Batched Dueling Bandits

Arpit Argarwal [1] **Rohan Ghuge** [2] **Viswanath Nagarajan** [2]

Abstract

The K-armed dueling bandit problem, where the feedback is in the form of noisy pairwise comparisons, has been widely studied. Previous works have only focused on the sequential setting where the policy adapts after every comparison. However, in many applications such as search ranking and recommendation systems, it is preferable to perform comparisons in a limited number of *parallel batches*. We study the *batched K-armed dueling bandit* problem under two standard settings: (i) existence of a Condorcet winner, and (ii) strong stochastic transitivity and stochastic triangle inequality. For both settings, we obtain algorithms with a smooth trade-off between the number of batches and regret. Our regret bounds match the best known sequential regret bounds (up to poly-logarithmic factors), using only a logarithmic number of batches. We complement our regret analysis with a nearly-matching lower bound. Finally, we also validate our theoretical results via experiments on synthetic and real data.

1. Introduction

The K-armed dueling bandits problem has been widely studied in machine learning due to its applications in search ranking, recommendation systems, sports ranking, etc. (Yue & Joachims, 2011; Yue et al., 2012; Urvoy et al., 2013; Ailon et al., 2014; Zoghi et al., 2014; 2015a;b; Dudik et al., 2015; Jamieson et al., 2015; Komiyama et al., 2015; 2016; Ramamohan et al., 2016; Chen & Frazier, 2017). It is a variation of the traditional stochastic bandit problem in which feedback is obtained in the form of pairwise preferences. This problem falls under the umbrella of *preference learning* (Wirth et al., 2017), where the goal is to learn from relative feedback (in our case, given two alternatives, which of the two is preferred). Designing learning algorithms for such relative feedback becomes crucial in domains where qualitative feedback is easily obtained, but real-valued feedback would be arbitrary or not interpretable. We illustrate this using the web-search ranking application.

Web-search ranking is an example of a complex information retrieval system, where the goal is to provide a list (usually *ranked*) of candidate documents to the user of the system in response to a query (Radlinski et al., 2008; Joachims, 2002; Yue & Joachims, 2009; Hofmann et al., 2013). Modern day search engines comprise hundreds of parameters which are used to output a ranked list in response to a query. However, manually tuning these parameters can sometimes be infeasible, and online learning frameworks (based on user feedback) have been invaluable in automatically tuning these parameters (Liu, 2009). These methods do not affect user experience, enable the system to continuously learn about user preferences, and thus continuously adapt to user behavior. For example, given two rankings ℓ_1 and ℓ_2, they can be interleaved and presented to the user in such a way that clicks indicate which of the two rankings is more preferable to the user (Radlinski et al., 2008). The availability of such pairwise comparison data motivates the study of learning algorithms that exploit such relative feedback.

Previous learning algorithms have focused on a *fully adaptive* setting; in the web-ranking application this corresponds to the learning algorithm updating its parameters after each query. Such updates might be impractical in large systems. For example, if the parameters are fine-tuned for each user and a user makes multiple queries in a short time, such continuous updates require a lot of computational power. Even if users are assigned to a small number of classes (and parameters are fine-tuned for each user-class), multiple users from the same class may simultaneously query the system, making it impractical to adapt after each interaction.

Motivated by this, we introduce the *batched K-armed dueling bandits problem* (or, batched dueling bandits), where the learning algorithm is only allowed to *adapt a limited number of times*. Specifically, the algorithm uses at most B *adaptive rounds* and in each round it commits to a fixed *batch* of pairwise comparisons. The feedback for a batch is received simultaneously, and the algorithm chooses the next

[1]Data Science Institute, Columbia University, New York, USA [2]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, USA. Research supported in part by NSF grants CMMI-1940766 and CCF-2006778.. Correspondence to: Rohan Ghuge <rghuge@umich.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

batch based on this (and previous) feedback.

1.1. Contributions

- We design three algorithms, namely PCOMP, SCOMP, and SCOMP2, for batched dueling bandits under a finite time-horizon T. We analyze the regret of PCOMP under the Condorcet assumption, and that of SCOMP and SCOMP2 under the strong stochastic transitivity (SST) and stochastic triangle inequality (STI) assumptions. In all cases, we obtain a smooth trade-off between the expected regret and the number of batches, B.
- Specifically, in $\log(T)+1$ batches, SCOMP has expected regret nearly matching the instance-dependent regret bound due to Yue et al. (2012), up to a $\sqrt{K}$ factor (K is the number of arms).
- Furthermore, in $O(\log(T))$ batches, SCOMP2 achieves a worst-case regret matching the best known result in the sequential setting (Yue & Joachims, 2011) up to a logarithmic factor.
- To complement our upper bound results, we provide a lower bound that shows that a $T^{1/B}$ factor in the expected regret is necessary, where B is the number of batches.
- Finally, we run computational experiments to validate our theoretical results.

1.2. Preliminaries

The *K-armed dueling bandits* problem (Yue et al., 2012) is an online optimization problem, where the goal is to find the best among K bandits $\mathcal{B} = \{b_1, \ldots, b_K\}$ using noisy pairwise comparisons with low *regret*. In the traditional multi-armed bandit problem (Auer et al., 2002), an *arm* (or equivalently, bandit) b_j can be pulled at each time-step t, which generates a random reward from an unknown stationary distribution with expected value μ_j. However, in the K-armed dueling bandits problem, each iteration comprises a noisy comparison between two bandits (possibly the same), say (b_i, b_j). The outcome of the comparison is an independent random variable, and the probability of picking b_i over b_j is a constant denoted $P_{i,j} = \frac{1}{2} + \epsilon_{i,j}$ where $\epsilon_{i,j} \in (-\frac{1}{2}, \frac{1}{2})$. Here $\epsilon_{i,j}$ can be thought of as a measure of distinguishability between the two bandits, and we use $b_i \succ b_j$ when $\epsilon_{i,j} > 0$. We also refer to $\epsilon_{i,j}$ as the *gap* between b_i and b_j.

Throughout the paper, we let b_1 refer to the best bandit. To further simplify notation, we define $\epsilon_j = \epsilon_{1,j}$; that is, the gap between b_1 and b_j. We define the *regret* per time-step as follows: suppose bandits b_{t_1} and b_{t_2} are chosen in iteration t, then the regret $r(t) = \frac{\epsilon_{t_1} + \epsilon_{t_2}}{2}$. The cumulative regret up to time T is $R(T) = \sum_{t=1}^{T} r(t)$, where T is the time horizon, and it's assumed that $K \leq T$. The cumulative regret can be equivalently stated as $R(T) = \frac{1}{2}\sum_{j=1}^{K} T_j \epsilon_j$, where T_j denotes the number comparisons involving b_j. We define $\epsilon_{\min} = \min_{j:\epsilon_j > 0} \epsilon_j$ to be the smallest non-zero gap of any bandit with b_1. We say that bandit b_i is a *Condorcet winner* if, and only if, $P_{i,j} \geq \frac{1}{2}$ for all $j \in \mathcal{B} \setminus \{i\}$. Furthermore, we say that the probabilistic comparisons exhibit *strong stochastic transitivity* (SST) if there exists a total ordering, denoted by $\succeq$, over arms such that for every triple $b_i \succeq b_j \succeq b_k$, we have $\epsilon_{i,k} \geq \max\{\epsilon_{i,j}, \epsilon_{j,k}\}$, and exhibits *stochastic triangle inequality* (STI) if for every triple $b_i \succeq b_j \succeq b_k$, $\epsilon_{i,k} \leq \epsilon_{i,j} + \epsilon_{j,k}$.

1.3. Batch Policies

In traditional bandit settings, actions are performed *sequentially*, utilizing the results of *all prior actions* in determining the next action. In the batched setting, the algorithm must commit to a round (or *batch*) of actions to be performed *in parallel*, and can only observe the results after all actions in the batch have been performed. More formally, in round $r = 1, 2, \ldots$, the algorithm must decide the comparisons to be performed; afterwards *all* outcomes of the comparisons in batch r are received. The algorithm can then, *adaptively*, select the next batch of comparisons. However, it can use at most a given number, B, of batches.

The batch sizes can be chosen *non-adaptively* (fixed upfront) or *adaptively*. In an adaptive policy the batch sizes may even depend on previous observations of the algorithm. An adaptive policy is more powerful than a non-adaptive policy, and may suffer a smaller regret. In this paper, we focus on such adaptive policies. Furthermore, note that the total number of comparisons (across all batches) must sum to T. We assume that the values of T and B are known. Observe that when $T = B$, we recover the fully sequential setting.

1.4. Results and Techniques

We provide a summary of our results in Table 1. Our first result is as follows.

Theorem 1.1. *For any integer $B > 1$, there is an algorithm for batched dueling bandits that uses at most B rounds, and if the instance admits a Condorcet winner, the expected regret is bounded by*

$$\mathbb{E}[R(T)] \leq 3KT^{1/B} \log\left(6TK^2B\right) \sum_{j:\epsilon_j > 0} \frac{1}{\epsilon_j}.$$

The above bound is an instance-dependent bound. To obtain an instance-independent bound, recall that $\epsilon_{\min} = \min_{j:\epsilon_j > 0} \epsilon_j$. We get that the expected worst-case regret

Table 1: A summary of our results

Setting	Fully Adaptive (prior work)	**Our Algorithms**		**Our Lower Bound**
		Regret	Rounds	(for B rounds)
Condorcet	$O\left(\frac{K\log T}{\epsilon_{\min}}\right)+O\left(\frac{K^2}{\epsilon_{\min}^2}\right)$	$O\left(\frac{K^2T^{1/B}\log(T)}{\epsilon_{\min}}\right)$	B	$\Omega\left(\frac{KT^{1/B}}{B^2\epsilon_{\min}}\right)$
SST + STI	$O\left(\frac{K\log(T)}{\epsilon_{\min}}\right)$	$O\left(\frac{KBT^{1/B}\log(T)}{\epsilon_{\min}}\right)$	$2B+1$	$\Omega\left(\frac{KT^{1/B}}{B^2\epsilon_{\min}}\right)$

is bounded by

$$\mathbb{E}[R(T)] \leq \frac{3K^2T^{1/B}\log\left(6TK^2B\right))}{\epsilon_{\min}}.$$

In the sequential setting, existing algorithms achieve a worst-case expected regret of $O\left(\frac{K\log T}{\epsilon_{\min}}\right)+O\left(\frac{K^2}{\epsilon_{\min}^2}\right)$ (Zoghi et al., 2014; Komiyama et al., 2015). When $B=\log(T)$, our worst-case regret is at most

$$\mathbb{E}[R(T)] \leq 3K^2\log(6TK^2B)/\epsilon_{\min} = O(K^2\log(T)/\epsilon_{\min}),$$

which nearly matches the best-known bound in the sequential setting. Our algorithm in Theorem 1.1 proceeds by performing all pairwise comparisons in an *active set* of bandits, and gradually eliminating sub-optimal bandits. This algorithm is straightforward, and its analysis follows that of (Esfandiari et al., 2021a) for batched stochastic multi-armed bandits. Although this is a simple result, it is an important step for our main results, described next.

Our main results are when the instance satisfies the SST and STI conditions. These conditions impose a structure on the pairwise preference probabilities, and we are able to exploit this additional structure to obtain improved bounds.

Theorem 1.2. *For any integer $B>1$, there is an algorithm for batched dueling bandits that uses at most $B+1$ rounds, and if the instance satisfies the SST and STI assumptions, the expected regret is bounded by*

$$\mathbb{E}[R(T)] = \sum_{j:\epsilon_j>0} O\left(\frac{\sqrt{K}T^{1/B}\log(T)}{\epsilon_j}\right).$$

The idea behind this algorithm is to first sample a "sufficiently small" *seed set*, and then to perform all pairwise comparisons between the seed set and the active set to eliminate sub-optimal arms. The idea is to exploit the structure of pairwise probabilities so that we do not need to perform *all* pairwise comparisons. Additionally, if the seed set is found to be sub-optimal, we can construct a *much smaller* active set; thus allowing us to switch to the pairwise comparison policy. In the sequential setting, (Yue et al., 2012) obtain instance-dependent regret bounded by $\sum_{j:\epsilon_j>0} O\left(\frac{\log(T)}{\epsilon_j}\right)$. Our result nearly matches this sequential bound (with an extra multiplicative factor of $\sqrt{K}$) when $B=\log(T)$. Observe that the worst-case regret of (Yue & Joachims, 2011) in the sequential setting is bounded by $O\left(\frac{K\log(T)}{\epsilon_{\min}}\right)$, while we obtain $\mathbb{E}[R(T)] \leq O\left(\frac{K\sqrt{K}T^{1/B}\log(T)}{\epsilon_{\min}}\right)$.

Next, we improve the worst-case regret by reducing the comparisons performed as follows. We first perform pairwise comparisons amongst bandits in the seed set, and pick a candidate bandit. This candidate bandit is used to eliminate sub-optimal arms from the active set. Although selecting a candidate bandit each time requires additional adaptivity, we get a better bound on the worst-case expected regret by exploiting the fact that there can be at most B candidate bandits.

Theorem 1.3. *For any integer $B>1$, there is an algorithm for batched dueling bandits that uses at most $2B+1$ rounds, and if the instance satisfies the SST and STI assumptions, the expected worst-case regret is bounded by*

$$\mathbb{E}[R(T)] = O\left(\frac{KBT^{1/B}\log(T)}{\epsilon_{\min}}\right).$$

Thus, in $B=\log(T)$ rounds, our expected worst-case regret is bounded by $E[R(T)] \leq O\left(\frac{K\log^2(T)}{\epsilon_{\min}}\right)$ matching the best known result in the sequential setting up to an additional logarithmic factor.

Finally, we complement our upper bound results with a lower bound for the batched K-armed dueling bandits problem, even under the SST and STI assumptions.

Theorem 1.4. *Given an integer $B>1$, and any algorithm that uses at most B batches, there exists an instance of the K-armed batched dueling bandit problem that satisfies the SST and STI condition such that the expected regret*

$$\mathbb{E}[R(T)] = \Omega\left(\frac{KT^{1/B}}{B^2\epsilon_{\min}}\right).$$

The above lower bound shows that the $T^{1/B}$ dependence in our upper bounds is necessary. Note that the above lower

bound also applies to the more general Condorcet winner setting. The proof is similar to the lower bound proof in (Gao et al., 2019) for batched multi-armed bandits. The main novelty in our proof is the design of a family of hard instances with different values of $\epsilon_{\min}$'s that satisfy the SST and STI conditions. We defer further discussion and the proof of Theorem 1.4 to Appendix C.

2. Related Work

The dueling bandits problem has been widely studied in recent years; we mention the most relevant works here and refer the reader to (Sui et al., 2018) for a more comprehensive survey. This problem was first studied by (Yue et al., 2012) under the SST and STI setting. The authors gave a worst-case regret upper bound of $\widetilde{O}(K \log T/\epsilon_{\min})$ and provided a matching lower bound. (Yue & Joachims, 2011) considered a slightly more general version of the SST and STI setting and achieved an instance-wise optimal regret upper bound of $\sum_{j:\epsilon_j>0} O\left(\frac{\log(T)}{\epsilon_j}\right)$. (Urvoy et al., 2013) studied this problem under the Condorcet winner setting and proved a $O(K^2 \log T/\epsilon_{\min})$ regret upper bound, which was improved by (Zoghi et al., 2014) to $O(K^2/\epsilon_{\min}^2) + \sum_{j:\epsilon_j>0} O(\log T/\epsilon_j^2)$. (Komiyama et al., 2015) achieved a similar but tighter KL divergence-based bound, which is shown to be *asymptotically instance-wise optimal* (even in terms constant factors). There are also other works that improve the dependence on K in the upper bound, but suffer a worse dependence on ϵ_js (Zoghi et al., 2015b). This problem has also been studied under other noise models such as utility based models (Ailon et al., 2014) and other notions of regret (Chen & Frazier, 2017). Alternate notions of winners such as Borda winner (Jamieson et al., 2015), Copeland winner (Zoghi et al., 2015a; Komiyama et al., 2016; Wu & Liu, 2016), and von Nuemann winner (Dudik et al., 2015) have also been considered. There are also several works on extensions of dueling bandits that allow multiple arms to be compared at once (Sui et al., 2017; Agarwal et al., 2020; Saha & Gopalan, 2019).

All of the aforementioned works on the dueling bandits problem are limited to the sequential setting. To the best of our knowledge, ours is the first work that considers the batched setting for dueling bandits. However, batched processing for the stochastic multi-armed bandit problem has been investigated in the past few years. A special case when there are two bandits was studied by (Perchet et al., 2016). They obtain a worst-case regret bound of $O\left(\left(\frac{T}{\log(T)}\right)^{1/B} \frac{\log(T)}{\epsilon_{\min}}\right)$. (Gao et al., 2019) studied the general problem and obtained a worst-case regret bound of $O\left(\frac{K \log(K) T^{1/B} \log(T)}{\epsilon_{\min}}\right)$, which was later improved by (Esfandiari et al., 2021a) to $O\left(\frac{K T^{1/B} \log(T)}{\epsilon_{\min}}\right)$. Furthermore, (Esfandiari et al., 2021a) obtained an instance-dependent regret bound of $\sum_{j:\epsilon_j>0} T^{1/B} O\left(\frac{\log(T)}{\epsilon_j}\right)$. Our results for batched dueling bandits are of a similar flavor; that is, we get a similar dependence on T and B. (Esfandiari et al., 2021a) also give batched algorithms for stochastic linear bandits and adversarial multi-armed bandits.

Adaptivity and batch processing has been recently studied for stochastic submodular cover (Golovin & Krause, 2017; Agarwal et al., 2019; Esfandiari et al., 2021b; Ghuge et al., 2021), and for various stochastic "maximization" problems such as knapsack (Dean et al., 2008; Bhalgat et al., 2011), matching (Bansal et al., 2012; Behnezhad et al., 2020), probing (Gupta & Nagarajan, 2013) and orienteering (Guha & Munagala, 2009; Gupta et al., 2015; Bansal & Nagarajan, 2015). Recently, there have also been several results examining the role of adaptivity in (deterministic) submodular optimization; e.g. (Balkanski & Singer, 2018a; Balkanski et al., 2018; Balkanski & Singer, 2018b; Balkanski et al., 2019; Chekuri & Quanrud, 2019).

3. Algorithms for Batched Dueling Bandits

In this section, we present three algorithms, namely PCOMP, SCOMP and SCOMP2, for the K-armed batched dueling bandits problem. Recall that given a set of K bandits (or arms) $\mathcal{B} = \{b_1, \ldots, b_K\}$, and a positive integer $B \leq T$, we wish to find a sequence of B batches of noisy comparisons with low regret. Given bandits b_i and b_j, $P_{i,j} = \frac{1}{2} + \epsilon_{i,j}$ denotes the probability of b_i winning over b_j. The first algorithm, termed PCOMP, proceeds by performing all-pairs comparisons amongst bandits in an *active* set, and gradually eliminating sub-optimal bandits. The other two algorithms, termed SCOMP and SCOMP2, first select a (sufficiently small) *seed* set $\mathcal{S} \subset \mathcal{B}$, and eliminate bandits in an *active* set by successively comparing them to (all or few) bandits in $\mathcal{S}$. If the seed set $\mathcal{S}$ is itself found to be *sub-optimal* in a subsequent round, then these algorithms call the all-pairs algorithm PCOMP over the remaining *active* arms.

Before describing our algorithms in detail we will set up some basic notation. We will denote by $\mathcal{A}$ the set of *active* arms, i.e. arms that have not been eliminated. We will use index r for rounds or batches. At the end of each round r, our algorithms compute a fresh estimate of the pairwise probabilities based on the feedback from comparisons in round r as:

$$\widehat{P}_{i,j} = \frac{\#b_i \text{ wins against } b_j \text{ in round } r}{\#\text{comparisons of } b_i \text{ and } b_j \text{ in round } r}. \quad (1)$$

If a pair (b_i, b_j) is compared in round r, it is compared $c_r = \lfloor q^r \rfloor$ times. In round r, the parameter $\gamma_r = \sqrt{\log\left(\frac{1}{\delta}\right)/2c_r}$ is used to eliminate bandits from the active set (the specific elimination criteria depends on the algorithm).

Algorithm 1 PCOMP(ALL PAIRS COMPARISONS)

Input: Bandits $\mathcal{B}$, time-horizon T, rounds B, comparison parameters q and τ
$K \leftarrow |\mathcal{B}|$, $\delta \leftarrow \frac{1}{6TK^2B}$, active bandits $\mathcal{A} \leftarrow \mathcal{B}$, $c_r \leftarrow \lfloor q^{r+\tau-1} \rfloor$, $\gamma_r \leftarrow \sqrt{\log(1/\delta)/2c_r}$, $r \leftarrow 1$
while number of comparisons $\leq T$ **do**
  for all $(b_i, b_j) \in \mathcal{A}^2$, perform c_r comparisons and compute $\widehat{P}_{i,j}$ using Eq(1).
  if $\exists\, b_i, b_j$ such that $\widehat{P}_{i,j} > \frac{1}{2} + \gamma_r$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{b_j\}$
  end if
  $r \leftarrow r + 1$
end while

3.1. All Pairs Comparisons

We first describe the PCOMP algorithm. This algorithm takes as input the set of bandits $\mathcal{B}$, time-horizon T, rounds B and comparison parameters q and τ. We will set the parameters $q = T^{1/B}$ and $\tau = 1$, unless otherwise specified.[1] In round $r \in [B]$, this algorithm compares each pair $(b_i, b_j) \in \mathcal{A}^2$ for c_r times. It then computes fresh estimates of the pairwise probabilities $\widehat{P}_{i,j}$ for all $(b_i, b_j) \in \mathcal{A}^2$. If, for some bandit b_j, there exists bandit b_i such that $\widehat{P}_{i,j} > \frac{1}{2} + \gamma_r$, then bandit b_j is eliminated from $\mathcal{A}$. We provide the pseudo-code in Algorithm 1.

The following theorem (see Appendix B for proof) describes the regret bound obtained by PCOMP under the Condorcet assumption, and formalizes Theorem 1.1.

Theorem 3.1. *Given any set $\mathcal{B}$ of K bandits, time-horizon T, rounds B, parameters $q = T^{1/B}$ and $\tau = 1$, the expected regret of* PCOMP *for the batched K-armed dueling bandits problem under the Condorcet assumption is at most*

$$\mathbb{E}[R(T)] \leq 3KT^{1/B} \log\left(6TK^2B\right) \sum_{j:\epsilon_j > 0} \frac{1}{\epsilon_j}.$$

Setting $\epsilon_{\min} := \min_{j:\epsilon_j>0} \epsilon_j$, we get

$$\mathbb{E}[R(T)] \leq \frac{3K^2T^{1/B} \log\left(6TK^2B\right)}{\epsilon_{\min}}.$$

3.2. Seeded Comparisons Algorithms

In this section, we present two algorithms for the batched dueling bandits problem, namely SCOMP and SCOMP2. The algorithms work in *two phases*:

- In the first phase, the algorithms sample a *seed set* $\mathcal{S}$ by including each bandit from $\mathcal{B}$ *independently* with probability $1/\sqrt{K}$. This seed set is used to eliminate bandits from the active set $\mathcal{A}$.
- Under certain *switching* criteria, the algorithms enter the second phase which involves running algorithm PCOMP on some of the remaining bandits.

The algorithms differ in how the candidate set is used to eliminate active bandits in the first phase.

In SCOMP, *all* pairwise comparisons between $\mathcal{S}$ (seed set) and $\mathcal{A}$ (active bandits) are performed. Specifically, in round r, every active bandit is compared with every bandit in $\mathcal{S}$ for c_r times. If, for some bandit b_j, there exists bandit b_i such that $\widehat{P}_{i,j} > \frac{1}{2} + 3\gamma_r$, then bandit b_j is eliminated (from $\mathcal{A}$ as well as $\mathcal{S}$); note that the elimination criteria here is stricter than in PCOMP. If, in some round r, there exists bandit b_j such that b_j eliminates *all* bandits $b_i \in \mathcal{S}$, then the algorithm constructs a set $\mathcal{A}^* = \{b_j \in \mathcal{A} \mid \widehat{P}_{j,i} > \frac{1}{2} + \gamma_r \text{ for all } b_i \in \mathcal{S}\}$, and invokes PCOMP on bandits $\mathcal{A}^*$ with starting batch r. This marks the beginning of the second phase, which continues until time T. We provide the pseudocode in Algorithm 2.

Algorithm 2 SCOMP(SEEDED COMPARISONS)

Input: Bandits $\mathcal{B}$, time-horizon T, rounds B
$q \leftarrow T^{1/B}$, $\delta \leftarrow \frac{1}{6TK^2B}$, active bandits $\mathcal{A} \leftarrow \mathcal{B}$, $c_r \leftarrow \lfloor q^r \rfloor$, $\gamma_r \leftarrow \sqrt{\log(1/\delta)/2c_r}$, $r \leftarrow 1$
$\mathcal{S} \leftarrow$ add elements from $\mathcal{B}$ into $\mathcal{S}$ w.p. $1/\sqrt{K}$
while number of comparisons $\leq T$ **do**
  for all $(b_i, b_j) \in \mathcal{S} \times \mathcal{A}$, compare b_i and b_j for c_r times and compute $\widehat{P}_{i,j}$
  if $\exists b_i \in \mathcal{S}, b_j \in \mathcal{A}, \widehat{P}_{i,j} > \frac{1}{2} + 3\gamma_r$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{b_j\}$, $\mathcal{S} \leftarrow \mathcal{S} \setminus \{b_j\}$
  end if
  if $\exists b_j$ such that $\widehat{P}_{j,i} > \frac{1}{2} + 3\gamma_r$ for all $b_i \in \mathcal{S}$ **then**
    construct set $\mathcal{A}^* = \{b_j \in \mathcal{A} \mid \widehat{P}_{j,i} > \frac{1}{2} + \gamma_r \text{ for all } b_i \in \mathcal{S}\}$
    $r^* \leftarrow r$, $T^* \leftarrow$ # comparisons until round r^*, **break**
  end if
  $r \leftarrow r + 1$
end while
run PCOMP$(\mathcal{A}^*, T - T^*, q, r^*)$

We obtain the following result, which formalizes Theorem 1.2, when the given instance satisfies SST and STI.

Theorem 3.2. *Given any set $\mathcal{B}$ of K bandits, time-horizon T, parameter B,* SCOMP *uses at most $B + 1$ batches, and has expected regret bounded by*

$$\mathbb{E}[R(T)] = \sum_{j:\epsilon_j > 0} O\left(\frac{\sqrt{K}T^{1/B} \log(T)}{\epsilon_j}\right)$$

[1] We allow general parameters q and τ in order to allow PCOMP to be used in conjunction with other policies.

under the strong stochastic transitivity and stochastic triangle inequality assumptions.

Observe that this gives a worst-case regret bound of $O\left(\frac{K\sqrt{K}T^{1/B}\log(T)}{\epsilon_{\min}}\right)$ for SCOMP under SST and STI. We can improve this by sampling each bandit from $\mathcal{B}$ independently into the seed set with probability $K^{-2/3}$: this gives us the following result.

Theorem 3.3. *Given any set $\mathcal{B}$ of K bandits, time-horizon T, parameter B, there is an algorithm that uses at most $B+1$ batches, and has worst-case regret bounded by*

$$\mathbb{E}[R(T)] = O\left(\frac{K^{4/3}T^{1/B}\log(T)}{\epsilon_{\min}}\right)$$

under the strong stochastic transitivity and stochastic triangle inequality assumptions.

To further improve this worst-case bound, we *add more rounds of adaptivity* in SCOMP to obtain SCOMP2. Specifically, each round r in the first phase is divided into two rounds of adaptivity.

- In the first round $r^{(1)}$, pairwise comparisons among the bandits in $\mathcal{S}$ are performed, and an undefeated $b_{i_r^*}$ is selected as a *candidate*. We say that b_i *defeats* b_j if $\widehat{P}_{i,j} > \frac{1}{2} + \gamma_r$
- In the second round $r^{(2)}$, the candidate $b_{i_r^*}$ is used to eliminate active bandits. A bandit b_j is eliminated if $\widehat{P}_{i_r^*,j} > \frac{1}{2} + 5\gamma_r$.

The switching criterion in SCOMP2 is different from that of SCOMP. Here, if in some round r, there is a bandit b_j such that b_j eliminates $b_{i_r^*}$, then the algorithm constructs set $\mathcal{A}^* = \{b_j \in \mathcal{A} \mid \widehat{P}_{j,i_r^*} > \frac{1}{2} + 3\gamma_r\}$, and invokes PCOMP on bandits $\mathcal{A}^*$ with starting batch r. See Algorithm 3 for a formal description.

We show that SCOMP2 obtains an improved worst-case regret bound (at the cost of additional adaptivity) over SCOMP when the given instance satisfies SST and STI, thus proving Theorem 1.3.

Theorem 3.4. *Given any set $\mathcal{B}$ of K bandits, time-horizon T and parameter B,* SCOMP2 *uses at most $2B+1$ batches, and has worst-case expected regret bounded by*

$$\mathbb{E}[R(T)] = O\left(\frac{KBT^{1/B}\log(T)}{\epsilon_{\min}}\right)$$

under strong stochastic transitivity and stochastic triangle inequality, where $\epsilon_{\min} := \min_{j:\epsilon_j>0}\epsilon_j$.

The proofs of Theorems 3.2 and 3.4 can be found in Appendix B.

Algorithm 3 SCOMP2 (SEEDED COMPARISONS 2)

Input: Bandits $\mathcal{B}$, time-horizon T, rounds B
$q \leftarrow T^{1/B}$, $\delta \leftarrow \frac{1}{6TK^2B}$, active bandits $\mathcal{A} \leftarrow \mathcal{B}$, $c_r \leftarrow \lfloor q^r \rfloor$, $\gamma_r \leftarrow \sqrt{\log(1/\delta)/2c_r}$, $r \leftarrow 1$
$\mathcal{S} \leftarrow$ add elements from $\mathcal{B}$ into $\mathcal{S}$ w.p. $1/\sqrt{K}$
while number of comparisons $\leq T$ **do**
  $r^{(1)}$: compare all pairs in $\mathcal{S}$ for c_r times; get $\widehat{P}_{i,j}$.
  candidate $b_{i_r^*} \leftarrow$ any bandit $i \in \mathcal{S}$ with $\max_{j\in\mathcal{S}} \widehat{P}_{j,i} \leq \frac{1}{2} + \gamma_r$.
  $r^{(2)}$: for all $b_j \in \mathcal{A}$, compare $b_{i_r^*}$ and b_j for c_r times and compute $\widehat{P}_{i_r^*,j}$.
  if $\exists b_j \in \mathcal{A}$, $\widehat{P}_{i_r^*,j} > \frac{1}{2} + 5\gamma_r$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{b_j\}$, $\mathcal{S} \leftarrow \mathcal{S} \setminus \{b_j\}$
  end if
  if $\exists b_j$ such that $\widehat{P}_{j,i_r^*} > \frac{1}{2} + 5\gamma_r$ **then**
    construct set $\mathcal{A}^* = \{b_j \in \mathcal{A} \mid \widehat{P}_{j,i_r^*} > \frac{1}{2} + 3\gamma_r\}$
    $r^* \leftarrow r$, $T^* \leftarrow$ # comparisons until round r^*, **break**
  end if
  $r \leftarrow r + 1$
end while
run PCOMP$(\mathcal{A}^*, T - T^*, q, r^*)$

4. Regret Analysis

We present a sketch of the regret analysis for the algorithms described in §3 in this section. Refer to Appendix B for complete proofs.

The following lemma follows from a direct application of Hoeffding's inequality.

Lemma 4.1. *For any batch $r \in [B]$, and for any pair b_i, b_j that are compared c_r times, we have*

$$\mathbf{P}\left(|P_{i,j} - \widehat{P}_{i,j}| > \gamma_r\right) \leq 2\delta,$$

where $\gamma_r = \sqrt{\log(\frac{1}{\delta})/2c_r}$.

We analyze the regret of our algorithms under a *good* event, G. We show that the G occurs with high probability; in the event that G does not occur (denoted $\overline{G}$), we incur a regret of T. Towards defining G, we say that an estimate $\widehat{P}_{i,j}$ at the end of batch r is *correct* if $|\widehat{P}_{i,j} - P_{i,j}| \leq \gamma_r$. We say that G occurs if every estimate in every batch is correct.

Lemma 4.2. *The probability that every estimate in every batch of* PCOMP, SCOMP, *and* SCOMP2 *is correct is at least $1 - 1/T$.*

Proof. Applying Lemma 4.1 and taking a union bound over all pairs and batches (note SCOMP2 has at most $2B+1 \leq 3B$ batches), we get that the probability that some estimate

is incorrect is at most $K^2 \times 3B \times 2\delta = \frac{1}{T}$ where $\delta = 1/6K^2BT$. Thus, $\mathbf{P}(\overline{G}) \leq \frac{1}{T}$. □

Using Lemma 4.2, the expected regret (of *any* algorithm) can be written as follows:

$$\begin{aligned}\mathbb{E}[R(T)] &= \mathbb{E}[R(T) \mid G] \cdot \mathbf{P}(G) + \mathbb{E}[R(T) \mid \overline{G}] \cdot \mathbf{P}(\overline{G}) \\ &\leq \mathbb{E}[R(T) \mid G] + T \cdot \frac{1}{T} = \mathbb{E}[R(T) \mid G] + 1 \qquad (2)\end{aligned}$$

The proof of Theorem 3.1 can be found in Appendix B.

4.1. Proofs of Theorems 3.2, 3.3 and 3.4

In this section, we discuss the proofs of Theorems 3.2, 3.3 and 3.4. Henceforth, we assume the SST and STI properties. We need the following definition. For a bandit b_j, let $E_j = \{b_i \in \mathcal{B} : \epsilon_{i,j} > 0\}$; that is, the set of bandits superior to bandit b_j. We define $rank(b_j) = |E_j|$. [2]

As before, we analyze the regret of SCOMP and SCOMP2 under event G. By Lemma 4.2 and (2), we only need to bound the expected regret under G; that is, we need to bound $\mathbb{E}[R(T) \mid G]$. Conditioned on event G, the following Lemmas 4.3,4.4 and 4.5 hold for both SCOMP and SCOMP2.

Lemma 4.3. *The best bandit b_1 is never deleted from $\mathcal{A}$ in the elimination step of phase I.*

Lemma 4.4. *When the algorithm switches to* PCOMP *on set $\mathcal{A}^*$, we have $b_1 \in \mathcal{A}^*$ and $|\mathcal{A}^*| \leq rank(b_{i^*_{\mathcal{S}}})$ where $b_{i^*_{\mathcal{S}}}$ is the best bandit in $\mathcal{S}$.*

Lemma 4.5. *We have $\mathbb{E}[rank(b_{i^*_{\mathcal{S}}})] \leq \sqrt{K}$ and $\mathbb{E}[rank(b_{i^*_{\mathcal{S}}})^2] \leq 2K$.*

Using Lemmas 4.3, 4.4 and 4.5, we complete the proof of Theorem 3.2.

Proof of Theorem 3.2. We bound the expected regret of SCOMP conditioned on G. Let R_1 and R_2 denote the regret incurred in phase I and II respectively.

Bounding R_1. Fix a bandit b_j. Let r denote the last round such that $b_j \in \mathcal{A}$ *and* switching does not occur (at the end of round r). Let $b_{i^*_{\mathcal{S}}}$ be the best bandit in $\mathcal{S}$. As j is not eliminated by $b_{i^*_{\mathcal{S}}}$, we have $\widehat{P}_{i^*_{\mathcal{S}},j} \leq \frac{1}{2} + 3\gamma_r$, which implies (by event G) $P_{i^*_{\mathcal{S}},j} \leq \frac{1}{2} + 4\gamma_r$. Moreover, as switching doesn't occur, we have $\min_{i\in\mathcal{S}} \widehat{P}_{1,i} \leq \frac{1}{2} + 3\gamma_r$ (by Lemma 4.3, b_1 is never deleted from $\mathcal{A}$). We now claim that $P_{1,i^*_{\mathcal{S}}} \leq \frac{1}{2} + 4\gamma_r$. Otherwise, by SST we have $\min_{i\in\mathcal{S}} P_{1,i} = P_{1,i^*_{\mathcal{S}}} > \frac{1}{2} + 4\gamma_r$, which (by event G) implies $\min_{i\in\mathcal{S}} \widehat{P}_{1,i} > \frac{1}{2} + 3\gamma_r$, a contradiction! It now follows that $\epsilon_{i^*_{\mathcal{S}},j} \leq 4\gamma_r$ and $\epsilon_{1,i^*_{\mathcal{S}}} \leq 4\gamma_r$. Consider now two cases:

1. $b_1 \succeq b_{i^*_{\mathcal{S}}} \succeq b_j$. Then, by STI, $\epsilon_{1,j} \leq 8\gamma_r$, and
2. $b_1 \succeq b_j \succeq b_{i^*_{\mathcal{S}}}$. Then, by SST $\epsilon_{1,j} \leq \epsilon_{i^*_{\mathcal{S}},j} \leq 4\gamma_r$.

In either case, we have $\epsilon_j = \epsilon_{1,j} \leq 8\gamma_r$, which implies $c_r \leq \frac{\log(1/\delta)}{2\gamma_r^2} \leq \frac{32\log(1/\delta)}{\epsilon_j^2}$.

Now, let T_j be a random variable denoting the number of comparisons of b_j with other bandits before switching. By definition of round r, bandit b_j will participate in at most one round after r (in phase I). So, we have

$$T_j \leq \begin{cases} |\mathcal{S}| \cdot \sum_{\tau=1}^{r+1} c_\tau & \text{if } b_j \notin \mathcal{S} \\ K \cdot \sum_{\tau=1}^{r+1} c_\tau & \text{if } b_j \in \mathcal{S} \end{cases}$$

Taking expectation over $\mathcal{S}$, we get

$$\begin{aligned}\mathbb{E}[T_j] &\leq \mathbb{E}\left[K \sum_{\tau=1}^{r+1} c_\tau \mid b_j \in \mathcal{S}\right] \cdot \mathbf{P}(b_j \in \mathcal{S}) \\ &\quad + \mathbb{E}\left[|\mathcal{S}| \sum_{\tau=1}^{r+1} c_\tau \mid b_j \notin \mathcal{S}\right] \cdot \mathbf{P}(b_j \notin \mathcal{S}) \\ &\leq \left(K \sum_{\tau=1}^{r+1} c_\tau\right) \cdot \frac{1}{\sqrt{K}} + \mathbb{E}[|\mathcal{S}| \mid b_j \notin \mathcal{S}] \cdot \sum_{\tau=1}^{r+1} c_\tau \\ &\leq 2\sqrt{K} \sum_{\tau=1}^{r+1} c_\tau,\end{aligned}$$

where the third inequality uses $\mathbb{E}[|\mathcal{S}| \mid b_j \notin \mathcal{S}] \leq \sqrt{K}$. Moreover,

$$\sum_{\tau=1}^{r+1} c_\tau \leq 2T^{1/B} \cdot c_r = O\left(\frac{T^{1/B}\log(1/\delta)}{\epsilon_j^2}\right).$$

$$\begin{aligned}\text{Thus, } \mathbb{E}[R_1] &= \sum_j \mathbb{E}[T_j] \cdot \epsilon_j \\ &= \sum_{j:\epsilon_j>0} O\left(\frac{T^{1/B}\sqrt{K}\log(6K^2TB)}{\epsilon_j}\right) \qquad (3)\end{aligned}$$

Bounding R_2. We now bound the regret after switching. From Lemmas 4.3 and 4.4, we know that b_1 is never deleted, $b_1 \in \mathcal{A}^*$, and $|\mathcal{A}^*| \leq rank(b_{i^*_{\mathcal{S}}})$. For any $\mathcal{A}^*$, applying Theorem 3.1 we get,

$$\begin{aligned}R_2 &\leq 3|\mathcal{A}^*|T^{1/B}\log(6T|\mathcal{A}^*|^2B) \sum_{j\in\mathcal{A}^*:\epsilon_j>0} \frac{1}{\epsilon_j} \qquad (4) \\ &\leq 3|\mathcal{A}^*|T^{1/B}\log(6TK^2B) \sum_{j\in\mathcal{B}:\epsilon_j>0} \frac{1}{\epsilon_j}\end{aligned}$$

[2] Note that SST and STI imposes a linear ordering on the bandits. So, we can assume $b_1 \succeq b_2 \succeq \cdots \succeq b_K$. Thus, $rank(b_j) \leq j$ and is at most the number of bandits strictly preferred over b_j.

By Lemma 4.5, $\mathbb{E}[|\mathcal{A}^*|] \leq \sqrt{K}$, hence

$$\mathbb{E}[R_2] \leq 3\sqrt{K}T^{1/B}\log(6TK^2B)\sum_{j:\epsilon_j>0}\frac{1}{\epsilon_j} \quad (5)$$

Combining (3) and (5), we get

$$\mathbb{E}[R(T)|G] = \sum_{j:\epsilon_j>0} O\left(\frac{T^{1/B}\sqrt{K}\log(6K^2TB)}{\epsilon_{1,j}^2}\right),$$

and by (2), this concludes the proof. □

Proof of Theorem 3.3. To prove Theorem 3.3, we use SCOMP with sampling probability (for sampling a bandit into the seed set) equal to $K^{-2/3}$. Note that Lemmas 4.3 and 4.4 continue to hold. The following lemma can be proved exactly as Lemma 4.5, where the sampling probability $p = K^{-2/3}$.

Lemma 4.6. *We have* $\mathbb{E}[rank(b_{i^*_S})] \leq K^{2/3}$ *and* $\mathbb{E}[rank(b_{i^*_S})^2] \leq K^{4/3}$.

We can bound R_1, as in the proof of Theorem 3.2, to obtain:

$$\mathbb{E}[R_1] \leq O\left(\frac{K^{4/3}T^{1/B}\log(6K^2TB)}{\epsilon_{\min}}\right) \quad (6)$$

From (4), we have

$$R_2 \leq 3|\mathcal{A}^*|T^{1/B}\log(6TK^2B)\cdot\frac{|\mathcal{A}^*|}{\epsilon_{\min}}$$

which gives

$$\mathbb{E}[R_2] \leq \frac{3\,\mathbb{E}[|\mathcal{A}^*|^2]\,T^{1/B}\log(6TK^2B)}{\epsilon_{\min}}.$$

Using Lemma 4.6, we get

$$\mathbb{E}[R_2] \leq O\left(\frac{K^{4/3}T^{1/B}\log(6TK^2B)}{\epsilon_{\min}}\right). \quad (7)$$

Combining (6) and (7) completes the proof. □

The proof of Theorem 3.4 follows along the same lines but requires additional ideas, and is deferred to Appendix B.

5. Experimental Results

We provide a summary of computational results of our algorithms for the batched dueling bandits problem. We conducted our computations using C++ and Python 2.7 with a 2.3 Ghz Intel Core $i5$ processor and 16 GB 2133 MHz LPDDR3 memory.

Experimental Setup. We compare all our algorithms, namely PCOMP, SCOMP, and SCOMP2 to a representative set of sequential algorithms for dueling bandits. Specifically, we use the dueling bandit library due to (Komiyama et al., 2015), and compare our algorithms to RUCB (Zoghi et al., 2014), RMED1 (Komiyama et al., 2015), and BEAT-THE-MEAN (Yue & Joachims, 2011). Henceforth, we refer to BEAT-THE-MEAN as BTM. We plot the cumulative regret $R(t)$ incurred by the algorithms against time t. Furthermore, to illustrate the dependence on B, we run another set of experiments on SCOMP2 and plot the cumulative regret $R(t)$ incurred by SCOMP2 against time t for varying values of B.[3] We perform these experiments using both real-world and synthetic data. We use the following datasets:

Six rankers. This real-world dataset is based on the 6 retrieval functions used in the engine of ArXiv.org.

Sushi. The Sushi dataset is based on the Sushi preference dataset (Kamishima, 2003) that contains the preference data regarding 100 types of Sushi. A preference dataset using the top-16 most popular types of sushi is obtained.

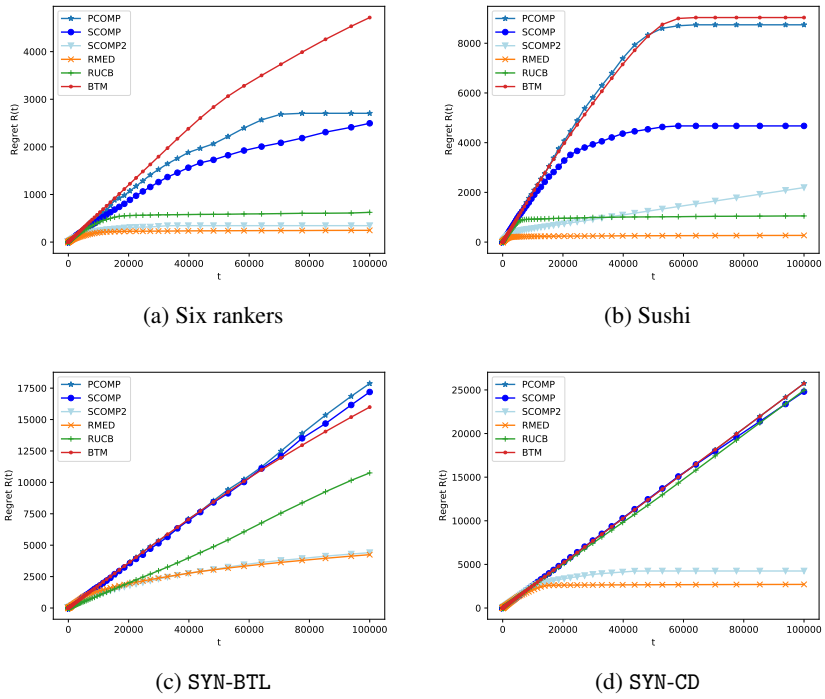
BTL-Uniform. We generate synthetic data using the Bradley-Terry-Luce (BTL) model. Under this model, each arm $b_i \in \mathcal{B}$ is associated with a weight $w_i > 0$ (sampled uniformly in the interval $(0, 1]$), and we set $P_{i,j} = w_i/(w_i + w_j)$. We set the number of arms $K = 100$. Note that the data generated in this way satisfies SST and STI (Yue et al., 2012). We refer to this data as SYN-BTL.

Hard-Instance. The last dataset is a synthetic dataset inspired by the hard instances that we construct for proving our lower bound (see Theorem 1.4). Again, we set $K = 100$, and pick $\ell \in [K]$ uniformly at random as the Condorcet winner. We select Δ uniformly in $(0, 0.5)$, and set $P_{\ell,i} = \frac{1}{2} + \Delta$ for $i \neq \ell$. Furthermore, for all $i, j \neq \ell$, we set $P_{i,j} = 1/2$. We refer to this data as SYN-CD.

Note that there exists a Condorcet winner in all datasets. Moreover, the SYN-BTL dataset satisfies SST and STI. We repeat each experiment 10 times and report the average regret. In our algorithms, we use the KL-divergence based confidence bound (as in RMED1) for elimination as it performs much better empirically (and our theoretical bounds continue to hold). In particular, we replace lines 5, 6 and 8 in PCOMP, SCOMP and SCOMP2, respectively, with KL-divergence based elimination criterion that eliminates an arm i if there exists another arm j if $\hat{P}_{ij} < \frac{1}{2}$ and $N_{ij} \cdot D_{\mathrm{KL}}(\hat{P}_{ij}, \frac{1}{2}) > \log(T\delta)$ where N_{ij} is the number of times arm i and j are played together. We report the average cumulative regret at each time step.

Comparison with sequential dueling bandit algorithms.

[3] We also conducted these experiment for PCOMP and SCOMP and the conclusions were similar.



(a) Six rankers (b) Sushi

(c) SYN-BTL (d) SYN-CD

Figure 1: Regret v/s t plots of algorithms

As mentioned earlier, we compare our algorithms against a representative set of sequential dueling bandits algorithms (RUCB (Zoghi et al., 2014), RMED1 (Komiyama et al., 2015), and BTM (Yue & Joachims, 2011)). Note that the purpose of these experiments is to perform a sanity check to ensure that our batched algorithms, using a small number of batches, perform well when compared with sequential algorithms. We set $\alpha = 0.51$ for RUCB, and $f(K) = 0.3K^{1.01}$ for RMED1, and $\gamma = 1.3$ for BTM. We chose these parameters as they are known to perform well both theoretically and empirically (Komiyama et al., 2015). We set $T = 10^5$, $\delta = 1/TK^2$ and $B = \lfloor \log(T) \rfloor = 16$. We plot the results in Figure 1. We observe that SCOMP2 performs comparably to RMED1 in all datasets, even outperforms RUCB in 3 out of the 4 datasets, and always beats BTM. Notice that both PCOMP and SCOMP considerably outperform BTM on the six rankers and sushi data; however their performance degrades on the synthetic data demonstrating the dependence on K.

Trade-off with number of batches B. We study the trade-off of cumulative regret against the number of batches using SCOMP2. We set $T = 10^5$, and vary $B \in \{2, 8, 16\}$. We also plot the regret incurred by RMED1 as it performs the best amongst all sequential algorithms (and thus serves as a good benchmark). We plot the results in Figure 2 in Appendix A. We observe that as we increase the number of batches, the (expected) cumulative regret decreases. Furthermore, we observe that on the synthetic datasets (where $K = 100$), the regret of SCOMP2 approaches that of RMED1; in fact, the regret incurred is almost identical for SYN-BTL dataset.

6. Conclusion

We introduced and studied the *batched dueling bandit* problem, where the learning algorithm is only allowed to adapt a limited number of times. Our main contribution was an algorithm for this problem under the SST and STI setting. This algorithm's regret (in a logarithmic number of batches) nearly matches the regret of the best known sequential algorithms. We also provided a lower bound demonstrating the dependence of the regret on the number of batches.

An avenue for future work is to obtain batched algorithms (with logarithmic number of batches) that *exactly* match the regret bounds of the best sequential algorithms for dueling bandits under SST and STI. Another direction concerns the batched dueling bandits problem under the more general Condorcet setting. Although we obtained a batched algorithm (PCOMP) for this setting, its regret (in a logarithmic number of batches) is still not asymptotically tight compared to known sequential algorithms.

References

Agarwal, A., Assadi, S., and Khanna, S. Stochastic submodular cover with limited adaptivity. In *Proceedings*

of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 323–342, 2019.

Agarwal, A., Johnson, N., and Agarwal, S. Choice bandits. In *NeurIPS*, 2020.

Ailon, N., Karnin, Z., and Joachims, T. Reducing Dueling Bandits to Cardinal Bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 05 2002. doi: 10.1023/A:1013689704352.

Balkanski, E. and Singer, Y. The adaptive complexity of maximizing a submodular function. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1138–1151, 2018a.

Balkanski, E. and Singer, Y. Approximation guarantees for adaptive sampling. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 393–402, 2018b.

Balkanski, E., Breuer, A., and Singer, Y. Non-monotone submodular maximization in exponentially fewer iterations. In *Advances in Neural Information Processing Systems*, pp. 2359–2370, 2018.

Balkanski, E., Rubinstein, A., and Singer, Y. An exponential speedup in parallel running time for submodular maximization without loss in approximation. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 283–302, 2019.

Bansal, N. and Nagarajan, V. On the adaptivity gap of stochastic orienteering. *Math. Program.*, 154(1-2):145–172, 2015.

Bansal, N., Gupta, A., Li, J., Mestre, J., Nagarajan, V., and Rudra, A. When LP is the cure for your matching woes: Improved bounds for stochastic matchings. *Algorithmica*, 63(4):733–762, 2012.

Behnezhad, S., Derakhshan, M., and Hajiaghayi, M. Stochastic matching with few queries: (1-ϵ) approximation. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1111–1124, 2020.

Bhalgat, A., Goel, A., and Khanna, S. Improved approximation results for stochastic knapsack problems. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1647–1665, 2011.

Chekuri, C. and Quanrud, K. Parallelizing greedy for submodular set function maximization in matroids and beyond. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 78–89, 2019.

Chen, B. and Frazier, P. I. Dueling Bandits with Weak Regret. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Dean, B. C., Goemans, M. X., and Vondrák, J. Approximating the stochastic knapsack problem: The benefit of adaptivity. *Math. Oper. Res.*, 33(4):945–964, 2008.

Dudik, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual Dueling Bandits. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

Esfandiari, H., Karbasi, A., Mehrabian, A., and Mirrokni, V. Regret bounds for batched bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8): 7340–7348, May 2021a.

Esfandiari, H., Karbasi, A., and Mirrokni, V. Adaptivity in adaptive submodularity. In *Proceedings of 34th Conference on Learning Theory*, volume 134, pp. 1823–1846. PMLR, 2021b.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. Batched multi-armed bandits problem. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 501–511, 2019.

Ghuge, R., Gupta, A., and Nagarajan, V. The power of adaptivity for stochastic submodular cover. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3702–3712. PMLR, 18–24 Jul 2021.

Golovin, D. and Krause, A. Adaptive submodularity: A new approach to active learning and stochastic optimization. *CoRR*, abs/1003.3967, 2017.

Guha, S. and Munagala, K. Multi-armed bandits with metric switching costs. In *Automata, Languages and Programming, 36th Internatilonal Colloquium (ICALP)*, pp. 496–507, 2009.

Gupta, A. and Nagarajan, V. A stochastic probing problem with applications. In *Integer Programming and Combinatorial Optimization - 16th International Conference*, pp. 205–216, 2013.

Gupta, A., Krishnaswamy, R., Nagarajan, V., and Ravi, R. Running errands in time: Approximation algorithms for stochastic orienteering. *Math. Oper. Res.*, 40(1):56–79, 2015.

Hofmann, K., Whiteson, S., and Rijke, M. Balancing exploration and exploitation in listwise and pairwise

online learning to rank for information retrieval. *Inf. Retr.*, 16(1):63–90, feb 2013. ISSN 1386-4564. doi: 10.1007/s10791-012-9197-9.

Jamieson, K., Katariya, S., Deshpande, A., and Nowak, R. Sparse Dueling Bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 133–142, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X. doi: 10.1145/775047.775067.

Kamishima, T. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pp. 583–588, 2003.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

Komiyama, J., Honda, J., and Nakagawa, H. Copeland Dueling Bandit Problem: Regret Lower Bound, Optimal Algorithm, and Computationally Efficient Algorithm. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Liu, T.-Y. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, mar 2009. ISSN 1554-0669. doi: 10.1561/1500000016.

Perchet, V., Rigollet, P., Chassang, S., and Snowberg, E. Batched bamdit problems. *The Annals of Statistics*, 44 (2):660–681, 2016. ISSN 00905364.

Radlinski, F., Kurup, M., and Joachims, T. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pp. 43–52, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939913. doi: 10.1145/1458082.1458092.

Ramamohan, S., Rajkumar, A., and Agarwal, S. Dueling Bandits : Beyond Condorcet Winners to General Tournament Solutions. In *Advances in Neural Information Processing Systems 29*, 2016.

Saha, A. and Gopalan, A. Combinatorial bandits with relative feedback. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 983–993, 2019.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multi-dueling Bandits with Dependent Arms. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

Sui, Y., Zoghi, M., Hofmann, K., and Yue, Y. Advancements in dueling bandits. In Lang, J. (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 5502–5510. ijcai.org, 2018.

Urvoy, T., Clerot, F., Feraud, R., and Naamane, S. Generic Exploration and K-armed Voting Bandits. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. A survey of preference-based reinforcement learning methods. *J. Mach. Learn. Res.*, 18(1):4945–4990, jan 2017. ISSN 1532-4435.

Wu, H. and Liu, X. Double thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 649–657, 2016.

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

Yue, Y. and Joachims, T. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

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

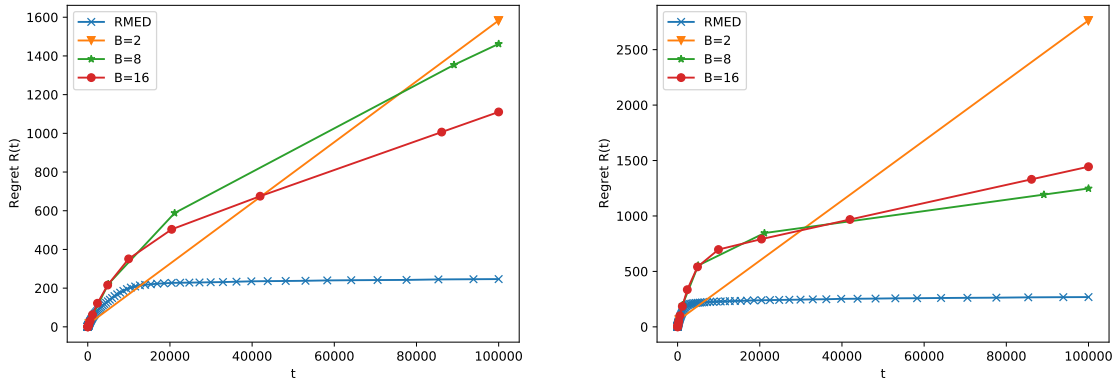
Zoghi, M., Whiteson, S., Munos, R., and de Rijke, M. Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Zoghi, M., Karnin, Z., Whiteson, S., and de Rijke, M. Copeland Dueling Bandits. In *Advances in Neural Information Processing Systems 28*, 2015a.

Zoghi, M., Whiteson, S., and de Rijke, M. MergeRUCB: A method for large-scale online ranker evaluation. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 2015b.

A. Additional Plots

In this section, we provide the missing plots from §5.



(a) Six rankers

(b) Sushi

(c) SYN-BTL

(d) SYN-CD

Figure 2: Regret v/s B for SCOMP2.

B. Regret Analysis

We present the regret analysis for the algorithms described in §3 in this section. We first prove the following lemma which will be used in the analysis of all three algorithms.

Lemma B.1. *For any batch $r \in [B]$, and for any pair b_i, b_j that are compared c_r times, we have*

$$\mathbf{P}\left(|P_{i,j} - \widehat{P}_{i,j}| > \gamma_r\right) \leq 2\delta,$$

where $\gamma_r = \sqrt{\log(\frac{1}{\delta})/2c_r}$.

Proof. Note that $\mathbb{E}[\widehat{P}_{i,j}] = P_{i,j}$, and applying Hoeffding's inequality gives

$$\mathbf{P}\left(|\widehat{P}_{i,j} - P_{i,j}| > \gamma_r\right) \leq 2\exp\left(-2c_r \cdot \gamma_r^2\right) = 2\delta.$$

□

We analyze the regret of our algorithms under a *good* event, G. We show that the G occurs with high probability; in the event that G does not occur (denoted $\overline{G}$), we incur a regret of T. Towards defining G, we say that an estimate $\widehat{P}_{i,j}$ at the end of batch r is *correct* if $|\widehat{P}_{i,j} - P_{i,j}| \leq \gamma_r$. We say that G occurs if every estimate in every batch is correct.

Lemma B.2. *The probability that every estimate in every batch of* PCOMP*,* SCOMP*, and* SCOMP2 *is correct is at least* $1 - 1/T$.

Proof. Applying Lemma B.1 and taking a union bound over all pairs and batches (note SCOMP2 has at most $2B + 1 \leq 3B$ batches), we get that the probability that some estimate is incorrect is at most $K^2 \times 3B \times 2\delta = \frac{1}{T}$ where $\delta = 1/6K^2BT$. Thus, $\mathbf{P}(\overline{G}) \leq \frac{1}{T}$. □

Using Lemma B.2, the expected regret (of *any* algorithm) can be written as follows:

$$\begin{aligned}\mathbb{E}[R(T)] &= \mathbb{E}[R(T) \mid G] \cdot \mathbf{P}(G) + \mathbb{E}[R(T) \mid \overline{G}] \cdot \mathbf{P}(\overline{G}) \\ &\leq \mathbb{E}[R(T) \mid G] + T \cdot \frac{1}{T} = \mathbb{E}[R(T) \mid G] + 1 \end{aligned} \tag{8}$$

Proof of Theorem 3.1. First, recall that in each batch of PCOMP every pair of active arms is compared c_r times where $c_r = \lfloor q^r \rfloor$ with $q = T^{1/B}$. Since, $q^B = T$, PCOMP uses at most B batches.

Following Lemma B.2 and (8), we only need to bound $\mathbb{E}[R(T) \mid G]$. Given G, whenever $P_{i,j} > \frac{1}{2} + 2\gamma_r$ (that is $\epsilon_{i,j} > 2\gamma_r$), we have $\widehat{P}_{i,j} > \frac{1}{2} + \gamma_r$: so bandit b_j will be eliminated by b_i. Furthermore, given bandits b_i and b_j such that $b_i \succeq b_j$, b_i will never be eliminated by b_j under event G. This implies that b_1 is never eliminated: this is crucial as we use b_1 as an anchor to eliminate sub-optimal bandits. Recall that the regret can be written as follows:

$$R(T) = \frac{1}{2} \sum_{j=1}^{K} T_j \epsilon_{1,j}$$

where T_j is the number of comparisons that b_j partakes in. We proceed by bounding T_j. Towards this end, let $T_{1,j}$ be a random variable denoting the number of comparisons performed between b_1 and b_j. As b_1 is never eliminated, $T_j \leq K \cdot T_{1,j}$. Let r denote the last round such that b_j survives round r, i.e., $b_j \in \mathcal{A}$ at the end of round r. We can then conclude that $\epsilon_j := \epsilon_{1,j} \leq 2\gamma_r$ (else b_1 would eliminate b_j in round r). We get

$$\epsilon_j \leq 2 \cdot \sqrt{\frac{\log(\frac{1}{\delta})}{2c_r}}$$

which on squaring and re-arranging gives:

$$c_r \leq \frac{2 \log\left(\frac{1}{\delta}\right)}{\epsilon_j^2} \tag{9}$$

Now, note that b_j could have been played for at most one more round. Thus, we have

$$T_{1,j} = \sum_{\tau=1}^{r+1} c_\tau \leq q \sum_{\tau=0}^{r} c_\tau \leq 2q \cdot c_r$$

where the final inequality follows from summing up $\sum_{\tau=1}^{r-1} c_\tau$, and using $B \leq \log(T)$. Then, we have $T_j \leq 2Kq \cdot c_r$. Using 9, and plugging in $q = T^{1/B}$ and $\delta = 1/6TK^2B$ we have

$$\begin{aligned} E[R(T) \mid G] &\leq \frac{1}{2} \sum_j \left(2KT^{1/B} \cdot \frac{2 \log\left(6TK^2B\right)}{\epsilon_j^2} \right) \cdot \epsilon_j \\ &= \sum_{j:\epsilon_j > 0} \frac{KT^{1/B} \log\left(6TK^2B\right)}{\epsilon_j} \\ &= 2KT^{1/B} \log\left(6TK^2B\right) \sum_{j:\epsilon_j > 0} \frac{1}{\epsilon_j}. \end{aligned}$$

Note that when $\epsilon_j = 0$ for $b_j \in \mathcal{B}$, we exclude the corresponding term in the regret bound. Combining this with (8) gives the first bound of Theorem 3.1. Plugging in $\epsilon_{\min} = \min_{j:\epsilon_j>0} \epsilon_j$ completes the proof. □

B.1. Proofs of Theorems 3.2 and 3.4

In this section, we provide the proofs of Theorem 3.2 and Theorem 3.4. Henceforth, we assume the SST and STI properties. We need the following definition. For a bandit b_j, let $E_j = \{b_i \in \mathcal{B} : \epsilon_{i,j} > 0\}$; that is, the set of bandits superior to bandit b_j. We define $rank(b_j) = |E_j|$. [4]

As before, we analyze the regret of SCOMP and SCOMP2 under event G. By Lemma B.2 and (8), we only need to bound the expected regret under G; that is, we need to bound $\mathbb{E}[R(T) \mid G]$. Conditioned on event G, the following Lemmas B.3,B.4 and B.5 hold for both SCOMP and SCOMP2.

Lemma B.3. *The best bandit b_1 is never deleted from $\mathcal{A}$ in the elimination step of phase I.*

Proof. In SCOMP, b_i deletes b_j in batch r if $\widehat{P}_{i,j} > \frac{1}{2} + 3\gamma_r$, and in SCOMP2 if $\widehat{P}_{i,j} > \frac{1}{2} + 5\gamma_r$. If b_1 is deleted due to some bandit b_j, then by applying Lemma B.1 (in either case), we get $P_{j,1} > \frac{1}{2} + 2\gamma_r$, a contradiction. □

Lemma B.4. *When the algorithm switches to* PCOMP *on set $\mathcal{A}^*$, we have $b_1 \in \mathcal{A}^*$ and $|\mathcal{A}^*| \leq rank(b_{i^*_{\mathcal{S}}})$ where $b_{i^*_{\mathcal{S}}}$ is the best bandit in $\mathcal{S}$.*

Proof. We first consider algorithm SCOMP. Here, the switching occurs when, in some batch r, there exists $b_{j^*} \in \mathcal{A}$ such that $\widehat{P}_{j^*,i} > \frac{1}{2} + 3\gamma_r$ for all $b_i \in \mathcal{S}$, Moreover, $\mathcal{A}^* = \{b_j \in \mathcal{A} \mid \widehat{P}_{j,i} > \frac{1}{2} + \gamma_r \text{ for all } b_i \in \mathcal{S}\}$. Consider any $b_i \in \mathcal{S}$. Given G, $\widehat{P}_{j^*,i} > \frac{1}{2} + 3\gamma_r$ implies that $P_{j^*,i} > \frac{1}{2} + 2\gamma_r$. By SST, $P_{1,i} \geq P_{j^*,i}$, and again using event G, $\widehat{P}_{1,i} > \frac{1}{2} + \gamma_r$. Thus, $b_1 \in \mathcal{A}^*$. We now bound $|\mathcal{A}^*|$. Let $b_{i^*_{\mathcal{S}}}$ be the best bandit in $\mathcal{S}$, i.e., the bandit of smallest rank. Consider any bandit $b_j \in \mathcal{A}^*$. We have $\widehat{P}_{j,i^*_{\mathcal{S}}} > \frac{1}{2} + \gamma_r$, which implies (by event G) that $P_{j,i^*_{\mathcal{S}}} > \frac{1}{2}$. So, we must have $b_j \succ b_{i^*_{\mathcal{S}}}$. Consequently, $\mathcal{A}^* \subseteq \{b_j \in \mathcal{B} : b_j \succ b_{i^*_{\mathcal{S}}}\}$, which implies $|\mathcal{A}^*| \leq rank(b_{i^*_{\mathcal{S}}})$.

We now consider SCOMP2. Here, we select an *undefeated* candidate bandit $b_{i^*_r}$ in batch r, and the algorithm switches if there exists $b_{j^*} \in \mathcal{A}$ such that $\widehat{P}_{j^*,i^*_r} > \frac{1}{2} + 5\gamma_r$. Moreover, $\mathcal{A}^* = \{b_j \in \mathcal{A} \mid \widehat{P}_{j,i^*_r} > \frac{1}{2} + 3\gamma_r\}$. Given G, we have $P_{j^*,i^*_r} > \frac{1}{2} + 4\gamma_r$. By SST and again applying G, we obtain $\widehat{P}_{1,i^*_r} > \frac{1}{2} + 3\gamma_r$. So, $b_1 \in \mathcal{A}^*$. We now argue that $|\mathcal{A}^*| \leq rank(b_{i^*_{\mathcal{S}}})$. Again, let $b_{i^*_{\mathcal{S}}}$ be the best bandit in $\mathcal{S}$. As $b_{i^*_r}$ is undefeated after round $r^{(1)}$, we have $\widehat{P}_{i^*_{\mathcal{S}},i^*_r} \leq \frac{1}{2} + \gamma_r$, which implies $P_{i^*_{\mathcal{S}},i^*_r} \leq \frac{1}{2} + 2\gamma_r$ (by event G). Now, consider any bandit $b_j \in \mathcal{A}^*$. We have $\widehat{P}_{j,i^*_{\mathcal{S}}} > \frac{1}{2} + 3\gamma_r$, which implies (by event G) that $P_{j,i^*_{\mathcal{S}}} > \frac{1}{2} + 2\gamma_r$. It follows that $b_j \succ b_{i^*_{\mathcal{S}}}$ for all $b_j \in \mathcal{A}^*$. Hence, $|\mathcal{A}^*| \leq rank(b_{i^*_{\mathcal{S}}})$. □

Lemma B.5. *We have $\mathbb{E}[rank(b_{i^*_{\mathcal{S}}})] \leq \sqrt{K}$ and $\mathbb{E}[rank(b_{i^*_{\mathcal{S}}})^2] \leq 2K$.*

Proof. Let R be a random variable denoting $rank(b_{i^*_{\mathcal{S}}})$. Note that $R = k$ if, and only if, the first $k-1$ bandits are not sampled into $\mathcal{S}$, and the k^{th} bandit is sampled into $\mathcal{S}$. Thus, R is a geometric random variable with success probability $p := \frac{1}{\sqrt{K}}$.[5] Recall that the mean and variance of a geometric random variable are $\frac{1}{p}$ and $\frac{1}{p^2} - \frac{1}{p}$ respectively. So, $\mathbb{E}[R] \leq \frac{1}{p} = \sqrt{K}$. Moreover, $\mathbb{E}[R^2] \leq \frac{2}{p^2} = 2K$. □

Using Lemmas B.3, B.4 and B.5, we complete the proofs of Theorems 3.2 and 3.4.

Proof of Theorem 3.2. We bound the expected regret of SCOMP conditioned on G. Let R_1 and R_2 denote the regret incurred in phase I and II respectively.

[4] Note that SST and STI imposes a linear ordering on the bandits. So, we can assume $b_1 \succeq b_2 \succeq \cdots \succeq b_K$. Thus, $rank(b_j) < j$; that is, it is at most the number of bandits strictly preferred over b_j.

[5] Strictly speaking, R is truncated at K.

Bounding R_1. Fix a bandit b_j. Let r denote the last round such that $b_j \in \mathcal{A}$ *and* switching does not occur (at the end of round r). Let $b_{i^*_{\mathcal{S}}}$ be the best bandit in $\mathcal{S}$. As b_j is not eliminated by $b_{i^*_{\mathcal{S}}}$, we have $\widehat{P}_{i^*_{\mathcal{S}},j} \le \frac{1}{2} + 3\gamma_r$, which implies (by event G) $P_{i^*_{\mathcal{S}},j} \le \frac{1}{2} + 4\gamma_r$. Moreover, as switching doesn't occur, we have $\min_{i\in\mathcal{S}} \widehat{P}_{1,i} \le \frac{1}{2} + 3\gamma_r$ (by Lemma B.3, b_1 is never deleted from $\mathcal{A}$). We now claim that $P_{1,i^*_{\mathcal{S}}} \le \frac{1}{2} + 4\gamma_r$. Otherwise, by SST we have $\min_{i\in\mathcal{S}} P_{1,i} = P_{1,i^*_{\mathcal{S}}} > \frac{1}{2} + 4\gamma_r$, which (by event G) implies $\min_{i\in\mathcal{S}} \widehat{P}_{1,i} > \frac{1}{2} + 3\gamma_r$, a contradiction! It now follows that $\epsilon_{i^*_{\mathcal{S}},j} \le 4\gamma_r$ and $\epsilon_{1,i^*_{\mathcal{S}}} \le 4\gamma_r$. Consider now two cases:

1. $b_1 \succeq b_{i^*_{\mathcal{S}}} \succeq b_j$. Then, by STI, $\epsilon_{1,j} \le 8\gamma_r$, and
2. $b_1 \succeq b_j \succeq b_{i^*_{\mathcal{S}}}$. Then, by SST $\epsilon_{1,j} \le \epsilon_{i^*_{\mathcal{S}},j} \le 4\gamma_r$.

In either case, we have $\epsilon_j = \epsilon_{1,j} \le 8\gamma_r$, which implies $c_r \le \frac{\log(1/\delta)}{2\gamma_r^2} \le \frac{32\log(1/\delta)}{\epsilon_j^2}$.

Now, let T_j be a random variable denoting the number of comparisons of b_j with other bandits before switching. By definition of round r, bandit b_j will participate in at most one round after r (in phase I). So, we have

$$T_j \le \begin{cases} |\mathcal{S}| \cdot \sum_{\tau=1}^{r+1} c_\tau & \text{if } b_j \notin \mathcal{S} \\ K \cdot \sum_{\tau=1}^{r+1} c_\tau & \text{if } b_j \in \mathcal{S} \end{cases}$$

Taking expectation over $\mathcal{S}$, we get

$$\begin{aligned} \mathbb{E}[T_j] &\le \mathbb{E}\left[K \sum_{\tau=1}^{r+1} c_\tau \,|\, b_j \in \mathcal{S}\right] \cdot \mathbf{P}(b_j \in \mathcal{S}) + \mathbb{E}\left[|\mathcal{S}| \sum_{\tau=1}^{r+1} c_\tau \,|\, b_j \notin \mathcal{S}\right] \cdot \mathbf{P}(b_j \notin \mathcal{S}) \\ &\le \left(K \sum_{\tau=1}^{r+1} c_\tau\right) \cdot \frac{1}{\sqrt{K}} + \mathbb{E}[|\mathcal{S}| \,|\, b_j \notin \mathcal{S}] \cdot \sum_{\tau=1}^{r+1} c_\tau \;\le\; 2\sqrt{K} \sum_{\tau=1}^{r+1} c_\tau, \end{aligned}$$

where the third inequality uses $\mathbb{E}[|\mathcal{S}| \,|\, b_j \notin \mathcal{S}] \le \sqrt{K}$. Moreover,

$$\sum_{\tau=1}^{r+1} c_\tau \le 2T^{1/B} \cdot c_r = O\left(\frac{T^{1/B}\log(1/\delta)}{\epsilon_j^2}\right).$$

Thus,

$$\mathbb{E}[R_1] = \sum_j \mathbb{E}[T_j] \cdot \epsilon_j = \sum_{j:\epsilon_j>0} O\left(\frac{T^{1/B}\sqrt{K}\log(6K^2TB)}{\epsilon_j}\right) \tag{10}$$

Bounding R_2. We now bound the regret after switching. From Lemmas B.3 and B.4, we know that b_1 is never deleted, $b_1 \in \mathcal{A}^*$, and $|\mathcal{A}^*| \le rank(b_{i^*_{\mathcal{S}}})$. For any $\mathcal{A}^*$, applying Theorem 3.1 we get,

$$R_2 \le 3|\mathcal{A}^*|T^{1/B}\log(6T|\mathcal{A}^*|^2B) \sum_{j\in\mathcal{A}^*:\epsilon_j>0} \frac{1}{\epsilon_j} \le 3|\mathcal{A}^*|T^{1/B}\log(6TK^2B) \sum_{j\in\mathcal{B}:\epsilon_j>0} \frac{1}{\epsilon_j}$$

By Lemma B.5, $\mathbb{E}[|\mathcal{A}^*|] \le \sqrt{K}$, hence

$$\mathbb{E}[R_2] \le 3\sqrt{K}T^{1/B}\log(6TK^2B) \sum_{j:\epsilon_j>0} \frac{1}{\epsilon_j} \tag{11}$$

Combining (10) and (11), we get

$$\mathbb{E}[R(T)|G] = \sum_{j:\epsilon_j>0} O\left(\frac{T^{1/B}\sqrt{K}\log(6K^2TB)}{\epsilon_j}\right),$$

and by (8), this concludes the proof. □

Proof of Theorem 3.4. We bound the expected regret conditioned on G. Let R_1 and R_2 denote the regret incurred in phase I and II respectively.

Bounding R_1. Fix a bandit b_j. Let r denote any round such that $b_j \in \mathcal{A}$ *and* switching does not occur (at the end of round r). As in the proof of Theorem 3.2, we first show that $c_r = O\left(\frac{\log(1/\delta)}{\epsilon_j^2}\right)$. Recall that $b_{i_r^*}$ is the candidate in round r. As b_j is not eliminated by $b_{i_r^*}$, we have $\widehat{P}_{i_r^*,j} \le \frac{1}{2} + 5\gamma_r$, which implies (by event G) $P_{i_{\mathcal{S}}^*,j} \le \frac{1}{2} + 6\gamma_r$. Moreover, as switching doesn't occur, we have $\widehat{P}_{1,i_r^*} \le \frac{1}{2} + 5\gamma_r$ (by Lemma B.3, b_1 is never deleted from $\mathcal{A}$). By event G, we get $P_{1,i_r^*} \le \frac{1}{2} + 6\gamma_r$. It now follows that $\epsilon_{i_r^*,j} \le 6\gamma_r$ and $\epsilon_{1,i_r^*} \le 6\gamma_r$. Consider now two cases:

1. $b_1 \succeq b_{i_r^*} \succeq b_j$. Then, by STI, $\epsilon_{1,j} \le 12\gamma_r$, and
2. $b_1 \succeq b_j \succeq b_{i_{\mathcal{S}}^*}$. Then, by SST $\epsilon_{1,j} \le \epsilon_{i_r^*,j} \le 6\gamma_r$.

In either case, we have $\epsilon_j = \epsilon_{1,j} \le 12\gamma_r$, which implies $c_r \le \frac{\log(1/\delta)}{2\gamma_r^2} = O\left(\frac{\log(1/\delta)}{\epsilon_j^2}\right)$.

We further divide R_1 into two kinds of regret: $R_1^{(c)}$ and $R_1^{(n)}$ where $R_1^{(c)}$ refers to the regret incurred by candidate arms and $R_1^{(n)}$ is the regret incurred by non-candidate arms.

Bounding $R_1^{(n)}$. For any bandit b_j, let T_j be a random variable denoting the number of comparisons of b_j (in phase I) when b_j is not a candidate. Also, let r be the last round such that $b_j \in \mathcal{A}$ and switching doesn't occur. So, b_j will participate in at most one round after r, and

$$T_j \le \begin{cases} \sum_{\tau=1}^{r+1} c_\tau & \text{if } b_j \notin \mathcal{S} \\ |\mathcal{S}| \cdot \sum_{\tau=1}^{r+1} c_\tau & \text{if } b_j \in \mathcal{S} \end{cases}$$

Taking expectation over $\mathcal{S}$, we get

$$\begin{aligned} \mathbb{E}[T_j] &\le \mathbb{E}\left[|\mathcal{S}| \sum_{\tau=1}^{r+1} c_\tau \,|\, b_j \in \mathcal{S}\right] \cdot \mathbf{P}(b_j \in \mathcal{S}) + \mathbb{E}\left[\sum_{\tau=1}^{r+1} c_\tau \,|\, b_j \notin \mathcal{S}\right] \cdot \mathbf{P}(b_j \notin \mathcal{S}) \\ &\le \sum_{\tau=1}^{r+1} c_\tau \cdot \left(\frac{1}{\sqrt{K}} \cdot \mathbb{E}[|\mathcal{S}| \,|\, b_j \in \mathcal{S}] + 1\right) \;\le\; (2 + \frac{1}{\sqrt{K}}) \cdot \sum_{\tau=1}^{r+1} c_\tau, \end{aligned}$$

where the third inequality uses $\mathbb{E}[|\mathcal{S}| \,|\, b_j \in \mathcal{S}] \le 1 + \sqrt{K}$.

Moreover, using $c_r = O\left(\frac{\log(1/\delta)}{\epsilon_j^2}\right)$, we have $\sum_{\tau=1}^{r+1} c_\tau = O\left(\frac{T^{1/B}\log(1/\delta)}{\epsilon_j^2}\right)$. Thus,

$$\mathbb{E}[R_1^{(n)}] = \sum_j \mathbb{E}[T_j] \cdot \epsilon_j \le \sum_{j:\epsilon_j > 0} O\left(\frac{T^{1/B}\log\left(\frac{1}{\delta}\right)}{\epsilon_{1,j}}\right) \le O\left(\frac{T^{1/B} K \log\left(\frac{1}{\delta}\right)}{\epsilon_{\min}}\right) \tag{12}$$

Bounding $R_1^{(c)}$. Observe that if b_j is a candidate in round r, then the regret incurred by b_j in round r is at most $Kc_r \cdot \epsilon_{1,j}$. Also, $c_{r-1} \le O\left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon_j^2}\right)$ because $b_j \in \mathcal{A}$ and switching hasn't occurred at end of round $r-1$. Thus, we have $c_r = T^{1/B} c_{r-1} \le O\left(\frac{T^{1/B}\log\left(\frac{1}{\delta}\right)}{\epsilon_j^2}\right)$. We can thus write

$$R_1^{(c)} = \sum_{r=1}^{B} \sum_j Kc_r \cdot \epsilon_j \cdot \mathbb{I}[i_r^* = j],$$

where $\mathbb{I}[i_r^* = j]$ is an indicator random variable denoting whether b_j was the candidate bandit in round r. Observe that there is exactly one candidate bandit, $b_{i_r^*}$, in each round. So,

$$\begin{aligned} R_1^{(c)} &= K \sum_{r=1}^{B} c_r \epsilon_{i_r^*} \le K \sum_{r=1}^{B} O\left(\frac{T^{1/B}\log\left(\frac{1}{\delta}\right)}{\epsilon_{i_r^*}^2}\right) \cdot \epsilon_{i_r^*} \\ &= K \sum_{r=1}^{B} O\left(\frac{T^{1/B}\log\left(\frac{1}{\delta}\right)}{\epsilon_{i_r^*}}\right) \le O\left(\frac{T^{1/B} K B \log\left(\frac{1}{\delta}\right)}{\epsilon_{\min}}\right) \end{aligned} \tag{13}$$

Combining (12) and (13), we get

$$\mathbb{E}[R_1] \leq O\left(\frac{T^{1/B}KB\log\left(\frac{1}{\delta}\right)}{\epsilon_{\min}}\right) \tag{14}$$

Bounding R_2. Finally, we bound the regret in phase II where we only have bandits $\mathcal{A}^*$. From Lemmas B.3 and B.4, we know that $b_1 \in \mathcal{A}^*$, and $|\mathcal{A}^*| \leq rank(b_{i_S^*})$. For any $\mathcal{A}^*$, applying Theorem 3.1 we get,

$$R_2 \leq 3|\mathcal{A}^*|T^{1/B}\log(6T|\mathcal{A}^*|^2B)\sum_{j\in\mathcal{A}^*:\epsilon_j>0}\frac{1}{\epsilon_j} \leq 3|\mathcal{A}^*|^2\cdot T^{1/B}\log(6TK^2B)\cdot\frac{1}{\epsilon_{\min}}$$

By Lemma B.5, $\mathbb{E}[|\mathcal{A}^*|^2] \leq 2K$, and so:

$$\mathbb{E}[R_2] \leq \frac{6T^{1/B}K\log(6TK^2B)}{\epsilon_{\min}} \tag{15}$$

Finally, combining (14) and (15) completes the proof. □

C. Lower Bound

In this section, we present a lower bound for the batched dueling bandits problem under the SST and STI setting. Note that this lower bound also applies to the more general Condorcet winner setting. The main result of this section is the following:

Theorem C.1. *Given an integer $B > 1$, and any algorithm that uses at most B batches, there exists an instance of the K-armed batched dueling bandit problem that satisfies the SST and STI conditions such that the expected regret*

$$\mathbb{E}[R_T] = \Omega\left(\frac{KT^{1/B}}{B^2\epsilon_{\min}}\right),$$

where $\epsilon_{\min}$ is defined with respect to the particular instance.

In order to prove this theorem, we will construct a family of instances such that any algorithm for batched dueling bandits cannot simultaneously beat the above regret lower bound over all instances in the family. We exploit the fact that the algorithm is unaware of the particular instance chosen from the family at run-time, and hence, is unaware of the gap $\epsilon_{\min}$ under that instance.

Family of Instances :

- Let F be an instance where $P_{i,j} = \frac{1}{2}$ for all $i, j \in \mathcal{B}$.
- For $j \in [B]$, let $\Delta_j = \frac{\sqrt{K}}{24B}\cdot T^{(j-1)/2B}$. For $j \in [B]$ and $k \in [K]$, let $E_{j,k}$ be an instance where bandit b_k is the Condorcet winner such that $P_{k,l} = \frac{1}{2} + \Delta_j$ for all $l \in [K] \setminus \{k\}$ and $P_{l,m} = \frac{1}{2}$ for all $l, m \in [K] \setminus \{k\}$.
- The family of instances $:= \{E_{j,k}\}_{j\in[B],k\in[K]} \cup \{F\}$.

C.1. Proof of Theorem C.1

Let us fix an algorithm $\mathcal{A}$ for this problem. Let $T_j = T^{j/B}$ for $j \in [B]$. Let t_j be the total (random) number of comparisons until the end of batch j during the execution of $\mathcal{A}$. We will overload notation and denote by I^t the distribution of observations seen by the algorithm when the underlying instance is I. We will sometimes use $P_{i,j}(I)$ for the probability of i beating j under an instance I to emphasize the dependence on I. We will also write $\epsilon_{\min}(I)$ to emphasize the dependence on the underlying instance I.

We define event A_j as follows:

$$A_j = \{t_{j'} < T_{j'}, \forall j' < j \text{ and } t_j \geq T_j\},$$

and denote by $E_{j,k}(A_j)$ the event that A_j occurs given that the instance selected is $E_{j,k}$. Similarly, $F(A_j)$ denotes the event that A_j occurs when the instance selected is F. Now, define

$$p_j = \frac{1}{K}\sum_{l=1}^{K} \mathbf{P}(E_{j,l}(A_j)).$$

Observe that p_j is the average probability of event A_j conditional on the instance having gap Δ_j.

Lemma C.2. $\sum_{j=1}^{B} p_j \geq \frac{1}{2}$.

Proof. Note that the event A_j is determined by observations until T_{j-1}. This is because $t_{j-1} < T_{j-1}$, and once the observations until t_{j-1} are seen: the next batch j determines whether or not A_j occurs. Hence, in order to bound the probability of A_j under two different instances F and $E_{j,l}$ we use the Pinsker's inequality as

$$|\mathbf{P}(F(A_j)) - \mathbf{P}(E_{j,l}(A_j))| \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(F^{T_{j-1}} || E_{j,l}^{T_{j-1}})}$$

for $l \in [K]$. Let τ_l be the random variable for the number of times arm l is played until T_{j-1}. We first bound $D_{\mathrm{KL}}(F^{T_{j-1}} || E_{j,l}^{T_{j-1}})$ as

$$\begin{aligned} D_{\mathrm{KL}}(F^{T_{j-1}} || E_{j,l}^{T_{j-1}}) &\overset{(a)}{=} \sum_{t=1}^{T_{j-1}} D_{\mathrm{KL}}\big(P_{t_1,t_2}(F) \,||\, P_{t_1,t_2}(E_{j,l})\big) \\ &\overset{(b)}{\leq} \sum_{t=1}^{T_{j-1}} \Pr_F(\text{arm } l \text{ is played in trial } t) \cdot D_{\mathrm{KL}}\left(\frac{1}{2} \,||\, \frac{1}{2} + \Delta_j\right) \\ &\overset{(c)}{\leq} \mathbb{E}_F[\tau_l] \cdot 4\Delta_j^2\,, \end{aligned} \tag{16}$$

where (a) follows from the fact that, given F, the outcome of comparisons are independent across trials, (b) follows from the fact that the KL-divergence between $P_{t_1,t_2}(F)$ and $P_{t_1,t_2}(E_{j,k})$ is non-zero only when arm l is played in trial t, and (c) follows from the fact that $D_{\mathrm{KL}}(p||q) \leq \frac{(p-q)^2}{q\cdot(1-q)}$. Using the above bounds, we have that

$$\begin{aligned} \frac{1}{K}\sum_{l=1}^{K} |\mathbf{P}(F(A_j)) - \mathbf{P}(E_{j,l}(A_j))| &\leq \frac{1}{K}\sum_{l=1}^{K} \sqrt{\frac{1}{2} D_{\mathrm{KL}}(F^{T_{j-1}} || E_{j,l}^{T_{j-1}})} \\ &\leq \frac{1}{K}\sum_{l=1}^{K} \sqrt{\frac{1}{2}\cdot 4\Delta_j^2 \mathbb{E}_F[\tau_l]} = \frac{1}{K}\sum_{l=1}^{K} \sqrt{2\Delta_j^2 \mathbb{E}_F[\tau_l]} \\ &\overset{(a)}{\leq} \sqrt{\frac{2\Delta_j^2 \mathbb{E}_F[\sum_{l=1}^{K} \tau_l]}{K}} \\ &\overset{(b)}{\leq} \sqrt{\frac{2\Delta_j^2 \cdot 2T_{j-1}}{K}} = \frac{1}{2B}\,, \end{aligned}$$

where (a) follows from the concavity of $\sqrt{x}$ and Jensen's inequality, and (b) follows from the fact that $\sum_{l=1}^{K} \tau_l \leq T_{j-1}$. We thus have

$$\begin{aligned} |\mathbf{P}(F(A_j)) - p_j| &= |\mathbf{P}(F(A_j)) - \frac{1}{K}\sum_{l=1}^{K} \mathbf{P}(E_{j,l}(A_j))| \\ &\leq \frac{1}{K}\sum_{l=1}^{K} |\mathbf{P}(F(A_j)) - \mathbf{P}(E_{j,l}(A_j))| \leq \frac{1}{2B}\,. \end{aligned}$$

Finally, we can write

$$\sum_{j=1}^{B} p_j \geq \sum_{j=1}^{B} (\mathbf{P}(F(A_j)) - \frac{1}{2B}) \geq \sum_{j=1}^{B} \mathbf{P}(F(A_j)) - \frac{1}{2} \geq \frac{1}{2}\,.$$

□

As a consequence of this lemma, we can conclude that there exists some $j \in [B]$ such that $p_j \geq \frac{1}{2B}$. We focus on the event where gap is Δ_j, and prove that when $p_j \geq \frac{1}{2B}$, $\mathcal{A}$ must suffer a high regret leading to a contradiction. The next lemma formalizes this.

Lemma C.3. *If, for some j, $p_j \geq \frac{1}{2B}$, then*

$$\sup_{I:\epsilon_{\min}(I)=\Delta_j} \mathbb{E}[R_T(I)] \geq \Omega\left(\frac{KT^{1/B}}{B^2\Delta_j}\right)$$

Proof. Fix $k \in [K]$. We will construct a family of instances $\{Q_{j,k,l}\}_{l\neq k}$ where $Q_{j,k,l}$ is defined as:

Instance $Q_{j,k,l}$: Arm l is the Condorcet winner and the pairwise preferences are defined as:

$$P_{lm} = \frac{1}{2} + 2\Delta_j, \forall m \in [K] \setminus \{l\}; \qquad P_{km} = \frac{1}{2} + \Delta_j, \forall m \in [K] \setminus \{l, k\};$$

and $P_{mm'} = \frac{1}{2}$ for remaining pairs (m, m').

We also let $Q_{j,k,k} := E_{j,k}$. Note that the regret is $\geq \Delta_j$ if the underlying instance is $Q_{j,k,l}$ and the pair played is not (b_l, b_l). We have that

$$\sup_{I:\epsilon_{\min}(I)=\Delta_j} \mathbb{E}[R_T(I)] \geq \Delta_j \sum_{t=1}^{T} \frac{1}{K} \sum_{l\neq k} Q^t_{j,k,l}\left((b_{t_1}, b_{t_2}) \neq (b_l, b_l)\right)\,,$$

where $Q^t_{j,k,l}$ denotes the distribution of observations available at time t under instance $Q_{j,k,l}$ and $Q^t_{j,k,l}\left((b_{t_1}, b_{t_2}) \neq (b_l, b_l)\right)$ is the probability that the algorithm does not play arm (b_l, b_l) at time t under $Q^t_{j,k,l}$. In order to bound the above quantity we will need the following lemma from (Gao et al., 2019).

Lemma C.4 (Lemma 3 of (Gao et al., 2019))**.** *Let $Q_1, \cdots Q_K$ be probability measures on some common probability space $(\Omega, \mathcal{F})$, and $\psi : \Omega \to [K]$ be any measurable function (i.e., test). Then, for any tree $\mathcal{T} = ([K], E)$ with vertex set $[K]$ and edge set E,*

$$\frac{1}{K}\sum_{i=1}^{K} Q_i(\psi \neq i) \geq \frac{1}{K} \sum_{(l,l')\in E} \int \min\{dQ_l, dQ_{l'}\}\,.$$

Using the above lemma for the star graph centered at k, we have that

$$\begin{aligned}
\sup_{I:\epsilon_{\min}(I)=\Delta_j} \mathbb{E}[R_T(I)] &\geq \Delta_j \sum_{t=1}^{T} \frac{1}{K} \sum_{l\neq k} \int \min\{dQ^t_{j,k,k}, dQ^t_{j,k,l}\} \\
&\overset{(a)}{\geq} \Delta_j \sum_{t=1}^{T_j} \frac{1}{K} \sum_{l\neq k} \int \min\{dQ^t_{j,k,k}, dQ^t_{j,k,l}\} \\
&\overset{(b)}{\geq} \Delta_j \sum_{t=1}^{T_j} \frac{1}{K} \sum_{l\neq k} \int \min\{dQ^{T_j}_{j,k,k}, dQ^{T_j}_{j,k,l}\} \\
&\geq \Delta_j \sum_{t=1}^{T_j} \frac{1}{K} \sum_{l\neq k} \int_{A_j} \min\{dQ^{T_j}_{j,k,k}, dQ^{T_j}_{j,k,l}\} \\
&\overset{(c)}{\geq} \Delta_j \sum_{t=1}^{T_j} \frac{1}{K} \sum_{l\neq k} \int_{A_j} \min\{dQ^{T_{j-1}}_{j,k,k}, dQ^{T_{j-1}}_{j,k,l}\}\,,
\end{aligned}$$

where (a) follows because $T_j \leq T$, (b) follows due to the fact that $\int \min\{dP, dQ\} = 1 - D_{\text{TV}}(P,Q)$ and the fact that $D_{\text{TV}}(Q^{T_j}_{j,k,k}, Q^{T_j}_{j,k,l})$ is at least $D_{\text{TV}}(Q^t_{j,k,k}, Q^t_{j,k,l})$ as the sigma algebra $\mathcal{F}_{Q^t_{j,k,k}}$ of $Q^t_{j,k,k}$ is a subset of the sigma algebra $\mathcal{F}_{Q^{T_j}_{j,k,k}}$ of $Q^{T_j}_{j,k,k}$, (c) follow from the fact that the event A_j is determined by observations until T_{j-1} as explained in the proof of Lemma C.2. We then have that

$$\begin{aligned}
\int_{A_j} \min\{dQ^{T_{j-1}}_{j,k,k}, dQ^{T_{j-1}}_{j,k,l}\} &= \int_{A_j} \frac{dQ^{T_{j-1}}_{j,k,k} + dQ^{T_{j-1}}_{j,k,l} - |dQ^{T_{j-1}}_{j,k,k} - dQ^{T_{j-1}}_{j,k,l}|}{2} \\
&= \frac{Q^{T_{j-1}}_{j,k,k}(A_j) + Q^{T_{j-1}}_{j,k,l}(A_j)}{2} - \int_{A_j} \frac{|dQ^{T_{j-1}}_{j,k,k} - dQ^{T_{j-1}}_{j,k,l}|}{2} \\
&\overset{(a)}{\geq} Q^{T_{j-1}}_{j,k,k}(A_j) - \frac{1}{2} D_{\text{TV}}\left(Q^{T_{j-1}}_{j,k,k}, Q^{T_{j-1}}_{j,k,l}\right) - D_{\text{TV}}\left(Q^{T_{j-1}}_{j,k,k}, Q^{T_{j-1}}_{j,k,l}\right) \\
&= Q^{T_{j-1}}_{j,k,k}(A_j) - \frac{3}{2} D_{\text{TV}}\left(Q^{T_{j-1}}_{j,k,k}, Q^{T_{j-1}}_{j,k,l}\right)\,,
\end{aligned}$$

where (a) follows from the fact that $D_{\text{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$. Let us define τ_l to be the random variable for the number of times arm l is played until T_{j-1} We also have that

$$\begin{aligned}
\frac{1}{K} \sum_{l\neq k} D_{\text{TV}}\left(Q^{T_{j-1}}_{j,k,k}, Q^{T_{j-1}}_{j,k,l}\right) &\leq \frac{1}{K} \sum_{l\neq k} \sqrt{\frac{1}{2} D_{\text{KL}}(Q^{T_{j-1}}_{j,k,k} || Q^{T_{j-1}}_{j,k,l})} \\
&\overset{(a)}{\leq} \frac{1}{K} \sum_{l\neq k} \sqrt{\frac{1}{2} \cdot 16\Delta_j^2 \mathbb{E}_{E_{j,k}}[\tau_l]} = \frac{1}{K} \sum_{l\neq k} \sqrt{8\Delta_j^2 \mathbb{E}_{E_{j,k}}[\tau_l]} \\
&\overset{(b)}{\leq} \sqrt{\frac{8\Delta_j^2 \mathbb{E}_{E_{j,k}}[\sum_{l\neq k} \tau_l]}{K}} \\
&\overset{(c)}{\leq} \sqrt{\frac{8\Delta_j^2}{K} 2T_{j-1}} = \frac{1}{6B}\,,
\end{aligned}$$

where (a) follows from a similar calculation as Equation (16) in the proof of Lemma C.2, (b) follows from the concavity of $\sqrt{x}$ and Jensen's inequality, and (c) follows from the fact that $\sum_{l=1}^{K} \tau_l \leq T_{j-1}$.

Combining **??????** we have that

$$\sup_{I:\epsilon_{\min}(I)=\Delta_j} \mathbb{E}[R_T(I)] \geq \Delta_j T_j \left(\mathbf{P}(E_{j,k}(A_j)) - \frac{1}{4B}\right)\,.$$

Since the above inequality holds for all $k \in [K]$, by averaging we get

$$\begin{aligned}\sup_{I:\epsilon_{\min}(I)=\Delta_j} \mathbb{E}[R_T(I)] &\geq \Delta_j T_j \left(\frac{1}{K} \sum_{k=1}^{K} \mathbf{P}(E_{j,k}(A_j)) - \frac{1}{4B} \right) \\ &= \Delta_j T_j \left(p_j - \frac{1}{4B} \right) \\ &\geq \Delta_j T_j \frac{1}{4B} \,.\end{aligned}$$

Substituting the value of $\Delta_j T_j$ we get

$$\begin{aligned}\sup_{I:\epsilon_{\min}(I)=\Delta_j} \mathbb{E}[R_T(I)] &\geq \Delta_j T_j \frac{1}{4B} = \frac{\sqrt{K}}{24B} T^{-(j-1)/2B} T^{j/B} \frac{1}{4B} \\ &= \frac{\sqrt{K}}{24B} T^{(j-1)/2B} T^{1/B} \frac{1}{4B} = \Omega\left(\frac{K T^{1/B}}{B^2 \Delta_j} \right) .\end{aligned}$$

□

Finally, $\sum_{j=1}^{B} p_j \geq \frac{1}{2}$ implies that there exists $j \in [B]$ with $p_j \geq 1/2B$. Combining the two lemmas above, we get that there exists $j \in [B]$ with $p_j \geq 1/2B$ such that the algorithm incurs a regret of $\Omega\left(\frac{KT^{1/B}}{B^2\Delta_j}\right)$. In this case, there must exist an instance $E_{j,k}$ with gap $\epsilon_{\min}(E_{j,k}) = \Delta_j$ such that the regret of the algorithm under $E_{j,k}$ is $\Omega\left(\frac{KT^{1/B}}{B^2\Delta_j}\right)$. This completes the proof of our lower bound.