Think Before You Duel: Understanding Complexities of Preference Learning under Constrained Resources

Rohan Deb
University of Illinois,
Urbana Champaign

Aadirupa Saha
Apple

Arindam Banerjee
University of Illinois,
Urbana Champaign

Abstract

We consider the problem of reward maximization in the dueling bandit setup along with constraints on resource consumption. As in the classic dueling bandits, at each round the learner has to choose a pair of items from a set of \( K \) items and observe a relative feedback for the current pair. Additionally, for both items, the learner also observes a vector of resource consumptions. The objective of the learner is to maximize the cumulative reward, while ensuring that the total consumption of any resource is within the allocated budget. We show that due to the relative nature of the feedback, the problem is more difficult than its bandit counterpart and that without further assumptions the problem is not learnable from a regret minimization perspective. Thereafter, by exploiting assumptions on the available budget, we provide an EXP3 based dueling algorithm that also considers the associated consumptions and show that it achieves an \( \tilde{O}\left(\left(\frac{\text{OPT}}{B} + 1\right)K^{1/3}T^{2/3}\right) \) regret, where \( \text{OPT} \) is the optimal value and \( B \) is the available budget. Finally, we provide numerical simulations to demonstrate the efficacy of our proposed method.

1 Introduction

The standard Multi-Armed Bandit (MAB) setting involves an agent learning from stochastic feedback, provided in the form of numerical rewards [Lai and Robbins, 1985, Auer et al., 2002a, Lattimore and Szepesvári, 2020]. At every round \( t \in [T] \) the learner pulls an arm from \( K \) arms and the environment provides a reward \( r_t \) drawn i.i.d. from a distribution unknown to the learner. The objective of the learner is to maximize the cumulative reward over the time horizon \( T \). In many real world scenarios, e.g., movie recommendations, ad placements, retail management, tournament ranking, search engine optimization, one does not receive a numerical reward but rather receives feedback in terms of pairwise comparisons or rankings. In the simplest Dueling Bandits setup, at each round \( t \in [T] \) the learner picks two items \( i, j \) from \( K \) arms, and receives the output of a duel between the two, i.e., whether item \( i \) is preferred over \( j \) or vice versa. The objective of the learner is to minimize regret as compared against a best arm in hindsight. This setting has garnered a fair amount of attention over several years [Yue and Joachims, 2011, Zoghi et al., 2014a, Gajane et al., 2015, Ailon et al., 2014b, Zoghi et al., 2014b]. The simplicity and ease of data collection techniques of preference based learning gained huge interest in the online learning community. As a result, Dueling Bandits has been generalized and studied in multiple practical scenarios, e.g., extending pairwise preferences to subsetwise preferences [Saha and Gopalan, 2018a, Ren et al. 2018, Saha and Gopalan, 2020], finite to large or potentially infinite decision spaces [Kumagai, 2017, Yue and Joachims, 2009, Saha et al. 2021a], stochastic to adversarial preferences [Gajane et al., 2015, Saha et al., 2021b] and contextual scenarios [Dudík et al., 2015, Saha, 2021, Bengs et al., 2022], item unavailability [Gaillard et al., 2023], non-stationary preferences [Gupta and Saha, 2022] or even in interdisciplinary fields of research like robotics [Bhatia et al., 2020] and assortment optimization [Agrawal et al., 2019, Déjazet et al., 2016]. However, in many of these real-world scenarios, the agent has to minimize regret while operating under certain constraints, e.g., a limited supply of resources or a cost associated with each item. For e.g., ad placement is often constrained by the available advertiser budget and user reach, recommendation of movies might have a cost associated with each recommendation, and retail management might need to worry about logistical or
Think Before You Duel: Understanding Complexities of Preference Learning under Constrained Resources

supply constraints. In recent times, preference based feedback is also used to train more complex systems like assistive robots and autonomous cars which are computationally demanding and often resource constraints may limit the available resources. Therefore from the perspective of actually deploying Dueling bandit algorithms in real-world, it is essential to study the problem in a more general constrained setup - Constrained Dueling Bandits.

Informal Problem Setup: At every round, the learner picks arms $x_t, y_t \in [K]$ to duel, where $K$ is the number of arms, and observes a Bernoulli output with parameter $P(x_t, y_t)$. Here $P(x_t, y_t)$ measures the probability of $x_t$ being preferred over $y_t$ and therefore the matrix $P = [P(i, j)]_{i,j\in[K]}$ is called the preference matrix. Further it also observes some consumption of resources associated with arms $x_t$ and $y_t$. The objective of the learner is to minimize the regret over a time horizon while ensuring that the total sum of consumptions is less than a pre-defined budget (see Section 3 for a formal description).

1.1 Our Contributions

We provide an outline of our main contributions here. Note in the dueling setup, the quality of an item is relative, so to estimate the quality of one arm we need to compare it with the rest. The primary challenge lies in ensuring we can actually query all the pairs, while staying within the budget constraint. A straightforward extension of standard algorithms from Bandit with Knapsacks fails in DB setting because they draw arms in an UCB manner which leads to selecting the same arm twice, thus revealing no statistical information. Further existing elimination based algorithms for DB [Yue et al., 2012] [Yue and Joachims, 2011] [Zoghi et al., 2014a] cannot be extended to the constrained setting since once an arm is eliminated, an unbiased estimate of the corresponding scores cannot be obtained (also see Remark 5.1). Precisely our contributions are as follows.

1. Formulation: We setup the Constrained Dueling Bandits (Constrained-DB) problem by defining two kinds of benchmarks, corresponding to two types of best arms/wonners - Condorcet winner and the Borda winner, such that the benchmarks also satisfy the given constraints (see Section 3).

2. Lower Bounds: We show that the ‘relative’ nature of feedback makes Constrained-DB a difficult problem to solve (in comparison to its MAB counterpart). Specifically, we provide lower bound results for both Condorcet Constrained-DB and Borda Constrained-DB and show that the most general setup has a regret of $\Omega(T)$ and therefore one needs to impose additional assumptions either on the structure of the preference matrix or the available budget to give meaningful regret bounds (see Section 4).

3. Algorithms and Upper Bounds: Under assumptions on the available budget, we provide an EXP3 based algorithm Vigilant D-EXP3 that also takes into account the associated resource consumptions before choosing two arms to duel. Thereafter we show it achieves sub-linear regret (see Section 5).

4. Empirical Evaluations: We also evaluate our proposed algorithms empirically on synthetic and real data and show that they outperform the existing DB algorithms when there are budget constraints associated with arm selections (see Section 6).

2 Related Works

We briefly discuss some related literature here; for a more detailed discussion see Appendix A.

Dueling Bandits: The Dueling Bandits setting has seen a lot of development in the past decade. The problem in its current form was introduced in [Yue et al., 2012] and upper and lower bounds on the regret were provided by assuming that the preference matrix had some specific structures such as total ordering, strong stochastic transitivity and strong triangle inequality (also see Section 3 for definitions). Subsequently [Yue and Joachims, 2011] proposed ‘Beat the Mean’ algorithm with improved regret bound while also relaxed the strong stochastic transitivity assumption to relaxed stochastic transitivity. [Zoghi et al., 2014a] further relaxed the total ordering assumption to the existence of a Condorcet winner (an arm that beats every other arm) and provided a upper confidence bound (UCB) based algorithm. [Allon et al., 2014a] studied the dueling bandit problem in an adversarial setup (where the preference matrix $P$ changes over time), introducing the sparring EXP3 idea, albeit without regret guarantees. Subsequent works [Gajane et al., 2015] [Saha et al., 2021b] utilized this concept to prove regret guarantees in adversarial environments.

Constrained Bandits: There is a body of literature that under the name Bandits with Knapsacks that looks at cumulative reward maximization under budget constraints. It was first introduced in [Badanidiyuru et al., 2013] for the MAB setting and the proposed algorithms BalancedExploration and PrimalDualBwK were shown to enjoy optimal regret bounds up to polylogarithmic factors. BalancedExploration however is not an efficient algorithm (see Remark 4.2 in [Badanidiyuru et al., 2013]).
We define the total order relation. We define them as follows.

Next we consider two notions of winners in the dueling setup, namely the Borda winner and the Condorcet winner. We define them as follows.

**Definition 3.3** (Borda Winner) [Saha et al. 2021b]. We define the Borda score of an arm \( x \in [K] \) as

\[
b(x) = \frac{1}{K-1} \sum_{x \neq y} P(x, y)
\]

The Borda winner \( x^{(b)} \) is defined as the arm that maximizes the Borda score, i.e., \( x^{(b)} \) = argmax\( x \) \( b(x) \).

**Definition 3.4** (Strong Stochastic Transitivity) [Yue et al. 2012] [Yue and Joachims 2011]. We say the preference matrix \( P \) satisfying TO condition further satisfies Strong Stochastic Transitivity (SST) if for every \( i, j, k \in [K] \), \( i \preceq j \preceq k \) implies \( P(i, k) \geq \max\{P(i, j), P(j, k)\} \) where \( \preceq \) is the underlying TO relation.

### 3.2 Our Problem Setup

While the problem of DB is well studied over the past two decades (see section 3), the literature lacks the practical setup of considering the aspects resource constraints in Dueling Bandits, which is often realizable in practical scenarios as we motivated in section 1. In this section, we define the constrained dueling bandit setup.

**Constrained Dueling Bandits.** At round \( t \in [T] \) when the learner selects two arms \( x_t, y_t \in [K] \) to duel, it also observes two consumption vectors \( u(x_t), v(y_t) \in [0, 1]^d \) associated with the pulled arms \( x_t \) and \( y_t \), drawn independent of the past history from an unknown distribution. The \( d \) elements of the vector are the consumptions associated with \( d \) different types of resources. We define \( u^*(\cdot) \) and \( v^*(\cdot) \) as the expected consumptions of the two arms respectively, i.e., \( u^*(x) = \mathbb{E}[u(x)] \), and \( v^*(x) = \mathbb{E}[v(x)] \). The learner also has the option of not choosing any arm at a round and see no feedback and incur no consumption. The total budget available to the learner is \( B \leq T \) and the interaction with the environment ends at \( t = \tau \) when either \( \tau = T \) (end of time horizon) or \( \sum_{t=1}^\tau u_t(x_t) + u_t(y_t) \rangle > B \) for some \( i \in [d] \) (the budget of some resource is exhausted).

**Benchmarks.** Next we describe the benchmarks in the two settings that our algorithms compete against. Suppose \( \pi_x(\cdot) \) and \( \pi_y(\cdot) \) represent the distribution of arms played in the first selection and second selection of the dual respectively. Then the two optimal solutions are defined as:

1. **Condorcet Optimal Solution** Consider the following Linear program (LP) with Condorcet score.

\[
\max_{\pi_x, \pi_y \in \Delta^K} \sum_{x, y \in [K]} \pi_x(x) c(x) + \pi_y(y)c(y),
\]

such that \( \sum_{x, y \in [K]} \pi_x(x) u^*(x) + \pi_y(y) u^*(y) \leq \frac{B}{T} \cdot \mathbb{1} \).

(LP − Condorcet)

where \( \Delta^K \) is the probability simplex over \( K \). Suppose \( \pi_x^{(c)}, \pi_y^{(c)} \) solve (LP − Condorcet) then we define the optimal value as

\[
\text{OPT}^{(c)} = T \sum_{x, y \in [K]} \pi_x^{(c)}(x) c(x) + \pi_y^{(c)}(y)c(y).
\]
2. Borda Optimal Solution Consider the following Linear program (LP) with Borda score.
\[
\begin{aligned}
\max_{\pi_x, \pi_y \in \Delta^K} & \quad \sum_{x, y \in [K]} \pi_x(x) b(x) + \pi_y(y) b(y), \\
\text{such that} & \quad \sum_{x, y \in [K]} \pi_x(x) w^*(x) + \pi_y(y) v^*(y) \leq \frac{B}{T},
\end{aligned}
\]
(LP – Borda)
where \(\Delta^K\) is the probability simplex over \(K\). Suppose \(\pi_x^{(b)}, \pi_y^{(b)}\) solve LP – Borda then we define the optimal value as
\[
\text{OPT}^{(b)} = T \sum_{x, y \in [K]} \pi_x^{(b)}(x) b(x) + \pi_y^{(b)}(y) b(y).
\]

Remark 3.1. Note that the benchmarks (LP – Condorcet) and (LP – Borda) compute a non-adaptive policy and follows the development in Agrawal and Devanur [2010]. It can be shown that \(\text{OPT}^{(c)}\) and \(\text{OPT}^{(b)}\) upper bounds the value of the corresponding optimal adaptive policy (e.g., Lemma 1 in Agrawal and Devanur [2010]).

Regret Next we define the total cumulative regret of an algorithm that chooses the sequence of arms \(\{(x_t, y_t)\}_{t=1}^T\), where \(\tau \leq T\) is the stopping time of the algorithm. We define the following types of regret.

1. Condorcet Regret. We define the cumulative Condorcet reward until the stopping time \(\tau\) as
\[
\text{REW}^{(c)} = E \sum_{t=1}^{\tau} c(x_t) + c(y_t), \tag{1}
\]
and the corresponding Condorcet regret as
\[
\text{REG}^{(c)}(T) = \text{OPT}^{(c)} - \text{REW}^{(c)} \tag{2}
\]
2. Borda Regret. We define the cumulative Borda reward until the stopping time \(\tau\) as
\[
\text{REW}^{(b)} = E \sum_{t=1}^{\tau} b(x_t) + b(y_t), \tag{3}
\]
and the corresponding Borda regret as
\[
\text{REG}^{(b)}(T) = \text{OPT}^{(b)} - \text{REW}^{(b)} \tag{4}
\]

Objective Informally the objective is to maximize the total sum of rewards while satisfying the budget constraint. Formally, the algorithm competes with the benchmarks defined in LP – Condorcet and LP – Borda and the performance is measured via the regret defined in (2) and (4).

Notation. For ease of exposition, we shall hide dependencies on constants and work with order notation. Towards that, we shall use the notation \(n_0 = O(t)\) to imply that there exists constant \(c\) (independent of \(t\)) such that \(\leq n_0 \leq ct\). The notation \(\tilde{O}(t)\) has a similar meaning but hides the dependence on logarithmic terms. Further, \(n_0 = \Omega(t)\) implies there exists such that \(n_0 \leq ct\) and \(n_0 = o(t)\) implies \(\lim \frac{n_0}{t} \rightarrow 0\).

4 Lower Bounds
We first analyze the lower bound for the Constrained-DB problem to analyze the problem complexity and achievable regret performance. Detailed proofs given in Appendix B.

4.1 Lower Bounds for Condorcet

Constrained-DB

We state two lower bound results.

(1) Lemma 4.1 states that in the most general setting, if the allocated budget \(B = o\left(\frac{K}{\epsilon^{(c)}_{\min}}\right)^2\) then the regret of any algorithm for Condorcet-constrained-DB is \(\Omega(T)\), where \(\epsilon^{(c)}_{\min}\) is the minimum gap in the Condorcet scores. Further the bound does not improve even if we assume that the preference matrix satisfies total ordering but does improve if we further assume that the preference matrix satisfies strong stochastic transitivity.

(2) Lemma 4.2 states that if the budget \(B = o\left(\frac{K}{\epsilon^{(c)}_{\min}}\right)^2\) then any algorithm for Condorcet-constrained-DB has regret \(\Omega(T)\).

Remark 4.1. The \(\Omega(T)\) regret in (2) cannot be improved with any structural assumptions on the preference matrix (such as TO or SST) and a similar result can be shown to hold in the Constrained MAB setting. The \(\Omega(T)\) regret in (1) is far more interesting because as we will show in the sequel, this arises precisely because of the interplay between relative feedback and budget constraints and could potentially be improved either by assuming some structure in the preference matrix (specifically SST in this case) or that the agent has enough budget \(B \geq \frac{K}{\epsilon^{(c)}_{\min}}\).

Lemma 4.1. Consider the Constrained-DB setting with preference matrix \(P\) and define the minimum gap in Condorcet scores \(\epsilon^{(c)}_{\min} = \min_{i, j \in [K]} \{ |c(i) - c(j)| \}\). Suppose the available budget \(B = o\left(\frac{K}{\epsilon^{(c)}_{\min}}\right)^2\) then there exists a preference matrix \(P\) such that \(\text{REG}^{(c)}(T) = \Omega(T)\).

Further we show that our \(\Omega(T)\) regret bound exists even
when $P$ satisfies total ordering (cf. Definition 3.3) but not when $P$ satisfies strong stochastic transitivity (cf. Definition 3.4).

**Proof sketch.** We outline the idea behind the creation of the lower bound example here. For ease of exposition we consider a simplified setup with $K = 3$. We start with the general setting without any assumption on the preference matrix and subsequently consider total ordering and strong stochastic transitivity.

**General setting:** Suppose the preference matrix $P$ is given by

$$P = \begin{bmatrix}
  \frac{1}{2} & \frac{1}{2} + \epsilon & \frac{1}{2} + 2\epsilon \\
  \frac{1}{2} - \epsilon & \frac{1}{2} & \frac{1}{2} \\
  \frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} 
\end{bmatrix}$$

and the true consumption of the three arms are given by $u^*(1) = v^*(1) = 1$, $u^*(2) = v^*(2) = 0$ and $u^*(3) = v^*(3) = 0$.

The optimal policy plays arms $x_t = 1, y_t = 1$ for $B$ rounds and thereafter plays arms $x_t = 2, y_t = 2$ for the remaining $T-B$ rounds. The total accumulated reward by the optimal policy is $\text{OPT}^{(c)} = B + (T-B)(\frac{1}{2} - \epsilon)$.

Note that arm $1$ is the Condorcet winner and any algorithm needs to play the pairs $(1,2)$ and $(1,3)$ at least $\frac{1}{T}$ number of times to determine if $c(2) > c(3)$. However since the budget $B = o\left(\frac{T}{2}\right)$, no algorithm can determine if $c(2) > c(3)$ or $c(2) < c(3)$ and hence would always end up playing the sub-optimal arm at least $\frac{1}{T-B}$ number of times after the initial $B$ rounds. Therefore

$$\text{REW}^{(c)} \leq B + \frac{(T-B)}{2} \left(\frac{1}{2} - \epsilon\right) + \frac{(T-B)}{2} \left(\frac{1}{2} - 2\epsilon\right)$$

which implies $\text{OPT}^{(c)} \geq \frac{(T-B)}{2} \epsilon = \Omega(T)$.

The key observation here is that although arm $2$ and arm $3$ have zero consumption playing the pair $(2,3)$ does not provide any information about whether $s(2) > s(3)$ or $s(3) > s(2)$. This is in contrast to the standard MAB setting where playing arms $2$ and $3$ does give information about whether $s(2) > s(3)$ or $s(2) < s(3)$.

**Total Ordering:** Next consider the preference matrix $P$ below that satisfies total ordering.

$$P = \begin{bmatrix}
  \frac{1}{2} & \frac{1}{2} + \epsilon & \frac{1}{2} + 2\epsilon \\
  \frac{1}{2} - \epsilon & \frac{1}{2} & \frac{1}{2} + \epsilon \\
  \frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} 
\end{bmatrix}$$

with the same consumptions as before. Does playing the pair $(2,3)$ give us any information about $s(2) > s(3)$? The answer is no. This is because we have another instance with

$$P' = \begin{bmatrix}
  \frac{1}{2} & \frac{1}{2} + \epsilon & \frac{1}{2} + 2\epsilon \\
  \frac{1}{2} - \epsilon & \frac{1}{2} & \frac{1}{2} - \epsilon \\
  \frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} 
\end{bmatrix}$$

such that although $s(2) > s(3)$, in the total ordering sense $3 > 2$. Therefore the algorithm cannot distinguish between the instances with preference matrices $P'$ and $P$.

**Strong Stochastic Transitivity:** Finally suppose we assume that the Preference matrix follows strong stochastic transitivity. With this assumption, notice that the instance $P'$ is not allowed and therefore the algorithm may learn about $s(2) > s(3)$ from the total order relation $2 > 3$ by playing the pair $(2,3)$ without consuming any resources.

The next Lemma shows that if the available budget is less than $\frac{K}{\epsilon_{\min}}$ then the regret of any algorithm is $\Omega(T)$.

**Lemma 4.2.** Consider the Constrained-DB setting with preference matrix $P$ and define the minimum gap in Condorcet scores $\epsilon_{\min} := \min_{i,j \in [K]} (|c(i) - c(j)|)$. Suppose the available budget $B = o\left(\frac{K}{\epsilon_{\min}}\right)$ then there exists a preference matrix $P$ such that $\text{REG}^{(c)}(T) = \Omega(T)$.

### 4.2 Lower Bounds for Borda Constrained-DB

We state similar lower bounds for the Borda Constrained Dueling Bandits setting.

**Lemma 4.3.** Consider the Constrained-DB setting with preference matrix $P$ and define the minimum gap in Borda scores $\epsilon_{\min} := \min_{i,j \in [K]} (|b(i) - b(j)|)$. Suppose the available budget $B = o\left(\frac{K}{\epsilon_{\min}}\right)$ then there exists a preference matrix $P$ such that $\text{REG}^{(b)}(T) = \Omega(T)$.

The above lower bound could potentially be removed by assuming total ordering.

**Lemma 4.4.** Consider the Constrained-DB setting with preference matrix $P$ and define the minimum gap in Borda scores $\epsilon_{\min} := \min_{i,j \in [K]} (|b(i) - b(j)|)$. Suppose the available budget $B = o\left(\frac{K}{\epsilon_{\min}}\right)$ then there exists a
Think Before You Duel: Understanding Complexities of Preference Learning under Constrained Resources

preference matrix $P$ such that $\text{REG}^{(b)} = \Omega(T)$.

The above lower bound cannot be removed even by assuming total ordering or strong stochastic transitivity.

5 Algorithm and Regret Bound

We consider the Borda Constrained Dueling Bandits under the assumption that the given budget $B = \mathcal{O}\left(\max\left\{\frac{K}{\varepsilon_{\min}}, T^{3/4}\right\}\right)$. Our algorithm Vigilant D-EXP3 (Dueling EXP3) is outlined in Algorithm 1. Before proceeding to a detailed description of our algorithm, we modify the Borda benchmark in two ways that do not change the benchmark value $\text{OPT}^{(b)}$ by more than a constant factor.

1. Shifted Borda Score [Saha et al., 2021b] We replace the Borda score $b(x)$ in $\text{LP - Borda}$ by the shifted Borda score $\tilde{b}(x)$ defined below.

**Definition 5.1 (Shifted Borda Score).** The shifted Borda score of item $i \in [K]$ is given by

$$\tilde{b}(x) := \frac{1}{K} \sum_{j \in [K]} P(i, j).$$

**Benchmark with shifted Borda score.** We define the benchmark LP with shifted Borda score as

$$\text{max}_{\pi_x, \pi_y \in \Delta^K} \sum_{x \in [K]} \pi_x(x) \tilde{b}(x) + \sum_{y \in [K]} \pi_y(y) \tilde{b}(y)$$

$$\text{such that } \sum_{x, y \in [K]} \pi_x(x) u^*(x) + \pi_y(y) v^*(y) \leq \frac{B}{T} 1$$

$$\text{(LP - Shifted-Borda-x)}$$

Let the solution to the above LP be $\tilde{\pi}^{(b)}$ and let

$$\tilde{\text{OPT}}^{(b)} = T \sum_{x, y \in [K]} \tilde{\pi}_x^{(b)} \tilde{b}(x) + \tilde{\pi}_y^{(b)} \tilde{b}(y)$$

$$\tilde{\text{REW}}^{(b)} = \mathbb{E} \sum_{t=1}^T \tilde{b}(x_t) + \tilde{b}(y_t).$$

Then in the following lemma we show that the original optimal value and the total reward is constant times the optimal value and total reward respectively, with shifted Borda score.

**Lemma 5.1.** For $\tilde{\text{OPT}}^{(b)}$ and $\tilde{\text{REW}}^{(b)}$ as defined in (5) and (6) we have

$$\text{REG}^{(b)} = \text{OPT}^{(b)} - \text{REW}^{(b)}$$

$$\leq \frac{K}{K-1} (\tilde{\text{OPT}}^{(b)} - \tilde{\text{REW}}^{(b)}).$$

Therefore bounding $\tilde{\text{OPT}}^{(b)} - \tilde{\text{REW}}^{(b)}$ bounds the final regret $\text{REG}^{(b)}$ up to a constant factor of $\frac{K}{K-1}$.

2. Define Separate LPs. Next we relax the benchmark [LP - Shifted-Borda-x] by separating the LPs associated with the two arms as defined below.

$$\text{max}_{\pi_x \in \Delta^K} \sum_{x \in [K]} \pi_x(x) \tilde{b}(x)$$

$$\text{such that } \sum_{x \in [K]} \pi_x(x) u^*(x) \leq \frac{B}{2T} 1$$

$$\text{(LP - Shifted-Borda-x)}$$

Following lemma proves that the sum of optimal values of the two LPs upper bounds the value of the optimal policy $\text{OPT}^{(b)}$.

**Lemma 5.2.** Let the optimal value of $\text{LP - Shifted-Borda-x}$ and $\text{LP - Shifted-Borda-y}$ be $\text{OPT}_x^{(b)}$ and $\text{OPT}_y^{(b)}$. Then

$$\tilde{\text{OPT}}^{(b)} + \tilde{\text{OPT}}^{(b)} \geq \tilde{\text{OPT}}^{(b)}.$$
Algorithm 1 Vigilant D-EXP3

1: Input: Item set indexed by \([K]\), learning rate \(\eta > 0\), exploration parameter \(\gamma \in (0, 1)\), and \(\mathcal{O}(\frac{\text{OPT}^{(b)}}{B}) \leq Z_w \leq \mathcal{O}(\frac{\text{OPT}^{(b)}}{B} + 1), w \in \{x, y\}\)

2: Initialize: Initial probability distribution \(q^0_i = (1/K, \forall i \in [K])\)

3: for \(t = 1, \ldots, T \) do

4: Sample \(x_t \sim q^x_t, y_t \sim q^y_t\) i.i.d.

5: Receive preference \(a_t(x_t, y_t) \sim \text{Ber}(P_t(x_t, y_t))\) and the consumption vectors \(u_t(x_t)\) and \(v_t(y_t)\).

6: Estimate the shifted Borda scores and the consumption vectors, for all \(i \in [K]\):

\[
\hat{b}_t(i) = \frac{1}{K} \sum_{j \in [K]} \frac{1}{q^y_t(j)} (1 - u_t(x_t)),
\]

\[
\hat{a}_t^x(i) = 1 - \frac{1}{q^x_t(i)} (1 - v_t(y_t)).
\]

7: Estimate the Lagrangians \(\forall i \in [K]\)

\[
\hat{\ell}_t^x(i) = \hat{b}_t(i) + Z_t \hat{\lambda}_t^x \left[ \frac{B}{2T} - \hat{a}_t^x(i) \right],
\]

\[
\hat{\ell}_t^y(i) = \hat{b}_t(i) + Z_t \hat{\lambda}_t^y \left[ \frac{B}{2T} - \hat{a}_t^y(i) \right].
\]

8: Update for all \(i \in [K]\):

\[
\hat{q}^x_{t+1}(i) = \frac{\exp(\eta_x \sum_{s=1}^K \hat{\ell}_s^x(i))}{\sum_{j=1}^K \exp(\eta_x \sum_{s=1}^K \hat{\ell}_s(j))},
\]

\[
q^x_{t+1}(i) = (1 - \gamma_x)q^x_{t+1}(i) + \frac{\gamma_x}{K},
\]

\[
\hat{q}^y_{t+1}(i) = \frac{\exp(\eta_y \sum_{s=1}^K \hat{\ell}_s^y(i))}{\sum_{j=1}^K \exp(\eta_y \sum_{s=1}^K \hat{\ell}_s(j))},
\]

\[
q^y_{t+1}(i) = (1 - \gamma_y)q^y_{t+1}(i) + \frac{\gamma_y}{K}.
\]

9: Update \(\lambda^x_t\) and \(\lambda^y_t\) using any online convex optimization on the following objective functions

\[
g^x(\lambda) = \left( \frac{B}{2T} - \hat{a}_t^x(x_t) \right),
\]

\[
g^y(\lambda) = \left( \frac{B}{2T} - \hat{a}_t^y(y_t) \right).
\]

Remark 5.1 (Overcoming Challenges). Although we follow the primal dual-approach from Agrawal et al. 2016, Agrawal and Devanur 2014, Sivakumar et al. 2022, we do not construct a UCB estimate of the lagrangian (by constructing UCB estimates of the rewards and LCB estimates of the consumptions) and draw the arm optimistically. This is because in the Dueling setting, such an approach would lead to choosing the same arm twice which does not reveal any statistical information since \(P(i, i) = 1/2, \forall i \in [K]\) is already known. Further the approach for RUCB Zoghi et al. 2014a. Beat the mean Yue and Joachims 2011 or Interleaved filter Yue et al. 2012 cannot be extended to the constrained setting since these algorithms are elimination algorithms, and once an arm is eliminated, an unbiased estimate of the Borda score cannot be constructed which is essential to do a trade-off between the Borda score and the consumptions of associated arms.

Remark 5.2 (Unknown \(Z\)). Note that although our algorithm assumes that the values \(\text{OPT}^{(x)}\) and \(\text{OPT}^{(y)}\) are known, in the case they are unknown, the standard trick of estimating \(Z_x\) and \(Z_y\) for the first \(\mathcal{O}(\sqrt{T})\) rounds can be used (see e.g., Section 3.3 in Agrawal and Devanur 2016). This requires the budget to be \(B = \Omega(T^{3/4})\).

Theorem 5.1. For \(\eta_x = (\frac{\log K}{T Z_y})^{2/3}, \eta_y = (\frac{\log K}{T Z_y})^{2/3}, \gamma_x = \sqrt{\eta_x K Z_x}, \gamma_y = \sqrt{\eta_y K Z_y}\), the regret of Vigilant D-EXP3 is bounded by

\[
\text{REG}^*(T) \leq \tilde{O}\left( (\frac{\text{OPT}^{(b)}}{B} + 1)(K \log K)^{1/3} T^{2/3} \right).
\]

Proof sketch Here we briefly outline the steps of the proof. For details see Appendix \(\Box\).

Step 1: We use an EXP-3 kind guarantee for the first arm to get the following bound for all \(a \in [K]\),

\[
\sum_{t=1}^\tau \hat{\ell}_t^x(a) - \sum_{t=1}^\tau \sum_{a=1}^K \hat{q}^x_t(a)\hat{\ell}_t^x(a) 
\leq \frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^\tau \sum_{a=1}^K \hat{q}^x_t(a)\hat{\ell}_t^x(a)^2.
\]

Since \(\hat{q}^x_t(i) = \frac{q^x_t(i) - \gamma_x}{1 - \gamma_x}\),

\[
\forall a, (1 - \gamma_x) \sum_{t=1}^\tau \hat{\ell}_t^x(a) - \sum_{t=1}^\tau \sum_{a=1}^K \hat{q}^x_t(a)\hat{\ell}_t^x(a) 
\leq \frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^\tau \sum_{a=1}^K \hat{q}^x_t(a)\hat{\ell}_t^x(a)^2 \quad (7)
\]

Step 2: Since the LHS in (7) holds for every \(a \in [K]\) we relate it to the regret \(\text{OPT}^{(b)} - \sum_{t=1}^\tau \hat{b}(x_t)\) using the following lemma.
Further, $\gamma_x(Z_x+1)T \leq O\left((K \log K)^{1/3}Z_xT^{2/3}\right)$

Next we upper bound the RHS in \((\ref{eq:opt-b-lemma})\) using the following lemma.

Lemma 5.4. For $\eta_x = \left(\frac{\log K}{T\sqrt{K}}\right)^{2/3} \frac{1}{2Z_x+1}$ and $\gamma_x = \sqrt{\eta_xZ_x}$

$$\frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^{\tau} \sum_{a=1}^{K} q^\tau_t(i)(\hat{p}_t^\tau(i))^2$$

$$\leq O\left(\frac{OPT^b_x}{B} + 1\right)(K \log K)^{1/3}T^{2/3}$$

Step 3: We repeat the same argument for the second arm choice and then combining with Lemma 5.1 and Lemma 5.2 completes the proof.

6 Experiments

We test our proposed algorithm Vigilant D-EXP3 on both synthetic and real world data in the constrained setting against existing Borda dueling bandit algorithms that do not factor in the associated consumptions. We briefly describe our datasets, benchmarks and results here (also see Appendix D).

Datasets. We run our experiments on two datasets.

1. Synthetic Data: We create a Constrained Dueling Bandits instance with $K = 6$ arms (see Appendix D for the exact description of the preference matrix). The vector of Borda scores $\bar{b} = (b(1), b(2), 6)$ is given by $(0.672, 0.646, 0.602, 0.582, 0.554, 0.544)^\top$. We experiment with three choices of consumptions. In all three cases the number of resources $d = 1$ and the true consumptions across both arms choices are given by the same function, i.e., $u^* = v^*$, and we add zero mean gaussian noise to each entry. The vector of consumptions for arms $\bar{u}^* = (u^*(1), u^*(2), 6)$ are given by $(0.9, 0.9, 0.1, 0.8, 0.8, 0.8)^\top$, $(0.1, 0.2, 0.3, 0.4, 0.5, 0.6)^\top$, and $(0, 0, 0, 0, 0, 0)^\top$. In the first case although arm 1 and 2 have high Borda scores, the associated consumptions are also high. In the second case the order of consumptions is the same as the order of Borda scores. In the last case all the consumptions are zero and our objective is to evaluate if our algorithm under performs in the absence of constraints. The experiments are run for $T = 2000$ rounds with $B = 1000$ and are run independently over 50 samples.

2. Car preference dataset: We consider the Car preference dataset from E. et al. [2013] where the preference matrix is generated by considering the user preferences for various models of cars. As in case 1, we consider three choices of consumptions that follow a similar structure (see Appendix D for more details). The experiments are run for $T = 5000$ rounds with $B = 4000$ and are run independently over 50 samples.

Benchmarks. We compare our algorithm against the following two choices of DB algorithms for Borda scores.

1. D-EXP3: Dueling EXP3 algorithm (Algorithm 1, Saha et al. [2021b]) runs an exponential weights algorithm with uniform exploration on the estimated Borda scores.

2. D-TS: Dueling Thompson Sampling (Algorithm 2, Lekang and Lamperski [2019]) runs a Thompson sampling algorithm by learning true parameter values, which can represent the preference matrix directly or by some other latent values for each action, by sampling the posterior distribution conditioned on the history. As in Lamperski [2019], we use $K^2 - K$ independent Beta(1, 1) as our prior.

Results. Figure 1 and Figure 2 plot the cumulative rewards across different rounds. For both datasets in the first two cases when it is not prudent to stick to the arms that have high Borda scores since they also have high associated consumptions, the benchmark algorithms earn more reward in the initial few rounds but drop early since they run out of budget. In contrast, our algorithm Vigilant D-EXP3, does take into account the associated consumptions and runs for far more number of rounds before exhausting its resources and therefore ends up acquiring far more reward. However in the unconstrained case although the performance of Vigilant D-EXP3 is almost same as D-EXP3, D-TS outperforms them both and as such an immediate direction of study would be to develop a constrained version of Dueling Thompson Sampling and compare its performance against Vigilant D-EXP3.
7 Conclusion

In this work we have developed the framework for learning from preference feedback under resource constraints and developed several lower bounds both with Borda scores and Condorcet scores that show that the setting is strictly more difficult than its multi-armed counterpart. Under the assumption that the resource budget is high, we developed an EXP3 based algorithm that via the lagrangian also takes into account the associated consumptions of a duel rather than just the associated scores. We show that the algorithm enjoys sub-linear regret bound and performs far better on both synthetic and real world data.

The Condorcet constrained DB problem is a strictly harder problem, since to be able to compute an estimate of the Condorcet score one needs to know the identity of the Condorcet winner. Further none of the elimination based algorithms for Condorcet winner can be used here since if the winner is eliminated based say on the lagrangian value, then further estimates of the scores cannot be computed and as such makes the problem quite challenging and developing algorithms for this setting is an interesting future work. Moreover as observed in Section 6 Duelling Thompson sampling appears to perform better in the unconstrained setting and as such it might be instructive to develop a constrained version of the algorithm. Finally, it would certainly be useful to consider more general and practical settings of dueling bandits.

Acknowledgement

The work supported in part by grants from the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award.
References


A. Saha and A. Gopalan. Active ranking with subset-wise preferences. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018b.


A. Slivkins. Introduction to multi-armed bandits, 2022.


Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. [Yes/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]
A Extended Related Works

In Section 2 we provided an overview of some existing works in Dueling Bandits and Constrained Bandits. Here we provide a more comprehensive study of these two areas.

- **Dueling Bandits:**
  
  Over the past decade, the Dueling Bandits setting has undergone significant advancements. The problem, in its current form, was introduced by Yue and Joachims (2012) [Yue et al., 2012], who established upper and lower bounds on regret. These bounds were established under the assumption that the preference matrix has specific structures, such as total ordering, strong stochastic transitivity, and strong triangle inequality (refer to Section 3 for definitions).

  Subsequently [Yue and Joachims, 2011] proposed ‘Beat the Mean’ algorithm with improved regret bound while also relaxed the strong stochastic transitivity assumption to relaxed stochastic transitivity. Zoghi et al. [2014a] further relaxed the total ordering assumption to the existence of a Condorcet winner (an arm that beats every other arm) and provided a upper confidence bound (UCB) based algorithm.

  Zoghi et al. [2015b] examine the same problem, but emphasizes learning situations where a vast number of arms are available. To minimize the number of comparisons, the authors introduce the MergeRUCB algorithm that employs a divide-and-conquer strategy akin to the merge sort algorithm. It begins by organizing the arms into predefined batches, processing each batch independently before combining the results. Zoghi et al. [2015a], Komiyama et al. [2016] study the dueling bandits problem focusing on Copeland winners and Li et al. [2020] suggest a Thompson sampling based algorithm to solve it.

  Ailon et al. [2014a] studied the dueling bandit problem in an adversarial setup (where the preference matrix $P$ changes over time), introducing the sparring EXP3 idea, albeit without regret guarantees. Subsequent works Gajane et al. [2015], Saha et al. [2021b], Saha and Gupta [2022] utilized this concept to prove regret guarantees in adversarial environments. Saha et al. [2021b] consider the Borda winner instead of Condorcet winner, which is known to always exist. Lekang and Lamperski [2019] provided both Thompson sampling and sparring EXP3 type algorithm for both maxmin and Borda winners along with regret guarantees and show impressive improvement on performance against existing benchmark algorithms.

  Over the last two decades the field of Dueling Bandits received significant attention due to the simplicity and effectiveness of the problem framework. Consequently Dueling Bandits has been generalized and studied for various real world problems, including but not limited to, extending pairwise preferences to subsetwise preferences Saha and Gopalan [2019, 2018b], Sui et al. [2017], Saha and Ghoshal [2022], large decision spaces Saha [2021], Saha et al. [2022], adversarial preferences Gupta and Sahai [2022], and contextual scenarios Saha and Krishnamurthy [2022], Balasubramani et al. [2016], item unavailability Saha and Gaillard [2021], non-stationary preferences Buining and Saha [2023]. Consequently, the framework of Dueling Bandit has also been adapted to other interdisciplinary fields of research, e.g., reinforcement learning Saha et al. [2023], Blum et al. [2023], robotics Lee et al. [2021], Li et al. [2023], language models Christiano et al. [2017], Ouyang et al. [2022] and assortment optimization Agrawal et al. [2019], Desir et al. [2016]. A detailed survey of Preference Bandit literature could be found in Bengs et al. [2021], Sui et al. [2018].

- **Constrained Bandits:**
  
  There is a body of literature that under the name Bandits with Knapsacks that looks at cumulative reward maximization under budget constraints. It was first introduced in Badanidiyuru et al. [2013] for the MAB setting and the proposed algorithms BalancedExploration and PrimalDualBwK were shown to enjoy optimal regret bounds up to polylogarithmic factors. BalancedExploration however is not an efficient algorithm (see Remark 4.2 in Badanidiyuru et al. [2013]) and we do not pursue this. We further show that in fact PrimalDualBwK attains an $\Omega(T)$ regret in the MAB setting, and as such we do not try to adapt this algorithm to the Constrained-DB setting.

  Subsequently, more general versions of the problem have been studied, e.g., in the linear contextual setting, Agrawal and Devanur [2016] provide an algorithm utilizing ideas from UCB and primal-dual methods with sub-linear regret bound. The idea is to maintain a UCB estimate of the associated lagrangian by maintaining a UCB estimate of the rewards and an LCB estimate of the consumptions and the arm being played is the one that maximized the lagrangian optimistically, while the lagrange multiplier is updated via a dual
optimization step. In the fully adversarial setting, Immorlica et al. [2022] show that regret minimization is not feasible and therefore provide guarantees on the competitive ratio of the proposed algorithm. In the smooth adversarial setting, where in the contexts and rewards are chosen by an adaptive adversary but nature perturbs it using a small gaussian noise, Sivakumar et al. [2022] provide sub-linear regret again using a primal-dual idea from Agrawal and Devanur [2016].

More general versions, where only realizability of the reward and consumptions by a function class is assumed have been studied in Slivkins et al. [2023], Han et al. [2023] where the Inverse gap weighting idea from Abe and Long [1999] has been employed to provide sub-linear regret bounds by computationally efficient algorithms. Finally recent works have also started study regret guarantees of algorithms in the reinforcement learning setting, where instead of a single state, the agent interacts with the environment through a sequence of states and actions Ding et al. [2021], Vaswani et al. [2022], Kalagarla et al. [2021], Wei et al. [2022].

B Proof of Lower Bounds

B.1 Lower Bounds for Condorcet Constrained-DB

Lemma 4.1. Consider the Constrained-DB setting with preference matrix $P$ and define the minimum gap in Condorcet scores $c_{\min}^{(c)} := \min_{i,j \in [K]} (c(i) - c(j))$. Suppose the available budget $B = o\left(\frac{K}{c_{\min}^{(c)}}\right)$ then there exists a preference matrix $P$ such that $REG^{(c)}(T) = \Omega(T)$. Further we show that our $\Omega(T)$ regret bound exists even when $P$ satisfies total ordering (cf. Definition 3.3) but not when $P$ satisfies strong stochastic transitivity (cf. Definition 3.4).

Proof. General setting: We start with the general setting without any assumption on the preference matrix and subsequently consider total ordering and strong stochastic transitivity. The proof relies on creating $K - 1$ problem instances that we denote by $I_1, \ldots, I_{K-1}$.

$I_1$ is defined by the following preference matrix:

$$P = \begin{bmatrix}
\frac{1}{2} & 1 + \epsilon & \frac{1}{2} + 2\epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} & 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\
\frac{1}{2} - \epsilon & \frac{1}{2} & \cdots & \frac{1}{2} \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \cdots & \frac{1}{2} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \cdots & \frac{1}{2} \\
\end{bmatrix}$$

with $u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0, \forall i \in \{2, \ldots, K\}$.

$I_2$ is defined by the following preference matrix:

$$P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} + 2\epsilon & \frac{1}{2} + \epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\
\frac{1}{2} - \epsilon & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\
\end{bmatrix}$$
with $u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0, \forall i \in \{2, \ldots, K\}$ and so on with $\mathcal{I}_{K-1}$ is defined by the following preference matrix:

$$P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} + 2\epsilon & \cdots & \frac{1}{2} + \epsilon & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \cdots & \frac{1}{2} & \frac{1}{2} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\frac{1}{2} - \epsilon & \frac{1}{2} & \cdots & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2}
\end{bmatrix}$$

with $u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0, \forall i \in \{2, \ldots, K\}$.

The optimal policy in $\mathcal{I}_k$ plays arms $x_t = 1, y_t = 1$ for $B$ rounds and thereafter plays arms $x_t = k+1, y_t = k+1$ for the remaining $T-B$ rounds. The total accumulated reward by the optimal policy is $\text{OPT}^{(c)} = B + (T-B)\left(\frac{1}{2} - \epsilon\right)$.

Note that to be able to switch to arm $k + 1$ in instance $\mathcal{I}_k$ the algorithm needs to identify the arm with the highest Condorcet score $c(i)$ among the arms $\{2, 3, \ldots, K\}$. This reduces to the best arm identification problem in multi armed bandits with $K-1$ arms with rewards $c(i), i \in [K]$. It is known that the sample complexity of best arm identification is $\Omega(K^{\frac{2}{3}})$ (see eg. [Slivkins, 2022]) and therefore any algorithm needs to play $(1, k), k \in [K]$ $\Omega(K^{\frac{2}{3}})$ number of times. However every time $(1, k)$ is played, 1 unit of resource is consumed and since the budget $B = o\left(\frac{K}{2}\right)$, no algorithm can differentiate between these instances and hence would always end up playing the sub-optimal arm at least $\frac{(T-B)}{2}$ number of times after the initial $B$ rounds. Therefore

$$\text{REW}^{(c)} \leq B + \frac{(T-B)}{2} \left(\frac{1}{2} - \epsilon\right) + \frac{(T-B)}{2} \left(\frac{1}{2} - 2\epsilon\right)$$

which implies $\text{OPT}^{(c)} \geq \frac{(T-B)}{2} \epsilon = \Omega(T)$

**Total Ordering:** Next consider the following instances:

$\mathcal{I}_1$ is defined by the following preference matrix:

$$P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} + \epsilon & \frac{1}{2} + 2\epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - \epsilon & \frac{1}{2} & \frac{1}{2} + \epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - \epsilon & \frac{1}{2} & \cdots & \frac{1}{2} + 2\epsilon \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \cdots & \frac{1}{2}
\end{bmatrix}$$

with $u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0, \forall i \in \{2, \ldots, K\}$.

$\mathcal{I}_2$ is defined by the following preference matrix:

$$P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} + 2\epsilon & \frac{1}{2} + \epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} + \epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - \epsilon & \frac{1}{2} - \epsilon & \frac{1}{2} & \cdots & \frac{1}{2} + 2\epsilon \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \cdots & \frac{1}{2}
\end{bmatrix}$$
with \( u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0, \forall i \in \{2, \ldots, K\} \) and so on. Corresponding to each of these instances we define the parallel set of instances given by:

\( \mathcal{I}'_1 \) is defined by the following preference matrix:

\[
P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} + \epsilon & \frac{1}{2} + 2\epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - \epsilon & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} + \epsilon & \frac{1}{2} & \cdots & \frac{1}{2} + 2\epsilon \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \cdots & \frac{1}{2}
\end{bmatrix}
\]

with \( u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0, \forall i \in \{2, \ldots, K\} \).

\( \mathcal{I}'_2 \) is defined by the following preference matrix:

\[
P' = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} + 2\epsilon & \frac{1}{2} + \epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} & \frac{1}{2} + \epsilon & \cdots & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - \epsilon & \frac{1}{2} + \epsilon & \frac{1}{2} & \cdots & \frac{1}{2} + 2\epsilon \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \frac{1}{2} - 2\epsilon & \cdots & \frac{1}{2}
\end{bmatrix}
\]

with \( u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0, \forall i \in \{2, \ldots, K\} \) and so on.

Notice that although both set of instances \( \mathcal{I}'_1, \ldots, \mathcal{I}'_K \) and \( \mathcal{I}''_1, \ldots, \mathcal{I}''_K \) have total ordering, the total ordering in \( \mathcal{I}'_1, \ldots, \mathcal{I}'_K \) does not match the ordering of the Condorcet scores in \( \mathcal{I}''_1, \ldots, \mathcal{I}''_K \). Therefore dueling the sub-optimal arms \( \{2, \ldots, K\} \) does not give any information about the order of Condorcet scores, and it essentially reduces to the previous case.

**Strong Stochastic Transitivity:** Finally suppose we assume that the Preference matrix follows strong stochastic transitivity, i.e., if \( i \succ j \) in the total ordering sense then \( P(i, k) \geq \max\{P(i, j), P(j, k)\} \). With this assumption, notice that the instances \( \mathcal{I}'_1, \ldots, \mathcal{I}'_K \) are not allowed and therefore the algorithm may learn about the Condorcet order from the total order relations.

**Lemma 4.2.** Consider the Constrained-DB setting with preference matrix \( P \) and define the minimum gap in Condorcet scores \( \epsilon^{(c)}_{\min} := \min_{i,j \in [K]} (|c(i) - c(j)|) \). Suppose the available budget \( B = o\left(\frac{K}{\epsilon^{(c)}_{\min}}\right) \) then there exists a preference matrix \( P' \) such that \( \text{REG}^{(c)}(T) = \Omega(T) \).

**Proof.** We provide the proof for \( K = 3 \); the general case can be proven by constructing \( K - 1 \) such instances as in the proof of Lemma 4.2. Consider the following preference matrix:

\[
P' = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} + \epsilon & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} - \epsilon & \frac{1}{2} + \epsilon & \frac{1}{2} \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - \epsilon & \frac{1}{2}
\end{bmatrix}
\]

with the following consumptions: \( u^*(1) = v^*(1) = 1, u^*(2) = v^*(2) = \epsilon \) and \( u^*(3) = v^*(3) = 0 \). The proof then follows as in the proof of Lemma 4.4. \( \square \)
B.2 Lower Bounds for Borda Constrained-DB

Lemma 4.3. Consider the Constrained-DB setting with preference matrix $P$ and define the minimum gap in Borda scores $\epsilon^{(b)}_{\min} := \min_{i,j \in [K]} (|b(i) - b(j)|)$. Suppose the available budget $B = o\left(\frac{K}{\epsilon^{(b)}_{\min}}\right)$ then there exists a preference matrix $P$ such that $\text{REG}^{(b)} = \Omega(T)$.

Proof. We provide the proof for $K = 3$; the general case can be proven by constructing $K - 1$ such instances as in the proof of Lemma 4.2. Consider the following preference matrix:

$$P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} + \epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - \epsilon & \frac{1}{2}
\end{bmatrix}$$

with the following consumptions: $u^*(1) = v^*(1) = 0, u^*(2) = v^*(2) = 0$ and $u^*(3) = v^*(3) = 1$. Note that the Borda scores are given by $b(1) = 1/2 + \epsilon, b(2) = 1/2 + \epsilon/2$ and $b(3) = 1/2 - 3\epsilon/2$. The Borda winner is arm 1 and the gap between second best arm and the Borda winner is $\Theta(\epsilon)$. Also note that the optimal policy always plays arm 1. To be able to differentiate between them we need to play $(1, 2)$ and $(1, 3)$ at least $\Theta(K\epsilon^2)$. Since $B = o\left(\frac{K}{\epsilon}\right)$, no algorithm can differentiate between arm 1 and arm 2 and therefore the regret of any algorithm will be $\Omega(T)$.

Lemma 4.4. Consider the Constrained-DB setting with preference matrix $P$ and define the minimum gap in Borda scores $\epsilon^{(c)}_{\min} := \min_{i,j \in [K]} (|b(i) - b(j)|)$. Suppose the available budget $B = o\left(\frac{K}{\epsilon^{(c)}_{\min}}\right)$ then there exists a preference matrix $P$ such that $\text{REG}^{(b)} = \Omega(T)$.

Proof. We provide the proof for $K = 3$; the general case can be proven by constructing $K - 1$ such instances as in the proof of Lemma 4.2. Consider the following preference matrix:

$$P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} + 2\epsilon \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} + \epsilon \\
\frac{1}{2} - 2\epsilon & \frac{1}{2} - \epsilon & \frac{1}{2}
\end{bmatrix}$$

with the following consumptions: $u^*(1) = v^*(1) = 0, u^*(2) = v^*(2) = \epsilon$ and $u^*(3) = v^*(3) = 1$. The optimal policy chooses arm 1 always. However, the gap between arm 1 and arm 2 is $\Theta(\epsilon)$ and therefore arms $(1, 2)$ and $(1, 3)$ have to be played $\Theta(\frac{K}{\epsilon})$ number of times. However, playing arm 2 consumes $\epsilon$ amount and therefore $\Theta(\frac{K}{\epsilon})\epsilon$ budget is needed to differentiate between arms 1 and 2. Since $B = o\left(\frac{K}{\epsilon}\right)$, any algorithm would incur a regret of $\Omega(T)$.

\[ \square \]

C Proof of Regret Bound

Lemma 5.1. For $\tilde{\text{OPT}}^{(b)}$ and $\tilde{\text{REW}}^{(b)}$ as defined in (5) and (6) we have

$$\text{REG}^{(b)} = \tilde{\text{OPT}}^{(b)} - \tilde{\text{REW}}^{(b)} \leq \frac{K}{K-1}(\tilde{\text{OPT}}^{(b)} - \tilde{\text{REW}}^{(b)}).$$
Proof. Observe that

\[ \tilde{\pi}^{*}(b) = \max_{\pi_x, \pi_y \in \Delta^K} \sum_{x \in [K]} \pi_x(x) \tilde{b}(x) + \sum_{y \in [K]} \pi_y(y) \tilde{b}(y) \]

such that \( \sum_{x,y \in [K]} \pi_x(x) u^*(x) + \pi_y(y) u^*(y) \leq \frac{B}{T} \mathbf{1} \)

\[ = \max_{\pi_x, \pi_y \in \Delta^K} \sum_{x \in [K]} \pi_x(x) \left( \frac{K-1}{K} b(x) + \frac{1}{2K} \right) + \sum_{y \in [K]} \pi_y(y) \left( \frac{K-1}{K} b(y) + \frac{1}{2K} \right) \]

such that \( \sum_{x,y \in [K]} \pi_x(x) u^*(x) + \pi_y(y) u^*(y) \leq \frac{B}{T} \mathbf{1} \)

\[ = \arg \max_{\pi_x, \pi_y \in \Delta^K} \sum_{x \in [K]} \pi_x(x) b(x) + \sum_{y \in [K]} \pi_y(y) b(y) \]

such that \( \sum_{x,y \in [K]} \pi_x(x) u^*(x) + \pi_y(y) u^*(y) \leq \frac{B}{T} \mathbf{1} \)

\[ = \pi^{*}(b), \]

i.e., the policy that solves both the LPs are same. Therefore

\[ \text{OPT}^{(b)} = T \sum_{x,y \in [K]} \tilde{\pi}_x^{*}(b) \tilde{b}(x) + \tilde{\pi}_y^{*}(b) \tilde{b}(y) \]

\[ = T \sum_{x,y \in [K]} \tilde{\pi}_x^{*}(b) \left( \frac{K-1}{K} b(x) + \frac{1}{2K} \right) + \tilde{\pi}_y^{*}(b) \left( \frac{K-1}{K} b(y) + \frac{1}{2K} \right) \]

\[ = \frac{K-1}{K} \left( T \sum_{x,y \in [K]} \pi_x^{*}(b)(x) b(x) + \pi_y^{*}(b)(y) b(y) \right) + \frac{T}{K} \]

Further

\[ \text{REW}^{(b)} = \sum_{t=1}^{\tau} \tilde{b}(x_t) + \tilde{b}(y_t) \]

\[ = \sum_{t=1}^{\tau} \left( \frac{K-1}{K} b(x_t) + \frac{1}{2K} \right) + \left( \frac{K-1}{K} b(y_t) + \frac{1}{2K} \right) \]

\[ = \sum_{t=1}^{\tau} b(x_t) + b(y_t) + \frac{\tau}{K} \]

\[ = \text{REW}^{(b)} + \frac{\tau}{K} \]

Therefore

\[ \text{OPT}^{(b)} - \text{REW}^{(b)} = \frac{K-1}{K} \text{OPT}^{(b)} - \text{REW}^{(b)} + \frac{T - \tau}{K} \geq \text{OPT}^{(b)} - \text{REW}^{(b)} \]

where the last line follows because \( \tau \leq T \).

Lemma 5.2. Let the optimal value of \( \text{LP} - \text{Shifted-Borda-x} \) and \( \text{LP} - \text{Shifted-Borda-y} \) be \( \text{OPT}_x^{(b)} \) and \( \text{OPT}_y^{(b)} \). Then

\[ \text{OPT}_x^{(b)} + \text{OPT}_y^{(b)} \geq \text{OPT}^{(b)}. \]
Think Before You Duel: Understanding Complexities of Preference Learning under Constrained Resources

Proof. We define the solution to \( \text{LP - Shifted-Borda} \) as
\[
\hat{\pi}_x^{(b)} = \arg\max_{\pi_x \in \Delta^K} \sum_{x \in [K]} \pi_x(x) \hat{b}(x)
\]
such that \( \sum_{x \in [K]} \pi_x(x) u^*(x) \leq \frac{B}{2T} 1 \)
and the solution to \( \text{LP - Shifted-Borda} \) as
\[
\hat{\pi}_y^{(b)} = \arg\max_{\pi_x \in \Delta^K} \sum_{x \in [K]} \pi_x(x) \hat{b}(x)
\]
such that \( \sum_{x \in [K]} \pi_x(x) u^*(x) \leq \frac{B}{2T} 1 \)

Now consider the following distribution \( \hat{\pi}(x) = \frac{1}{2} \hat{\pi}_x^{(b)} + \frac{1}{2} \hat{\pi}_y^{(b)} \). Note that \( \sum_{x,y \in [K]} \hat{\pi}(x) u^*(x) + \hat{\pi}(y) v^*(y) \leq \frac{B}{2T} 1 + \frac{B}{2T} 1 = \frac{B}{T} 1 \) and therefore \( \hat{\pi} \) is a feasible solution to \( \text{LP - Shifted-Borda} \) and therefore
\[
\overline{\text{OPT}}_x^{(b)} + \overline{\text{OPT}}_y^{(b)} \geq \text{OPT}^{(b)}.
\]

Theorem 5.1. For \( \eta_x = \left( \frac{\log K}{2\sqrt{K}} \right)^{2/3} \frac{1}{2x + 1}, \eta_y = \left( \frac{\log K}{2\sqrt{K}} \right)^{2/3} \frac{1}{2y + 1} \) and \( \gamma_x = \sqrt{\eta_x K} x, \gamma_y = \sqrt{\eta_y K} y, \) the regret of Vigilant D-EXP3 is bounded by
\[
\text{REG}^{(b)}(T) \leq \tilde{O}\left( \left( \text{OPT}^{(b)} \right) B + 1 \right) (K \log K)^{1/3} T^{2/3}
\]

Proof. The proof of the theorem follows along the following three steps:

Step-1: Note that
\[
|\hat{\ell}_i^x(a)| = \left| \hat{b}_i(i) + Z_x \lambda^T \left( \frac{B}{2T} I - \hat{u}_i(i) \right) \right|
\leq |\hat{b}_i(i)| + Z_x \| \lambda \|_1 \left( \frac{B}{2T} \| I \|_\infty + \| \hat{u}_i \|_\infty \right)
\leq 1 + Z_x (1 + 1)
= 1 + 2Z_x
\]
Therefore \( \frac{1}{2Z_x + 1} \hat{\ell}_i^x(a) \leq 1 \) and using the regret guarantee of Exponential Weights algorithm [Auer et al., 2002b], [Lattimore and Szepesvári, 2020] Chapter 11] we get for all \( a \in [K] \)
\[
\sum_{i=1}^\tau \hat{\ell}_i^x(a) - \sum_{i=1}^\tau \sum_{i=1}^K q_i^x(i) \hat{\ell}_i^x(i) \leq \frac{\log K}{\eta_x} + \eta_x \sum_{i=1}^\tau \sum_{i=1}^K q_i^x(i) (\hat{\ell}_i^x(i))^2
\]
Since \( \frac{q_i^x(i) - \gamma_x}{1 - \gamma_x} \), we have
\[
\forall a \in [K], (1 - \gamma_x) \sum_{i=1}^\tau \hat{\ell}_i^x(a) - \sum_{i=1}^\tau \sum_{i=1}^K q_i^x(i) \hat{\ell}_i^x(i) \leq \frac{\log K}{\eta_x} + \eta_x \sum_{i=1}^\tau \sum_{i=1}^K q_i^x(i) (\hat{\ell}_i^x(i))^2
\]
and therefore,
\[
(1 - \gamma_x) \sum_{i=1}^\tau \sum_{a=1}^K \hat{\ell}_i^x(a) \pi_x^{(b)}(a) - \sum_{i=1}^\tau \sum_{i=1}^K q_i^x(i) \hat{\ell}_i^x(i) \leq \frac{\log K}{\eta_x} + \eta_x \sum_{i=1}^\tau \sum_{i=1}^K q_i^x(i) (\hat{\ell}_i^x(i))^2
\]
Step-2: Next we relate the LHS of \( \text{[8]} \) to the regret using the following lemma.

**Lemma 5.3.** For any \( a \in [K] \)

\[
(1 - \gamma_x) \sum_{t=1}^{T} \sum_{a=1}^{K} \hat{\ell}_t^a(a) \tilde{\pi}_x^a(b)(a) - \sum_{t=1}^{T} \sum_{a} q_t^a(\hat{\ell}_t^a(a)) \\
\geq \text{OPT}_x(b) - B \sum_{t=1}^{T} \bar{b}(x_t) - O(Z + 1)\sqrt{T \log T} \\
- \gamma_x (Z_x + 1)T
\]

Further, \( \gamma_x (Z_x + 1)T \leq O \left( (K \log K)^{1/3} Z_x T^{2/3} \right) \)

**Proof.** Let \( \tilde{\pi}_x^a(b) \) be the solution of \( \text{[LP - Shifted-Borda]} \). Next define \( \mathcal{H}_{t-1} = \sigma\{ x_i, y_i, o_i(x_i, y_i)\}_{i=1}^{t-1} \) be the sigma algebra generated by \{ \( x_i, y_i, o_i(x_i, y_i) \) \}_{i=1}^{t-1} \) and \( \mathbb{E}_{\mathcal{H}_{t-1}} \) be the conditional expectation with respect to \( \mathcal{H}_{t-1} \). Consider term I and observe that

\[
(1 - \gamma_x) \sum_{t=1}^{T} \sum_{a} \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \hat{\ell}_t^a(a) \tilde{\pi}_x^a(b)(a) \right] = (1 - \gamma_x) \sum_{t=1}^{T} \sum_{a} \tilde{\pi}_x^a(b)(a) \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \bar{b}_t(a) + Z_x \lambda_t^\top \left( B \frac{1}{2T} \mathbb{I} - \tilde{u}_t^a(a) \right) \right].
\]

Note that \( \mathbb{E}[\bar{b}_t(a)] = \tilde{b}(a) \) (see Lemma 4 in Saha et al. [2021b]) and \( \mathbb{E}[\tilde{u}_t^a(a)] = u^*(a) \). Using Azuma-Hoeffding inequality with probability at least \( 1 - O(1/T^2) \)

\[
(1 - \gamma_x) \sum_{t=1}^{T} \sum_{a} \hat{\ell}_t^a(a) \tilde{\pi}_x^a(b)(a) \geq (1 - \gamma_x) \sum_{t=1}^{T} \sum_{a} \tilde{\pi}_x^a(b)(a) \mathbb{E}_{\mathcal{H}_{t-1}} \left[ \bar{b}_t(a) + Z_x \lambda_t^\top \left( B \frac{1}{2T} \mathbb{I} - \bar{u}_t^a(a) \right) \right] \\
- (Z_x + 1)\sqrt{T \log T}
\]

Next observe that with \( D_t = \bar{b}(a) + Z_x \lambda_t^\top \left( B \frac{1}{2T} \mathbb{I} - \bar{u}_t^a(a) \right) \) and \( \mathbb{E}_{\mathcal{H}_{t-1}}[D_t] \geq 0 \) and therefore using Azuma-Hoeffding with probability at least \( 1 - O(1/T^2) \)

\[
\sum_{t=1}^{T} \sum_{a} \hat{\ell}_t^a(a) \tilde{\pi}_x^a(b)(a) \geq \frac{T}{T} \text{OPT}_x(b) - 2(Z_x + 1)\sqrt{T \log T}
\]

Therefore

\[
(1 - \gamma_x) \sum_{t=1}^{T} \sum_{a} \hat{\ell}_t^a(a) \tilde{\pi}_x^a(b)(a) \geq \frac{T}{T} \text{OPT}_x(b) - O(Z + 1)\sqrt{T \log T} - \gamma_x (Z_x + 1)T
\]  \( \text{(9)} \)

Next consider term II from \( \text{[8]} \):

\[
\sum_{t=1}^{T} \sum_{a} q_t^a(\hat{\ell}_t^a(a) \leq \sum_{t=1}^{T} \sum_{a} q_t^a(\hat{\ell}_t^a(a) \left[ \bar{b}_t(a) + Z_x \lambda_t^\top \left( B \frac{1}{2T} \mathbb{I} - \tilde{u}_t^a(a) \right) \right]
\]

Since \( x_t \sim q_t^a \), therefore \( \mathbb{E}_{\mathcal{H}_{t-1}}[\sum_a q_t^a(\bar{b}(a)] = \bar{b}(x_t) \) and \( \mathbb{E}_{\mathcal{H}_{t-1}}[\sum_a q_t^a(\tilde{u}_t^a(a)] = u^*(x_t) \), therefore using Azuma-Hoeffding with probability at least \( 1 - O(1/T^2) \)

\[
\sum_{t=1}^{T} \sum_{a} q_t^a(\hat{\ell}_t^a(a) \leq \sum_{t=1}^{T} b(x_t) + Z_x \lambda_t^\top \left( B \frac{1}{2T} \mathbb{I} - u^*(x_t) \right) + O(Z_x + 1)\sqrt{T \log T}
\]
Combining with \([9]\) we get

\[
(1 - \gamma_x) \sum_{t=1}^{\tau} \hat{\ell}_t^*(a) - \frac{\tau}{T} \sum_{i=1}^{K} q_t^*(i) \hat{\ell}_t^*(i) \geq \frac{\tau}{T} \operatorname{OPT}_x^{(b)} - \frac{\tau}{T} \sum_{t=1}^{\tau} b(x_t) - Z_x \lambda^T \left( \frac{B_{2T}}{T} I - u^*(x_t) \right) - O(Z_x + 1) \sqrt{T \log T} - \gamma_x(Z_x + 1) T
\]  

From the regret guarantee of OCO on \(q_t^*(\lambda)\) we have that for any \(\lambda \in [0, 1]^d\),

\[
(1 - \gamma_x) \sum_{t=1}^{\tau} \hat{\ell}_t^*(a) - \frac{\tau}{T} \sum_{i=1}^{K} q_t^*(i) \hat{\ell}_t^*(i) \geq \frac{\tau}{T} \operatorname{OPT}_x^{(b)} - \frac{\tau}{T} \sum_{t=1}^{\tau} b(x_t) - Z_x \lambda^T \left( \frac{B_{2T}}{T} I - u^*(x_t) \right) - O(Z_x + 1) \sqrt{T \log T} - \gamma_x(Z_x + 1) T.
\]

Next if \(\tau = T\), choosing \(\lambda = 0\) gives

\[
(1 - \gamma_x) \sum_{t=1}^{\tau} \hat{\ell}_t^*(a) - \frac{\tau}{T} \sum_{i=1}^{K} q_t^*(i) \hat{\ell}_t^*(i) \geq \operatorname{OPT}_x^{(b)} - \frac{\tau}{T} \sum_{t=1}^{\tau} b(x_t) - O(Z_x + 1) \sqrt{T \log T} - \gamma_x(Z_x + 1) T.
\]

If \(\tau < T\) then \(\exists j\) such that \(\sum_{t=1}^{T} u^*(x_t)_j > B/2\) (i.e., one of the resources is exhausted). Choose \(\lambda = Z_x \epsilon_j\) and observe that

\[
\sum_{t=1}^{\tau} Z_x \lambda^T \left( \frac{B_{2T}}{T} I - u^*(x_t) \right) \leq Z_x \left( \frac{\tau}{2T} B - B/2 \right).
\]

Combining with \([11]\) we get with probability \((1 - O(\frac{1}{T^2}))\)

\[
(1 - \gamma_x) \sum_{t=1}^{\tau} \hat{\ell}_t^*(a) - \frac{\tau}{T} \sum_{i=1}^{K} q_t^*(i) \hat{\ell}_t^*(i) \geq \frac{\tau}{T} \operatorname{OPT}_x^{(b)} - \frac{\tau}{T} \sum_{t=1}^{\tau} b(x_t) - \frac{2 \operatorname{OPT}_x^{(b)}}{B} \left( \frac{\tau}{2T} B - B/2 \right) - O(Z_x + 1) \sqrt{T \log T} - \gamma_x(Z_x + 1) T
\]

which implies

\[
(1 - \gamma_x) \sum_{t=1}^{\tau} \hat{\ell}_t^*(a) - \frac{\tau}{T} \sum_{i=1}^{K} q_t^*(i) \hat{\ell}_t^*(i) \geq \operatorname{OPT}_x^{(b)} - \sum_{t=1}^{\tau} b(x_t) - O(Z_x + 1) \sqrt{T \log T} - \gamma_x(Z_x + 1) T.
\]

Further,

\[
\gamma_x(Z_x + 1) T \leq \sqrt{\eta_x K} \sqrt{Z_x(Z_x + 1) T} = O \left( (K \log K)^{1/3} Z_x T^{2/3} \right)
\]

Next we upper bound the RHS of \((8)\) using the following lemma.

**Lemma 5.4.** For \(\eta_x = \left( \frac{\log K}{T \sqrt{K}} \right)^{2/3} \frac{1}{2Z_x + 1}\) and \(\gamma_x = \sqrt{\eta_x K Z_x}\)

\[
\frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^{\tau} \sum_{i=1}^{K} q_t^*(i) \hat{\ell}_t^*(i)^2 \leq O \left( \left( \frac{\operatorname{OPT}_x^{(b)}}{B} + 1 \right) (K \log K)^{1/3} T^{2/3} \right)
\]
Proof. Consider the following term:

\[
\sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(a) \hat{s}_t^\tau(a) = \sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(a) \left( \hat{s}_t(a) + Z_{ax} \lambda_t \left( \frac{B}{2T} - \bar{\hat{u}}_t(x_t) \right) \right)^2 \\
\leq \sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(a) \left( \hat{s}_t(a)^2 + \left( Z_{ax} \lambda_t \left( \frac{B}{2T} - \bar{\hat{u}}_t(x_t) \right) \right)^2 \right) \\
\leq \sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(a) \left( \hat{s}_t(a)^2 + \left( Z_{ax} \lambda_t \left( \frac{B}{2T} - \bar{\hat{u}}_t(x_t) \right) \right)^2 \right) \\
\leq \sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(a) \hat{s}_t(a)^2 + \sum_{t=1}^{\tau} \sum_{a=1}^{K} 2Z_{ax}^2 \alpha_T + Z_{ax}^2 \sum_{t=1}^{\tau} \sum_{a=1}^{K} 4q_t^\tau(a) \lambda_t^2 \bar{\hat{u}}_t(a)^2
\]

We have \(\sum_a E[ q_t^\tau(a) \hat{s}_t(a)^2] \leq \frac{K}{\eta_x} \) and \(\sum_a E[ q_t^\tau(a) \lambda_t^2 \bar{\hat{u}}_t(x_t)^2] \leq \frac{K}{\eta_x} \) (see \cite[saha et al. 2021b, Lemma 6, Lattimore and Szepesvári 2020, Chapter 11]).

\[
\sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(a) \hat{s}_t^\tau(a)^2 \leq \frac{K}{\eta_x} T + \frac{Z_{ax}^2 B \alpha_T}{\eta_x^2} + \frac{K}{\eta_x} Z_{ax}^2 T \\
\leq \frac{K}{\eta_x} T (Z_{ax}^2 + 2)
\]

Therefore

\[
\frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(i) \hat{s}_t^\tau(i)^2 \leq \frac{\log K}{\eta_x} + \eta_x \frac{K}{\eta_x} T (Z_{ax}^2 + 2)
\]

Choosing \(\gamma_x = \sqrt{\eta_x K \sqrt{Z_x}}\) we get

\[
\frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(i) \hat{s}_t^\tau(i)^2 \leq \frac{\log K}{\eta_x} + T \sqrt{\eta_x K (Z_{ax}^2 + 2) / \sqrt{Z_x}}
\]

Finally choosing \(\eta_x = \left( \frac{\log K}{T \sqrt{K}} \right)^{2/3} \frac{1}{2Z_{ax}^{1/3}}\) we have

\[
\frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^{\tau} \sum_{a=1}^{K} q_t^\tau(i) \hat{s}_t^\tau(i)^2 \leq O \left( (K \log K)^{1/3} T^{2/3} Z_x \right) = O \left( \left( \frac{\OPT_x^{(b)}}{B} + 1 \right) (K \log K)^{1/3} T^{2/3} \right)
\]

\(\square\)

Step-3: We repeat the same argument for the second arm choice \(y_t\) and then combining with Lemma 5.1 and Lemma 5.2 we get

\[
\OPT_x^{(b)} + \OPT_y^{(b)} - \left( \sum_{t=1}^{\tau} b(x_t) + b(y_t) \right) \leq O \left( \left( \frac{\OPT_x^{(b)}}{B} + 1 \right) (K \log K)^{1/3} T^{2/3} \right)
\]

which completes the proof.

\(\square\)

**D Details of Experiments**

We provide more detailed descriptions of our experiments in this section.

**Datasets.** We run our experiments on two datasets.
1. **Synthetic Data:** We create a Constrained Dueling Bandits instance with $K = 6$ arms where the preference matrix is given by

$$P = \begin{pmatrix}
0.5 & 0.55 & 0.55 & 0.54 & 0.61 & 0.61 \\
0.45 & 0.5 & 0.55 & 0.55 & 0.58 & 0.6 \\
0.45 & 0.45 & 0.5 & 0.54 & 0.51 & 0.56 \\
0.46 & 0.45 & 0.46 & 0.5 & 0.54 & 0.5 \\
0.39 & 0.42 & 0.49 & 0.46 & 0.5 & 0.51 \\
0.39 & 0.4 & 0.44 & 0.5 & 0.49 & 0.5 \\
\end{pmatrix}.$$ 

The vector of Borda scores $\bar{b} = (b(1) \ b(2) \ldots \ b(6))^\top$ is given by

$$(0.672 \ 0.646 \ 0.602 \ 0.582 \ 0.554 \ 0.544)^\top.$$ 

We experiment with three choices of consumptions. In all three cases the number of resources $d = 1$ and the true consumptions across both arms choices are given by the same function, i.e., $u^* = v^*$, and we add zero mean gaussian noise to each entry. The vector of consumptions for arms $\bar{u}^* = (u^*(1) \ u^*(2) \ldots \ u^*(6))^\top$ are given by:

(a) $(0.9 \ 0.9 \ 0.1 \ 0.8 \ 0.8 \ 0.8)^\top$

(b) $(0.6 \ 0.5 \ 0.4 \ 0.3 \ 0.2 \ 0.1)^\top$

(c) $(0 \ 0 \ 0 \ 0 \ 0 \ 0)^\top$

In the first case although arm 1 and 2 have high Borda scores, the associated consumptions are also high. In the second case the order of consumptions is the same as the order of Borda scores. In the last case all the consumptions are zero and our objective is to evaluate if our algorithm under performs in the absence of constraints. The experiments are run for $T = 2000$ rounds with $B = 1000$ and are run independently over 50 samples.

2. **Car preference dataset:** We consider the Car preference dataset from [E et al. 2013](#) where the preference matrix is generated by considering the user preferences for various models of cars. The dataset utilized 10 items to generate all 45 possible preferences. The study was conducted in two phases, with data collected from 40 and 20 users separately. Participants in the initial experiment were presented with cars featuring specific attributes given by (1) Body type, (2) Transmission, (3) Engine capacity and (4) Fuel consumed. We use the dataset to compute the preference matrix. As in case 1, we consider three choices of consumptions that follow a similar structure as given below:

(a) $(0.9 \ 0.9 \ 0.01 \ 0.02 \ 0.7 \ 0.3 \ 0.6 \ 0.7 \ 0.7 \ 0.8)$

(b) $(0.7 \ 0.9 \ 0.9 \ 0.8 \ 0.6 \ 0.1 \ 0.4 \ 0.3 \ 0.5 \ 0.2)$

(c) $(0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0)$

The experiments are run for $T = 5000$ rounds with $B = 4000$ and are run independently over 50 samples.