

Finally, we bound the second moment of our estimates.

**Lemma 6.** For all $t \in [T]$ it holds that $E[\sum_{i=1}^K q_t(i)\tilde{s}_t(i)^2] \leq K/\gamma$.

**Proof overview.** We upper bound $R_T^\ast$, the shifted Borda score regret, and recall that $R_T^\ast = \frac{1}{K}R_T$. Note that $E_{\mathcal{H}_T}[s_t(x_t) + s_t(y_t)] = E_{\mathcal{H}_{t-1}}[E_{x \sim q_t}[2s_t(x) | \mathcal{H}_{t-1}]]$, since $x_t$ and $y_t$ are i.i.d. Further note that we can write

$$E_{\mathcal{H}_T}[R_T^\ast] = E_{\mathcal{H}_T} \left[ \sum_{t=1}^T [s_t(i^*) - \frac{1}{2}(s_t(x_t) + s_t(y_t))] \right] = \max_{k \in [K]} E_{\mathcal{H}_T} \left[ \sum_{t=1}^T s_t(k) - \sum_{t=1}^T [E_{x \sim q_t}[s_t(x) | \mathcal{H}_{t-1}]] \right].$$

Next, as $\eta \tilde{s}_t(i) \leq \eta K/\gamma^2$ (from Lemma 3), for any $\gamma \geq \sqrt{\eta K}$ and $\eta > 0$ we have $\eta \tilde{s}_t(i) \in [0, 1]$. From the regret guarantee of standard Multiplicative Weights algorithm (Bubeck & Cesa-Bianchi, 2012) over the completely observed fixed sequence of reward vectors $\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_T$ we have for any $k \in [K]$: 

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

Note that $\hat{q}_t := (q_t - \frac{\eta}{K})/(1 - \gamma)$. Let $i^* = \arg\max_{k \in [K]} \sum_{t=1}^T s_t(k) = \arg\max_{k \in [K]} \sum_{t=1}^T b_t(k)$. Taking expectation on both sides of the above inequality for $k = i^*$, we get:

$$(1 - \gamma) \sum_{t=1}^T E_{\mathcal{H}_T}[\tilde{s}_t(i^*)] - \sum_{t=1}^T E_{\mathcal{H}_T}[\tilde{q}_t^\top \tilde{s}_t] \leq \frac{\log K}{\eta} + E_{\mathcal{H}_T} \left[ \eta \sum_{t=1}^T \sum_{i=1}^K q_t(i)\tilde{s}_t(i)^2 \right],$$

which by applying Lemma 4, Lemma 5 and Lemma 6 and the fact that $s_t(k^*) \leq 1$, $\gamma = \sqrt{\eta K}$, we have

$$E_{\mathcal{H}_T}[R_T^\ast] \leq 2T \sqrt{\eta K} + \frac{\log K}{\eta} \leq 3(K \log K)^{1/3}T^{2/3},$$

where the second inequality follows by optimizing over $\eta$. The theorem follows since $R_T = \frac{1}{K}R_T^\ast \leq 2R_T^\ast$. A complete proof is given in the supplementary material.

### 3.2. High Probability Regret Analysis

We can show that a slightly modified version of Dueling-EXP3 can lead to a high probability regret bound for
the same setup. (This is inspired by the EXP3.P algorithm (Auer et al., 2002b).) The modified algorithm runs almost identically to that of Algorithm 1, except we now use a different score estimate \( s'_t(i) \) in place of \( \tilde{s}_t(i) \), where \( s'_t(i) = \tilde{s}_t(i) + \beta/q_t(i) \), where \( \beta \in (0, 1) \) is a tuning parameter. The items weights \( q_t \in \Delta[K] \) are now similarly updated using an exponential weight update on these modified score estimates along with an \( \gamma \)-uniform exploration. The complete algorithm is described in Algorithm 2.

**Algorithm 2 Dueling-EXP3 (High Probability)**

1. **Input:** Item set: \([K]\), learning rate \( \eta > 0 \), parameters \( \beta \in (0, 1), \gamma \in (0, 1) \)
2. **Initialize:** Initial distribution \( q_1(i) = \frac{1}{K}, \forall i \in [K] \)
3. **while** \( t = 1, 2, \ldots \) **do**
   4. Sample \( x_t, y_t \sim q_t \) (i.i.d., with replacement)
   5. Receive preference \( \alpha_t(x_t, y_t) \sim \text{Ber}(P_t(x_t, y_t)) \)
   6. Compute \( \forall i \in [K]: \)
      \[
      s'_t(i) = \frac{1}{Kq_t(i)} \sum_{j \in [K]} 1(y_t = j)\alpha_t(x_t, y_t) + \beta/q_t(i) \]
   7. Update \( \forall i \in [K]: \)
      \[
      \tilde{q}_{t+1}(i) = \frac{\exp(\eta \sum_{j=1}^{t} s'_t(j))}{\sum_{j=1}^{K} \exp(\eta \sum_{j=1}^{t} s'_t(j))} q_t(i) \\
      q_{t+1}(i) = (1 - \gamma)\tilde{q}_{t+1}(i) + \gamma/K \]
8. **end while**

We now prove a high probability regret bound for Algorithm 2:

**Theorem 7.** Given any \( T \) and \( \delta > 0 \), there exists a setting of \( \gamma, \beta \) and \( \eta \), such that with probability at least \( 1 - \delta \), the regret of the modified D-EXP3 algorithm is \( R_T = \tilde{O}(K^{1/3}T^{2/3}) \).

The proof builds on the following steps. Similarly to our estimates \( \tilde{s}_t(i) \) above, we can show the following properties.

**Lemma 8.** For any item \( i \) and round \( t \in [T] \), we have \( s'_t(i) \leq K/\gamma^2 + K\beta/\gamma \).

**Lemma 9.** For any item \( i \) and round \( t \in [T] \), it holds that \( \mathbb{E}[s'_t(i) | H_{t-1}] = s_t(i) + \beta/q_t(i) \).

However, unlike \( \tilde{s}_t(i) \), the adjusted score estimates \( s'_t(i) \) are no longer unbiased for the true scores \( s_t(i) \), and are larger in expectation by \( \beta \). Nevertheless, this does not hurt the regret analysis as its key element lies in showing that for any item \( i \in [K] \), the cumulative estimated scores are not too far from the accumulated true scores. Precisely, the next lemma ensures a high confidence upper bound on the cumulative scores \( \sum_{t=1}^{T} s_t(i) \) and thus we can upper bound the learners performance in terms of estimated scores \( s'_t \) (instead of \( s_t \)).

**Lemma 10.** For any \( i \in [K], \beta \in (0, 1) \) and \( \gamma, \beta \in (0, 1) \), with probability at least \( 1 - \delta \), we have
\[
\sum_{t=1}^{T} s'_t(i) \geq \sum_{t=1}^{T} s_t(i) - \frac{1}{\gamma\beta} \log \frac{1}{\delta}.
\]

Incorporating this idea, the rest of the analysis closely follows that of Theorem 2. See complete proof in the supplementary material.

### 4. Fixed-Gap Adversarial Dueling Bandits

In this section we study an adversarial setting with a fixed-gap of \( \Delta > 0 \), and give an algorithm with regret \( O((K \log(KT))/\Delta^2) \). In this case, our algorithm is based on using confidence intervals of the estimated average Borda-scores. The algorithm has two phases. In the first phase, it samples uniformly at random two different items, and observes the outcome of their duel; in the second phase, it has a specific single item \( i \), which it uses in all rounds (for both items). The algorithm moves to its second phase when it detects an item \( i \) whose lower confidence bound (LCB) is larger than the upper confidence bound (UCB) of any other item \( j \). The complete description is given in Algorithm 3.

Because of the non-stationary nature of the item preferences, and unlike classical action-elimination algorithms (Auer, 2000; Even-Dar et al., 2006), we still need to maintain an unbiased estimate of the Borda-score for every item at every round. (In contrast, in the stochastic dueling bandit problem (Zoghi et al., 2014a), for any fixed item \( i \in [K] \), the unbiased estimate of its Borda score at round \( t \) is also an unbiased estimate for any other round \( s \neq t \); this simplifying condition does not hold in our fixed-gap adversarial model.) Towards this, we maintain an estimate of the Borda score of any item \( i \) in \([K]\) at any round \( t \) as \( b_t(i) := K1(x_t = i)\alpha_t(x_t, y_t) \), and show that it is an unbiased estimator.

**Lemma 11.** At any round \( t \), we have \( \mathbb{E}_{H_t}[\hat{b}_t(i)] = b_t(i) \) for all \( i \in [K] \).

Thus, an unbiased estimate for the \( t \)-step average Borda score \( \hat{b}_t(i) \), is \( b_t(i) := \frac{1}{t} \sum_{\tau=1}^{t} b_t(i) \). We further maintain confidence intervals \([LCB(i; t), UCB(i; t)]\) around each \( \hat{b}_t(i) \), within which the means \( b_t(i) \) lie with high probability.

**Lemma 12.** With probability \( \geq 1 - \delta \), we have \( \hat{b}_t(i) \in [LCB(i; t), UCB(i; t)] \) for all \( i \in [K] \) and \( t \in [T] \).

The proof uses Bernstein’s inequality to show that the estimates \( \hat{b}_t(i) \) are concentrated around their means \( b_t(i) \), within the respective confidence intervals. Assuming these confidence bounds hold, as soon as we find an item \( i \in [K] \) such that \( LCB(i; t) > UCB(j; t) \) for any other item \( j \neq i \),
we are guaranteed that \( \hat{i} \) is the best item (in hindsight), i.e., \( \hat{i} = i^* \). In the remaining rounds, \( t + 1, \ldots, T \), we play only item \( i \) (for both items) and suffer no regret. This results with the algorithm detailed in Algorithm 3

**Theorem 13.** Given any \( \delta > 0 \), with probability at least \( 1 - \delta \), the regret of Algorithm 3 (with parameter \( \delta \)) is upper bounded by \( 64(K/\Delta^2) \log(2KT/\delta) \).

We remark that unlike most MAB algorithms, we do not gain by incremental elimination. The reason is that we need to sample a second random item, \( y_t \), which would have an expected Borda score which equals the average Borda score. This random item implies a constant regret per round until we identify \( i \). After we identify \( i \), with high probability, we do not incur any regret.

**Algorithm 3 Borda-Confidence-Bound (BCB)**

1. **Input:** item set indexed by \([K]\), confidence \( \delta > 0 \)
2. **for** \( t = 1, \ldots, T \) **do**
   3. Select \( x_t, y_t \in [K] \), \( x_t \neq y_t \) uniformly at random
   4. Receive preference \( a_t(x_t, y_t) \sim \text{Ber}(P_t(x_t, y_t)) \)
   5. Estimate: \( \hat{b}_t(i) = K a_t(x_t, y_t) 1(x_t = i), \forall i \in [K] \)
   6. Compute: \( \bar{b}_t(i) = \frac{1}{t} \sum_{\tau=1}^{t} \hat{b}_\tau(i), \forall i \in [K] \)
   7. Compute:
      \[
      LCB(i; t) = \bar{b}_t(i) - 2 \sqrt{\frac{K}{t} \log \frac{2KT}{\delta}},
      \]
      \[
      UCB(i; t) = \bar{b}_t(i) + 2 \sqrt{\frac{K}{t} \log \frac{2KT}{\delta}}.
      \]
8. **if** \( \exists \hat{i} \in [K] \) s.t. \( LCB(\hat{i}; t) > UCB(j; t) \ \forall j \neq \hat{i} \), **then** break
9. **end for**
10. Play \( (\hat{i}, \hat{y}) \) for rest of the rounds \( t + 1, \ldots, T \).

**5. Lower Bounds**

This section derives lower bounds for the adversarial dueling bandit settings. Theorem 15 and Theorem 16 respectively give the regret lower bound for fixed gap and general adversarial setting. We first prove the following key lemma before proceeding to the individual lower bounds:

**Lemma 14. For the problem of Adversarial Dueling Bandits with Borda Score objective, for any learning algorithm \( A \) and any \( \epsilon \in (0, 0.1) \), there exists a problem instance (sequence of preference matrices \( P_1, P_2, \ldots, P_T \)) such that the expected regret incurred by \( A \) on that instance is at least \( \Omega(\min(\epsilon T, K/\epsilon^2)) \), for any \( K \geq 4 \).**

**Proof outline.** The proof of the lemma has the following outline. We initially construct a stochastic preference matrix \( P_0 \), and later we consider perturbations of it. We start by describing \( P_0 \). We split the items to two equal size subsets \( K_g \) and \( K_b \). For any two items \( i, j \in K_g \), they are equally likely to win or lose in \( P_0 \), i.e., \( P_0(i, j) = 1/2 \). Similarly, for any \( i, j \in K_b \) we have \( P_0(i, j) = 1/2 \). When we pick item \( i \in K_g \) and item \( j \in K_b \), then item \( i \) wins with probability \( 0.9 \), i.e., \( P_0(i, j) = 0.9 \). This implies that the Borda score of any \( i \in K_g \) is \( s(i) = 0.7 \) and for any \( j \in K_b \) it is \( s(j) = 0.3 \). Note that in \( P_0 \) all the items in \( K_g \) have the highest Borda score.

The main idea of the proof is that we will introduce a perturbation that will make one item \( i^* \in K_g \) to have the highest Borda score. Formally, for each \( i \in K_g \) we have a preference matrix \( P_i \). The only difference between \( P_i \) and \( P_0 \) is in the entries of \( i \in K_g \), where for any \( j \in K_b \) we have \( P_i(i, j) = 0.9 + \epsilon \). We select our stochastic preference matrix at random from all the \( P_i \) where \( i \in K_g \), and denote by \( i^* \) the selected index. More explicitly following shows the form of \( P_1 \):

\[
\begin{bmatrix}
0.5 & \ldots & 0.9 + \epsilon & \ldots & 0.9 + \epsilon \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
0.5 & \ldots & 0.9 & \ldots & 0.9 \\
0.5 - \epsilon & 0.1 & 0.5 & \ldots & 0.5 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
0.5 - \epsilon & 0.1 & 0.5 & \ldots & 0.5
\end{bmatrix}
\]

A key observation is that in order to determine the best Borda score item, we need to match items \( i \in K_g \) with items \( j \in K_b \), since the expected outcome of other comparisons is known. However, each time we match an item \( i \in K_g \) with an item \( j \in K_b \), we have a constant regret of about \( 0.2 - O(\epsilon) = \Theta(1) \). We will need to have \( \Omega(|K_g|/\epsilon^2) \) samples to distinguish a bias of \( \epsilon \) in the Borda score of \( i^* \in K_g \) compared to other items \( i \in K_g \). This leads to a regret of \( \Omega(K/\epsilon^2) \). If, with some constant probability, we do not identify the item with the best Borda score, we will have a regret of at least \( \Omega(\epsilon T) \). This follows since any sub-optimal item has regret at least \( \Omega(\epsilon) \) per time step.

We remark that the lower bound holds for \( K = 3 \) with an almost an identical proof. (Technically, our lower bound requires that \( K \) is even, but this is only for ease of presentation.) On the other hand, for \( K = 2 \) the true regret bound scales \( \Theta(1/\Delta) \), since when we match the (only) two items we have a regret of only \( \Delta/2 \). Finally, there is an additional logarithmic dependency on the time horizon, which our lower bound does not capture.

**Lower bound for the fixed-gap setting.** In this case, given any fixed \( \Delta > 0 \), Theorem 15 shows a lower bound of \( \Omega(K/\Delta^2) \). The proof follows from Lemma 14 setting \( \epsilon = \Delta \).

**Theorem 15.** Fix any \( \Delta \in (0, 0.1) \) and \( K \geq 4 \). For the fixed gap setting, for any learning algorithm \( A \), there
exists an instance with fixed gap $\Delta$, such that the expected regret incurred by $A$ on that instance is at least $\Omega(\min(\Delta T, K/\Delta^2))$.

The regret bound in this scales as $K/\Delta^2$ compared to $K/\Delta$ for MAB. The reason is that in order to distinguish between near-optimal items, the learner must compare them to significantly suboptimal items, which leads to the increase in the regret. Essentially, the regret bound is identical to the sample complexity bound in our lower bound instance.

**Lower bound for the general adversarial setup.** In this general case, since $\{P_t\}_{t\in[T]}$ could be any arbitrary sequence, the adversary has the provision to tune $\epsilon$ based on $T$. Precisely, given any $K$ and $T$, the adversary here can set $\epsilon = \Theta(K^{1/3}/T^{1/3})$. For any $T \geq K$ we guarantee that $\epsilon \in (0, 0.1]$ and apply Lemma 14. For $T < K$ we clearly have a lower bound of $\Omega(T)$, since we need to sample each item at least once. Therefore, for this general setup, we derive the following lower bound of $\Omega(K^{1/3}T^{2/3})$.

**Theorem 16.** For the problem of Adversarial Dueling Bandits with Borda Score objective, for any learning algorithm $A$, there exists a problem instance Adv-Borda($K, T$) with $T \geq K$, $K \geq 4$, and sequence of preference matrices $P_1, P_2, \ldots, P_T$, such that the expected regret incurred by $A$ on that Adv-Borda($K, T$) is at least $\Omega(K^{1/3}T^{2/3})$.

Note that the lower bound of $\Omega(T^{2/3})$ steams from the fact that we can essentially cannot mix exploration and exploitation, at least in our lower bound instance. Namely, while we are searching for the best Borda score item, we have a constant regret per time step. If we settle on any sub-optimal item, we get a regret of $\Omega(\epsilon T) = \Omega(T^{2/3})$, due to the selection of $\epsilon$.

**6. Experiments**

In this section we evaluate the empirical evaluation of our proposed algorithm Dueling-EXP3 and compare its performances with the only other existing adversarial dueling bandit algorithm, REX3, although it is known to work only under the restricted class of linear utility based preferences (Gajane et al., 2015; Ailon et al., 2014).

In more detail, we run our experiments with the following setup:

**Algorithms.** (1) Dueling-EXP3: As introduced in Section 3 with parameters tuned according to Theorem 2. (2) REX3: As introduced in Gajane et al. (2015). Note that their suggested optimal tuning parameters, i.e., the uniform exploration rate $\gamma$ as well as the learning rate $\eta$ requires the knowledge of problem dependent parameters $\tau$—the algorithm’s expected loss regret with respect to a random strategy (see Thm. 1 of Gajane et al., 2015), which is unknown to the learner. We used $T$ in place of $\tau$ henceforth. However, other settings of $\tau$ give similar outcomes. (3) Random: A naive baseline that draws any arbitrary duel at each round.

**Performance Measures.** In all cases, we report the cumulative regret of the algorithms averaged over 500 runs.

**Environments: Adversarial preferences.**

We consider $K = 20$ and generate the sequence of adversarial probability matrices as follows:

1. **Switching Borda or SB($t$).** We generate the preference sequence such that the best performing Borda winner changes after every $t$ length epochs by appropriate tweaking of the entries of the current preference matrix at time $t$: Precisely, we manipulated the entries carefully to make sure the new Borda winner is always selected from one of the first 10 arms and different from the latest Borda winner (of the matrix $P_{t-1}$). Towards this, upon swapping the matrix entries if needed, we randomly select a row $i$ from $[10]$ (such that $i \neq$ Borda-winner($P_{t-1}$)), and iteratively increase the row entries $P_t(i, j)$ for all $j \neq i$ in a round robin fashion (up to a threshold of 1) with subsequently resetting $P_t(j, i) = 1 - P_t(i, j)$, until $i$ becomes the new Borda winner of $P_t$.

2. **Random-walk preferences or RW($\nu$).** In the literature of adversarial Multi-armed Bandits, one popular technique to generate adversarial loss sequence is through random walk (Neu & Valko, 2014; Saha et al., 2020). Taking cues, we generate the sequence of preferences $P_t(i, j)$ for each pair of arm $(i, j)$ as random walks with increments $\nu$ with some randomly chosen probability $q \in (0.2, 0.8)$, where each $P_t(i, j)$ is initialized uniformly on $[0, 1]$ for all $i, j$. Any values that fall outside $[0, 1]$ in the process are truncated back to $[0, 1]$.

3. **Lower Bound instance or LB($\epsilon$).** Our lower bound preference instance $P_1$ parameterized by $\epsilon \in (0, 0.5)$ (see Section 5). The explicit values used for $\tau, \nu, \epsilon$ are specified in the corresponding figures.

**6.1. Cumulative regret over time**

We first conduct a set of experiments to compare the regret performance of the three algorithms over the two problem instances, SB(500) and RW(0.01), as shown in Fig. 1.

**Remark.** As shown in Fig. 1, our algorithm Dueling-EXP3 outperforms REX3 in both the instances. This is expected since the later is guaranteed to work only under linear utility based adversarial preference models, whereas we have constructed completely adversarial preference matrices through SB and RW instances. Also, both of the above algorithms perform better than the naive Random duel selection baseline.
We also conduct a set of experiments changing the item set size $K$ over a range ($K = 10$ to $100$). We report the final cumulative regret of all algorithms vs. $K$ on the LB(0.1) instance as specified in Fig. 2.

Remark. In terms of the comparative regret performances of three algorithms, Fig. 2 shows the same trend as reflected in the first set of experiments (Fig. 1), where Dueling-EXP3 performs best, then REX3 and the worse is Random. Additionally, Fig. 2 shows that with increasing $K$ but fixed gap ($\epsilon$)—that we ensured with our LB(0.1) instance construction keeping the gap $\epsilon = 0.1$ fixed for all $K$—we see the regret of all the algorithms scales up with increasing $K$, as expected and also justified by Theorem 2.
References


Adversarial Dueling Bandits


A. Appendix for Section 3

Lemma 3. For all $t \in [T]$ and $i \in [K]$ it holds that $\tilde{s}_t(i) \leq K/\gamma^2$.

Proof. The claim simply follows from the definition of $\tilde{s}_t(i)$, and the fact that $q_t(i) \geq \frac{1}{K}$, for all $i \in [K]$ and $t \in [T]$. □

Lemma 4. For all $t \in [T]$ and $i \in [K]$ it holds that $E[\tilde{s}_t(i)] = s_t(i)$.

Proof. Note that:

$$E[\tilde{s}_t(i)] = E_{H_t}[\frac{1(x_t = i)}{q_t(i)K} \sum_{j \in [K]} \frac{1(y_t = j)q_t(j)}{q_t(j)}]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{(x_t, y_t, o_t)} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

$$= \frac{1}{K} E_{H_{t-1}} \left[ \sum_{j \in [K]} E_{x_t} \left[ \frac{1(x_t = i)1(y_t = j)q_t(j)}{q_t(j)} \right] \right]$$

which concludes the proof. □

Lemma 5. For all $t \in [T]$ it holds that $E_{H_t}[q_t \tilde{s}_t] = E_{H_{t-1}}[E_{x \sim q_t}[s_t(x) \mid H_{t-1}]]$.

Proof. Following the same techniques from the proof of Lemma 4, we have:

$$E_{H_t}[q_t \tilde{s}_t] = E_{H_t} \left[ \sum_{i=1}^{K} q_t(i)\tilde{s}_t(i) \right] = E_{H_{t-1}} \left[ \sum_{i=1}^{K} q_t(i)E_{(x_t, y_t, o_t)} \left[ \tilde{s}_t(i) \mid H_{t-1} \right] \right]$$

$$\overset{(1)}{=} E_{H_{t-1}} \left[ \sum_{i=1}^{K} q_t(i)s_t(i) \right] = E_{H_{t-1}}[E_{x \sim q_t}[s_t(x) \mid H_{t-1}]]$$

where (1) follows from the proof of Lemma 4, and hence the result follows. □
Lemma 6. For all $t \in [T]$ it holds that $\mathbb{E}\left[\sum_{i=1}^{K} q_t(i) \tilde{s}_t(i)^2\right] \leq K/\gamma$.

Proof. Recall $\tilde{s}_t(i) := \frac{1(x_t = i)}{q_t(i)} \sum_{j \in [K]} \frac{1(y_t = j) \omega_t}{q_t(j)}$. The argument follows similar to the proof of Lemma 4:

$$\mathbb{E}\left[\sum_{i=1}^{K} q_t(i) \tilde{s}_t(i)^2\right] = \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \sum_{i=1}^{K} q_t(i) \mathbb{E}_{(x_t,y_t,\omega_t)^{\mathcal{H}_{t-1}}} \left[ \sum_{j \in [K]} \frac{1(x_t = i, y_t = j) \omega_t}{K q_t(i) q_t(j)} \mathbb{I}(x_t = i) \mathbb{I}(y_t = j) \mid \mathcal{H}_{t-1} \right] \right]$$

$$= \frac{1}{K^2} \left( \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \sum_{i=1}^{K} q_t(i) \mathbb{E}_{(x_t,y_t)^{\mathcal{H}_{t-1}}} \left[ \sum_{j \in [K]} \frac{1}{q_t(j)^2} \mathbb{I}(x_t = i) \mid \mathcal{H}_{t-1} \right] \right] \right)$$

$$\leq \frac{1}{K^2} \left( \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \sum_{i=1}^{K} q_t(i) \mathbb{E}_{(x_t,y_t)^{\mathcal{H}_{t-1}}} \left[ \sum_{j \in [K]} \frac{1}{q_t(j)^2} (q_t(i) q_t(j)) \right] \right] \right)$$

$$= \frac{1}{K^2} \left( \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \sum_{i=1}^{K} \frac{1}{q_t(i)} \right] \right) = \frac{K}{\gamma},$$

where last inequality follows since $q_t(j) \geq \frac{\gamma}{K}$, and the claim follows. \hfill \square

Theorem 2. Let $\eta = ((\log K) / (T \sqrt{K}))^{2/3}$ and $\gamma = \sqrt{\eta K}$. For any $T$, the expected regret of Algorithm 1 satisfies $\mathbb{E}[R_T] \leq 6(K \log K)^{1/3} T^{2/3}$.

Proof. Recall the definition of ‘shifted Borda score’ $s_t$ and the regret $R_T^s$ defined in terms of $s_t$. It would be convenient to upper bound $R_T^s$ and recall that $R_T = (K/(K - 1)) R_T^s$.

Note that $\mathbb{E}_{\mathcal{H}_T}[s_t(x_t) + s_t(y_t)] = \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \mathbb{E}_{x_t,y_t \sim q_t} \left[ s_t(x_t) + s_t(y_t) \mid \mathcal{H}_{t-1} \right] \right] = \mathbb{E}_{\mathcal{H}_{t-1}} [\mathbb{E}_{x \sim q_t}[s_t(x) \mid \mathcal{H}_{t-1}]]$, since $x_t$ and $y_t$ are i.i.d. Further note that we can write

$$\mathbb{E}_{\mathcal{H}_T}[R_T^s] := \mathbb{E}_{\mathcal{H}_T} \left[ \sum_{t=1}^{T} \left[ s_t(i^*) - \frac{s_t(x_t) + s_t(y_t)}{2} \right] \right] = \max_{k \in [K]} \mathbb{E}_{\mathcal{H}_T} \left[ \sum_{t=1}^{T} \left[ s_t(k) - \frac{s_t(x_t) + s_t(y_t)}{2} \right] \right],$$

where the last equality holds since $P_t$s are chosen obliviously, and hence $s_t$ and $i^*$ are independent of the randomness of the algorithm. Thus we get:

$$\mathbb{E}_{\mathcal{H}_T}[R_T^s] = \max_{k \in [K]} \left[ \frac{1}{T} \sum_{t=1}^{T} s_t(k) - \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_{t-1}} [\mathbb{E}_{x \sim q_t}[s_t(x) \mid \mathcal{H}_{t-1}]] \right],$$

(2)

First, since $\eta \tilde{s}_t(i) \leq \frac{\eta K}{T}$ (from Lemma 3), for any $\gamma \geq \sqrt{\eta K}$ and $\eta > 0$, we have $\eta \tilde{s}_t(i) \in [0, 1]$ for any $i \in [K], t \in [T]$.

From the regret guarantee of standard Exponential Weight algorithm (Auer et al., 2002a) over the completely observed fixed sequence of reward vectors $\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_T$ we have for any $k \in [K]$:

$$\sum_{t=1}^{T} \tilde{s}_t(k) - \sum_{t=1}^{T} \tilde{q}_t(k) \tilde{s}_t \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} \tilde{q}_t(i) \tilde{s}_t(i)^2$$
where \( \tilde{q}_t(i) := \frac{e^{\eta \sum_{s=1}^{t-1} \tilde{s}_s(i)}}{\sum_{j=1}^{K} e^{\eta \sum_{s=1}^{t-1} \tilde{s}_s(j)}}, \forall i \in [K]. \)

Since \( \tilde{q}_t = \left( \frac{(\eta - \gamma)}{\gamma} \right) \) and \( \gamma \in (0, 1) \), from above inequality we get for any \( k \in [K] \):

\[
(1 - s_k) \sum_{t=1}^{T} \tilde{s}_t(k) - \sum_{t=1}^{T} q_t^\top \tilde{s}_t \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_t(i) \tilde{s}_t(i)^2.
\]

Since \( i^* = \arg \max_{k \in [K]} \sum_{t=1}^{T} s_t(k) = \arg \max_{k \in [K]} \sum_{t=1}^{T} b_t(k) \), using the above inequality for \( k = i^* \), and taking expectation on both sides, we have

\[
(1 - s_k) \sum_{t=1}^{T} \mathbb{E}_H[s_t(i^*)] - \sum_{t=1}^{T} \mathbb{E}_H[q_t^\top \tilde{s}_t] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_t(i) \tilde{s}_t(i)^2
\]

\[
\Rightarrow (1 - s_k) \sum_{t=1}^{T} \mathbb{E}_{H_{t-1}} [s_t(i^*)] - \sum_{t=1}^{T} \mathbb{E}_{H_{t-1}} [E_{x \sim q_t}[s_t(x) | H_{t-1}]] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{E}_{H_{t-1}} [s_t(k) \tilde{s}_t(k)] \leq \frac{s_k}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}_{H_{t}} [R_T^2] \leq 3(K \log K)^{1/3} T^{2/3}
\]

where (1) follows from Lemma 4, Lemma 5 and Lemma 6, (2) follows since \( s_k(i^*) \leq 1 \), and (3) follows from Eq. (2), (4) follows since \( \gamma = \sqrt{\eta K} \), and 5 follows by optimizing over \( \eta \) which gives \( \eta = \frac{\log K}{K} \) as desired. Finally from Definition 1 since \( R_T = (K/(K-1)) R_T^2 \), this concludes the proof.

\]

### B. Appendix for Section 3.2

#### B.1. Proofs for Section 3.2

**Lemma 8.** For any item \( i \) and round \( t \in [T] \), we have \( s_t(i) \leq K/\gamma^2 + K \beta/\gamma \).

**Proof.** The claim simply follows from the definition of \( s_t(i) \), and the fact that \( q_t(i) \geq \frac{K}{\gamma} \), for all \( i \in [K] \) and \( t \in [T] \).

**Lemma 9.** For any item \( i \) and round \( t \in [T] \), it holds that \( \mathbb{E}[s_t(i) | H_{t-1}] = s_t(i) + \beta/q_t(i) \).

**Proof.** Note that for any \( i \in [K] \),

\[
s_t(i) = \frac{1}{K} \sum_{j \in [K]} \frac{1}{q_t(i)} \mathbb{1}(y_t = j, x_t = i) \mathbb{1}(x_t = i) + \beta/q_t(i).
\]

Recalling the definition of \( \tilde{s}_t(i) := \frac{1}{K} \sum_{j \in [K]} \frac{1}{q_t(i)} \mathbb{1}(x_t = i) \mathbb{1}(x_t = j) \) from Algorithm 1, we further note:

\[
\mathbb{E}[s_t(i) | H_{t-1}] = \mathbb{E}_{(y_t, x_t, \omega_t)} \left[ \tilde{s}_t(i) + \frac{\beta}{q_t(i)} | H_{t-1} \right]
\]
we have which proves the claim.

Lemma 10. For any \( i \in [K], \delta \in (0, 1) \) and \( \beta, \gamma \in (0, 1) \), with probability at least \( 1 - \delta \), we have

\[
\sum_{t=1}^{T} s_t(i) \geq \sum_{t=1}^{T} s_t(i) - \frac{1}{\gamma \beta} \log \frac{1}{\delta}.
\]

Proof. Let \( \beta' = \gamma \beta \). Then note \( \beta' \in (0, 1) \) by the choice of \( \beta, \gamma \). Thus using Markov Inequality:

\[
Pr\left( \sum_{t=1}^{T} s_t(i) \leq \sum_{t=1}^{T} s_t(i) - \frac{\log(1/\delta)}{\beta'} \right) = Pr_{\mathcal{H}_T}\left( \exp(\beta' \sum_{t=1}^{T} (s_t(i) - s'_t(i))) \geq \frac{1}{\delta} \right)
\]

\[
= \delta E_{\mathcal{H}_T}\left[ \exp\left( \beta' \sum_{t=1}^{T} (s_t(i) - s'_t(i)) \right) \right] = \delta E_{\mathcal{H}_T}\left[ \prod_{t=1}^{T} \exp\left( \beta' \left( s_t(i) - s'_t(i) \right) \right) \right]
\]

Now for any fixed \( t \in [T] \), for any \( i \in [K] \), note that \( s'_t(i) \geq \frac{\beta}{q_t(i)} \) (due to Lemma 9). Thus since \( s_t(i) \in [0, 1] \) by definition, we have \( s_t(i) - (s'_t(i) - \frac{\beta}{q_t(i)}) \leq 1 \). Moreover as \( \beta' \in (0, 1) \), using \( e^x \leq (1 + x + x^2) \) for any \( x \leq 1 \), we get:

\[
E_{\mathcal{H}_t}\left[ \exp\left( \beta' \left( s_t(i) - s'_t(i) \right) \right) \mid \mathcal{H}_{t-1} \right] = E_{\mathcal{H}_t}\left[ \exp\left( \beta' \left( s_t(i) - s'_t(i) + \frac{\beta}{q_t(i)} \right) \right) \mid \mathcal{H}_{t-1} \right] \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right)
\]

\[
\leq E_{\mathcal{H}_t}\left[ 1 + \beta' \left( s_t(i) - s'_t(i) + \frac{\beta}{q_t(i)} \right)^2 \mid \mathcal{H}_{t-1} \right] \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right), \quad \text{(as } E_{\mathcal{H}_t}\left[ s_t(i) - s'_t(i) + \frac{\beta}{q_t(i)} \mid \mathcal{H}_{t-1} \right] = 0 \text{)}
\]

\[
= 1 + \beta' \left( s_t(i) - s'_t(i) + \frac{\beta}{q_t(i)} \right)^2 \mid \mathcal{H}_{t-1} \right] \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right), \quad \text{(as } q_t(i) \text{ is } \mathcal{H}_t \text{ measurable)}
\]

\[
= 1 + \beta^2 \text{Var}_{\mathcal{H}_t}\left( s'_t(i) \mid \mathcal{H}_{t-1} \right) \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right)
\]

\[
\leq 1 + \beta'^2 \left[ E_{\mathcal{H}_t}\left[ \left( \frac{1}{q_t(i) K} \sum_{j=1}^{K} 1(y_t = j) o_t \right)^2 \mid \mathcal{H}_{t-1} \right] \right] \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right), \quad \text{(since } \frac{\beta}{q_t(i)} \text{ is constant given } \mathcal{H}_{t-1} \text{)}
\]

\[
\leq 1 + \beta^2 \left[ E_{\mathcal{H}_t}\left[ \frac{1}{q_t(i) K} \sum_{j=1}^{K} 1(y_t = j) o_t \mid \mathcal{H}_{t-1} \right] \right] \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right), \quad \text{(since } o_t \leq 1 \text{)}
\]

\[
\leq 1 + \beta'^2 \left[ \left( \frac{1}{q_t(i) K} \sum_{j=1}^{K} 1(y_t = j) \right) \right] \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right)
\]

\[
\leq 1 + \left( \frac{\beta' q_t(i)}{q_t(i)} \right) \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right), \quad \text{(since } q_t(j) \geq \frac{\gamma}{K}, \forall j \in [K] \text{)}
\]

\[
\leq 1 + \frac{\beta' q_t(i)}{q_t(i)} \exp\left( - \frac{\beta' q_t(i)}{q_t(i)} \right), \quad \text{(since } \beta = \frac{\beta'}{\gamma} \text{)}
\]
Applying the above result for all \( t \in [T] \) in Eq. (3) concludes the proof.

**Theorem 7.** Given any \( T \) and \( \delta > 0 \), there exists a setting of \( \gamma, \beta \) and \( \eta \), such that with probability at least \( 1 - \delta \), the regret of the modified D-EXP3 algorithm is \( R_T = \tilde{O}(K^{1/3}T^{2/3}) \).

**Proof.** We set \( \gamma = \sqrt{2}\eta K, \beta = \frac{T^{-1/2}}{(2\eta)^{1/2}K^{1/2}}, \) and \( \eta = \frac{(\log K)^{2/3}}{T^{1/3}2K} \). We will prove a regret bound of

\[
R_T \leq 2 \left( 3(2 \log K)^{1/3} + 2^{5/6} \sqrt{(\log K)^{1/6}} \right) K^{1/3}T^{2/3} = \tilde{O}(K^{1/3}T^{2/3}).
\]

Note that for \( T < 2T \log T \), the bound regret bound is more than \( T \) and therefore holds trivially. For the remainder of the proof we assume that \( T \geq 2T \log T \).

We start by recalling the definition of `shifted borda score` \( s_t \) and the regret \( R_T^0 \) from Section 2. Same as Theorem 2, we find it convenient to first upper bound \( R_T^0 \).

Note that by Lemma 8, if we set \( \eta \leq \left( \frac{K^2}{\gamma} + \frac{K^2}{\beta} \right)^{-1} \), we have \( \eta s_t(i) \leq \frac{K^2}{\gamma} + \frac{K^2}{\beta} \in (0, 1), \forall i \in [K], t \in [T] \). Then again from the regret guarantee of standard Exponential Weight algorithm (Auer et al., 2002a) over the fully observed fixed sequence of reward vectors \( s_1, s_2, \ldots, s_T \) we have for any \( k \in [K] \):

\[
\sum_{t=1}^{T} s_t^i(k) \leq \sum_{t=1}^{T} [q_t^s s_t^i] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_t(i)s_t^i(i)^2,
\]

where \( q_t(i) := \frac{e^{\eta \sum_{r=t}^{t-1} s_r^i(i)}}{K e^{\eta \sum_{r=t}^{t-1} s_r^i(j)}}, \forall i \in [K] \). Further by definition since \( q_t = \frac{(q_t - \mathbb{I})}{1 - \gamma} \) and \( \gamma \in (0, 1) \), from above inequality we get for any \( k \in [K] \):

\[
(1 - \gamma) \sum_{t=1}^{T} s_t^i(k) \leq \sum_{t=1}^{T} q_t^s s_t^i + \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_t(i)s_t^i(i)^2,
\]

Here note that for a given \( x_t \),

\[
q_t^s s_t^i = \sum_{i=1}^{K} q_t(i) \left( \frac{1_{(x_t = i)}}{K} \sum_{i \in [K]} 1_{(y_t = j) \alpha_t(i, j)} q_t(i) \beta \right) = \left( \sum_{j \in [K]} 1_{(y_t = j) \alpha_t(x_t, j)} q_t(j)K + K\beta \right),
\]

where \( \alpha_t(i, j) \sim \text{Ber}(P_t(i, j)), \forall i, j \in [K] \). Thus taking expectation:

\[
E_{y_t, \alpha_t} \left[q_t^s s_t^i | \mathcal{H}_{t-1}, x_t\right] = E_{y_t, \alpha_t} \left[\sum_{j \in [K]} 1_{(y_t = j) \alpha_t(x_t, j)} q_t(j)K \Bigg| \mathcal{H}_{t-1}, x_t\right] + K\beta = s_t(x_t) + K\beta
\]

Further noting \( q_t(i)s_t^i(i) \leq \left( \beta + \gamma^{-1} \right) \),

\[
\eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_t(i)s_t^i(i)^2 \leq \eta \sum_{t=1}^{T} \left( \beta + \gamma^{-1} \right) \sum_{i=1}^{K} s_t^i(i) \leq \eta K \left( \beta + \gamma^{-1} \right) \sum_{t=1}^{T} s_t^i(i^*),
\]

Thus we have

\[
\sum_{t=1}^{T} s_t^i(k) \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_t(i)s_t^i(i)^2 \leq \frac{\log K}{\eta} + \eta K \left( \beta + \gamma^{-1} \right) \sum_{t=1}^{T} s_t^i(i^*),
\]

which concludes the proof.
where we denote by $i^* = \arg\max_{k \in [K]} \sum_{t=1}^{T} s_t^i(k)$. Combining the results of Eq. (7) to Eq. (4), and the fact that we set $\eta \leq \left( \frac{K}{\gamma^2} + \frac{K\beta}{\gamma} \right)^{-1}$,

$$
(1 - \gamma) \sum_{t=1}^{T} s_t^i(i^*) \leq \sum_{t=1}^{T} q_t^i s_t^i + \frac{\log K}{\eta} + \gamma \sum_{t=1}^{T} s_t^i(i^*) \quad \text{(since } \eta \leq \left( \frac{K}{\gamma^2} + \frac{K\beta}{\gamma} \right)^{-1})
$$

$$
(1 - 2\gamma) \left[ \sum_{t=1}^{T} s_t^i(i^*) \right] \leq \sum_{t=1}^{T} q_t^i s_t^i + \frac{\log K}{\eta}
$$

$$
(1 - 2\gamma) \max_{k \in [K]} \left[ \sum_{t=1}^{T} s_t(k) - \frac{\log(K/\delta)}{\gamma\beta} \right] \leq \sum_{t=1}^{T} \left[ \sum_{j \in [K]} 1(y_t = j) a_t(x_t, j) q_t(j) K + K\beta \right] + \frac{\log K}{\eta}
$$

$$
(1 - 2\gamma) \max_{k \in [K]} \left[ \sum_{t=1}^{T} s_t(k) \right] - \sum_{t=1}^{T} s_t(x_t) \leq 2\gamma T + K\beta T + \frac{\log K}{\eta} + \frac{\log(K/\delta)}{\gamma\beta} \quad \text{(since } \max_{k \in [K]} \sum_{t=1}^{T} s_t(k) \leq T)
$$

$$
\max_{k \in [K]} \sum_{t=1}^{T} s_t(k) - \sum_{t=1}^{T} s_t(x_t) \leq 2\sqrt{2\eta KT} + K\beta T + \frac{\log K}{\eta} + \frac{\log(K/\delta)}{\gamma\beta} \quad \text{(since } \max_{k \in [K]} \sum_{t=1}^{T} s_t(k) \leq T)
$$

$$
\max_{k \in [K]} \sum_{t=1}^{T} s_t(k) - \sum_{t=1}^{T} s_t(x_t) \leq 2\sqrt{2\eta KT} + \frac{\log K}{\eta} + \frac{\log(K/\delta)}{\gamma\beta}
$$

$$
\max_{k \in [K]} \sum_{t=1}^{T} s_t(k) - \sum_{t=1}^{T} s_t(x_t) \leq 3(2KT^2 \log K)^{1/3} + 2^{5/6}(KT^2)^{1/3} \sqrt{\frac{\log K/\delta}{(\log K)^{1/6}}}
$$

$$
R_T^\ast \leq \left(3(2\log K)^{1/3} + 2^{5/6} \sqrt{\frac{\log K/\delta}{(\log K)^{1/6}}} \right) K^{1/3} T^{2/3},
$$

where (1) follows from Eq. (5) and taking an union bound over all $i \in [K]$ for the claim Lemma 10, (2) holds from Eq. (6), (3) follows by setting $\gamma = \sqrt{2\eta K}$ (note since if we can ensure $\beta \gamma \leq 1$, this ensures $\eta \left( \frac{K}{\gamma^2} + \frac{K\beta}{\gamma} \right) \leq \left( \frac{2K\eta}{\gamma^2} \right) \leq 1$ as desired), (4) follows by setting $\beta = \frac{\sqrt{\log(K/\delta)}}{(2\eta)^{3/4} K^{3/4} \sqrt{T}}$. (5) follows by optimizing over $\eta$ which gives $\eta = \left( \frac{\log(K/\delta)}{T\sqrt{2K}} \right)^{2/3}$. Further note that any $T \geq 2K \log K$ implies $\eta = \sqrt{2\eta K} \in [0, 1]$, and any $T \geq \frac{\log(K/\delta)^{1/3}}{K^2 \sqrt{2\log K}}$ implies $\beta \in (0, 1)$ as desired. Finally from Definition 1 since $R_T \leq 2R_T^\ast$, this concludes the proof. 

\[\square\]

C. Appendix for Section 4

**Lemma 11.** At any round $t$, we have $E_{H_t} \left[ \hat{b}_t(i) \right] = b_t(i)$ for all $i \in [K]$.

**Proof.** Note that by definition for any $i \in [K]$, $\hat{b}_t(i) = K1(x_t = i) \sum_{j=1}^{K} 1(y_t = j) a_t(i, j)$. It is easy to see that for any $i \in [K], t \in [T]$,

$$
E_{H_t} \left[ \hat{b}_t(i) \right] = E_{H_{t-1}} \left[ E_{x_t,y_t,a_t} \left[ K1(x_t = i) \sum_{j=1}^{K} 1(y_t = j) a_t(i, j) \mid H_{t-1} \right] \right] = E_{H_{t-1}} \left[ \hat{b}_t(i) \right] = b_t(i),
$$

where the last equality is simply due to the fact that $P_t$ is chosen obliviously w.r.t. the history $H_{t-1}$, and the second last equality follows since for any $i \in K$:
We have \( \exists \) Let us fix any item \( \text{Proof.} \) For any round \( K \)

\[ K_{\tau/\delta}^{*} = \frac{1}{K} \left( \frac{K}{T} \right) \]

\[ \bar{b}(i) \leq 
\]

which concludes the proof.

**Lemma 12.** With probability \( \geq 1 - \delta \), we have \( \bar{b}(t) \in [\text{LCB}(i; t), \text{UCB}(i; t)] \) for all \( i \in [K] \) and \( t \in [T] \).

**Proof.** For any round \( t \leq 4K \log(2KT/\delta) \) we have that \( 2\sqrt{(K/t) \log(2KT/\delta)} \geq 1 \) and therefore, for any item \( i \in [K] \) we have \( \text{LCB}(i, t) < 0 \) and \( \text{UCB}(i; t) > 1 \), and hence the lemma holds trivially.

Let us fix any item \( i \in [K] \), and some \( t \geq 4K \log \frac{KT}{\delta} \). Note that owning to our ‘random arm-pair \( (x_t, y_t) \) selection strategy’, the random variables \( \bar{b}_1(i), \bar{b}_2(i), \ldots, \bar{b}_t(i) \) are independent. Let us denote denote by \( \epsilon = \frac{1}{t} \). Let us also define for any \( \tau \in [t] \), \( z_t(\tau) = \frac{1}{t} \left( \bar{b}_t(\tau) - \bar{b}_t(i) \right) \). Then note: (i) \( z_1(i), z_2(i), \ldots, z_t(i) \) are independent (ii) \( E_{\bar{H}_t}[z_t(\tau)] = 0 \) (see Lemma 11), (iii) \( |z_t(\tau)| < \frac{\delta}{K} \), and (iv) \( \sum_{t=1}^{T} E_{\bar{H}_t}[z_t^2(i)] \leq \frac{1}{K} \frac{K^2}{t^2} + \frac{K-1}{K} \frac{1}{t^2} \leq \frac{K+1}{K} \frac{1}{t^2} \) (as \( Pr(x_t = i) = \frac{1}{K} \)) for any \( K \geq 2 \). Hence applying Bernstein’s inequality we get:

\[ Pr \left( \frac{t}{2} \sum_{t=1}^{T} z_t(i) \geq \epsilon \right) \leq 2 \exp \left( - \frac{\epsilon^2 / 2}{K + 1 + \frac{\epsilon}{2}} \right) \leq 2 \exp \left( - \frac{\epsilon^2 / 2}{4K} \right) = \frac{\delta}{KT}. \]

where the second inequality follows since for any \( t \geq \frac{16K \log(2KT/\delta)}{\delta^2} \), we have \( \frac{4K}{\delta} \). The proof follows taking union bound over all \( i \in [K] \) and \( t \in [T] \).

**Theorem 13.** Given any \( \delta > 0 \), with probability at least \( 1 - \delta \), the regret of Algorithm 3 (with parameter \( \delta \)) is upper bounded by \( 64(K/\Delta^2) \log(2KT/\delta) \).

**Proof.** We would first assume the good event of Lemma 12: \( \forall i \in [K], \forall t \in [T] \), we have \( \bar{b}(t) \in [\text{LCB}(i; t), \text{UCB}(i; t)]. \)

Recall that by problem setup: \( \exists i^* \in [K], \forall t \in [T] \) such that \( \bar{b}_t(i^*) > \bar{b}_t(j) + \Delta, \forall j \in [K] \setminus \{i^*\} \). Then if \( t > \frac{64K \log(2KT/\delta)}{\Delta^2} \), this implies \( \sqrt{\frac{4K \log(2KT/\delta)}{t}} \leq \Delta/4. \) Thus for any \( j \in K \setminus \{i^*\} \), at any \( t > \frac{64K \log(2KT/\delta)}{\Delta^2} \),

\[ \text{UCB}(j; t) = \bar{b}_t(j) + \sqrt{\frac{4K \log(2KT/\delta)}{t}} \leq \bar{b}_t(j) + 2 \sqrt{\frac{4K \log(2KT/\delta)}{t}} \leq \bar{b}_t(j) + \Delta/2 \]

On the other hand, for \( i^* \) we have

\[ \bar{b}_t(i^*) - \Delta/2 < \bar{b}_t(i^*) - \Delta + 2 \sqrt{\frac{4K \log(2KT/\delta)}{t}} < \bar{b}_t(i^*) - 2 \sqrt{\frac{4K \log(2KT/\delta)}{t}} < \text{LCB}(i^*; j). \]
Since \( \tilde{b}_t(i^*) \geq \tilde{b}_t(j) + \Delta \), it implies that \( UCB(j; t) < LCB(i^*; t) \) for \( t > \frac{64K \log(2KT/\delta)}{\Delta^2} \). Thus for any \( t > \frac{64K \log(2KT/\delta)}{\Delta^2} \), the algorithm would detect \( \hat{i} = \{i^*\} \), and hence the regret at \( \tau = r \) is \( r = 0 \) for the remaining rounds \( \tau = t + 1, \ldots, T \). The final high probability regret upper bound now follows from the statement of Lemma 12 and the fact that instantaneous regret at any round \( t \) such that \( (x_t, y_t) \neq (i^*, i^*) \) is at most 1.

\[ \square \]

\section*{D. Appendix for Sec. 5}

**Lemma 14.** For the problem of Adversarial Dueling Bandits with Borda Score objective, for any learning algorithm \( A \) and any \( \epsilon \in (0, 0.1] \), there exists a problem instance (sequence of preference matrices \( P_1, P_2, \ldots, P_T \)) such that the expected regret incurred by \( A \) on that instance is at least \( \Omega(\min(\epsilon T, K/\epsilon^2)) \), for any \( K \geq 4 \).

\[ \text{Proof.} \] We will show specifically that for \( T \leq \frac{K}{1440} \) we have \( R_T = \Omega(\epsilon T) \) and for \( T > \frac{K}{1440} \) we have \( R_T = \Omega\left(\frac{K}{\epsilon T}\right) \).

The proof relies on constructing a ‘hard enough’ problem instance for the learning framework and showing no algorithm can achieve a smaller rate of regret on that instance than the claimed lower bounds. 

For simplicity of notation we assume \( K \) is even (similar technique could also be used to prove the same bound when \( K \) is odd, and show that the lemma also applies to \( K = 3 \)). We denote by \( \tilde{K} := \frac{K}{2} \). Let us construct \( \tilde{K} + 1 \) problem instances \( \mathcal{T}^1, \mathcal{T}^2, \ldots, \mathcal{T}^{\tilde{K}} \) and \( \mathcal{T}^0 \), where each instance is uniquely identified by its underlying preference matrix as defined below:

**Problem instance** (\( \mathcal{T}^0 \)): For all \( t \in [T], P_t(i, j) = \begin{cases} 0.5, & \forall i, j \in [\tilde{K}] \text{ or } i, j \in [K] \setminus [\tilde{K}] \\ 0.9, & \forall i \in [\tilde{K}] \text{ and } \forall j \in [K] \setminus [\tilde{K}] \end{cases} \), or more explicitly:

\[
P_t = \begin{bmatrix}
0.5 & \ldots & 0.5 & 0.9 & \ldots & 0.9 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0.5 & \ldots & 0.5 & 0.9 & \ldots & 0.9 \\
0.1 & \ldots & 0.1 & 0.5 & \ldots & 0.5 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0.1 & \ldots & 0.1 & 0.5 & \ldots & 0.5 \\
\end{bmatrix}, \quad \forall t \in [T].
\]

Note for \( \mathcal{T}^0, \forall t \in [T], \) for any item \( i \in [\tilde{K}], s_t(i) = 0.7, \) and for any item \( i \in [K] \setminus [\tilde{K}], s_t(i) = 0.3. \) Thus for the instance \( \mathcal{T}^0 \), any item \( i \in [\tilde{K}] \) is an optimal arm. Now let us consider \( \tilde{K} \) alternative problem instances \( \mathcal{T}^m \) \( \forall m \in [\tilde{K}] \):

**Problem instance** (\( \mathcal{T}^m \)): For all \( t \in [T], P_t(i, j) = \begin{cases} 0.5, & \forall i, j \in [\tilde{K}] \text{ or } i, j \in [K] \setminus [\tilde{K}] \\ 0.9, & \forall i \in [\tilde{K}] \text{ and } \forall j \in [K] \setminus [\tilde{K}] \end{cases} \), for some \( \epsilon \in (0, 0.1] \). For example \( \mathcal{T}^1 \) would be:

\[
P_t = \begin{bmatrix}
0.5 & \ldots & 0.5 & 0.9 + \epsilon & \ldots & 0.9 + \epsilon \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0.5 & \ldots & 0.5 & 0.9 & \ldots & 0.9 \\
0.1 - \epsilon & \ldots & 0.1 & 0.5 & \ldots & 0.5 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0.1 - \epsilon & \ldots & 0.1 & 0.5 & \ldots & 0.5 \\
\end{bmatrix}, \quad \forall t \in [T].
\]
We now make the following two key observations:

We now turn our attention to proving the main result. We will break it into the following two case analyses:

would be convenient to first lower bound

would also hold over a randomized class of algorithms.

Then taking average over $T^m$s for all $m \in [\hat{K}]:$

$$\mathbf{E}[R^* T^m(A)] = \sum_{m \in [\hat{K}]} \mathbf{E}_{T^m}[R^*_T(A)] \geq \epsilon \left( T - \frac{\sum_{m \in [\hat{K}]} \mathbf{E}_{T^m}[N_T(m, m)]}{\hat{K}} \right)$$

(8)

since $i^m_\ast = m$. Now note that:

$$\mathbf{E}_{T^m}[N_T(m, m)] - \mathbf{E}_{T^0}[N_T(m, m)]$$
where $D_{TV}(\mathcal{I}_0, \mathcal{I}_m)$ denotes the total variation distance between the probability distribution of $\mathcal{I}_0$ and $\mathcal{I}_m$ with respect to $\mathcal{H}_T$, i.e., $D_{TV}(\mathcal{I}_0, \mathcal{I}_m) := \sup_{E \in \mathcal{H}_T} |P_{\mathcal{I}_0}(E) - P_{\mathcal{I}_m}(E)|$, $\mathcal{H}_T = \sigma\left(\{P_t(x_t, y_t)\}_{t \in T}\right)$ being the sigma algebra generated by the observed history till time $T$.

Further using Pinsker’s inequality we have

$$D_{TV}(\mathcal{I}_0, \mathcal{I}_m) \leq \sqrt{\frac{1}{2} D_{KL}(\mathcal{I}_0, \mathcal{I}_m)},$$

where $D_{KL}(\mathcal{I}_0, \mathcal{I}_m)$ denotes the KL-divergence between the probability distribution induced on the observed history $\mathcal{H}_T$ by the problem instance $\mathcal{I}_0$ and $\mathcal{I}_m$. Thus averaging over $\mathcal{I}_m$s for all $m \in [K]$:

$$\sum_{m \in [\hat{K}]} E_{\mathcal{I}_m}[N_T(m, m)] \leq \frac{\sum_{m \in [\hat{K}]} \left( E_{\mathcal{I}_0}[N_T(m, m)] + T D_{TV}(\mathcal{I}_0, \mathcal{I}_m) \right)}{K}$$

$$= \sum_{m \in [\hat{K}]} E_{\mathcal{I}_0}[N_T(m, m)] + T \sum_{m \in [\hat{K}]} \frac{1}{K} \left( \sqrt{\frac{1}{2} D_{KL}(\mathcal{I}_0, \mathcal{I}_m)} \right)$$

$$= \sum_{m \in [\hat{K}]} E_{\mathcal{I}_0}[N_T(m, m)] + T \sqrt{\frac{1}{2K} \sum_{m \in [\hat{K}]} D_{KL}(\mathcal{I}_0, \mathcal{I}_m)}$$

Now with slight abuse of notation, by denoting $\mathcal{I}_t^0 := P_{\mathcal{I}_0}(P_t(x_t, y_t) \mid \mathcal{H}_{t-1})$ and $\mathcal{I}_t^m := P_{\mathcal{I}_m}(P_t(x_t, y_t) \mid \mathcal{H}_{t-1})$, we note that:

$$D_{KL}(\mathcal{I}_t^0, \mathcal{I}_t^m) \sim KL(\text{Ber}(0.9), \text{Ber}(0.9 + \epsilon)), \text{ if } D_t = \{m, n\} \text{ for any } n \in [K] \setminus \hat{K}$$

Further using chain rule of KL-divergence we get:

$$D_{KL}(\mathcal{I}_0^0, \mathcal{I}_m) = \sum_{t=1}^{T} D_{KL}(\mathcal{I}_t^0, \mathcal{I}_t^0) = \sum_{t=1}^{T} \sum_{n=K+1}^{K} P_{\mathcal{I}_0}(D_t = \{m, n\}) D_{KL}(\text{Ber}(0.9), \text{Ber}(0.9 + \epsilon))$$

$$\leq D_{KL}(\text{Ber}(0.9), \text{Ber}(0.9 + \epsilon)) \sum_{t=1}^{T} \sum_{n=K+1}^{K} P_{\mathcal{I}_0}(D_t = \{m, n\}) \leq 90K^2 \sum_{n=K+1}^{K} E_{\mathcal{I}_0}[N_T(m, n)],$$

where the last inequality follows by noting $D_{KL}(\text{Ber}(0.9), \text{Ber}(0.9 + \epsilon)) \leq 90\epsilon^2$ for any $\epsilon \in (0, 0.1)$. Further averaging over $\mathcal{I}_m$s for all $m \in [\hat{K}]$:

$$\frac{\sum_{m \in [\hat{K}]} D_{KL}(\mathcal{I}_0^0, \mathcal{I}_m)}{K} \leq \frac{90K^2 \sum_{m \in [\hat{K}]} \sum_{n=K+1}^{K} E_{\mathcal{I}_0}[N_T(m, n)]}{K} \leq 90\epsilon^2 T$$

where the last inequality follows due to Observation 2. Now combining above with Eqn. (8) and 11 we get:

$$E[R^*_T(A)] \geq \frac{\epsilon}{2} \left( T - \frac{\sum_{m \in [\hat{K}]} E_{\mathcal{I}_0}[N_T(m, m)]}{K} \right)$$

$$\geq \frac{\epsilon}{2} \left( T - \left( \frac{\sum_{m \in [\hat{K}]} E_{\mathcal{I}_0}[N_T(m, m)]}{K} + T \sqrt{\frac{1}{2K} \sum_{m \in [\hat{K}]} D_{KL}(\mathcal{I}_0^0, \mathcal{I}_m)} \right) \right)$$

so that:

$$E[R^*_T(A)] \geq \frac{\epsilon}{2} \left( T - \left( \frac{\sum_{m \in [\hat{K}]} E_{\mathcal{I}_0}[N_T(m, m)]}{K} + T \sqrt{\frac{1}{2K} \sum_{m \in [\hat{K}]} D_{KL}(\mathcal{I}_0^0, \mathcal{I}_m)} \right) \right)$$
where (1) holds since by the assumption of Case 1 we have $T \leq \frac{K}{1440\epsilon^3}$, and the last inequality follows for any $K \geq 4$. This gives a regret lower bound for Case 1.

**Case 2.** ($T > \frac{K}{1440\epsilon^3}$): Let us denote by $T_0 = \frac{K}{1440\epsilon^3}$, and first assume that there exist a $T' > T_0$ such that $R_s^{T'} \leq \frac{K}{115200\epsilon^2}$. However this implies $R_s^{T_0} \leq R_s^{T'} \leq \frac{K}{115200\epsilon^2} = \frac{2T_0}{8}$. But this is a contradiction as per the lower bound of Case 1. Thus for any $\epsilon \in (0, 1)$ and $T > \frac{K}{1440\epsilon^3}$, any learning algorithm must incur at least an expected regret of $R_s^{T} \geq \frac{K}{115200\epsilon^2}$.

The final regret lower bound now follows combining the lower bounds of Case 1 and 2, and from the fact that $R_T \geq R_s^{T}$ (as per Rem. 1).