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 $\Omega(K^{1/3}T^{2/3})$ 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 Δ 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 Δ 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 & Hüllermeier, 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 & Gopalan, 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

[1]Microsoft Research, New York City [2]Blavatnik School of Computer Science, Tel Aviv University, and Google Research Tel Aviv. Correspondence to: Aadirupa Saha <aadirupa.saha@microsoft.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

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 Δ 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 $\tilde{\Theta}(K/\Delta^2)$, versus the well-known $\tilde{\Theta}(K/\Delta)$ 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 $\tilde{\Theta}(\sqrt{KT})$. 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., $\tilde{\Theta}(T^{2/3})$ 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 $o_t \sim \mathrm{Ber}(P_t(x_t, y_t))$ for the selected pair is revealed, where $\mathrm{Ber}(p)$ 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

$$\forall\, i \in [K] \;:\; b_t(i) := \frac{1}{K-1}\sum_{j\neq i} P_t(i,j),$$

$$\text{and} \quad i^* := \operatorname*{arg\,max}_{i\in[K]} \sum_{t=1}^{T} b_t(i),$$

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:

$$R_T := \sum_{t=1}^{T} r_t, \text{ where}$$
$$r_t := b_t(i^*) - \tfrac{1}{2}(b_t(x_t) + b_t(y_t)). \tag{1}$$

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 $\bar{b}_t(i^*) \geq \bar{b}_t(j) + \Delta$ for any other $j \neq i^*$, where $\bar{b}_t(j) := \frac{1}{t}\sum_{\tau=1}^{t} b_\tau(j)$ 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 $\tilde{O}(K^{1/3}T^{2/3}\sqrt{\log(K/\delta)})$ 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 $R_T^s := \sum_{t=1}^{T}[s_t(i^*) - \frac{1}{2}(s_t(x_t) + s_t(y_t))]$.

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

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 $\tilde{s}_t(i)$ for each item i, being an unbiased estimate of $s_t(i)$ (Lemma 4). Thus, the cumulative estimated score $\sum_{\tau=1}^{t}\tilde{s}_t(i)$ 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 an γ-uniform exploration.

We now state the expected regret guarantee we establish for

Algorithm 1 **Dueling-EXP3 (D-EXP3)**

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

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 $q_t(j) \ge \gamma/K$.

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

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 $\mathbf{E}[\tilde{s}_t(i)] = s_t(i)$.*

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 Borsda scores at time t.

Lemma 5. *For all $t \in [T]$ it holds that $\mathbf{E}_{\mathcal{H}_t}[q_t^\top \tilde{s}_t] = \mathbf{E}_{\mathcal{H}_{t-1}}\big[\mathbf{E}_{x \sim q_t}[s_t(x) \mid \mathcal{H}_{t-1}]\big]$.*

Finally, We bound the second moment of our estimates.

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

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

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

where the last equality holds since we assume the P_t are chosen obliviously and so i^* does not depend on the learning algorithm. Thus we can rewrite:

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

Now, as $\eta \tilde{s}_t(i) \le \eta K/\gamma^2$ (from Lemma 3), for any $\gamma \ge \sqrt{\eta K}$ 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 $k \in [K]$:

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

Note that $\tilde{q}_t := (q_t - \frac{\gamma}{K})/(1-\gamma)$. 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:

$$\begin{aligned}(1-\gamma) \sum_{t=1}^{T} \mathbf{E}_{\mathcal{H}_T}[\tilde{s}_t(i^*)] &- \sum_{t=1}^{T} \mathbf{E}_{\mathcal{H}_T}[q_t^\top \tilde{s}_t] \\ &\le \frac{\log K}{\eta} + \mathbf{E}_{\mathcal{H}_T}\left[\eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_t(i) \tilde{s}_t(i)^2\right],\end{aligned}$$

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

$$\begin{aligned}\mathbf{E}_{\mathcal{H}_T}[R_T^s] &\le 2T\sqrt{\eta K} + \frac{\log K}{\eta} \\ &\le 3(K \log K)^{1/3} T^{2/3},\end{aligned}$$

where the second inequality follows by optimizing over η. The theorem follows since $R_T = \frac{K}{K-1} R_T^s \le 2 R_T^s$. 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 γ-uniform exploration. The complete algorithm is described in Algorithm 2.

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

```
Input: Item set: $[K]$, learning rate $\eta > 0$, parameters $\beta \in (0,1), \gamma \in (0,1)$
Initialize: Initial distribution $q_1(i) = \frac{1}{K}, \ \forall i \in [K]$
while $t = 1, 2, \ldots$ do
    Sample $x_t, y_t \sim q_t$ (i.i.d., with replacement)
    Receive preference $o_t(x_t, y_t) \sim \text{Ber}(P_t(x_t, y_t))$
    Compute $\forall \ i \in [K]$:
        $$s'_t(i) = \frac{\mathbf{1}(x_t = i)}{Kq_t(i)} \sum_{j \in [K]} \frac{\mathbf{1}(y_t = j) o_t(x_t, y_t)}{q_t(j)} + \frac{\beta}{q_t(i)}$$
    Update $\forall \ i \in [K]$:
        $$\tilde{q}_{t+1}(i) = \frac{\exp(\eta \sum_{\tau=1}^{t} s'_\tau(i))}{\sum_{j=1}^{K} \exp(\eta \sum_{\tau=1}^{t} s'_\tau(j))}$$
        $$q_{t+1}(i) = (1-\gamma)\tilde{q}_{t+1}(i) + \frac{\gamma}{K}$$
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 γ, β and η, 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 $s'_t(i) \le K/\gamma^2 + K\beta/\gamma$.*

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

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 β. 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 $i \in [K]$, $\delta \in (0,1)$ and $\beta, \gamma \in (0,1)$, with probability at least $1-\delta$, we have*

$$\sum_{t=1}^{T} s'_t(i) \ge \sum_{t=1}^{T} s_t(i) - \frac{1}{\gamma\beta} \log \frac{1}{\delta}.$$

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 $\hat{i}$, which it uses in all rounds (for both items). The algorithm moves to its second phase when it detects an item $\hat{i}$ 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 $s \ne t$; 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 $\hat{b}_i(t) := K\mathbf{1}(x_t = i)o_t(x_t, y_t)$, and show that it is an *unbiased estimator*.

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

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

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

The proof uses Bernstein's inequality to show that the estimates $\bar{b}_i(t)$ are concentrated around their means $\tilde{b}_t(i)$, within the respective confidence intervals. Assuming these confidence bounds hold, as soon as we find an item $\hat{i} \in [K]$ such that $LCB(\hat{i};t) > UCB(j;t)$ for any other item $j \ne \hat{i}$, 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 $\hat{i}$ (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 δ) 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 $\hat{i}$. After we identify $\hat{i}$, with high probability, we do not incur any regret.

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

```
Input: item set indexed by $[K]$, confidence $\delta > 0$
for $t = 1, \ldots, T$ do
    Select $x_t, y_t \in [K]$, $x_t \neq y_t$ uniformly at random
    Receive preference $o_t(x_t, y_t) \sim \text{Ber}(P_t(x_t, y_t))$
    Estimate: $\hat{b}_i(t) = K\, o_t(x_t, y_t)\, \mathbf{1}(x_t = i)$, $\forall i \in [K]$
    Compute: $\tilde{b}_t(i) \leftarrow \frac{1}{t}\sum_{\tau=1}^{t} \hat{b}_\tau(i)$, $\forall i \in [K]$
    Compute:
        $LCB(i;t) = \tilde{b}_t(i) - 2\sqrt{\frac{K}{t}\log\frac{2KT}{\delta}}$,
        $UCB(i;t) = \tilde{b}_t(i) + 2\sqrt{\frac{K}{t}\log\frac{2KT}{\delta}}$
    if $\exists\, \hat{i} \in [K]$ s.t. $LCB(\hat{i};t) > UCB(j;t) \;\; \forall j \neq \hat{i}$, then break
end for
Play $(\hat{i}, \hat{i})$ 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 $\mathcal{A}$ 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 $\mathcal{A}$ 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:

$$P_1 = \begin{bmatrix} 0.5 & ... & 0.5 & 0.9+\epsilon & ... & 0.9+\epsilon \\ . & ... & . & . & ... & . \\ . & ... & . & . & ... & . \\ 0.5 & ... & 0.5 & 0.9 & ... & 0.9 \\ 0.1-\epsilon & ... & 0.1 & 0.5 & ... & 0.5 \\ . & ... & . & . & ... & . \\ . & ... & . & . & ... & . \\ 0.1-\epsilon & ... & 0.1 & 0.5 & ... & 0.5 \end{bmatrix}.$$

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 ϵ 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 $\mathcal{A}$, there* *exists an instance with fixed gap Δ, such that the expected regret incurred by $\mathcal{A}$ on that instance is at least* $\Omega(\min(\Delta T, K/\Delta^2))$.

The regret bound in this scales as K/Δ^2 compared to K/Δ 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 ϵ 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 $\mathcal{A}$, 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 $\mathcal{A}$ on that Adv-Borda(K, T) is atleast $\Omega(K^{1/3}T^{2/3})$.*

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 ϵ.

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 γ as well as the learning rate η requires the knowledge of problem dependent parameters τ—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 τ henceforth. However, other settings of τ 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(ν).* 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 ν with some randomly chosen probability $q \in (0.2, 0.8)$, where each $P_1(i, j)$ 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(ϵ).* Our lower bound preference instance P_1 parameterized by $\epsilon \in (0, 0.5)$ (see Section 5). The explicit values used for τ, ν, ϵ 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.

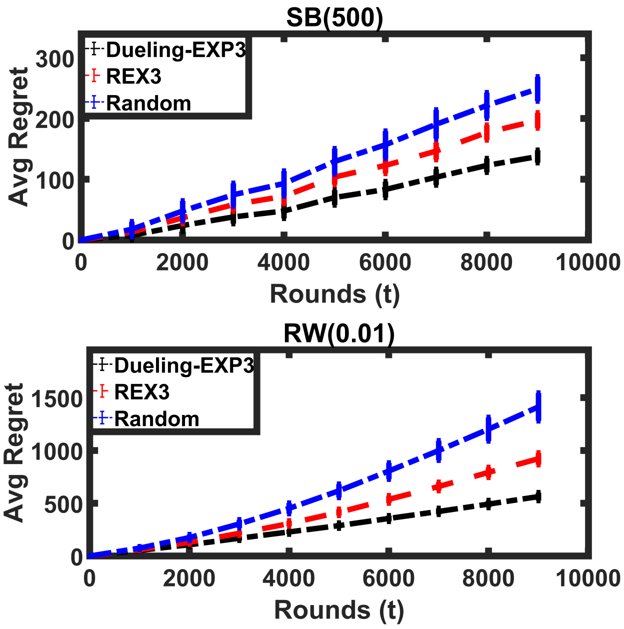

Figure 1. Averaged cumulative regret over time

6.2. Regret vs Varying Item-size (K)

We also conduct a set of experiment 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.

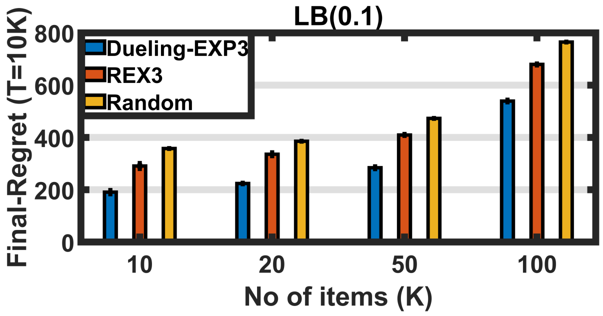

Figure 2. Final regret (at $T = 10K$) with increasing arm size K

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 (ϵ)—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.

7. Conclusion and Future Scopes

We considered the problem of dueling bandits with any adversarial preferences, i.e., *adversarial dueling bandits*. To the best of our knowledge, this work is the first to consider the dueling bandit problem for fully adversarial setup. (The work of (Gajane et al., 2015) introduced adversarial utility-based dueling bandits with a transfer function, which has very different characteristics, as we discussed earlier.)

We proposed algorithms for online regret minimization with Borda scores. We gave an $\tilde{O}(K^{1/3}T^{2/3})$ regret algorithm (Dueling-EXP3), for the problem, and also shown optimality of our bounds with a matching $\Omega(K^{1/3}T^{2/3})$ lower bound analysis. We also proved a similar high probability regret bound. Finally, for an intermediate *fixed-gap* adversarial setup—which bridges the gap between stochastic and adversarial dueling bandits—we gave an $\tilde{O}((K/\Delta^2)\log T)$ regret algorithm, Borda-Confidence-Bound, and also a corresponding regret lower bound of $\Omega(K/\Delta^2)$.

Moving forward, one can potentially address many open threads along this direction; for example, considering other general notions of regret performances, considering the problem on larger (potentially infinite) arm-spaces, or even analyzing dynamic regret for adversarial preferences (Besbes et al., 2019; Luo et al., 2017). Few more open questions to answer here are: In case of more strcutured utility based preferences (e.g., Plackett-Luce preference model (Azari et al., 2012), etc.), where the item utility scores are chosen adversarially at every round, is it possible to show an improved performance limit of $\Theta(\sqrt{KT})$? In such cases, how does the learning rate varies with K and T for general subsetwise preferences (i.e., where more than two items can be compared at every round and the learner receives a winner feedback of the subset played) (Brost et al., 2016; Ren et al., 2018)? Another interesting direction would be to understand the connection of this problem with other bandit setups, e.g., learning with feedback graphs (Alon et al., 2015; 2017) or other side information (Mannor & Shamir, 2011; Kocak et al., 2014).

Acknowledgments

The work of YM has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17) and the Yandex Initiative for Machine Learning at Tel Aviv University. TK was supported in part by the Israeli Science Foundation (ISF) grant no. 2549/19, by the Len Blavatnik and the Blavatnik Family foundation, and by the Yandex Initiative in Machine Learning.

References

Ailon, N., Karnin, Z. S., and Joachims, T. Reducing dueling bandits to cardinal bandits. In *ICML*, volume 32, pp. 856–864, 2014.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. In *Annual Conference on Learning Theory*, volume 40. Microtome Publishing, 2015.

Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Auer, P. Using upper confidence bounds for online learning. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pp. 270–279. IEEE, 2000.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

Azari, H., Parkes, D., and Xia, L. Random utility theory for social choice. In *Advances in Neural Information Processing Systems*, pp. 126–134, 2012.

Balsubramani, A., Karnin, Z., Schapire, R. E., and Zoghi, M. Instance-dependent regret bounds for dueling bandits. In *Conference on Learning Theory*, pp. 336–360, 2016.

Besbes, O., Gur, Y., and Zeevi, A. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

Brost, B., Seldin, Y., Cox, I. J., and Lioma, C. Multi-dueling bandits and their application to online ranker evaluation. *CoRR*, abs/1608.06253, 2016.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1, 2012.

Busa-Fekete, R. and Hüllermeier, E. A survey of preference-based online learning with bandit algorithms. In *International Conference on Algorithmic Learning Theory*, pp. 18–39. Springer, 2014.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

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

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, 2006.

Falahatgar, M., Hao, Y., Orlitsky, A., Pichapati, V., and Ravindrakumar, V. Maxing and ranking with few assumptions. In *Advances in Neural Information Processing Systems*, pp. 7063–7073, 2017.

Gajane, P., Urvoy, T., and Clérot, F. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 218–227, 2015.

Jamieson, K. G., Katariya, S., Deshpande, A., and Nowak, R. D. Sparse dueling bandits. In *AISTATS*, 2015.

Kocak, T., Neu, G., Valko, M., and Munos, R. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, pp. 613–621, 2014.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. Regret lower bound and optimal algorithm in dueling bandit problem. In *COLT*, pp. 1141–1154, 2015.

Komiyama, J., Honda, J., and Nakagawa, H. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. *arXiv preprint arXiv:1605.01677*, 2016.

Lattimore, T. and Szepesvári, C. Bandit algorithms. *preprint*, 2018.

Luo, H., Wei, C.-Y., Agarwal, A., and Langford, J. Efficient contextual bandits in non-stationary worlds. *arXiv preprint arXiv:1708.01799*, 2017.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pp. 684–692, 2011.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 3168–3176, 2015.

Neu, G. and Valko, M. Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, pp. 2780–2788, 2014.

Ramamohan, S. Y., Rajkumar, A., and Agarwal, S. Dueling bandits: Beyond condorcet winners to general tournament solutions. In *Advances in Neural Information Processing Systems*, pp. 1253–1261, 2016.

Ren, W., Liu, J., and Shroff, N. B. PAC ranking from pairwise and listwise queries: Lower bounds and upper bounds. *arXiv preprint arXiv:1806.02970*, 2018.

Saha, A. and Gopalan, A. Battle of bandits. In *Uncertainty in Artificial Intelligence*, 2018.

Saha, A. and Gopalan, A. PAC battling bandits in the plackett-luce model. In *Algorithmic Learning Theory*, pp. 700–737, 2019.

Saha, A., Gaillard, P., and Valko, M. Improved sleeping bandits with stochastic action sets and adversarial rewards. In *International Conference on Machine Learning*, pp. 8357–8366. PMLR, 2020.

Seldin, Y. and Lugosi, G. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. *arXiv preprint arXiv:1702.06103*, 2017.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *ICML*, pp. 1287–1295, 2014.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2):1–286, 2019.

Sui, Y., Zhuang, V., Burdick, J., and Yue, Y. Multi-dueling bandits with dependent arms. In *Conference on Uncertainty in Artificial Intelligence*, UAI'17, 2017.

Wu, H. and Liu, X. Double Thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 649–657, 2016.

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*, pp. 1201–1208. ACM, 2009.

Yue, Y. and Joachims, T. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 241–248, 2011.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Zoghi, M., Whiteson, S., Munos, R., Rijke, M. d., et al. Relative upper confidence bound for the k-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pp. 10–18. JMLR, 2014a.

Zoghi, M., Whiteson, S. A., De Rijke, M., and Munos, R. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 73–82. ACM, 2014b.

Zoghi, M., Karnin, Z. S., Whiteson, S., and De Rijke, M. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 307–315, 2015a.

Zoghi, M., Whiteson, S., and de Rijke, M. Mergerucb: A method for large-scale online ranker evaluation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 17–26. ACM, 2015b.