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 $\tilde{\mathcal{O}}\left(\left(\frac{\mathrm{OPT}^{(b)}}{B}+1\right)K^{1/3}T^{2/3}\right)$ regret, where $\mathrm{OPT}^{(b)}$ 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 et al., 2012, 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], Désir 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

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

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 Knapsack 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/winners - *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 take 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. [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] 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]) 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. [Agrawal and Devanur, 2016] study the natural extension of the problem - the linear contextual bandits with knapsacks and provide an algorithm utilizing ideas from UCB and primal-dual methods with sub-linear regret bound. Further extensions have been studied such as regret minimization with concave reward and convex objective [Agrawal et al., 2016, Agrawal and Devanur, 2014], adversarial setting [Immorlica et al., 2022] and smoothed adversarial setting [Sivakumar et al., 2022] that all use a version of the central primal-dual idea from [Agrawal and Devanur, 2016]. In Section 5 we discuss why a direct extension of the algorithm from [Agrawal and Devanur, 2016] does not work in the Constrained-DB setting (see Remark 5.1).

3 Preliminaries

3.1 Existing Concepts from Dueling Bandits

A learner at round $t \in [T]$ is presented with K arms to choose from. It then selects two arms $x_t, y_t \in [K] := \{1, 2, \dots, K\}$ to duel, and receives a feedback $o_t \sim \text{Ber}(P(x_t, y_t))$. Here, $\text{Ber}(p)$ denotes the Bernoulli distribution with parameter p and $P(x_t, y_t)$ measures the probability of x_t being preferred over y_t, that satisfies $P(x, y) = 1 - P(y, x)$ for all $x, y \in [K]$. Further we define the matrix $P = [P(i, j)]_{i,j\in[K]}$ and call it the preference matrix.

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 3.1 (**Condorcet Winner** Zoghi et al. [2014a])**.** We define the Condorcet winner $x^{(c)}$ as the arm that is preferred over all the other arms, i.e., $x^{(c)} = i$ iff $P(i, j) > 1/2,\ \forall j \in [K] \setminus \{i\}$. Further the Condorcet score of arm x is defined as

$$c(x) = P(x, x^{(c)}).$$

Definition 3.2 (**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., $x^{(b)} = \text{argmax}_x\, b(x)$.

Definition 3.3 (**Total Ordering** [Yue et al., 2012, Yue and Joachims, 2011])**.** We say the preference matrix P satisfies Total Ordering(TO) if there exists a binary total order relation $\succ$ with $\forall\ i, j \in [K]$, $i \succ j$ implies $P(i, j) > \frac{1}{2}$.

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 $i, j, k \in [K]$, $i \succ j \succ k$ implies $P(i, k) \geq \max\{P(i, j), P(j, k)\}$ where $\succ$ 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 2), 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 $(\sum_{t=1}^{\tau} u_t(x_t) + v_t(x_t))_i > B$ 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.

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

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

where Δ^K is the probability simplex over K. Suppose $\pi_x^{*(c)}, \pi_y^{*(c)}$ solve $(\texttt{LP} - \texttt{Condorcet})$ 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.

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

$$\text{such that} \sum_{x,y\in[K]} \pi_x(x) u^*(x) + \pi_y(y) v^*(y) \le \frac{B}{T}\mathbb{1} \ . \qquad (\texttt{LP}-\texttt{Borda})$$

where Δ^K is the probability simplex over K. Suppose $\pi_x^{*(b)}, \pi_y^{*(b)}$ solve $(\texttt{LP}-\texttt{Borda})$ 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 $(\texttt{LP}-\texttt{Condorcet})$ and $(\texttt{LP}-\texttt{Borda})$ compute a non-adaptive policy and follows the development in [Agrawal and Devanur, 2016]. 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 [2016]).

Regret Next we define the total cumulative regret of an algorithm that chooses the sequence of arms $\{(x_t, y_t)_{t=1}^{\tau}\}$, where $\tau \le T$ 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 τ as

$$\text{REW}^{(c)} = \mathbb{E}\sum_{t=1}^{\tau} c(x_t) + c(y_t), \qquad (1)$$

and the corresponding Condorcet regret as

$$\text{REG}^{(c)}(T) = \text{OPT}^{(c)} - \text{REW}^{(c)} \qquad (2)$$

2. **Borda Regret.** We define the cumulative Borda reward until the stopping time τ as

$$\text{REW}^{(b)} = \mathbb{E}\sum_{t=1}^{\tau} b(x_t) + b(y_t), \qquad (3)$$

and the corresponding Borda regret as

$$\text{REG}^{(b)}(T) = \text{OPT}^{(b)} - \text{REW}^{(b)} \qquad (4)$$

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 $(\texttt{LP}-\texttt{Condorcet})$ and $(\texttt{LP}-\texttt{Borda})$ and the performance is measured via the regret defined in (2) and (2).

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

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* $B = o\left(\frac{K}{{\epsilon_{\min}^{(c)}}^2}\right)$ *then the regret of any algorithm for Condorcet-constrained-DB is* $\Omega(T)$*, where* $\epsilon_{\min}^{(c)}$ *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* $B = o\left(\frac{K}{\epsilon_{\min}^{(c)}}\right)$ *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 $B \ge \frac{K}{\epsilon_{\min}^2}$.

Lemma 4.1. *Consider the Constrained-DB setting with preference matrix* P *and define the minimum gap in Condorcet scores* $\epsilon_{\min}^{(c)} := \min_{i,j\in[K]} (|c(i) - c(j)|)$*. Suppose the available budget* $B = o\left(\frac{K}{{\epsilon_{\min}^{(c)}}^2}\right)$ *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 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 $\mathrm{OPT}^{(c)} = B + (T-B)\left(\frac{1}{2}-\epsilon\right)$.

Note that arm 1 is the Condorcet winner and any algorithm needs to play the pairs (1,2) and (1,3) at least $\frac{1}{\epsilon^2}$ number of times to determine if $c(2) > c(3)$. However since the budget $B = o\left(\frac{1}{\epsilon^2}\right)$, 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 $\frac{(T-B)}{2}$ number of times after the initial B rounds. Therefore

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

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

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.

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

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

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

such that although $s(2) > s(3)$, in the total ordering sense $3 \succ 2$. 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 $2 \succ 3$ 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 $\epsilon^{(c)}_{\min} := \min_{i,j\in[K]}(|c(i)-c(j)|)$. Suppose the available budget $B = o\left(\frac{K}{\epsilon^{(c)}_{\min}}\right)$ then there exists a preference matrix P such that $REG^{(c)}(T) = \Omega(T)$.*

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 $\epsilon^{(b)}_{\min} := \min_{i,j\in[K]}(|b(i)-b(j)|)$. Suppose the available budget $B = o\left(\frac{K}{{\epsilon^{(b)}_{\min}}^2}\right)$ then there exists a preference matrix P such that $REG^{(b)} = \Omega(T)$.*

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 $\epsilon^{(c)}_{\min} := \min_{i,j\in[K]}(|b(i)-b(j)|)$. Suppose the available budget $B = o\left(\frac{K}{\epsilon^{(b)}_{\min}}\right)$ then there exists a*

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

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 $B = \mathcal{O}\left(\max\{\frac{K}{\epsilon_{\min}^2}, T^{3/4}\}\right)$. 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 $\mathrm{OPT}^{(b)}$ by more than a constant factor.

1. **Shifted Borda Score [Saha et al., 2021b]** We replace the Borda score $b(x)$ in $(\mathtt{LP-Borda})$ 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

   $$\max_{\pi_x, \pi_y \in \Delta^K} \sum_{x \in [K]} \pi_x(x)\tilde{b}(x) + \sum_{y \in [K]} \pi_y(y)\tilde{b}(y)$$
   $$\text{such that} \sum_{x,y \in [K]} \pi_x(x)u^*(x) + \pi_y(y)v^*(y) \le \frac{B}{T}\mathbf{1} \qquad (\mathtt{LP-Shifted\text{-}Borda})$$

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

   $$\widetilde{\mathrm{OPT}^{(b)}} = T \sum_{x,y \in [K]} \widetilde{\pi}_x^{*(b)}\, \tilde{b}(x) + \widetilde{\pi}_y^{*(b)}\, \tilde{b}(y) \quad (5)$$

   $$\widetilde{\mathrm{REW}^{(b)}} = \mathbb{E}\sum_{t=1}^{\tau} \tilde{b}(x_t) + \tilde{b}(y_t). \quad (6)$$

   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 $\widetilde{OPT^{(b)}}$ and $\widetilde{REW^{(b)}}$ as defined in (5) and (6) we have*

   $$REG^{(b)} = OPT^{(b)} - REW^{(b)} \le \frac{K}{K-1}(\widetilde{OPT^{(b)}} - \widetilde{REW^{(b)}}).$$

   Therefore bounding $\widetilde{\mathrm{OPT}^{(b)}} - \widetilde{\mathrm{REW}^{(b)}}$ bounds the final regret $\mathrm{REG}^{(b)}$ upto a constant factor of $\frac{K}{K-1}$.

2. **Define Separate LPs.** Next we relax the benchmark $(\mathtt{LP-Shifted\text{-}Borda})$ by separating the LPs associated with the two arms as defined below.

   $$\max_{\pi_x \in \Delta^K} \sum_{x \in [K]} \pi_x(x)\tilde{b}(x)$$
   $$\text{such that} \sum_{x \in [K]} \pi_x(x)u^*(x) \le \frac{B}{2T}\mathbf{1} \qquad (\mathtt{LP-Shifted\text{-}Borda\text{-}x})$$

   $$\max_{\pi_y \in \Delta^K} \sum_{y \in [K]} \pi_y(y)\tilde{b}(y)$$
   $$\text{such that} \sum_{y \in [K]} \pi_y(y)v^*(y) \le \frac{B}{2T}\mathbf{1} \qquad (\mathtt{LP-Shifted\text{-}Borda\text{-}y})$$

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

   Lemma 5.2. *Let the optimal value of $(\mathtt{LP-Shifted\text{-}Borda\text{-}x})$ and $(\mathtt{LP-Shifted\text{-}Borda\text{-}y})$ be $\widetilde{OPT_x^{(b)}}$ and $\widetilde{OPT_y^{(b)}}$. Then*

   $$\widetilde{OPT_x^{(b)}} + \widetilde{OPT_y^{(b)}} \ge \widetilde{OPT^{(b)}}.$$

   Therefore bounding $\widetilde{\mathrm{OPT}_x^{(b)}} - \mathbb{E}\sum_{t=1}^{\tau} \tilde{b}(x_t)$ and $\widetilde{\mathrm{OPT}_y^{(b)}} - \mathbb{E}\sum_{t=1}^{\tau} \tilde{b}(y_t)$ separately bounds the final regret $\widetilde{\mathrm{OPT}^{(b)}} - \widetilde{\mathrm{REW}^{(b)}}$.

`Vigilant D-EXP3`(Vigilant Dueling EXP3) In Algorithm 1 we maintain two distributions q_t^x and q_t^y to sample the arms x_t and y_t at time t. Initially both distributions are initialized to the uniform distribution (Step 2). At time $t \in [T]$ the arms x_t and y_t are sample from the distributions q_t^x and q_t^y respectively and observe the preference output $o_t(x_t, y_t) \sim \mathrm{Ber}(P_t(x_t, y_t))$ and the consumption vectors $u_t(x_t)$ and $v_t(x_t)$ (Step 4 and 5). Next we compute unbiased estimates of the shifted Borda score and the two consumption vectors (Step 6) and the empirical lagrangians for the two arms (Step 7). Next we update the arm distributions q_t^x and q_t^y using exponential weights on the estimated cumulative lagrangians along with an γ-uniform exploration (Step 8). Finally we update the lagrange multipliers on the dual objectives (Step 9).

Algorithm 1 `Vigilant D-EXP3`

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

Initialize: Initial probability distribution $q_1^x(i) = q_1^y(i) = 1/K,\ \forall i \in [K]$

for $t = 1, \ldots, T$ **do**

Sample $x_t \sim q_t^x, y_t \sim q_t^y$ i.i.d.

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

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

$$\hat{b}_t(i) = \frac{\mathbb{1}(x_t = i)}{K q_t^x(i)} \sum_{j \in [K]} \frac{\mathbb{1}(y_t = j) o_t(x_t, y_t)}{q_t^y(j)},$$

$$\hat{u}_t^x(i) = 1 - \frac{\mathbb{1}(x_t = i)}{q_t(i)}(1 - u_t(x_t)),$$

$$\hat{u}_t^y(i) = 1 - \frac{\mathbb{1}(y_t = i)}{q_t(i)}(1 - v_t(y_t)).$$

Estimate the Lagrangians $\forall\ i \in [K]$

$$\hat{\ell}_t^x(i) = \hat{b}_t(i) + Z_x \lambda_t^{x\top} \left[\frac{B}{2T}\mathbb{1} - \hat{u}_t^x(i)\right],$$

$$\hat{\ell}_t^y(i) = \hat{b}_t(i) + Z_y \lambda_t^{y\top} \left[\frac{B}{2T}\mathbb{1} - \hat{v}_t(i)\right]$$

Update for all $i \in [K]$:

$$\tilde{q}_{t+1}^x(i) = \frac{\exp(\eta_x \sum_{s=1}^t \hat{\ell}_s^x(i))}{\sum_{j=1}^K \exp(\eta_x \sum_{s=1}^t \hat{\ell}_s^x(j))}$$

$$q_{t+1}^x(i) = (1 - \gamma_x)\tilde{q}_{t+1}^x(i) + \frac{\gamma_x}{K}$$

$$\tilde{q}_{t+1}^y(i) = \frac{\exp(\eta_y \sum_{s=1}^t \hat{\ell}_s^y(i))}{\sum_{j=1}^K \exp(\eta_y \sum_{s=1}^t \hat{\ell}_s^y(j))}$$

$$q_{t+1}^y(i) = (1 - \gamma_y)\tilde{q}_{t+1}^y(i) + \frac{\gamma_y}{K}$$

Update λ_t^x and λ_t^y using any online convex optimization on the following objective functions

$$g_t^x(\lambda) = \left\langle \frac{B}{2T}\mathbb{1} - \hat{u}_t^x(x_t) \right\rangle,\ g_t^y(\lambda) = \left\langle \frac{B}{2T}\mathbb{1} - \hat{v}_t^x(y_t) \right\rangle.$$

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 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 fom 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 $OPT_x^{(b)}$ and $OPT_y^{(b)}$ 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* $\eta_x = \left(\frac{\log K}{T\sqrt{K}}\right)^{2/3} \frac{1}{2Z_x+1}, \eta_y = \left(\frac{\log K}{T\sqrt{K}}\right)^{2/3} \frac{1}{2Z_y+1}$ *and* $\gamma_x = \sqrt{\eta_x K Z_x}, \gamma_y = \sqrt{\eta_y K Z_y}$, *the regret of* `Vigilant D-EXP3` *is bounded by*

$$REG^{(b)}(T) \leq \tilde{\mathcal{O}}\Big(\big(\frac{OPT^{(b)}}{B} + 1\big)(K \log K)^{1/3} T^{2/3}\Big)$$

Proof sketch Here we briefly outline the steps of the proof. For details see Appendix C.

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

$$\sum_{t=1}^{\tau} \hat{\ell}_t^x(a) - \sum_{t=1}^{\tau} \sum_a \tilde{q}_t^x(a) \hat{\ell}_t^x(a) \leq \frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^{\tau} \sum_{i=1}^{K} \tilde{q}_t^x(i) (\hat{\ell}_t^x(i))^2.$$

Since $\tilde{q}_t^x(i) = \dfrac{q_t^x(i) - \frac{\gamma_x}{K}}{1 - \gamma_x}$,

$$\forall a,\ (1 - \gamma_x) \sum_{t=1}^{\tau} \hat{\ell}_t^x(a) - \sum_{t=1}^{\tau} \sum_a q_t^x(i) \hat{\ell}_t^x(a) \leq \frac{\log K}{\eta_x} + \eta_x \sum_{t=1}^{\tau} \sum_{i=1}^{K} q_t^x(i) (\hat{\ell}_t^x(i))^2 \quad (7)$$

Step 2: Since the LHS in (7) holds for every $a \in [K]$ we relate it to the regret $\widetilde{\text{OPT}}_x^{(b)} - \sum_{t=1}^{\tau} \tilde{b}(x_t)$ using the following lemma.

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

$$(1-\gamma_x)\sum_{t=1}^{\tau}\sum_{a=1}^{K}\hat{\ell}_t^x(a)\tilde{\pi}_x^{*(b)}(a) - \sum_{t=1}^{\tau}\sum_{a} q_t^x(i)\hat{\ell}_t^x(a)$$
$$\geq \widetilde{OPT}_x^{(b)} - \mathbb{E}\sum_{t=1}^{\tau}\tilde{b}(x_t) - \mathcal{O}(Z+1)\sqrt{T\log T}$$
$$- \gamma_x(Z_x+1)T$$

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

Next we upper bound the RHS in (7) 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* $\gamma_x = \sqrt{\eta_x K Z_x}$

$$\frac{\log K}{\eta_x} + \eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K} q_t^x(i)(\hat{\ell}_t^x(i))^2$$
$$\leq \mathcal{O}\left(\left(\frac{OPT_x^{(b)}}{B}+1\right)(K\log K)^{1/3}T^{2/3}\right)$$

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 $\bar{b} = \begin{pmatrix} b(1) & b(2)6 & \dots & b(6)\end{pmatrix}^\top$ is given by $\begin{pmatrix}0.672 & 0.646 & 0.602 & 0.582 & 0.554 & 0.544\end{pmatrix}^\top$. 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 $\bar{u}^* = \begin{pmatrix} u^*(1) & u^*(2)6 & \dots & u^*(6)\end{pmatrix}^\top$ are given by $\begin{pmatrix}0.9 & 0.9 & 0.1 & 0.8 & 0.8 & 0.8\end{pmatrix}^\top$, $\begin{pmatrix}0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6\end{pmatrix}^\top$, and $\begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}^\top$. 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 Lekang and 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 stop 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`.

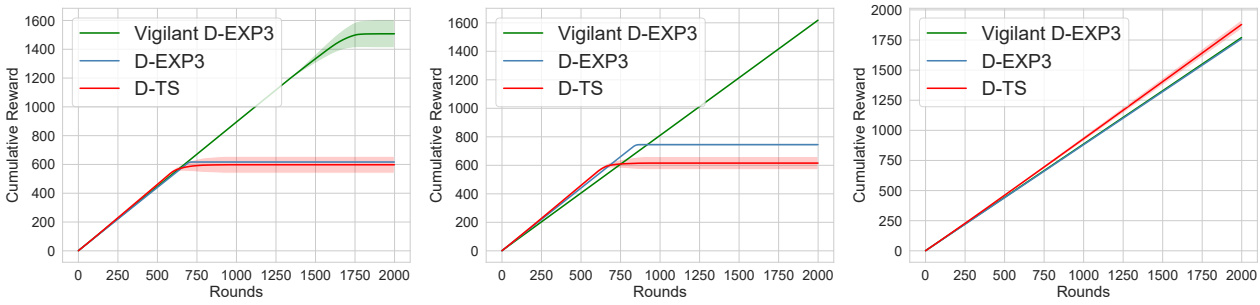

Figure 1: Cumulative Reward across Rounds on synthetic data for three choices of consumptions. The first corresponds to the case when the consumption of an arm with intermediate Borda score is lowest, second to when the order of consumptions is the same as the order of Borda score and third corresponds to zero consumption.

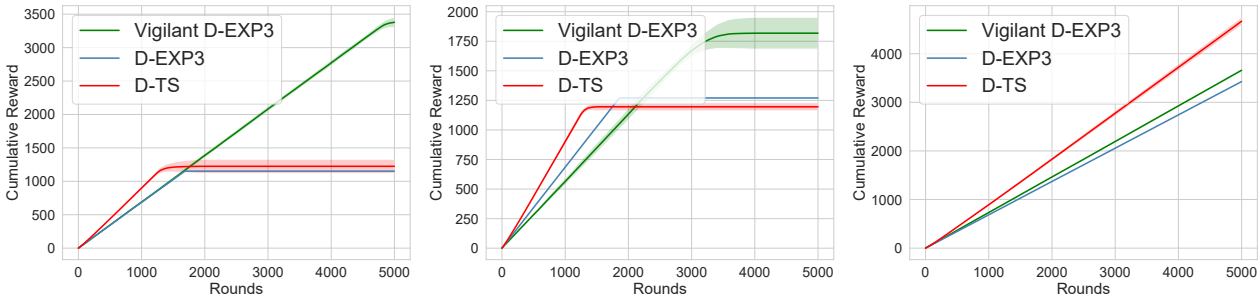

Figure 2: Cumulative Reward across Rounds on car preference dataset for three choices of consumptions. The first corresponds to the case when the consumption of an arm with intermediate Borda score is lowest, second to when the order of consumptions is the same as the order of Borda score and third corresponds to zero consumption.

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 Dueling 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

N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

S. Agrawal and N. R. Devanur. Bandits with concave rewards and convex knapsacks. EC '14, page 989–1006, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325653. doi: 10.1145/2600057.2602844. URL https://doi.org/10.1145/2600057.2602844.

S. Agrawal and N. R. Devanur. Linear contextual bandits with knapsacks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3458–3467, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

S. Agrawal, N. R. Devanur, and L. Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 4–18, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/agrawal16.html.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 856–864, Bejing, China, 22–24 Jun 2014a. PMLR. URL https://proceedings.mlr.press/v32/ailon14.html.

N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864. PMLR, 2014b.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002a. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

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

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, oct 2013. doi: 10.1109/focs.2013.30. URL https://doi.org/10.1109%2Ffocs.2013.30.

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

V. Bengs, R. Busa-Fekete, A. El Mesaoudi-Paul, and E. Hüllermeier. Preference-based online learning with dueling bandits: A survey. *J. Mach. Learn. Res.*, 22 (1), jan 2021. ISSN 1532-4435.

V. Bengs, A. Saha, and E. Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pages 1764–1786. PMLR, 2022.

K. Bhatia, A. Pananjady, P. Bartlett, A. Dragan, and M. J. Wainwright. Preference learning along multiple criteria: A game-theoretic perspective. *Advances in neural information processing systems*, 33:7413–7424, 2020.

A. Blum, M. Gupta, G. Li, N. S. Manoj, A. Saha, and Y. Yang. Dueling optimization with a monotone adversary. *Neural Information Processing Systems (NeurIPS), 2023*, 2023.

T. K. Buening and A. Saha. Anaconda: An improved dynamic regret algorithm for adaptive non-stationary dueling bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3854–3878. PMLR, 2023.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

A. Désir, V. Goyal, S. Jagabathula, and D. Segev. Assortment optimization under the mallows model. In *Advances in Neural Information Processing Systems*, pages 4700–4708, 2016.

D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3304–3312. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/ding21d.html.

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

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

P. Gaillard, A. Saha, and S. Dan. One arrow, two kills: A unified framework for achieving optimal regret guarantees in sleeping bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7755–7773. PMLR, 2023.

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

S. Gupta and A. Saha. Optimal and efficient dynamic regret algorithms for non-stationary dueling bandits. In *International Conference on Machine Learning*, pages 19027–19049. PMLR, 2022.

Y. Han, J. Zeng, Y. Wang, Y. Xiang, and J. Zhang. Optimal contextual bandits with knapsacks under realizability via regression oracles. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5011–5035. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/han23b.html`.

N. Immorlica, K. Sankararaman, R. Schapire, and A. Slivkins. Adversarial bandits with knapsacks. *J. ACM*, 69(6), nov 2022. ISSN 0004-5411. doi: 10.1145/3557045. URL `https://doi.org/10.1145/3557045`.

K. C. Kalagarla, R. Jain, and P. Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8030–8037, May 2021. doi: 10.1609/aaai.v35i9.16979. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16979`.

J. Komiyama, J. Honda, and H. Nakagawa. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1235–1244, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/komiyama16.html`.

W. Kumagai. Regret analysis for continuous dueling bandit. In *Advances in Neural Information Processing Systems*, 2017.

T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

K. Lee, L. Smith, and P. Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

T. Lekang and A. G. Lamperski. Simple algorithms for dueling bandits. *ArXiv*, abs/1906.07611, 2019. URL `https://api.semanticscholar.org/CorpusID:189999589`.

C. li, I. Markov, M. Rijke, and M. Zoghi. Mergedts: A method for effective large-scale online ranker evaluation. *ACM Transactions on Information Systems*, 38:1–28, 10 2020. doi: 10.1145/3411753.

Z. Li, Z. Yang, and M. Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

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

A. Saha. Optimal algorithms for stochastic contextual dueling bandits. In *Advances in Neural Information Processing Systems*, 2021.

A. Saha and P. Gaillard. Dueling bandits with adversarial sleeping. *Advances in Neural Information Processing Systems*, 34, 2021.

A. Saha and S. Ghoshal. Exploiting correlation to achieve faster learning rates in low-rank preference bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 456–482. PMLR, 2022.

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

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

A. Saha and A. Gopalan. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, 2019.

A. Saha and A. Gopalan. Best-item learning in random utility models with subset choices. In *International Conference on Artificial Intelligence and Statistics*, pages 4281–4291. PMLR, 2020.

A. Saha and S. Gupta. Optimal and efficient dynamic regret algorithms for non-stationary dueling bandits. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19027–19049. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saha22b.html.

A. Saha and A. Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pages 968–994. PMLR, 2022.

A. Saha, T. Koren, and Y. Mansour. Dueling convex optimization. In *International Conference on Machine Learning*, pages 9245–9254. PMLR, 2021a.

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

A. Saha, T. Koren, and Y. Mansour. Dueling convex optimization with general preferences. *arXiv preprint arXiv:2210.02562*, 2022.

A. Saha, A. Pacchiano, and J. Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 6263–6289. PMLR, 2023.

V. Sivakumar, S. Zuo, and A. Banerjee. Smoothed adversarial linear contextual bandits with knapsacks. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20253–20277. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sivakumar22a.html.

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

A. Slivkins, K. A. Sankararaman, and D. J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression, 2023.

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

Y. Sui, M. Zoghi, K. Hofmann, and Y. Yue. Advancements in dueling bandits. In *IJCAI*, pages 5502–5510, 2018.

S. Vaswani, L. Yang, and C. Szepesvari. Near-optimal sample complexity bounds for constrained mdps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3110–3122. Curran Associates, Inc., 2022.

H. Wei, X. Liu, and L. Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (4):3868–3876, Jun. 2022. doi: 10.1609/aaai.v36i4.20302. URL https://ojs.aaai.org/index.php/AAAI/article/view/20302.

Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208, 2009.

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

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78 (5):1538–1556, 2012. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2011.12.028. URL https://www.sciencedirect.com/science/article/pii/S0022000012000281. JCSS Special Issue: Cloud Computing 2011.

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

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

M. Zoghi, Z. Karnin, S. Whiteson, and M. d. Rijke. Copeland dueling bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 307–315, Cambridge, MA, USA, 2015a. MIT Press.

M. Zoghi, S. Whiteson, and M. de Rijke. Mergeucb: A method for large-scale online ranker evaluation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 17–26, New York, NY, USA, 2015b. Association for Computing Machinery. ISBN 9781450333177. doi: 10.1145/2684822.2685290. URL https://doi.org/10.1145/2684822.2685290.

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 Canstrained 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 Saha [2022] and contextual scenarios Saha and Krishnamurthy [2022], Balsubramani et al. [2016], item unavailability Saha and Gaillard [2021], non-stationary preferences Buening 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], Désir 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, eg., 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 gurantees 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 $\epsilon_{\min}^{(c)} := \min_{i,j\in[K]}(|c(i)-c(j)|)$. Suppose the available budget $B = o\left(\frac{K}{{\epsilon_{\min}^{(c)}}^2}\right)$ 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 $\mathcal{I}_1, \ldots, \mathcal{I}_{K-1}$.

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

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

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

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

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

with $u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0,\ \forall i \in \{2, \dots, K\}$ and so on with $\mathcal{I}_{K-1}$ is defined by the following preference matrix:

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

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

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 $\text{OPT}^{(c)} = B + (T-B)\Big(\frac{1}{2}-\epsilon\Big)$.

Note that to be able to switch to arm $k+1$ in instance $\mathcal{I}_k$ the algorithm needs to identify the the arm with the highest Condorcet score $c(\cdot)$ among the arms $\{2, 3, \dots, K\}$. 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 $\Omega(\frac{K}{\epsilon^2})$ (see eg. [Slivkins, 2022]) and therefore any algorithm needs to play $(1, k),\ k \in [K]$ $\Omega(\frac{K}{\epsilon^2})$ number of times. However every time $(1, k)$ is played, 1 unit of resource is consumed and since the budget $B = o\left(\frac{K}{\epsilon^2}\right)$, 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

$$\text{REW}^{(c)} \le B + \frac{(T-B)}{2}\Big(\frac{1}{2}-\epsilon\Big) + \frac{(T-B)}{2}\Big(\frac{1}{2}-2\epsilon\Big)$$

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

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

$\mathcal{I}_2$ 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 $u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0,\ \forall i \in \{2, \dots, K\}$ 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:

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

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

$\mathcal{I}'_2$ 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 $u^*(1) = v^*(1) = 1, u^*(i) = v^*(i) = 0,\ \forall i \in \{2, \dots, K\}$ and so on.

Notice that although both set of instances $\mathcal{I}_1, \dots, \mathcal{I}_K$ and $\mathcal{I}'_1, \dots, \mathcal{I}'_K$ have total ordering, the total ordering in $\mathcal{I}'_1, \dots, \mathcal{I}'_K$ does not match the ordering of the Condorcet scores in $\mathcal{I}'_1, \dots, \mathcal{I}'_K$. Therefore dueling the sub-optimal arms $\{2, \dots, K\}$ 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 $\mathcal{I}'_1, \dots, \mathcal{I}'_K$ 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 $REG^{(c)}(T) = \Omega(T)$.*

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:

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

with the following consumptions: $u^*(1) = v^*(1) = 1, u^*(2) = v^*(2) = \epsilon$ and $u^*(3) = v^*(3) = 0$. The proof then follows as in the proof of Lemma 4.4. □

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 $B = o\left(\frac{K}{\epsilon^{(b)\ 2}_{\min}}\right)$ then there exists a preference matrix P such that $REG^{(b)} = \Omega(T)$.*

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 $\mathcal{O}(\frac{K}{\epsilon^2})$. Since $B = o(\frac{K}{\epsilon^2})$, 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 $B = o\left(\frac{K}{\epsilon^{(b)}_{\min}}\right)$ then there exists a preference matrix P such that $REG^{(b)} = \Omega(T)$.*

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 $\Theta(\frac{K}{\epsilon^2})$ number of times. However, playing arm 2 consumes ϵ amount and therefore $\Theta(\frac{K}{\epsilon^2})\epsilon$ budget is needed to differentiate between arms 1 and 2. Since $B = o(\frac{K}{\epsilon})$, any algorithm would incur a regret of $\Omega(T)$. □

C Proof of Regret Bound

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

$$\begin{aligned} REG^{(b)} &= OPT^{(b)} - REW^{(b)} \\ &\leq \frac{K}{K-1}(\widetilde{OPT^{(b)}} - \widetilde{REW^{(b)}}). \end{aligned}$$

Proof. Observe that

$$\begin{aligned}
\widetilde{\pi}^{*(b)} &= \underset{\pi_x,\pi_y\in\Delta^K}{\operatorname{argmax}} \sum_{x\in[K]} \pi_x(x)\tilde{b}(x) + \sum_{y\in[K]} \pi_y(y)\tilde{b}(y) \\
&\text{such that} \sum_{x,y\in[K]} \pi_x(x)u^*(x) + \pi_y(y)v^*(y) \le \frac{B}{T}\mathbf{1} \\
&= \underset{\pi_x,\pi_y\in\Delta^K}{\operatorname{argmax}} \sum_{x\in[K]} \pi_x(x)\left(\frac{K-1}{K}b(x) + \frac{1}{2K}\right) + \sum_{y\in[K]} \pi_y(y)\left(\frac{K-1}{K}b(y) + \frac{1}{2K}\right) \\
&\text{such that} \sum_{x,y\in[K]} \pi_x(x)u^*(x) + \pi_y(y)v^*(y) \le \frac{B}{T}\mathbf{1} \\
&= \underset{\pi_x,\pi_y\in\Delta^K}{\operatorname{argmax}} \sum_{x\in[K]} \pi_x(x)b(x) + \sum_{y\in[K]} \pi_y(y)b(y) \\
&\text{such that} \sum_{x,y\in[K]} \pi_x(x)u^*(x) + \pi_y(y)v^*(y) \le \frac{B}{T}\mathbf{1} \\
&= \pi^{*(b)},
\end{aligned}$$

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

$$\begin{aligned}
\widetilde{\mathrm{OPT}^{(b)}} &= T \sum_{x,y\in[K]} \widetilde{\pi}_x^{*(b)}\,\tilde{b}(x) + \widetilde{\pi}_y^{*(b)}\,\tilde{b}(y) \\
&= T \sum_{x,y\in[K]} \widetilde{\pi}_x^{*(b)}\left(\frac{K-1}{K}b(x) + \frac{1}{2K}\right) + \widetilde{\pi}_y^{*(b)}\left(\frac{K-1}{K}b(y) + \frac{1}{2K}\right) \\
&= \frac{K-1}{K}\left(T \sum_{x,y\in[K]} \pi_x^{*(b)}(x)\,b(x) + \pi_y^{*(b)}(y)\,b(y)\right) + \frac{T}{K} \\
&= \frac{K-1}{K}\mathrm{OPT}^{(b)} + \frac{T}{K}
\end{aligned}$$

Further

$$\begin{aligned}
\widetilde{\mathrm{REW}^{(b)}} &= \sum_{t=1}^{\tau} \tilde{b}(x_t) + \tilde{b}(y_t) \\
&= \sum_{t=1}^{\tau} \left(\frac{K-1}{K}b(x_t) + \frac{1}{2K}\right) + \left(\frac{K-1}{K}b(y_t) + \frac{1}{2K}\right) \\
&= \sum_{t=1}^{\tau} b(x_t) + b(y_t) + \frac{\tau}{K} \\
&= \mathrm{REW}^{(b)} + \frac{\tau}{K}
\end{aligned}$$

Therefore

$$\begin{aligned}
\widetilde{\mathrm{OPT}^{(b)}} - \widetilde{\mathrm{REW}^{(b)}} &= \frac{K-1}{K}\mathrm{OPT}^{(b)} - \mathrm{REW}^{(b)} + \frac{T-\tau}{K} \\
&\ge \mathrm{OPT}^{(b)} - \mathrm{REW}^{(b)}
\end{aligned}$$

where the last line follows because $\tau \le T$. □

Lemma 5.2. *Let the optimal value of* $(\mathtt{LP-Shifted\text{-}Borda\text{-}x})$ *and* $(\mathtt{LP-Shifted\text{-}Borda\text{-}y})$ *be* $\widetilde{OPT_x^{(b)}}$ *and* $\widetilde{OPT_y^{(b)}}$. *Then*

$$\widetilde{OPT_x^{(b)}} + \widetilde{OPT_y^{(b)}} \ge \widetilde{OPT^{(b)}}.$$

Proof. We define the solution to ($\texttt{LP} - \texttt{Shifted-Borda-x}$) as

$$\tilde{\pi}_x^{(b)} = \underset{\pi_x \in \Delta^K}{\text{argmax}} \sum_{x \in [K]} \pi_x(x)\tilde{b}(x)$$
$$\text{such that} \sum_{x \in [K]} \pi_x(x)u^*(x) \leq \frac{B}{2T}\mathbf{1}$$

and the solution to ($\texttt{LP} - \texttt{Shifted-Borda-x}$) as

$$\tilde{\pi}_y^{(b)} = \underset{\pi_x \in \Delta^K}{\text{argmax}} \sum_{x \in [K]} \pi_x(x)\tilde{b}(x)$$
$$\text{such that} \sum_{x \in [K]} \pi_x(x)u^*(x) \leq \frac{B}{2T}\mathbf{1}$$

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

$$\widetilde{\text{OPT}_x^{(b)}} + \widetilde{\text{OPT}_y^{(b)}} \geq \widetilde{\text{OPT}^{(b)}}.$$

□

Theorem 5.1. *For* $\eta_x = \left(\frac{\log K}{T\sqrt{K}}\right)^{2/3}\frac{1}{2Z_x+1}, \eta_y = \left(\frac{\log K}{T\sqrt{K}}\right)^{2/3}\frac{1}{2Z_y+1}$ *and* $\gamma_x = \sqrt{\eta_x K Z_x}, \gamma_y = \sqrt{\eta_y K Z_y}$*, the regret of* $\texttt{Vigilant D-EXP3}$ *is bounded by*

$$REG^{(b)}(T) \leq \tilde{\mathcal{O}}\Big(\big(\frac{OPT^{(b)}}{B} + 1\big)(K\log K)^{1/3}T^{2/3}\Big)$$

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

Step-1: Note that

$$\begin{aligned}
|\ \hat{\ell}_t^x(a)\ | &= \left| \hat{b}_t(i) + Z_x\lambda_t^\top\left(\frac{B}{2T}\mathbb{1} - \hat{u}_t(i)\right)\right| \\
&\leq |\hat{b}_t(i)| + Z_x\|\lambda_t\|_1\left(\frac{B}{2T}\|\mathbb{1}\|_\infty + \|\hat{u}_t\|_\infty\right) \\
&\leq 1 + Z_x(1+1) \\
&= 1 + 2Z_x
\end{aligned}$$

Therefore $\frac{1}{2Z_x+1}\hat{\ell}_t^x(a) \leq 1$ and using the regret guarantee of Exponential Weights algorithm Auer et al. [2002b], [Lattimore and Szepesvári, 2020, Chapter 11] we get for all $a \in [K]$

$$\sum_{t=1}^{\tau}\hat{\ell}_t^x(a) - \sum_{t=1}^{\tau}\sum_i \tilde{q}_t^x(i)\hat{\ell}_t^x(i) \leq \frac{\log K}{\eta_x} + \eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K}\tilde{q}_t^x(i)(\hat{\ell}_t^x(i))^2$$

Since $\tilde{q}_t^x(i) = \dfrac{q_t^x(i) - \frac{\gamma_x}{K}}{1-\gamma_x}$, we have

$$\forall a \in [K],\ (1-\gamma_x)\sum_{t=1}^{\tau}\hat{\ell}_t^x(a) - \sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)\hat{\ell}_t^x(i) \leq \frac{\log K}{\eta_x} + \eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)(\hat{\ell}_t^x(i))^2$$

and therefore,

$$\underbrace{(1-\gamma_x)\sum_{t=1}^{\tau}\sum_{a=1}^{K}\hat{\ell}_t^x(a)\tilde{\pi}_x^{*(b)}(a)}_{I} - \underbrace{\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)\hat{\ell}_t^x(i)}_{II} \leq \frac{\log K}{\eta_x} + \eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)(\hat{\ell}_t^x(i))^2 \tag{8}$$

Step-2: Next we relate the LHS of (8) to the regret using the following lemma.

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

$$\begin{aligned}(1-\gamma_x)\sum_{t=1}^{\tau}\sum_{a=1}^{K}\hat{\ell}_t^x(a)\tilde{\pi}_x^{*(b)}(a) - \sum_{t=1}^{\tau}\sum_a q_t^x(i)\hat{\ell}_t^x(a)\\ \geq \widetilde{OPT}_x^{(b)} - \mathbb{E}\sum_{t=1}^{\tau}\tilde{b}(x_t) - \mathcal{O}(Z+1)\sqrt{T\log T}\\ -\gamma_x(Z_x+1)T\end{aligned}$$

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

Proof. Let $\tilde{\pi}_x^{*(b)}$ be the solution of $(\texttt{LP}-\texttt{Shifted-Borda-x})$. Next define $\mathcal{H}_{t-1} = \sigma\left(\{x_i, y_i, o_i(x_i,y_i)\}_{i=1}^{t-1}\right)$ 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

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

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

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

Next observe that with $D_t = \tilde{b}(a) + Z_x\lambda_t^{x\top}\left(\frac{B}{2T}\mathbb{1} - u^*(a)\right) - \frac{\widetilde{\mathrm{OPT}}_x^{(b)}}{T}$ is adapted to $\mathcal{H}_t$, $|D_t| \leq 2(Z_x+1)$ and $\mathbb{E}_{\mathcal{H}_{t-1}}[D_t] \geq 0$ and therefore using Azuma-Hoeffding with probability at least $1-\mathcal{O}(1/T^2)$

$$\sum_{t=1}^{\tau}\sum_a\hat{\ell}_t^x(a)\tilde{\pi}_x^{*(b)}(a) \geq \frac{\tau}{T}\widetilde{\mathrm{OPT}}_x^{(b)} - 2(Z_x+1)\sqrt{T\log T}$$

Therefore

$$(1-\gamma_x)\sum_{t=1}^{\tau}\sum_a\hat{\ell}_t^x(a) \geq \frac{\tau}{T}\widetilde{\mathrm{OPT}}_x^{(b)} - \mathcal{O}(Z+1)\sqrt{T\log T} - \gamma_x(Z_x+1)T \tag{9}$$

Next consider term II from (8):

$$\sum_{t=1}^{\tau}\sum_a q_t^x(a)\hat{\ell}_t^x(a) = \sum_{t=1}^{\tau}\sum_a q_t^x(a)\left[\hat{b}_t(a) + Z_x\lambda_t^{x\top}\left(\frac{B}{2T}\mathbb{1} - \hat{u}_t^x(a)\right)\right]$$

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

$$\sum_{t=1}^{\tau}\sum_a q_t^x(a)\hat{\ell}_t^x(a) \leq \sum_{t=1}^{\tau}b(x_t) + Z_x\lambda_t^{x\top}\left(\frac{B}{2T}\mathbb{1} - u^*(x_t)\right) + \mathcal{O}(Z_x+1)\sqrt{T\log T}$$

Combining with (9) we get

$$(1-\gamma_x)\sum_{t=1}^{\tau}\hat{\ell}_t^x(a)-\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)\hat{\ell}_t^x(i)\geq\frac{\tau}{T}\widetilde{\mathrm{OPT}_x^{(b)}}-\sum_{t=1}^{\tau}b(x_t)-Z_x\lambda_t^{x\top}\left(\frac{B}{2T}\mathbb{1}-u^*(x_t)\right)\\-\mathcal{O}(Z_x+1)\sqrt{T\log T}-\gamma_x(Z_x+1)T \tag{10}$$

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

$$(1-\gamma_x)\sum_{t=1}^{\tau}\hat{\ell}_t^x(a)-\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)\hat{\ell}_t^x(i)\geq\frac{\tau}{T}\widetilde{\mathrm{OPT}_x^{(b)}}-\sum_{t=1}^{\tau}b(x_t)-Z_x\lambda^{\top}\left(\frac{B}{2T}\mathbb{1}-u^*(x_t)\right)\\-\mathcal{O}(Z_x+1)\sqrt{T\log T}-\gamma_x(Z_x+1)+\mathcal{O}(\sqrt{T}) \tag{11}$$

Next if $\tau=T$, choosing $\lambda=0$ gives

$$(1-\gamma_x)\sum_{t=1}^{\tau}\hat{\ell}_t^x(a)-\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)\hat{\ell}_t^x(a)\geq\mathrm{OPT}_x^{(b)}-\sum_{t=1}^{\tau}b(x_t)-\mathcal{O}(Z_x+1)\sqrt{T\log T}-\gamma_x(Z_x+1)T.$$

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 $\lambda=Z_xe_j$ and observe that

$$\sum_{t=1}^{T}Z_x\lambda^{\top}\left(\frac{B}{2T}\mathbb{1}-u^*(x_t)\right)\leq Z_x\left(\frac{\tau}{2T}B-B/2\right)$$

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

$$\begin{aligned}(1-\gamma_x)\sum_{t=1}^{\tau}\hat{\ell}_t^x(a)-\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)\hat{\ell}_t^x(i)&\geq\frac{\tau}{T}\widetilde{\mathrm{OPT}_x^{(b)}}-\sum_{t=1}^{\tau}b(x_t)-\frac{2\widetilde{\mathrm{OPT}_x^{(b)}}}{B}\left(\frac{\tau}{2T}B-B/2\right)\\&\quad-\mathcal{O}(Z_x+1)\sqrt{T\log T}-\gamma_x(Z_x+1)+\mathcal{O}(\sqrt{T})\\&\geq\widetilde{\mathrm{OPT}_x^{(b)}}-\sum_{t=1}^{\tau}b(x_t)-\mathcal{O}(Z_x+1)\sqrt{T\log T}-\gamma_x(Z_x+1)T\end{aligned}$$

which implies

$$(1-\gamma_x)\sum_{t=1}^{\tau}\hat{\ell}_t^x(a)-\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)\hat{\ell}_t^x(i)\geq\widetilde{\mathrm{OPT}_x^{(b)}}-\mathbb{E}\sum_{t=1}^{\tau}b(x_t)-\mathcal{O}(Z_x+1)\sqrt{T\log T}-\gamma_x(Z_x+1)T$$

Further,

$$\gamma_x(Z_x+1)T\leq\sqrt{\eta_xK}\sqrt{Z_x}(Z_x+1)T=\mathcal{O}\left((K\log K)^{1/3}Z_xT^{2/3}\right)$$

□

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* $\gamma_x=\sqrt{\eta_xKZ_x}$

$$\frac{\log K}{\eta_x}+\eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K}q_t^x(i)(\hat{\ell}_t^x(i))^2\\\leq\mathcal{O}\left(\left(\frac{OPT_x^{(b)}}{B}+1\right)(K\log K)^{1/3}T^{2/3}\right)$$

Proof. Consider the following term:

$$\begin{aligned}
\sum_{t=1}^{\tau}\sum_{a=1}^{K} q_t^x(a)\hat{\ell}_t^x(a)^2 &= \sum_{t=1}^{\tau}\sum_{a=1}^{K} q_t^x(a)\left(\hat{s}_t(a) + Z_x\lambda_t^{x\top}\left(\frac{B}{2T}\mathbb{1} - \hat{u}_t(x_t)\right)\right)^2 \\
&\le \sum_{t=1}^{\tau}\sum_{a=1}^{K} q_t^x(a)\left(\hat{s}_t(a)^2 + \left[Z_x\lambda_t^{x\top}\left(\frac{B}{2T}\mathbb{1} - \hat{u}_t(x_t)\right)\right]^2\right) \\
&\le \sum_{t=1}^{\tau}\sum_{a=1}^{K} q_t^x(a)\left(\hat{s}_t(a)^2 + \left[Z_x\lambda_t^{x\top}\left(\frac{B}{2T}\mathbb{1} - \hat{u}_t(x_t)\right)\right]^2\right) \\
&\le \sum_{t=1}^{\tau}\sum_{a=1}^{K} q_t^x(a)\hat{s}_t(a)^2 + \sum_{t=1}^{\tau}\sum_{a=1}^{K} 4Z_x^2\frac{B^2}{4T^2} + Z_x^2\sum_{t=1}^{\tau}\sum_{a=1}^{K} 4q_t^x(a)[\lambda_t^{x\top}\hat{u}_t^x(a)]^2
\end{aligned}$$

We have $\sum_a \mathbb{E}[q_t^x(a)\hat{s}_t(a)^2] \le \frac{K}{\gamma_x}$ and $\sum_a \mathbb{E}\, q_t^x(a)[\lambda_t^{x\top}\hat{u}_t(x_t)]^2 \le \frac{K}{\gamma_x}$ (see [Saha et al., 2021b, Lemma 6], [Lattimore and Szepesvári, 2020, Chapter 11]).

$$\begin{aligned}
\sum_{t=1}^{\tau}\sum_{a=1}^{K} q_t^x(a)\hat{\ell}_t^x(a)^2 &\le \frac{K}{\gamma_x}T + \frac{Z_x^2B^2}{T^2}\frac{K}{\gamma_x}T + \frac{K}{\gamma_x}Z_x^2T \\
&\le \frac{K}{\gamma_x}T(Z_x^2+2)
\end{aligned}$$

Therefore

$$\frac{\log K}{\eta_x} + \eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K} q_t^x(i)(\hat{\ell}_t^x(i))^2 \le \frac{\log K}{\eta_x} + \eta_x\frac{K}{\gamma_x}T(Z_x^2+2)$$

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

$$\frac{\log K}{\eta_x} + \eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K} q_t^x(i)(\hat{\ell}_t^x(i))^2 \le \frac{\log K}{\eta_x} + T\sqrt{\eta_x K}(Z_x^2+2)/\sqrt{Z_x}$$

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

$$\frac{\log K}{\eta_x} + \eta_x\sum_{t=1}^{\tau}\sum_{i=1}^{K} q_t^x(i)(\hat{\ell}_t^x(i))^2 \le \mathcal{O}\left((K\log K)^{1/3}T^{2/3}Z_x)\right) = \mathcal{O}\left(\left(\frac{\mathrm{OPT}_x^{(b)}}{B}+1\right)(K\log K)^{1/3}T^{2/3}\right)$$

□

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

$$\widetilde{\mathrm{OPT}_x^{(b)}} + \widetilde{\mathrm{OPT}_y^{(b)}} - \left(\sum_{t=1}^{\tau} b(x_t) + b(y_t)\right) \le \mathcal{O}\left(\left(\frac{\mathrm{OPT}_x^{(b)}}{B}+1\right)(K\log K)^{1/3}T^{2/3}\right)$$

which completes the proof.

□

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 $\bar{b} = \begin{pmatrix} b(1) & b(2)6 & \dots & b(6) \end{pmatrix}^\top$ is given by

$$\begin{pmatrix} 0.672 & 0.646 & 0.602 & 0.582 & 0.554 & 0.544 \end{pmatrix}^\top.$$

   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 $\bar{u}^* = \begin{pmatrix} u^*(1) & u^*(2)6 & \dots & u^*(6) \end{pmatrix}^\top$ are given by:

   (a) $\begin{pmatrix} 0.9 & 0.9 & 0.1 & 0.8 & 0.8 & 0.8 \end{pmatrix}^\top$

   (b) $\begin{pmatrix} 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \end{pmatrix}^\top$

   (c) $\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^\top$

   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) $\begin{pmatrix} 0.9 & 0.9 & 0.01 & 0.02 & 0.7 & 0.3 & 0.6 & 0.7 & 0.7 & 0.8 \end{pmatrix}$

   (b) $\begin{pmatrix} 0.7 & 0.9 & 0.9 & 0.8 & 0.6 & 0.1 & 0.4 & 0.3 & 0.5 & 0.2 \end{pmatrix}$

   (c) $\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

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