Proceedings of Machine Learning Research vol 98:1–38, 2019 30th International Conference on Algorithmic Learning Theory

PAC Battling Bandits in the Plackett-Luce Model

Aadirupa Saha AADIRUPA@IISC.AC.IN
Indian Institute of Science, Bangalore 560012

Aditya Gopalan ADITYA@IISC.AC.IN
Indian Institute of Science, Bangalore 560012

Editors: Aurélien Garivier and Satyen Kale

Abstract

We introduce the probably approximately correct (PAC) *Battling-Bandit* problem with the Plackett-Luce (PL) subset choice model–an online learning framework where at each trial the learner chooses a subset of k arms from a fixed set of n arms, and subsequently observes a stochastic feedback indicating preference information of the items in the chosen subset, e.g., the most preferred item or ranking of the top m most preferred items etc. The objective is to identify a near-best item in the underlying PL model with high confidence. This generalizes the well-studied PAC *Dueling-Bandit* problem over n arms, which aims to recover the *best-arm* from pairwise preference information, and is known to require $O(\frac{n}{\epsilon^2}\ln\frac{1}{\delta})$ sample complexity (Szörényi et al., 2015; Busa-Fekete et al., 2013). We study the sample complexity of this problem under various feedback models: (1) Winner of the subset (WI), and (2) Ranking of top-m items (TR) for $2 \le m \le k$. We show, surprisingly, that with winner information (WI) feedback over subsets of size $2 \le k \le n$, the best achievable sample complexity is still $O\left(\frac{n}{\epsilon^2}\ln\frac{1}{\delta}\right)$, independent of k, and the same as that in the Dueling Bandit setting ($k = 2$). For the more general top-m ranking (TR) feedback model, we show a significantly smaller lower bound on sample complexity of $\Omega\left(\frac{n}{m\epsilon^2}\ln\frac{1}{\delta}\right)$, which suggests a multiplicative reduction by a factor m owing to the additional information revealed from preferences among m items instead of just 1. We also propose two algorithms for the PAC problem with the TR feedback model with optimal (upto logarithmic factors) sample complexity guarantees, establishing the increase in statistical efficiency from exploiting rank-ordered feedback.

Keywords: Online Learning, Probably Approximate Correct (PAC) Learning, Subsetwise Preference, Sample Complexity, Lower Bound, Plackett-Luce, Random Utility Models, Dueling Bandit.

1. Introduction

The dueling bandit problem has recently gained attention in the machine learning community (Yue et al., 2012; Ailon et al., 2014; Zoghi et al., 2014; Szörényi et al., 2015). This is a variant of the multi-armed bandit problem (Auer et al., 2002) in which the learner needs to learn an 'best arm' from pairwise comparisons between arms. In this work, we consider a natural generalization of the dueling bandit problem where the learner can adaptively select a subset of k arms ($k \ge 2$) in each round, and observe relative preferences in the subset following a Plackett-Luce (PL) feedback model (Marden, 1996), with the objective of learning the 'best arm'. We call this the *battling bandit* problem with the Plackett-Luce model.

The battling bandit decision framework (Saha and Gopalan, 2018; Chen et al., 2018) models several application domains where it is possible to elicit feedback about preferred options from among a general set of offered options, instead of being able to compare only two options at a time as in the dueling setup. Furthermore, the phenomenon of competition – that an option's utility or attractiveness is often assessed relative to that of other items in the offering – is captured effectively by a subset-dependent stochastic choice model such as Plackett-Luce. Common examples of learning settings with such feedback include recommendation systems and search engines, medical interviews, tutoring systems–any applications where relative preferences from a chosen pool of options are revealed.

We consider a natural probably approximately correct (PAC) learning problem in the battling bandit setting: Output an ϵ-approximate best item (with respect to its Plackett-Luce parameter) with probability at least $(1-\delta)$, while keeping the total number of adaptive exploration rounds small. We term this the (ϵ,δ)-PAC objective of searching for an approximate winner or top-1 item.

Our primary interest lies in understanding how the subset size k influences the sample complexity of achieving (ϵ,δ)-PAC objective in subset choice models for various feedback information structures, e.g., winner information (WI), which returns only a single winner of the chosen subset, or the more general top ranking (TR) information structure, where an ordered tuple of m 'most-preferred' items is observed. More precisely, we ask: Does being able to play size-k subsets help learn optimal items faster than in the dueling setting ($k=2$)? How does this depend on the subset size k, and on the feedback information structure? How much, if any, does rank-ordered feedback accelerate the rate of learning, compared to only observing winner feedback? This paper takes a step towards resolving such questions within the context of the Plackett-Luce choice model. Among the contributions of this paper are:

1. We frame a *PAC version of Battling Bandits* with n arms – a natural generalization of the PAC-Dueling-Bandits problem (Szörényi et al., 2015) – with the objective of finding an ϵ-approximate best item with probability at least $1-\delta$ with minimum possible sample complexity, termed as the (ϵ,δ)-PAC objective (Section 3.2).

2. We consider learning with winner information (WI) feedback, where the learner can play a subsets $S_t \subseteq [n]$ of exactly $|S_t| = k$ distinct elements at each round t, following which a winner of S_t is observed according to an underlying, unknown, Plackett-Luce model. We show an information-theoretic lower bound on sample complexity for (ϵ,δ)-PAC of $\Omega\left(\frac{n}{\epsilon^2}\ln\frac{1}{\delta}\right)$ rounds (Section 4.1), which is of the same order as that for the dueling bandit ($k=2$) (Yue and Joachims, 2011). This implies that, despite the increased flexibility of playing sets of potentially large size k, with just winner information feedback, one cannot hope for a faster rate of learning than in the case of pairwise selections. Intuitively, competition among a large number (k) of elements vying for the top spot at each time exactly offsets the potential gain that being able to test more alternatives together brings. On the achievable side, we design two algorithms (Section 4.2) for the (ϵ,δ)-PAC objective, and derive sample complexity guarantees which are optimal within a logarithmic factor of the lower bound derived earlier. When the learner is allowed to play subsets of sizes $1,2,\ldots$ upto k, which is a slightly more flexible setting than above, we design a median elimination-based algorithm with order-optimal $O\left(\frac{n}{\epsilon^2}\ln\frac{1}{\delta}\right)$ sample complexity which, when specialized to $k=2$, improves upon existing sample complexity bounds for PAC-dueling bandit algorithms, e.g. Yue and Joachims (2011); Szörényi et al. (2015) under the PL model (Section. 4.3).

3. We next study the (ϵ,δ)-PAC problem in a more general top-ranking (TR) feedback model where the learner gets to observe the ranking of top m items drawn from the Plackett-Luce

distribution, $2 \leq m \leq k$ (Section 3.1), departing from prior work. For $m = 1$, the setting simply boils down to WI feedback model. In this case, we are able to prove a sample complexity lower bound of $\Omega\left(\frac{n}{m\epsilon^2} \ln \frac{1}{\delta}\right)$ (Theorem 10), which suggests that with top-m ranking (TR) feedback, it may be possible to aggregate information m times faster than with just winner information feedback. We further present two algorithms (Section 5.2) for this problem which, are shown to enjoy optimal (upto logarithmic factors) sample complexity guarantees. This formally shows that the m-fold increase in statistical efficiency by exploiting richer information contained in top-m ranking feedback is, in fact, algorithmically achievable.

4. From an algorithmic point of view, we elucidate how the structure of the Plackett-Luce choice model, such as its independent of irrelevant attributes (IIA) property, play a crucial role in allowing the development of parameter estimates, together with tight confidence sets, which form the basis for our learning algorithms. It is indeed by leveraging this property (Lemma 1) that we afford to maintain consistent pairwise preferences of the items by applying the concept of *Rank Breaking* to subsetwise preference data. This significantly alleviates the combinatorial explosion that could otherwise result if one were to keep more general subset-wise estimates.

Related Work: Statistical parameter estimation in Plackett-Luce models has been studied in detail in the offline batch (non-adaptive) setting (Chen and Suh, 2015; Khetan and Oh, 2016; Jang et al., 2017).

In the online setting, there is a fairly mature body of work concerned with PAC best-arm (or top-ℓ arm) identification in the classical multi-armed bandit (Even-Dar et al., 2006; Audibert and Bubeck, 2010; Kalyanakrishnan et al., 2012; Karnin et al., 2013; Jamieson et al., 2014), where absolute utility information is assumed to be revealed upon playing a single arm or item. Though most work on dueling bandits has focused on the regret minimization goal (Zoghi et al., 2014; Ramamohan et al., 2016), there have been recent developments on the PAC objective for different pairwise preference models, such as those satisfying stochastic triangle inequalities and strong stochastic transitivity (Yue and Joachims, 2011), general utility-based preference models (Urvoy et al., 2013), the Plackett-Luce model (Szörényi et al., 2015), the Mallows model (Busa-Fekete et al., 2014a), etc. Recent work in the PAC setting focuses on learning objectives other than identifying the single (near) best arm, e.g. recovering a few of the top arms (Busa-Fekete et al., 2013; Mohajer et al., 2017; Chen et al., 2017), or the true ranking of the items (Busa-Fekete et al., 2014b; Falahatgar et al., 2017).

The work which is perhaps closest in spirit to ours is that of Chen et al. (2018), which addresses the problem of learning the top-ℓ items in Plackett-Luce battling bandits. Even when specialized to $\ell = 1$ (as we consider here), however, this work differs in several important aspects from what we attempt. Chen et al. (2018) develop algorithms for the probably *exactly* correct objective (recovering a near-optimal arm is not favored), and, consequently, show instance-dependent sample complexity bounds, whereas we allow a tolerance of ϵ in defining best arms, which is often natural in practice Szörényi et al. (2015); Yue and Joachims (2011). As a result, we bring out the dependence of the sample complexity on the specified tolerance level ϵ, rather than on purely instance-dependent measures of hardness. Also, their work considers only winner information (WI) feedback from the subsets chosen, whereas we consider, for the first time, general m top-ranking information feedback.,

A related battling-type bandit setting has been studied as the MNL-bandits assortment-optimization problem by Agrawal et al. (2016), although it takes prices of items into account when defining their utilities. As a result, their work optimizes for a subset with highest expected revenue (price), whereas we search for a best item (Condorcet winner). and the two settings are in general incomparable.

2. Preliminaries

Notation. We denote by $[n]$ the set $\{1, 2, ..., n\}$. For any subset $S \subseteq [n]$, let $|S|$ denote the cardinality of S. When there is no confusion about the context, we often represent (an unordered) subset S as a vector, or ordered subset, S of size $|S|$ (according to, say, a fixed global ordering of all the items $[n]$). In this case, $S(i)$ denotes the item (member) at the ith position in subset S. $\mathbf{\Sigma}_S = \{\sigma \mid \sigma$ is a permutation over items of $S\}$, where for any permutation $\sigma \in \Sigma_S$, $\sigma(i)$ denotes the element at the i-th position in $\sigma, i \in [|S|]$. $\mathbf{1}(\varphi)$ is generically used to denote an indicator variable that takes the value 1 if the predicate φ is true, and 0 otherwise. $x \vee y$ denotes the maximum of x and y, and $Pr(A)$ is used to denote the probability of event A, in a probability space that is clear from the context.

2.1. Discrete Choice Models and Plackett-Luce (PL)

A discrete choice model specifies the relative preferences of two or more discrete alternatives in a given set. A widely studied class of discrete choice models is the class of *Random Utility Models* (RUMs), which assume a ground-truth utility score $\theta_i \in \mathbb{R}$ for each alternative $i \in [n]$, and assign a conditional distribution $\mathcal{D}_i(\cdot|\theta_i)$ for scoring item i. To model a winning alternative given any set $S \subseteq [n]$, one first draws a random utility score $X_i \sim \mathcal{D}_i(\cdot|\theta_i)$ for each alternative in S, and selects an item with the highest random score.

One widely used RUM is the *Multinomial-Logit (MNL)* or *Plackett-Luce model (PL)*, where the $\mathcal{D}_i$s are taken to be independent Gumbel distributions with location parameters θ'_i and scale parameter 1 (Azari et al., 2012), which result to probability densities $\mathcal{D}_i(x_i|\theta'_i) = e^{-(x_j-\theta'_j)}e^{-e^{-(x_j-\theta'_j)}}$, $\theta'_i \in R$, $\forall i \in [n]$. Moreover assuming $\theta'_i = \ln \theta_i$, $\theta_i > 0\ \forall i \in [n]$, in this case the probability that an alternative i emerges as the winner in the set $S \ni i$ becomes proportional to its parameter value:

$$Pr(i|S) = \frac{\theta_i}{\sum_{j \in S} \theta_j}. \tag{1}$$

We will henceforth refer the above choice model as PL model with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. Clearly the above model induces a total ordering on the arm set $[n]$: If $p_{ij} = P(i \succ j) = Pr(i|\{i,j\}) = \frac{\theta_i}{\theta_i+\theta_j}$ denotes the pairwise probability of item i being preferred over item j, then $p_{ij} \ge \frac{1}{2}$ if and only if $\theta_i \ge \theta_j$, or in other words if $p_{ij} \ge \frac{1}{2}$ and $p_{jk} \ge \frac{1}{2}$ then $p_{ik} \ge \frac{1}{2}, \forall i,j,k \in [n]$ (Ramamohan et al., 2016).

Other families of discrete choice models can be obtained by imposing different probability distributions over the utility scores X_i, e.g. if $(X_1, \ldots X_n) \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{\Lambda})$ are jointly normal with mean $\boldsymbol{\theta} = (\theta_1, \ldots \theta_n)$ and covariance $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$, then the corresponding RUM-based choice model reduces to the *Multinomial Probit (MNP)*. Unlike MNL, though, the choice probabilities $Pr(i|S)$ for the MNP model do not admit a closed-form expression (Vojacek et al., 2010).

2.2. Independence of Irrelevant Alternatives

A choice model Pr is said to possess the *Independence of Irrelevant Alternatives (IIA)* property if the ratio of probabilities of choosing any two items, say i_1 and i_2 from within any choice set

$S \ni i_1, i_2$ is independent of a third alternative j present in S (Benson et al., 2016). More specifically, $\frac{Pr(i_1|S_1)}{Pr(i_2|S_1)} = \frac{Pr(i_1|S_2)}{Pr(i_2|S_2)}$ for any two distinct subsets $S_1, S_2 \subseteq [n]$ that contain i_1 and i_2. One example of such a choice model is Plackett-Luce.

Remark 1 *IIA turns out to be very valuable in estimating the parameters of a PL model, with high confidence, via* Rank-Breaking *– the idea of extracting pairwise comparisons from (partial) rankings and applying estimators on the obtained pairs, treating each comparison independently. Although this technique has previously been used in batch (offline) PL estimation (Khetan and Oh, 2016), we show that it can be used in online problems for the first time. We crucially exploit this property of the PL model in the algorithms we design (Algorithms 1-3), and in establishing their correctness and sample complexity guarantees.*

Lemma 1 (Deviations of pairwise win-probability estimates for PL model) *Consider a Plackett-Luce choice model with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ (see Eqn. (1)), and fix two distinct items $i, j \in [n]$. Let $S_1, \ldots, S_T$ be a sequence of (possibly random) subsets of $[n]$ of size at least 2, where T is a positive integer, and $i_1, \ldots, i_T$ a sequence of random items with each $i_t \in S_t$, $1 \le t \le T$, such that for each $1 \le t \le T$, (a) S_t depends only on $S_1, \ldots, S_{t-1}$, and (b) i_t is distributed as the Plackett-Luce winner of the subset S_t, given $S_1, i_1, \ldots, S_{t-1}, i_{t-1}$ and S_t, and (c) $\forall t : \{i,j\} \subseteq S_t$ with probability 1. Let $n_i(T) = \sum_{t=1}^{T} \mathbf{1}(i_t = i)$ and $n_{ij}(T) = \sum_{t=1}^{T} \mathbf{1}(\{i_t \in \{i,j\}\})$. Then, for any positive integer v, and $\eta \in (0,1)$,*

$$Pr\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \ge \eta, n_{ij}(T) \ge v\right) \vee Pr\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \le -\eta, n_{ij}(T) \ge v\right) \le e^{-2v\eta^2}.$$

Proof (sketch). The proof uses a novel coupling argument to work in an equivalent probability space for the PL model with respect to the item pair i, j, as follows. Let $Z_1, Z_2, \ldots$ be a sequence of iid Bernoulli random variables with success parameter $\theta_i/(\theta_i + \theta_j)$. A counter C is first initialized to 0. At each time t, given $S_1, i_1, \ldots, S_{t-1}, i_{t-1}$ and S_t, an independent coin is tossed with probability of heads $(\theta_i + \theta_j)/\sum_{k \in S_t} \theta_k$. If the coin lands tails, then i_t is drawn as an independent sample from the Plackett-Luce distribution over $S_t \setminus \{i,j\}$, else, the counter is incremented by 1, and i_t is returned as i if $Z_C = 1$ or j if $Z_C = 0$. This construction yields the correct joint distribution for the sequence $i_1, S_1, \ldots, i_T, S_T$, because of the IIA property of the PL model:

$$Pr(i_t = i | i_t \in \{i,j\}, S_t) = \frac{Pr(i_t = i|S_t)}{Pr(i_t \in \{i,j\}|S_t)} = \frac{\theta_i/\sum_{k \in S_t} \theta_k}{(\theta_i + \theta_j)/\sum_{k \in S_t} \theta_k} = \frac{\theta_i}{\theta_i + \theta_j}.$$

The proof now follows by applying Hoeffding's inequality on prefixes of the sequence $Z_1, Z_2, \ldots$. ■

3. Problem Setup

We consider the PAC version of the sequential decision-making problem of finding the best item in a set of n items by making subset-wise comparisons. Formally, the learner is given a finite set $[n]$ of $n > 2$ arms. At each decision round $t = 1, 2, \ldots$, the learner selects a subset $S_t \subseteq [n]$ of k distinct items, and receives (stochastic) feedback depending on (a) the chosen subset S_t, and (b) a Plackett-Luce (PL) choice model with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ a priori unknown to the learner. The nature of the feedback can be of several types as described in Section 3.1. Without loss of generality, we will henceforth assume $\theta_i \in [0,1]$, $\forall i \in [n]$, since the PL choice probabilities are positive scale-invariant by (1). We also let $\theta_1 > \theta_i \, \forall i \in [n] \setminus \{1\}$ for ease of exposition[1]. We call this decision-making model, parameterized by a PL instance $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ and a playable subset size k, as Battling Bandits (BB) with the Plackett-Luce (PL), or *BB-PL* in short. We define a *best item* to be one with the highest score parameter: $i^* \in \underset{i \in [n]}{\text{argmax}}\, \theta_i$. Under the assumptions above, $i^* = 1$ uniquely. Note that here we have $p_{1i} = P(1 \succ i) > \frac{1}{2}, \forall i \in [n] \setminus \{1\}$, so item 1 is the *Condorcet Winner* (Ramamohan et al., 2016) of the PL model.

3.1. Feedback models

By feedback model, we mean the information received (from the 'environment') once the learner plays a subset $S \subseteq [n]$ of k items. We define three types of feedback in the PL battling model:

- **Winner of the selected subset (WI):** The environment returns a single item $I \in S$, drawn independently from the probability distribution $Pr(I = i|S) = \frac{\theta_i}{\sum_{j \in S} \theta_j} \;\; \forall i \in S.$
- **Full ranking selected subset of items (FR):** The environment returns a full ranking $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_S$, drawn from the probability distribution $Pr(\boldsymbol{\sigma} = \sigma|S) = \prod_{i=1}^{|S|} \frac{\theta_{\sigma(i)}}{\sum_{j=i}^{|S|} \theta_{\sigma(j)}}, \;\; \sigma \in \boldsymbol{\Sigma}_S$. In fact, this is equivalent to picking $\boldsymbol{\sigma}(1)$ according to the winner (WI) feedback from S, then picking $\boldsymbol{\sigma}(2)$ according to WI feedback from $S \setminus \{\boldsymbol{\sigma}(1)\}$, and so on, until all elements from S are exhausted, or, in other words, successively sampling $|S|$ winners from S according to the PL model, without replacement.

  A feedback model that generalizes the types of feedback above is:

- **Top-m ranking of items (TR-m or TR):** The environment returns a ranking of only m items from among S, i.e., the environment first draws a full ranking $\boldsymbol{\sigma}$ over S according to Plackett-Luce as in **FR** above, and returns the first m rank elements of $\boldsymbol{\sigma}$, i.e., $(\boldsymbol{\sigma}(1), \ldots, \boldsymbol{\sigma}(m))$. It can be seen that for each permutation σ on a subset $S_m \subset S$, $|S_m| = m$, we must have $Pr(\boldsymbol{\sigma} = \sigma|S) = \prod_{i=1}^{m} \frac{\theta_{\sigma(i)}}{\sum_{j=i}^{m} \theta_{\sigma(j)} + \sum_{j \in S \setminus S_m} \theta_{\sigma(j)}}$. Generating such a $\boldsymbol{\sigma}$ is also equivalent to successively sampling m winners from S according to the PL model, without replacement. It follows that **TR** reduces to **FR** when $m = k = |S|$ and to **WI** when $m = 1$.

3.2. Performance Objective: Correctness and Sample Complexity

Suppose $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_n)$ and $k \leq n$ define a BB-PL instance with best arm $i^* = 1$, and $0 < \epsilon \leq \frac{1}{2}, 0 < \delta \leq 1$ are given constants. An arm $i \in [n]$ is said to be ϵ-optimal[2] if the probability that i beats 1 is over $\frac{1}{2} - \epsilon$, i.e., if $p_{i1} := Pr(i|\{1, i\}) > \frac{1}{2} - \epsilon$. A sequential algorithm that operates in this BB-PL instance, using feedback from an appropriate subset-wise feedback model (e.g., WI, FR or TR), is said to be (ϵ, δ)-PAC if (a) it stops and outputs an arm $I \in [n]$ after a finite number of decision rounds (subset plays) with probability 1, and (b) the probability that its output I is an ϵ-optimal arm is at least $1 - \delta$, i.e, $Pr(I \text{ is } \epsilon\text{-optimal}) \geq 1 - \delta$. Furthermore, by *sample complexity* of the algorithm, we mean the expected time (number of decision rounds) taken by the algorithm to stop.

1. We naturally assume that this knowledge is not known to the learning algorithm, and note that extension to the case where several items have the same highest parameter value is easily accomplished.
2. informally, a 'near-best' arm

Note that $p_{ij} > \frac{1}{2} + \epsilon \Leftrightarrow \frac{\theta_i}{\theta_j} > \frac{1/2+\epsilon}{1/2-\epsilon}$, $\forall i, j \in [n]$, so the score parameter θ_i of a near-best item must be at least $\frac{1/2-\epsilon}{1/2+\epsilon}$ times θ_1.

4. Analysis with Winner Information (WI) feedback

In this section we consider the PAC-WI goal with the WI feedback information model in BB-PL instances of size n with playable subset size k. We start by showing that a sample complexity-lower bound for any (ϵ, δ)-PAC algorithm with WI feedback is $\Omega\bigg(\frac{n}{\epsilon^2} \ln \frac{1}{\delta}\bigg)$ (Theorem 2). This bound is independent of k, implying that playing a dueling game ($k = 2$) is as good as the battling game as the extra flexibility of k-subsetwise feedback does not result in a faster learning rate. We next propose two algorithms for (ϵ, δ)-PAC, with WI feedback, with optimal (upto a logarithmic factor) sample complexity of $O(\frac{n}{\epsilon^2} \ln \frac{k}{\delta})$ (Section 4.2). We also analyze a slightly different setting allowing the learner to play subsets S_t of any size $1, 2, \ldots, k$, rather than a fixed size k – this gives somewhat more flexibility to the learner, resulting in algorithms with improved sample complexity guarantees of $O(\frac{n}{\epsilon^2} \ln \frac{1}{\delta})$, without the $\ln k$ dependency as before (Section 4.3).

4.1. Lower Bound for Winner Information (WI) feedback

Theorem 2 (Lower bound on Sample Complexity with WI feedback) *Given $\epsilon \in (0, \frac{1}{\sqrt{8}}]$ and $\delta \in (0, 1]$, and an (ϵ, δ)-PAC algorithm A for BB-PL with feedback model WI, there exists a PL instance ν such that the sample complexity of A on ν is at least $\Omega\bigg(\frac{n}{\epsilon^2} \ln \frac{1}{2.4\delta}\bigg)$.*

Proof (sketch). The argument is based on a change-of-measure argument (Lemma 1) of Kaufmann et al. (2016), restated below for convenience:

Consider a multi-armed bandit (MAB) problem with n arms or actions $\mathcal{A} = [n]$. At round t, let A_t and Z_t denote the arm played and the observation (reward) received, respectively. Let $\mathcal{F}_t = \sigma(A_1, Z_1, \ldots, A_t, Z_t)$ be the sigma algebra generated by the trajectory of a sequential bandit algorithm upto round t.

Lemma 3 (Lemma 1, Kaufmann et al. (2016)) *Let ν and ν' be two bandit models (assignments of reward distributions to arms), such that ν_i (resp. ν'_i) is the reward distribution of any arm $i \in \mathcal{A}$ under bandit model ν (resp. ν'), and such that for all such arms i, ν_i and ν'_i are mutually absolutely continuous. Then for any almost-surely finite stopping time τ with respect to $(\mathcal{F}_t)_t$,*

$$\sum_{i=1}^{n} \mathbf{E}_\nu[N_i(\tau)] KL(\nu_i, \nu'_i) \geq \sup_{\mathcal{E} \in \mathcal{F}_\tau} kl(Pr_\nu(\mathcal{E}), Pr_{\nu'}(\mathcal{E})),$$

where $kl(x, y) := x \log(\frac{x}{y}) + (1 - x) \log(\frac{1-x}{1-y})$ is the binary relative entropy, $N_i(\tau)$ denotes the number of times arm i is played in τ rounds, and $Pr_\nu(\mathcal{E})$ and $Pr_{\nu'}(\mathcal{E})$ denote the probability of any event $\mathcal{E} \in \mathcal{F}_\tau$ under bandit models ν and ν', respectively.

To employ this result, note that in our case, each bandit instance corresponds to an instance of the BB-PL problem with the arm set containing all subsets of $[n]$ of size k: $\mathcal{A} = \{S = (S(1), \ldots S(k)) \subseteq [n] \mid S(i) < S(j), \forall i < j\}$. The key part of our proof relies on carefully crafting a true instance, with optimal arm 1, and a family of slightly perturbed alternative instances $\{\boldsymbol{\nu}^a : a \neq 1\}$, each with optimal arm $a \neq 1$.

We choose the true problem instance $\boldsymbol{\nu}^1$ as the Plackett-Luce model with parameters

$$\theta_j = \theta\left(\frac{1}{2} - \epsilon\right), \forall j \in [n] \setminus \{1\}, \text{ and } \theta_1 = \theta\left(\frac{1}{2} + \epsilon\right), \quad \text{(true instance)}$$

for some $\theta \in \mathbb{R}_+,\ \epsilon > 0$. Corresponding to each suboptimal item $a \in [n] \setminus \{1\}$, we now define an alternative problem instance $\boldsymbol{\nu}^a$ as the Plackett-Luce model with parameters

$$\theta'_j = \theta\left(\frac{1}{2} - \epsilon\right)^2, \forall j \in [n] \setminus \{a, 1\},\ \theta'_1 = \theta\left(\frac{1}{4} - \epsilon^2\right), \theta'_a = \theta\left(\frac{1}{2} + \epsilon\right)^2 \quad \text{(alternative instance)}.$$

The result of Theorem 2 is now obtained by applying Lemma 3 on pairs of problem instances $(\nu, \nu'^{(a)})$, with suitable upper bounds on the KL-divergence terms, and the observation that $kl(\delta, 1-\delta) \geq \ln \frac{1}{2.4\delta}$. The complete proof is given in Appendix B.1. ■

Remark 2 *Theorem 2 shows, rather surprisingly, that the PAC sample complexity of identifying a near-optimal item with only winner feedback information from k-size subsets, does not reduce with k, implying that there is no reduction in hardness of learning from the pairwise comparisons case ($k = 2$). On one hand, one may expect to see improved sample complexity as the number of items being simultaneously tested in each round is large (k). On the other hand, the sample complexity could also worsen, since it is intuitively 'harder' for a good (near-optimal) item to win and show itself, in just a single winner draw, against a large population of $k-1$ other competitors. The result, in a sense, formally establishes that the former advantage is nullified by the latter drawback. A somewhat more formal, but heuristic, explanation for this phenomenon is that the number of bits of information that a single winner draw from a size-k subset provides is $O(\ln k)$, which is not significantly larger than when $k > 2$, thus an algorithm cannot accumulate significantly more information per round compared to the pairwise case.*

4.2. Algorithms for Winner Information (WI) feedback model

This section describes our proposed algorithms for the (ϵ, δ)-PAC objective with winning item (WI) feedback.

Principles of algorithm design. The key idea on which all our learning algorithms are based is that of maintaining estimates of the pairwise win-loss probabilities $p_{ij} = Pr(i|i,j)$ in the Plackett-Luce model. This helps circumvent an $O(n^k)$ combinatorial explosion that would otherwise result if we directly attempted to estimate probability distributions for each possible k-size subset. However, it is not obvious if consistent and tight pairwise estimates can be constructed in a general subset-wise choice model, but the special form of the Plackett-Luce model again comes to our rescue. The IIA property that the PL model enjoys, allows for accurate pairwise estimates via interpretation of partial preference feedback as a set of pairwise preferences, e.g., a winner a sampled from among a, b, c is interpreted as the pairwise preferences $a \succ b$, $a \succ c$. Lemma 1 formalizes this property and allows us to use pairwise win/loss probability estimators with explicit confidence intervals for them.

Algorithm 1: (*Trace-the-Best*). Our first algorithm *Trace-the-Best* is based on the simple idea of tracing the empirical best item–specifically, it maintains a running winner r_ℓ at every iteration ℓ, making it battle with a set of $k-1$ arbitrarily chosen items. After battling long enough (precisely, for $\frac{2k}{\epsilon^2} \ln \frac{2n}{\delta}$ many rounds), if the empirical winner c_ℓ turns out to be more than $\frac{\epsilon}{2}$-favorable than the running winner r_ℓ, in term of its pairwise preference score: $\hat{p}_{c_\ell, r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}$, then c_ℓ replaces r_ℓ, or else r_ℓ retains its place and status quo ensues.

Theorem 4 (*Trace-the-Best*: Correctness and Sample Complexity with WI) *Trace-the-Best (Algorithm 1) is (ϵ, δ)-PAC with sample complexity $O(\frac{n}{\epsilon^2}\log\frac{n}{\delta})$.*

Proof (sketch). The main idea is to retain an estimated best item as a 'running winner' r_ℓ, and compare it with the 'empirical best item' c_ℓ of $\mathcal{A}$ at every iteration ℓ. The crucial observation lies in noting that at any iteration ℓ, r_ℓ gets updated as follows:

Lemma 5 *At any iteration $\ell = 1, 2 \ldots \lfloor \frac{n}{k-1} \rfloor$, with probability at least $(1-\frac{\delta}{2n})$, Algorithm 1 retains $r_{\ell+1} \leftarrow r_\ell$ if $p_{c_\ell r_\ell} \le \frac{1}{2}$, and sets $r_{\ell+1} \leftarrow c_\ell$ if $p_{c_\ell r_\ell} \ge \frac{1}{2} + \epsilon$.*

This leads to the claim that between any two successive iterations ℓ and $\ell+1$, we must have, with high probability, that $p_{r_{\ell+1} r_\ell} \ge \frac{1}{2}$ and, $p_{r_{\ell+1} c_\ell} \ge \frac{1}{2} - \epsilon$, showing that the estimated 'best' item r_ℓ can only get improved per iteration as $p_{r_{\ell+1} r_\ell} \ge \frac{1}{2}$ (with high probability at least $1 - \frac{(k-1)\delta}{2n}$). Repeating this above argument for each iteration $\ell \in \lfloor \frac{n}{k-1} \rfloor$ results in the desired correctness guarantee of $p_{r_* 1} \ge \frac{1}{2} - \epsilon$. The sample complexity bound follows easily by noting the total number of possible iterations can be at most $\lceil \frac{n}{k-1} \rceil$, with the per-iteration sample complexity being $t = \frac{2k}{\epsilon^2}\ln\frac{2n}{\delta}$. ∎

Algorithm 1 *Trace-the-Best*

Input:
  Set of items: $[n]$, Subset size: $n \ge k > 1$
  Error bias: $\epsilon > 0$, Confidence parameter: $\delta > 0$
Initialize:
  $r_1 \leftarrow$ Any (random) item from $[n]$, $\mathcal{A} \leftarrow$ Randomly select $(k-1)$ items from $[n] \setminus \{r_1\}$
  Set $\mathcal{A} \leftarrow \mathcal{A} \cup \{r_1\}$, and $S \leftarrow [n] \setminus \mathcal{A}$
while $\ell = 1, 2, \ldots$ **do**
  Play the set $\mathcal{A}$ for $t := \frac{2k}{\epsilon^2}\ln\frac{2n}{\delta}$ rounds
  $w_i \leftarrow$ Number of times i won in t plays of $\mathcal{A}$, $\forall i \in \mathcal{A}$
  Set $c_\ell \leftarrow \underset{i \in \mathcal{A}}{\text{argmax}}\, w_i$, and $\hat{p}_{ij} \leftarrow \frac{w_i}{w_i + w_j}$, $\forall i, j \in \mathcal{A}, i \neq j$
  if $\hat{p}_{c_\ell, r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}$, **then** set $r_{\ell+1} \leftarrow c_\ell$; **else** $r_{\ell+1} \leftarrow r_\ell$
  if $(S == \emptyset)$ **then**
    Break (exit the while loop)
  else if $|S| < k-1$ **then**
    $\mathcal{A} \leftarrow$ Select $(k-1-|S|)$ items from $\mathcal{A} \setminus \{r_\ell\}$ uniformly at random
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{r_\ell\} \cup S$; and $S \leftarrow \emptyset$
  else
    $\mathcal{A} \leftarrow$ Select $(k-1)$ items from S uniformly at random
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{r_\ell\}$ and $S \leftarrow S \setminus \mathcal{A}$
  end if
end while
Output: $r_* = r_\ell$ as the ϵ-optimal item

Remark 3 *The sample complexity of Trace-the-Best, is order wise optimal when $\delta < \frac{1}{n}$, as follows from our derived lower bound guarantee (Theorem 2).*

When $\delta > \frac{1}{n}$, the sample complexity guarantee of *Trace-the-Best* is off by a factor of $\ln n$. We now propose another algorithm, *Divide-and-Battle* (Algorithm 2) that enjoys an (ϵ, δ)-PAC sample complexity of $O\left(\frac{n}{\epsilon^2} \ln \frac{k}{\delta}\right)$.

Algorithm 2: (*Divide-and-Battle*). *Divide-and-Battle* first divides the set of n items into groups of size k, and plays each group long enough so that a good item in the group stands out as the empirical winner with high probability (Line 11). It then retains the empirical winner per group (Line 13) and recurses on the retained set of the winners, until it is left with only a single item, which is finally declared as the ϵ-optimal item. *The pseudo code of Divide-and-Battle is given in Appendix B.3.*

Theorem 6 (*Divide-and-Battle*: Correctness and Sample Complexity with WI) *Divide-and-Battle (Algorithm 2) is (ϵ, δ)-PAC with sample complexity $O(\frac{n}{\epsilon^2} \log \frac{k}{\delta})$.*

Proof (sketch). The crucial observation here is that at any iteration ℓ, for any set $\mathcal{G}_g$ $(g = 1, 2, \ldots G)$, the item c_g retained by the algorithm is likely to be not more than ϵ_ℓ-worse than the best item of the set $\mathcal{G}_g$, with probability at least $(1 - \delta_\ell)$. Precisely, we show that:

Lemma 7 *At any iteration ℓ, for any $\mathcal{G}_g$, if $i_g := \arg\max_{i \in \mathcal{G}_g} \theta_i$, then with probability at least $(1 - \delta_\ell)$, $p_{c_g i_g} > \frac{1}{2} - \epsilon_\ell$.*

This guarantees that, between any two successive rounds ℓ and $\ell + 1$, we do not lose out by more than an additive factor of ϵ_ℓ in terms of highest score parameter of the remaining set of items. Aggregating this claim over all iterations can be made to show that $p_{r_* 1} > \frac{1}{2} - \epsilon$, as desired. The sample complexity bound follows by carefully summing the total number of times ($t = \frac{k}{2\epsilon_\ell^2} \ln \frac{k}{\delta_\ell}$) a set $\mathcal{G}_g$ is played per iteration ℓ, with the maximum number of possible iterations being $\lceil \ln_k n \rceil$. ■

Remark 4 *The sample complexity of Divide-and-Battle is order-wise optimal in the 'small-δ' regime $\delta \ll \frac{1}{k}$ by the lower bound result (Theorem 2). However, for the 'moderate-δ' regime $\delta \gtrapprox \frac{1}{k}$, we conjecture that the lower bound is loose by an additive factor of $\frac{n \ln k}{\epsilon^2}$, i.e., that a improved lower bound of $\Omega(\frac{n}{\epsilon^2} \log \frac{k}{\delta})$ holds. This is primarily because we believe that the error probability δ of any typical, label-invariant PAC algorithm ought to be distributed roughly uniformly across misidentification of all the items, allowing us to use δ/k instead of δ on the right hand side of the change-of-measure inequalities of Lemma 3, resulting in the improved quantity $\ln(k/2.4\delta)$. This is perhaps in line with recent work in multi-armed bandits (Simchowitz et al., 2017) that points to an increased difficulty of PAC identification in the moderate-confidence regime.*

We now consider a variant of the BB-PL decision model which allows the learner to play sets of any size $1, 2, \ldots, k$, instead of a fixed size k. In this setting, we are indeed able to design an (ϵ, δ)-PAC algorithm that enjoys an order-optimal $O(\frac{n}{\epsilon^2} \ln \frac{1}{\delta})$ sample-complexity.

4.3. BB-PL2: A slightly different battling bandit decision model

The new winner information feedback model BB-PL-2 is formally defined as follows: At each round t, here the learner is allowed to select a set $S_t \subseteq [n]$ of size $2, 3, \ldots,$ upto k. Upon receiving any set S_t, the environment returns the index of the winning item as $I \in [|S|]$ such that, $\mathbf{P}(I = i|S) = \frac{\theta_{S(i)}}{\sum_{j=1}^{|S|} \theta_{S(j)}} \;\; \forall i \in [|S|]$.

On applying existing PAC-Dueling-Bandit strategies. Note that given the flexibility of playing sets of any size, one might as well hope to apply the PAC-Dueling Bandit algorithm *PLPAC*(ϵ, δ)

of Szörényi et al. (2015) which plays only pairs of items per round. However, their algorithm is shown to have a sample complexity guarantee of $O\left(\frac{n}{\epsilon^2}\ln\frac{n}{\epsilon\delta}\right)$, which is suboptimal by an additive $O\left(\frac{n}{\epsilon^2}\ln\frac{n}{\epsilon}\right)$ as our results will show. A similar observation holds for the *Beat-the-Mean* (BTM) algorithm of Yue and Joachims (2011), which in fact has a even worse sample complexity guarantee of $O\left(\frac{n}{\epsilon^2}\ln\left(\frac{n}{\epsilon^2\delta}\ln\frac{n}{\delta}\right)\right)$.

Algorithm 3 *Halving-Battle*

Input:
  Set of items: $[n]$, Maximum subset size: $n \ge k > 1$
  Error bias: $\epsilon > 0$, Confidence parameter: $\delta > 0$
Initialize:
  $S \leftarrow [n]$, $\epsilon_0 \leftarrow \frac{\epsilon}{4}$, and $\delta_0 \leftarrow \delta$
  Divide S into $G := \lceil\frac{n}{k}\rceil$ sets $\mathcal{G}_1, \mathcal{G}_2, \cdots \mathcal{G}_G$ such that $\cup_{j=1}^{G}\mathcal{G}_j = S$ and $\mathcal{G}_j \cap \mathcal{G}_{j'} = \emptyset,\ \forall j, j' \in [G]$, where $|G_j| = k,\ \forall j \in [G-1]$
while $\ell = 1, 2, \ldots$ **do**
  $S \leftarrow \emptyset$, $\delta_\ell \leftarrow \frac{\delta_{\ell-1}}{2}$, $\epsilon_\ell \leftarrow \frac{3}{4}\epsilon_{\ell-1}$
  for $g = 1, 2, \cdots G$ **do**
    Play $\mathcal{G}_g$ for $t := \frac{k}{2\epsilon_\ell^2}\ln\frac{4}{\delta_\ell}$ rounds
    $w_i \leftarrow$ Number of times i won in t plays of $\mathcal{G}_g$, $\forall i \in \mathcal{G}_g$
    Set $h_g \leftarrow$ Median$(\{w_i \mid i \in \mathcal{G}_g\})$, and $S \leftarrow S \cup \{i \in \mathcal{G}_g \mid w_i \ge w_{h_g}\}$
  end for
  if $|S| == 1$ **then**
    Break (exit the while loop)
  else
    Divide S into $G := \lceil\frac{|S|}{k}\rceil$ sets $\mathcal{G}_1, \mathcal{G}_2, \cdots \mathcal{G}_G$ such that $\cup_{j=1}^{G}\mathcal{G}_j = S$ and $\mathcal{G}_j \cap \mathcal{G}_{j'} = \emptyset,\ \forall j, j' \in [G]$, where $|G_j| = k,\ \forall j \in [G-1]$
  end if
end while
Output: r_* as the ϵ-optimal item, where $S = \{r_*\}$

Algorithm 3: ***Halving-Battle*.** We here propose a Median-Elimination-based approach (Even-Dar et al., 2006) which is shown to run with *optimal sample complexity* $O(\frac{n}{\epsilon^2}\ln\frac{1}{\delta})$ rounds (Theorem 8). (Note that an $\Omega(\frac{n}{\epsilon^2}\ln\frac{1}{\delta})$ fundamental limit on PAC sample complexity for BB-PL2-WI can easily be derived using an argument along the lines of Theorem 2; we omit the explicit derivation.) The name *Halving-Battle* for the algorithm is because it is based on the idea of dividing the set of items into two partitions with respect to the *empirical median item* and retaining the 'better half'. Specifically, it first divides the entire item set into groups of size k, and plays each group for a fixed number of times. After this step, only the items that won more than the empirical median h_g are retained and rest are discarded. The algorithm recurses until it is left with a single item. The intuition here is that some *ϵ-best item* is always likely to beat the group median and can never get wiped off.

Theorem 8 **(*Halving-Battle*: Correctness and Sample Complexity with WI)** *Halving-Battle (Algorithm 3) is (ϵ, δ)-PAC with sample complexity $O\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$.*

Proof (sketch). The sample complexity bound follows by carefully summing the total number of times ($t = \frac{k}{2\epsilon_\ell^2}\ln\frac{1}{\delta_\ell}$) a set $\mathcal{G}_g$ is played per iteration ℓ, with the maximum number of possible iterations being $\lceil \ln n \rceil$ (this is because the size of the set S of remaining items gets halved at each iteration as it is pruned with respect to its median). The key intuition in proving the correctness property of *Halving-Battle* lies in showing that at any iteration ℓ, *Halving-Battle* always carries forward at least one 'near-best' item to the next iteration $\ell + 1$.

Lemma 9 *At any iteration ℓ, for any set $\mathcal{G}_g$, let $i_g \leftarrow \arg\max_{i \in \mathcal{G}_g} \theta_i$, and consider any suboptimal item $b \in \mathcal{G}_g$ such that $p_{bi_g} < \frac{1}{2} - \epsilon_\ell$. Then with probability at least $\big(1 - \frac{\delta_\ell}{4}\big)$, the empirical win count of i_g lies above that of b, i.e. $w_{i_g} \ge w_b$ (equivalently $\hat{p}_{i_g b} = \frac{w_{i_g}}{w_{i_g} + w_b} \ge \frac{1}{2}$).*

Using the property of the median element h_g along with Lemma 9 and Markov's inequality, we show that we do not lose out more than an additive factor of ϵ_ℓ in terms of highest score θ_i of the remaining set of items between any two successive iterations ℓ and $\ell + 1$. This finally leads to the desired (ϵ, δ)-PAC correctness of *Halving-Battle*. ■

Remark 5 *Theorem 8 shows that the sample complexity guarantee of Halving-Battle improves over the that of existing PLPAC algorithm for the same objective in dueling bandit setup ($k = 2$), which was shown to be $O\big(\frac{n}{\epsilon^2}\log\frac{n}{\epsilon\delta}\big)$ (see Theorem 3, Szörényi et al. (2015)), and also the $O\Big(\frac{n}{\epsilon^2}\ln\big(\frac{n}{\epsilon^2\delta}\ln\frac{n}{\delta}\big)\Big)$ complexity of BTM algorithm (Yue and Joachims, 2011) for dueling feedback from any pairwise preference matrix with relaxed stochastic transitivity and stochastic triangle inequality (of which PL model is a special case).*

5. Analysis with Top Ranking (TR) feedback

We now proceed to analyze the BB-PL problem with *Top-m Ranking* (TR) feedback (Section 3.1). We first show that unlike WI feedback, the sample complexity lower bound here scales as $\Omega\bigg(\frac{n}{m\epsilon^2}\ln\frac{1}{\delta}\bigg)$ (Theorem 10), which is a factor m smaller than that in Thm. 2 for the WI feedback model. At a high level, this is because TR reveals the preference information of m items per feedback step (round of battle), as opposed to just a single (noisy) information sample of the winning item (WI). Following this, we also present two algorithms for this setting which are shown to enjoy an optimal (upto logarithmic factors) sample complexity guarantee of $O\bigg(\frac{n}{m\epsilon^2}\ln\frac{k}{\delta}\bigg)$ (Section 5.2).

5.1. Lower Bound for Top-m Ranking (TR) feedback

Theorem 10 (Sample Complexity Lower Bound for TR) *Given $\epsilon \in (0, \frac{1}{\sqrt{8}}]$ and $\delta \in (0, 1]$, and an (ϵ, δ)-PAC algorithm A with top-m ranking (TR) feedback ($2 \le m \le k$), there exists a PL instance ν such that the expected sample complexity of A on ν is at least $\Omega\bigg(\frac{n}{m\epsilon^2}\ln\frac{1}{2.4\delta}\bigg)$.*

Remark 6 *The sample complexity lower for PAC-WI objective for BB-PL with top-m ranking (TR) feedback model is $\frac{1}{m}$-times that of the WI model (Thm. 2). Intuitively, revealing a ranking on m items in a k-set provides about $\ln\Big(\binom{k}{m}m!\Big) = O(m\ln k)$ bits of information per round, which is about m times as large as that of revealing a single winner, yielding an acceleration of m.*

Corollary 11 *Given $\epsilon \in (0, \frac{1}{\sqrt{8}}]$ and $\delta \in (0,1]$, and an (ϵ,δ)-PAC algorithm A with full ranking (FR) feedback ($m = k$), there exists a PL instance ν such that the expected sample complexity of A on ν is at least $\Omega\left(\frac{n}{k\epsilon^2}\ln\frac{1}{2.4\delta}\right)$.*

5.2. Algorithms for Top-m Ranking (TR) feedback model

This section presents two algorithms for (ϵ,δ)-PAC objective for *BB-PL* with top-m ranking feedback. We achieve this by generalizing our earlier two proposed algorithms (see Algorithm 1 and 2, Sec. 4.2 for WI feedback) to the top-m ranking (TR) feedback mechanism. [3]

Rank-Breaking. The main trick we use in modifying the above algorithms for TR feedback is *Rank Breaking* (Soufiani et al., 2014), which essentially extracts pairwise comparisons from multiwise (subsetwise) preference information. Formally, given any set S of size k, if $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_{S_m}$, ($S_m \subseteq S$, $|S_m| = m$) denotes a possible top-m ranking of S, the *Rank Breaking* subroutine considers each item in S to be beaten by its preceding items in $\boldsymbol{\sigma}$ in a pairwise sense. For instance, given a full ranking of a set of 4 elements $S = \{a,b,c,d\}$, say $b \succ a \succ c \succ d$, Rank-Breaking generates the set of 6 pairwise comparisons: $\{(b \succ a),(b \succ c),(b \succ d),(a \succ c),(a \succ d),(c \succ d)\}$. Similarly, given the ranking of only 2 most preferred items say $b \succ a$, it yields the 5 pairwise comparisons $(b, a \succ c),(b,a \succ d)$ and $(b \succ a)$ etc. See Algorithm 4 for detailed description of the Rank-Breaking procedure.

Lemma 12 (Rank-Breaking Update) *Consider any subset $S \subseteq [n]$ with $|S| = k$. Let S be played for t rounds of battle, and let $\boldsymbol{\sigma}_\tau \in \boldsymbol{\Sigma}_{S_m^\tau}$, ($S_m^\tau \subseteq S$, $|S_m^\tau| = m$), denote the TR feedback at each round $\tau \in [t]$. For each item $i \in S$, let $q_i := \sum_{\tau=1}^t \mathbf{1}(i \in S_m^\tau)$ be the number of times i appears in the top-m ranked output in t rounds. Then, the most frequent item(s) in the top-m positions must appear at least $\frac{mt}{k}$ times, i.e.* $\max_{i\in S} q_i \geq \frac{mt}{k}$.

Algorithm 4 *Rank-Breaking* (for updating the pairwise win counts w_{ij} for TR feedback)

1: **Input:** STATE Subset $S \subseteq [n]$, $|S| = k$ ($n \geq k$)
2: A top-m ranking $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_{S_m}$, $S_m \subseteq [n]$, $|S_m| = m$
3: Pairwise (empirical) win-count w_{ij} for each item pair $i,j \in S$
4: **while** $\ell = 1,2,\ldots m$ **do**
5: Update $w_{\sigma(\ell)i} \leftarrow w_{\sigma(\ell)i} + 1$, for all $i \in S \setminus \{\sigma(1),\ldots,\sigma(\ell)\}$
6: **end while**

Proposed Algorithms for TR feedback. The formal descriptions of our two algorithms, *Trace-the-Best* and *Divide-and-Battle*, generalized to the setting of TR feedback, are given as Algorithm 5 and Algorithm 6 respectively. They essentially maintain the empirical pairwise preferences $\hat{p}_{ij}$ for each pair of items i,j by applying *Rank Breaking* on the TR feedback $\boldsymbol{\sigma}$ after each round of battle. Of course in general, *Rank Breaking* may lead to arbitrarily inconsistent estimates of the underlying model parameters (Azari et al., 2012). However, owing to the *IIA property* of the Plackett-Luce model, we get clean concentration guarantees on p_{ij} using Lemma 1. This is precisely the idea

3. Our third algorithm *Halving-Battle* is not applicable to TR feedback as it allows the learner to play sets of sizes $1,2,3,\ldots$ upto k, whereas the TR feedback is defined only when the size of the subset played is at least m. The lower bound analysis of Theorem 10 also does not apply if sets of size less than m is allowed.

used for obtaining the $\frac{1}{m}$ factor improvement in the sample complexity guarantees of our proposed algorithms along with Lemma 12 (see proofs of Theorem 13 and 14).

Theorem 13 (*Trace-the-Best*: **Correctness and Sample Complexity with TR**) *With top-m ranking (TR) feedback model, Trace-the-Best (Algorithm 5) is (ϵ, δ)-PAC with sample complexity $O(\frac{n}{m\epsilon^2}\log\frac{n}{\delta})$.*

Theorem 14 (*Divide-and-Battle*: **Correctness and Sample Complexity with TR**) *With top-m ranking (TR) feedback, Divide-and-Battle (Algorithm 6) is (ϵ, δ)-PAC with sample complexity $O(\frac{n}{m\epsilon^2}\log\frac{k}{\delta})$.*

Remark 7 *The sample complexity bounds of the above two algorithms are $\frac{1}{m}$ fraction lesser than their corresponding counterparts for WI feedback, as follows comparing Theorem 4 vs. 13, or Theorem 6 vs. 14, which admit a faster learning rate with TR feedback. Similar to the case with WI feedback, sample complexity of Divide-and-Battle is still orderwise optimal for any $\delta \le \frac{1}{k}$, as follows from the lower bound guarantee (Theorem 10). However, we believe that the above lower bound can be tightened by a factor of $\ln k$ for 'moderate' $\delta \gtrapprox \frac{1}{k}$, for reasons similar to those stated in Remark 4.*

Algorithm 5 *Trace-the-Best* (for TR feedback)

Input:
  Set of items: $[n]$, and subset size: $k > 2$ $(n \ge k \ge m)$
  Error bias: $\epsilon > 0$, and confidence parameter: $\delta > 0$
Initialize:
  $r_1 \leftarrow$ Any (random) item from $[n]$, $\mathcal{A} \leftarrow$ Randomly select $(k-1)$ items from $[n] \setminus \{r_1\}$
  Set $\mathcal{A} \leftarrow \mathcal{A} \cup \{r_1\}$, and $S \leftarrow [n] \setminus \mathcal{A}$
while $\ell = 1, 2, \ldots$ **do**
  Initialize pairwise (empirical) win-count $w_{ij} \leftarrow 0$, for each item pair $i, j \in \mathcal{A}$
  for $\tau = 1, 2, \ldots t$ $(:= \frac{2k}{m\epsilon^2}\ln\frac{2n}{\delta})$ **do**
    Play the set $\mathcal{A}$ (one round of battle)
    Receive TR feedback: $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_{\mathcal{A}_m^\tau}$, where $\mathcal{A}_m^\tau \subseteq \mathcal{A}$ such that $|\mathcal{A}_m^\tau| = m$
    Update pairwise win-counts w_{ij} of each item pair $i, j \in \mathcal{A}$ using *Rank-Breaking*$(\mathcal{A}, \boldsymbol{\sigma})$
  end for
  $B_\ell \leftarrow \text{argmax}\{i \in \mathcal{A} \mid \sum_{j \in \mathcal{A}\setminus\{i'\}} \mathbf{1}\big(w_{ij} \ge w_{ji}\big)\}$,
  $\hat{p}_{ij} \leftarrow \frac{w_{ij}}{w_{ij}+w_{ji}}, \forall i, j \in \mathcal{A}, i \ne j$
  if $\exists c_\ell \in B_\ell$ such that $\hat{p}_{c_\ell, r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}$, **then** set $r_{\ell+1} \leftarrow c_\ell$; **else** set $r_{\ell+1} \leftarrow r_\ell$
  if $(S == \emptyset)$ **then**
    Break (go out of the while loop)
  else if $|S| < k-1$ **then**
    $\mathcal{A} \leftarrow$ Randomly select $(k-1-|S|)$ items from $\mathcal{A} \setminus \{r_\ell\}$
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{r_\ell\} \cup S$; and $S \leftarrow \emptyset$
  else
    $\mathcal{A} \leftarrow$ Randomly select $(k-1)$ items from S
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{r_\ell\}$ and $S \leftarrow S \setminus \mathcal{A}$
  end if
end while
Output: $r_* = r_\ell$ as the ϵ-optimal item

Algorithm 6 *Divide-and-Battle* (for TR feedback)

Input:
  Set of items: $[n]$, and subset size: $k > 2$ ($n \geq k \geq m$)
  Error bias: $\epsilon > 0$, and confidence parameter: $\delta > 0$
Initialize:
  $S \leftarrow [n]$, $\epsilon_0 \leftarrow \frac{\epsilon}{8}$, and $\delta_0 \leftarrow \frac{\delta}{2}$
  Divide S into $G := \lceil \frac{n}{k} \rceil$ sets $\mathcal{G}_1, \mathcal{G}_2, \cdots \mathcal{G}_G$ such that $\cup_{j=1}^{G} \mathcal{G}_j = S$ and $\mathcal{G}_j \cap \mathcal{G}_{j'} = \emptyset$, $\forall j, j' \in [G]$, $|G_j| = k$, $\forall j \in [G-1]$. **If** $|\mathcal{G}_G| < k$, **then** set $\mathcal{R}_1 \leftarrow \mathcal{G}_G$ and $G = G - 1$.
while $\ell = 1, 2, \ldots$ **do**
  Set $S \leftarrow \emptyset$, $\delta_\ell \leftarrow \frac{\delta_{\ell-1}}{2}$, $\epsilon_\ell \leftarrow \frac{3}{4}\epsilon_{\ell-1}$
  for $g = 1, 2, \cdots G$ **do**
    Initialize pairwise (empirical) win-count $w_{ij} \leftarrow 0$, for each item pair $i, j \in \mathcal{G}_g$
    for $\tau = 1, 2, \ldots t$ $(:= \frac{4k}{m\epsilon_\ell^2} \ln \frac{2k}{\delta_\ell})$ **do**
      Play the set $\mathcal{G}_g$ (one round of battle)
      Receive feedback: The top-m ranking $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_{\mathcal{G}_{gm}^\tau}$, where $\mathcal{G}_{gm}^\tau \subseteq \mathcal{G}_g$, $|\mathcal{G}_{gm}^\tau| = m$
      Update win-count w_{ij} of each item pair $i, j \in \mathcal{G}_g$ using *Rank-Breaking*$(\mathcal{G}_g, \boldsymbol{\sigma})$
    end for
    Define $\hat{p}_{i,j} = \frac{w_{ij}}{w_{ij} + w_{ji}}$, $\forall i, j \in \mathcal{G}_g$
    If $\exists i \in \mathcal{G}_g$ such that $\hat{p}_{ij} + \frac{\epsilon_\ell}{2} \geq \frac{1}{2}$, $\forall j \in \mathcal{G}_g$, then set $c_g \leftarrow i$, else select $c_g \leftarrow$ uniformly at random from $\mathcal{G}_g$, and set $S \leftarrow S \cup \{c_g\}$
  end for
  $S \leftarrow S \cup \mathcal{R}_\ell$
  if $(|S| == 1)$ **then**
    Break (go out of the while loop)
  else if $|S| \leq k$ **then**
    $S' \leftarrow$ Randomly sample $k - |S|$ items from $[n] \setminus S$, and $S \leftarrow S \cup S'$, $\epsilon_\ell \leftarrow \frac{2\epsilon}{3}$, $\delta_\ell \leftarrow \delta$
  else
    Divide S into $G := \lceil \frac{|S|}{k} \rceil$ sets $\mathcal{G}_1, \cdots \mathcal{G}_G$ such that $\cup_{j=1}^{G} \mathcal{G}_j = S$, $\mathcal{G}_j \cap \mathcal{G}_{j'} = \emptyset$, $\forall j, j' \in [G]$, $|G_j| = k$, $\forall j \in [G-1]$. **If** $|\mathcal{G}_G| < k$, **then** set $\mathcal{R}_{\ell+1} \leftarrow \mathcal{G}_G$ and $G = G - 1$.
  end if
end while
Output: r_* as the ϵ-optimal item, where $S = \{r_*\}$ (i.e. r_* is the only item remaining in S)

6. Conclusion and Future Directions

We have developed foundations for probably approximately correct (PAC) online learning with subset choices: introducing the problem of *Battling-Bandits (BB)* with subset choice models – a novel generalization of the well-studied Dueling-Bandit problem, where the objective is to find the 'best item' by successively choosing subsets of k alternatives from n items, and subsequently receiving a set-wise feedback information in an online fashion. We have specifically studied the Plackett-Luce (PL) choice model along with winner information (WI) and top ranking (TR) feedback, with the goal of finding an (ϵ, δ)-*PAC item*: an ϵ-approximation of the best item with probability at least $(1 - \delta)$. Our results show that with just the WI feedback, playing a battling game is just as good as that of a dueling game ($k = 2$), as in this case the required sample complexity of the PAC learning problem is independent of the subset set k. However with TR feedback, the battling framework provides a $\frac{1}{m}$-times faster learning rate, leading to an improved performance guarantee owing to the information gain with top-m ranking feedback, as intuitively well justified as well.

Future Directions. Our proposed framework of *Battling Bandits* opens up a set of new directions to pursue - with different feedback mechanisms, choice models (e.g. Multinomial Probit, Mallows, nested logit, generalized extreme-value models etc.), other learning objectives, etc. It is an interesting open problem to analyse the trade-off between the subset size k and the learning rate for other choice models with different feedback mechanisms. Another relevant direction to pursue within battling bandits could be to extend it to more general settings such as revenue maximization (Agrawal et al., 2016), learning with cost budgets (Xia et al., 2016; Zhou and Tomlin, 2017), feature-based preference information and adversarial choice feedback (Gajane et al., 2015).

Acknowledgments

The authors are grateful to the anonymous reviewers for valuable feedback. This work is supported by a Qualcomm Innovation Fellowship 2018, and an Indigenous 5G Test Bed project grant from the Dept. of Telecommunications, Government of India.

References

Shipra Agrawal, Vashist Avandhanula, Vineet Goyal, and Assaf Zeevi. A near-optimal exploration-exploitation approach for assortment selection. 2016.

Nir Ailon, Zohar Shay Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *ICML*, volume 32, pages 856–864, 2014.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

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

Hossein Azari, David Parkes, and Lirong Xia. Random utility theory for social choice. In *Advances in Neural Information Processing Systems*, pages 126–134, 2012.

Austin R Benson, Ravi Kumar, and Andrew Tomkins. On the relevance of irrelevant alternatives. In *Proceedings of the 25th International Conference on World Wide Web*, pages 963–973. International World Wide Web Conferences Steering Committee, 2016.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Róbert Busa-Fekete, Balazs Szorenyi, Weiwei Cheng, Paul Weng, and Eyke Hüllermeier. Top-k selection based on adaptive sampling of noisy preferences. In *International Conference on Machine Learning*, pages 1094–1102, 2013.

Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of mallows. In *Proceedings of The 31st International Conference on Machine Learning*, volume 32, 2014a.

Róbert Busa-Fekete, Balázs Szörényi, and Eyke Hüllermeier. Pac rank elicitation through adaptive sampling of stochastic pairwise preferences. In *AAAI*, pages 1701–1707, 2014b.

Xi Chen, Sivakanth Gopi, Jieming Mao, and Jon Schneider. Competitive analysis of the top-k ranking problem. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1245–1264. SIAM, 2017.

Xi Chen, Yuanzhi Li, and Jieming Mao. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2504–2522. SIAM, 2018.

Yuxin Chen and Changho Suh. Spectral mle: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pages 371–380, 2015.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. In *Advances in Neural Information Processing Systems*, pages 7063–7073, 2017.

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

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sbastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvri, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 423–439. PMLR, 2014.

Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Optimal sample complexity of m-wise data for top-k ranking. In *Advances in Neural Information Processing Systems*, pages 1685–1695, 2017.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016.

John I. Marden. *Analyzing and Modeling Rank Data*. Chapman and Hall/CRC, 1996.

Soheil Mohajer, Changho Suh, and Adel Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *International Conference on Machine Learning*, pages 2488–2497, 2017.

Pantelimon G Popescu, Silvestru Dragomir, Emil I Slusanschi, and Octavian N Stanasila. Bounds for Kullback-Leibler divergence. *Electronic Journal of Differential Equations*, 2016, 2016.

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

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

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Proceedings of The 30th Conference on Learning Theory*, 2017.

Hossein Azari Soufiani, David C Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *ICML*, pages 360–368, 2014.

Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for plackett-luce: A dueling bandits approach. In *Advances in Neural Information Processing Systems*, pages 604–612, 2015.

Tanguy Urvoy, Fabrice Clerot, Raphael Féraud, and Sami Naamane. Generic exploration and k-armed voting bandits. In *International Conference on Machine Learning*, pages 91–99, 2013.

Ondrej Vojacek, Iva Pecakova, et al. Comparison of discrete choice models for economic environmental research. *Prague Economic Papers*, 19(1):35–53, 2010.

Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits with multiple plays. In *IJCAI*, pages 2210–2216, 2016.

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

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Datong P Zhou and Claire J Tomlin. Budget-constrained multi-armed bandits with multiple plays. *arXiv preprint arXiv:1711.05928*, 2017.

Masrour Zoghi, Shimon Whiteson, Remi Munos, Maarten de 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, 2014.

Supplementary for PAC Battling Bandits in the Plackett-Luce Model

Appendix A. Appendix for Section 2.2

A.1. Proof of Lemma 1

Lemma 1 (Deviations of pairwise win-probability estimates for PL model) *Consider a Plackett-Luce choice model with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ (see Eqn. (1)), and fix two distinct items $i, j \in [n]$. Let $S_1, \ldots, S_T$ be a sequence of (possibly random) subsets of $[n]$ of size at least 2, where T is a positive integer, and $i_1, \ldots, i_T$ a sequence of random items with each $i_t \in S_t$, $1 \le t \le T$, such that for each $1 \le t \le T$, (a) S_t depends only on $S_1, \ldots, S_{t-1}$, and (b) i_t is distributed as the Plackett-Luce winner of the subset S_t, given $S_1, i_1, \ldots, S_{t-1}, i_{t-1}$ and S_t, and (c) $\forall t : \{i,j\} \subseteq S_t$ with probability 1. Let $n_i(T) = \sum_{t=1}^{T} \mathbf{1}(i_t = i)$ and $n_{ij}(T) = \sum_{t=1}^{T} \mathbf{1}(\{i_t \in \{i,j\}\})$. Then, for any positive integer v, and $\eta \in (0,1)$,*

$$Pr\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \ge \eta, n_{ij}(T) \ge v\right) \vee Pr\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \le -\eta, n_{ij}(T) \ge v\right) \le e^{-2v\eta^2}.$$

Proof We prove the lemma by using a coupling argument. Consider the following 'simulator' or probability space for the Plackett-Luce choice model that specifically depends on the item pair i, j, constructed as follows. Let $Z_1, Z_2, \ldots$ be a sequence of iid Bernoulli random variables with success parameter $\theta_i/(\theta_i + \theta_j)$. A counter is first initialized to 0. At each time t, given $S_1, i_1, \ldots, S_{t-1}, i_{t-1}$ and S_t, an independent coin is tossed with probability of heads $(\theta_i + \theta_j)/\sum_{k \in S_t} \theta_k$. If the coin lands tails, then i_t is drawn as an independent sample from the Plackett-Luce distribution over $S_t \setminus \{i,j\}$, else, the counter is incremented by 1, and i_t is returned as i if $Z_C = 1$ or j if $Z_C = 0$ where C is the present value of the counter.

It may be checked that the construction above indeed yields the correct joint distribution for the sequence $i_1, S_1, \ldots, i_T, S_T$ as desired, due to the independence of irrelevant alternatives (IIA) property of the Plackett-Luce choice model:

$$Pr(i_t = i | i_t \in \{i,j\}, S_t) = \frac{Pr(i_t = i | S_t)}{Pr(i_t \in \{i,j\} | S_t)} = \frac{\theta_i / \sum_{k \in S_t} \theta_k}{(\theta_i + \theta_j)/\sum_{k \in S_t} \theta_k} = \frac{\theta_i}{\theta_i + \theta_j}.$$

Furthermore, $i_t \in \{i,j\}$ if and only if C is incremented at round t, and $i_t = i$ if and only if C is incremented at round t and $Z_C = 1$. We thus have

$$\begin{aligned}
Pr\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \ge \eta,\ n_{ij}(T) \ge v\right) &= Pr\left(\frac{\sum_{\ell=1}^{n_{ij}(T)} Z_\ell}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \ge \eta,\ n_{ij}(T) \ge v\right) \\
&= \sum_{m=v}^{T} Pr\left(\frac{\sum_{\ell=1}^{n_{ij}(T)} Z_\ell}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \ge \eta,\ n_{ij}(T) = m\right) \\
&= \sum_{m=v}^{T} Pr\left(\frac{\sum_{\ell=1}^{m} Z_\ell}{m} - \frac{\theta_i}{\theta_i + \theta_j} \ge \eta,\ n_{ij}(T) = m\right) \\
&\overset{(a)}{=} \sum_{m=v}^{T} Pr\left(\frac{\sum_{\ell=1}^{m} Z_\ell}{m} - \frac{\theta_i}{\theta_i + \theta_j} \ge \eta\right) Pr\left(n_{ij}(T) = m\right)
\end{aligned}$$

$$\overset{(b)}{\leq} \sum_{m=v}^{T} Pr\left(n_{ij}(T) = m\right) e^{-2m\eta^2} \leq e^{-2v\eta^2},$$

where (a) uses the fact that $S_1, \ldots, S_T, X_1, \ldots, X_T$ are independent of $Z_1, Z_2, \ldots,$, and so $n_{ij}(T) \in \sigma(S_1, \ldots, S_T, X_1, \ldots, X_T)$ is independent of $Z_1, \ldots, Z_m$ for any fixed m, and (b) uses Hoeffding's concentration inequality for the iid sequence Z_i.

Similarly, one can also derive

$$Pr\left(\frac{n_i(T)}{n_{ij}(T)} - \frac{\theta_i}{\theta_i + \theta_j} \leq -\eta,\ n_{ij}(T) \geq v\right) \leq e^{-2v\eta^2},$$

which concludes the proof. ■

Appendix B. Appendix for Section 4

B.1. Proof of Theorem 2

Theorem 2 (Lower bound on Sample Complexity with WI feedback) *Given $\epsilon \in (0, \frac{1}{\sqrt{8}}]$ and $\delta \in (0, 1]$, and an (ϵ, δ)-PAC algorithm A for BB-PL with feedback model WI, there exists a PL instance ν such that the sample complexity of A on ν is at least $\Omega\left(\frac{n}{\epsilon^2} \ln \frac{1}{2.4\delta}\right)$.*

Proof We will apply Lemma 3 to derive the desired lower bounds of Theorem 2 for BB-PL with WI feedback model.

Let us consider a bandit instance with the arm set containing all subsets of size k: $\mathcal{A} = \{S = (S(1), \ldots, S(k)) \subseteq [n] \mid S(i) < S(j), \forall i < j\}$. Let $\boldsymbol{\nu}^1$ be the true distribution associated with the bandit arms, given by the Plackett-Luce parameters:

$$\textbf{True Instance } (\boldsymbol{\nu}^1) : \theta_j^1 = \theta\left(\frac{1}{2} - \epsilon\right), \forall j \in [n] \setminus \{1\}, \text{ and } \theta_1^1 = \theta\left(\frac{1}{2} + \epsilon\right),$$

for some $\theta \in \mathbb{R}_+,\ \epsilon > 0$. Now for every suboptimal item $a \in [n] \setminus \{1\}$, consider the modified instances $\boldsymbol{\nu}^a$ such that:

$$\textbf{Instance–a } (\boldsymbol{\nu}^a) : \theta_j^a = \theta\left(\frac{1}{2} - \epsilon\right)^2, \forall j \in [n] \setminus \{a, 1\},\ \theta_1^a = \theta\left(\frac{1}{4} - \epsilon^2\right), \text{ and } \theta_a^a = \theta\left(\frac{1}{2} + \epsilon\right)^2.$$

For problem instance $\boldsymbol{\nu}^a$, $a \in [n] \setminus \{1\}$, the probability distribution associated with arm $S \in \mathcal{A}$ is given by

$$\nu_S^a \sim Categorical(p_1, p_2, \ldots, p_k), \text{ where } p_i = Pr(i|S),\ \ \forall i \in [k], \forall S \in \mathcal{A},$$

where $Pr(i|S)$ is as defined in Section 3.1. Note that the only ϵ-optimal arm for **Instance-a** is arm a. Now applying Lemma 3, for some event $\mathcal{E} \in \mathcal{F}_\tau$ we get,

$$\sum_{\{S \in \mathcal{A} : a \in S\}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)] KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) \geq kl(Pr_\nu(\mathcal{E}), Pr_{\nu'}(\mathcal{E})). \tag{2}$$

The above result holds from the straightforward observation that for any arm $S \in \mathcal{A}$ with $a \notin S$, $\boldsymbol{\nu}_S^1$ is same as $\boldsymbol{\nu}_S^a$, hence $KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) = 0, \forall S \in \mathcal{A},\, a \notin S$. For notational convenience, we will henceforth denote $S^a = \{S \in \mathcal{A} : a \in S\}$.

Now let us analyse the right hand side of (2), for any set $S \in S^a$. We further denote $r = \mathbf{1}(1 \in S)$, $q = (k-1-r)$, and $R = \frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}$. Note that

$$\nu_S^1(i) = \begin{cases} \frac{\theta(\frac{1}{2}+\epsilon)}{r\theta(\frac{1}{2}+\epsilon)+(k-r)\theta(\frac{1}{2}-\epsilon)} = \frac{R}{rR+(k-r)}, \forall i \in [k], \text{ such that } S(i) = 1, \\ \frac{\theta(\frac{1}{2}-\epsilon)}{r\theta(\frac{1}{2}+\epsilon)+(k-r)\theta(\frac{1}{2}-\epsilon)} = \frac{1}{rR+(k-r)}, \text{ otherwise.} \end{cases}$$

On the other hand, for problem **Instance-a**, we have that:

$$\nu_S^a(i) = \begin{cases} \frac{R}{rR+R^2+q}, \forall i \in [k], \text{ such that } S(i) = 1, \\ \frac{R^2}{rR+R^2+q}, \forall i \in [k], \text{ such that } S(i) = a, \\ \frac{1}{rR+R^2+q}, \text{ otherwise.} \end{cases}$$

Now using the following upper bound on $KL(\mathbf{p}_1, \mathbf{p}_2) \le \sum_{x \in \mathcal{X}} \frac{p_1^2(x)}{p_2(x)} - 1$, $\mathbf{p}_1$ and $\mathbf{p}_2$ be two probability mass functions on the discrete random variable $\mathcal{X}$ (Popescu et al., 2016) we get:

$$KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) \le \frac{rR+R^2+q}{(rR+k-r)^2}(rR + \frac{1}{R} + q) - 1.$$

Replacing q by $(k-1-r)$ and re-arranging terms, we get

$$\begin{aligned} KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) &\le \frac{(rR+(k-r)+(R^2-1))(rR+(k-r)+(R^{-2}-1))}{(rR+k-r)^2} - 1 \\ &= \frac{(rR+k-r-1)}{(rR+k-r)^2}\Big(R - \frac{1}{R}\Big)^2 \le \frac{1}{k}\Big(R - \frac{1}{R}\Big)^2 \;\; [\text{since } s \ge 0, \text{ and } R > 1]. \end{aligned} \tag{3}$$

Note that the only ϵ-optimal arm for any **Instance-a** is arm a, for all $a \in [n]$. Now, consider $\mathcal{E}_0 \in \mathcal{F}_\tau$ be an event such that the algorithm A returns the element $i = 1$, and let us analyse the left hand side of (2) for $\mathcal{E} = \mathcal{E}_0$. Clearly, A being an (ϵ, δ)-PAC algorithm, we have $Pr_{\boldsymbol{\nu}^1}(\mathcal{E}_0) > 1-\delta$, and $Pr_{\boldsymbol{\nu}^a}(\mathcal{E}_0) < \delta$, for any suboptimal arm $a \in [n] \setminus \{1\}$. Then we have

$$kl(Pr_{\boldsymbol{\nu}^1}(\mathcal{E}_0), Pr_{\boldsymbol{\nu}^a}(\mathcal{E}_0)) \ge kl(1-\delta, \delta) \ge \ln \frac{1}{2.4\delta} \tag{4}$$

where the last inequality follows from Kaufmann et al. (2016, Equation (3)).

Now applying (2) for each modified bandit **Instance-$\boldsymbol{\nu}^a$**, and summing over all suboptimal items $a \in [n] \setminus \{1\}$ we get,

$$\sum_{a=2}^{n} \sum_{\{S \in \mathcal{A} | a \in S\}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)] KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) \ge (n-1)\ln \frac{1}{2.4\delta}. \tag{5}$$

Moreover, using (3), the term of the right hand side of (5) can be further upper bounded as

$$\sum_{a=2}^{n}\sum_{\{S\in\mathcal{A}|a\in S\}}\mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)]KL(\boldsymbol{\nu}_S^1,\boldsymbol{\nu}_S^a) \le \sum_{S\in\mathcal{A}}\mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)]\sum_{\{a\in S|a\neq 1\}}\frac{1}{k}\bigg(R-\frac{1}{R}\bigg)^2$$
$$=\sum_{S\in\mathcal{A}}\mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)]\frac{k-\big(\mathbf{1}(1\in S)\big)}{k}\bigg(R-\frac{1}{R}\bigg)^2$$
$$\le \sum_{S\in\mathcal{A}}\mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)](256\epsilon^2)\ \bigg[\text{since } \bigg(R-\frac{1}{R}\bigg)=\frac{8\epsilon}{(1-4\epsilon^2)}\le 16\epsilon, \forall \epsilon\in[0,\frac{1}{\sqrt{8}}]\bigg]. \quad (6)$$

Finally noting that $\tau_A=\sum_{S\in\mathcal{A}}[N_S(\tau_A)]$, combining (6) and (5), we get

$$(256\epsilon^2)\mathbf{E}_{\boldsymbol{\nu}^1}[\tau_A]=\sum_{S\in\mathcal{A}}\mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)](256\epsilon^2)\ge(n-1)\ln\frac{1}{2.4\delta}.$$

Thus above construction shows the existence of a problem instance $\boldsymbol{\nu}=\boldsymbol{\nu}^1$, such that $\mathbf{E}_{\boldsymbol{\nu}^1}[\tau_A]=\Omega(\frac{n}{\epsilon^2}\ln\frac{1}{2.4\delta})$, which concludes the proof.

■

B.2. Proof of Theorem 4

Theorem 4 (*Trace-the-Best*: Correctness and Sample Complexity with WI) *Trace-the-Best (Algorithm 1) is (ϵ,δ)-PAC with sample complexity $O(\frac{n}{\epsilon^2}\log\frac{n}{\delta})$.*

Proof We start by analyzing the required sample complexity first. Note that the 'while loop' of Algorithm 1 always discards away $k-1$ items per iteration. Thus, n being the total number of items the loop can be executed is at most for $\lceil\frac{n}{k-1}\rceil$ many number of iterations. Clearly, the sample complexity of each iteration being $t=\frac{2k}{\epsilon^2}\ln\frac{n}{2\delta}$, the total sample complexity of the algorithm thus becomes $\big(\lceil\frac{n}{k-1}\rceil\big)\frac{2k}{\epsilon^2}\ln\frac{n}{2\delta}\le\big(\frac{n}{k-1}+1\big)\frac{2k}{\epsilon^2}\ln\frac{n}{2\delta}=\big(n+\frac{n}{k-1}+k\big)\frac{2}{\epsilon^2}\ln\frac{n}{2\delta}=O(\frac{n}{\epsilon^2}\ln\frac{n}{\delta})$.

We now prove the (ϵ,δ)-PAC correctness of the algorithm. As argued before, the 'while loop' of Algorithm 1 can run for maximum $\lceil\frac{n}{k-1}\rceil$ many number of iterations. We denote the iterations by $\ell=1,2,\ldots,\lceil\frac{n}{k-1}\rceil$, and the corresponding set $\mathcal{A}$ of iteration ℓ by $\mathcal{A}_\ell$.

Note that our idea is to retain the estimated best item in 'running winner' r_ℓ and compare it with the 'empirical best item' c_ℓ of $\mathcal{A}_\ell$ at every iteration ℓ. The crucial observation lies in noting that at any iteration ℓ, r_ℓ gets updated as follows:

Lemma 5 *At any iteration $\ell=1,2\ldots\lfloor\frac{n}{k-1}\rfloor$, with probability at least $(1-\frac{\delta}{2n})$, Algorithm 1 retains $r_{\ell+1}\leftarrow r_\ell$ if $p_{c_\ell r_\ell}\le\frac{1}{2}$, and sets $r_{\ell+1}\leftarrow c_\ell$ if $p_{c_\ell r_\ell}\ge\frac{1}{2}+\epsilon$.*

Proof Consider any set $\mathcal{A}_\ell$, by which we mean the state of $\mathcal{A}$ in the algorithm at iteration ℓ. The crucial observation to make is that since c_ℓ is the empirical winner of t rounds of battle, then $w_{c_\ell}\ge\frac{t}{k}$. Thus $w_{c_\ell}+w_{r_\ell}\ge\frac{t}{k}$. Let $n_{ij}:=w_i+w_j$ denotes the total number of pairwise comparisons between item i and j in t rounds, for any $i, j \in \mathcal{A}_\ell$. Then clearly, $0 \le n_{ij} \le t$ and $n_{ij} = n_{ji}$. Specifically we have $\hat{p}_{r_\ell c_\ell} = \frac{w_{r_\ell}}{w_{r_\ell} + w_{c_\ell}} = \frac{w_{r_\ell}}{n_{r_\ell c_\ell}}$. We prove the claim by analyzing the following cases:

Case 1. (If $p_{c_\ell r_\ell} \le \frac{1}{2}$, *Trace-the-Best* retains $r_{\ell+1} \leftarrow r_\ell$): Note that *Trace-the-Best* replaces $r_{\ell+1}$ by c_ℓ only if $\hat{p}_{c_\ell, r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}$, but this happens with probability:

$$
\begin{aligned}
&Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}\bigg\}\Bigg)\\
&= Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{t}{k}\bigg\}\Bigg) + \cancelto{0}{Pr\bigg\{n_{c_\ell r_\ell} < \frac{t}{k}\bigg\}} Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}\bigg\} \bigg| \bigg\{n_{c_\ell r_\ell} < \frac{t}{k}\bigg\}\Bigg)\\
&\le Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} - p_{c_\ell r_\ell} > \frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{t}{k}\bigg\}\Bigg) \le \exp\Big(-2\frac{t}{k}\Big(\frac{\epsilon}{2}\Big)^2\Big) = \frac{\delta}{2n},
\end{aligned}
$$

where the first inequality follows as $p_{c_\ell r_\ell} \le \frac{1}{2}$, and the second inequality is by applying Lemma 1 with $\eta = \frac{\epsilon}{2}$ and $v = \frac{t}{k}$. We now proceed to the second case:

Case 2. (If $p_{c_\ell r_\ell} \ge \frac{1}{2} + \epsilon$, *Trace-the-Best* sets $r_{\ell+1} \leftarrow c_\ell$): Recall again that *Trace-the-Best* retains $r_{\ell+1} \leftarrow r_\ell$ only if $\hat{p}_{c_\ell, r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}$. This happens with probability:

$$
\begin{aligned}
&Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}\bigg\}\Bigg)\\
&= Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{t}{k}\bigg\}\Bigg) + \cancelto{0}{Pr\bigg\{n_{c_\ell r_\ell} < \frac{t}{k}\bigg\}} Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}\bigg\} \bigg| \bigg\{n_{c_\ell r_\ell} < \frac{t}{k}\bigg\}\Bigg)\\
&= Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \epsilon - \frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{t}{k}\bigg\}\Bigg)\\
&\le Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} - p_{c_\ell r_\ell} \le -\frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{t}{k}\bigg\}\Bigg) \le \exp\Big(-2\frac{t}{k}\Big(\frac{\epsilon}{2}\Big)^2\Big) = \frac{\delta}{2n},
\end{aligned}
$$

where the first inequality holds as $p_{c_\ell r_\ell} \ge \frac{1}{2} + \epsilon$, and the second one by applying Lemma 1 with $\eta = \frac{\epsilon}{2}$ and $v = \frac{t}{k}$. The proof follows combining the above two cases. ■

Given Algorithm 1 satisfies Lemma 5, and taking union bound over $(k-1)$ elements in $\mathcal{A}_\ell \setminus \{r_\ell\}$, we get that with probability at least $(1 - \frac{(k-1)\delta}{2n})$,

$$
p_{r_{\ell+1} r_\ell} \ge \frac{1}{2} \text{ and, } p_{r_{\ell+1} c_\ell} \ge \frac{1}{2} - \epsilon. \tag{7}
$$

Above suggests that for each iteration ℓ, the estimated 'best' item r_ℓ only gets improved as $p_{r_{\ell+1} r_\ell} \ge \frac{1}{2}$. Let, ℓ_* denotes the specific iteration such that $1 \in \mathcal{A}_\ell$ for the first time, i.e. $\ell_* = \min\{\ell \mid 1 \in \mathcal{A}_\ell\}$. Clearly $\ell_* \le \lceil \frac{n}{k-1} \rceil$. Now (7) suggests that with probability at least $(1 - \frac{(k-1)\delta}{2n})$, $p_{r_{\ell_*+1} 1} \ge \frac{1}{2} - \epsilon$. Moreover (7) also suggests that for all $\ell > \ell_*$, with probability at least $(1 - \frac{(k-1)\delta}{2n})$, $p_{r_{\ell+1} r_\ell} \ge \frac{1}{2}$, which implies for all $\ell > \ell_*$, $p_{r_{\ell+1} 1} \ge \frac{1}{2} - \epsilon$ as well – This holds due to the following transitivity property of the Plackett-Luce model: For any three items $i_1, i_2, i_3 \in [n]$, if $p_{i_1 i_2} \geq \frac{1}{2}$ and $p_{i_2 i_3} \geq \frac{1}{2}$, then we have $p_{i_1 i_3} \geq \frac{1}{2}$ as well.

This argument finally leads to $p_{r_* 1} \geq \frac{1}{2} - \epsilon$. Since failure probability at each iteration ℓ is at most $\frac{(k-1)\delta}{2n}$, and Algorithm 1 runs for maximum $\lceil \frac{n}{k-1} \rceil$ many number of iterations, using union bound over ℓ, the total failure probability of the algorithm is at most $\lceil \frac{n}{k-1} \rceil \frac{(k-1)\delta}{2n} \leq (\frac{n}{k-1} + 1)\frac{(k-1)\delta}{2n} = \delta\Big(\frac{n+k-1}{2n}\Big) \leq \delta$ (since $k \leq n$). This concludes the correctness of the algorithm showing that it indeed satisfies the (ϵ, δ)-PAC objective. ■

Algorithm 2 *Divide-and-Battle*

Input:
  Set of items: $[n]$, Subset size: $n \geq k > 1$
  Error bias: $\epsilon > 0$, Confidence parameter: $\delta > 0$
Initialize:
  $S \leftarrow [n]$, $\epsilon_0 \leftarrow \frac{\epsilon}{8}$, and $\delta_0 \leftarrow \frac{\delta}{2}$
  Divide S into $G := \lceil \frac{n}{k} \rceil$ sets $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G$ such that $\cup_{j=1}^{G}\mathcal{G}_j = S$ and $\mathcal{G}_j \cap \mathcal{G}_{j'} = \emptyset,\ \forall j, j' \in [G]$, where $|G_j| = k,\ \forall j \in [G-1]$
  If $|\mathcal{G}_G| < k$, **then** set $\mathcal{R}_1 \leftarrow \mathcal{G}_G$ and $G = G - 1$
while $\ell = 1, 2, \ldots$ **do**
  Set $S \leftarrow \emptyset$, $\delta_\ell \leftarrow \frac{\delta_{\ell-1}}{2}$, $\epsilon_\ell \leftarrow \frac{3}{4}\epsilon_{\ell-1}$
  for $g = 1, 2, \ldots, G$ **do**
    Play the set $\mathcal{G}_g$ for $t := \frac{k}{2\epsilon_\ell^2} \ln \frac{k}{\delta_\ell}$ rounds
    $w_i \leftarrow$ Number of times i won in t plays of $\mathcal{G}_g$, $\forall i \in \mathcal{G}_g$
    Set $c_g \leftarrow \arg\max_{i \in \mathcal{A}} w_i$ and $S \leftarrow S \cup \{c_g\}$
  end for
  $S \leftarrow S \cup \mathcal{R}_\ell$
  if $(|S| == 1)$ **then**
    Break (go out of the while loop)
  else if $|S| \leq k$ **then**
    $S' \leftarrow$ Randomly sample $k - |S|$ items from $[n] \setminus S$, and $S \leftarrow S \cup S'$, $\epsilon_\ell \leftarrow \frac{2\epsilon}{3}$, $\delta_\ell \leftarrow \delta$
  else
    Divide S into $G := \lceil \frac{|S|}{k} \rceil$ sets $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G$, such that $\cup_{j=1}^{G}\mathcal{G}_j = S$, and $\mathcal{G}_j \cap \mathcal{G}_{j'} = \emptyset,\ \forall j, j' \in [G]$, where $|G_j| = k,\ \forall j \in [G-1]$
    If $|\mathcal{G}_G| < k$, **then** set $\mathcal{R}_{\ell+1} \leftarrow \mathcal{G}_G$ and $G = G - 1$
  end if
end while
Output: r_* as the ϵ-optimal item, where $S = \{r^*\}$

B.3. Proof of Theorem 6

Theorem 6 (*Divide-and-Battle*: **Correctness and Sample Complexity with WI)** *Divide-and-Battle (Algorithm 2) is (ϵ, δ)-PAC with sample complexity $O(\frac{n}{\epsilon^2} \log \frac{k}{\delta})$.*

Proof For the notational convenience we will use $\tilde{p}_{ij} = p_{ij} - \frac{1}{2},\ \forall i, j \in [n]$. We start by proving the following lemma which would be used crucially in the analysis:

Lemma 15 *For any three items $a, b, c \in [n]$ such that $\theta_a > \theta_b > \theta_c$. If $\tilde{p}_{ba} > -\epsilon_1$, and $\tilde{p}_{cb} > -\epsilon_2$, where $\epsilon_1, \epsilon_2 > 0$, and $(\epsilon_1 + \epsilon_2) < \frac{1}{2}$, then $\tilde{p}_{ca} > -(\epsilon_1 + \epsilon_2)$.*

Proof Note that $\tilde{p}_{ba} > -\epsilon_1 \implies \frac{\theta_b - \theta_a}{2(\theta_b + \theta_a)} > -\epsilon_1 \implies \frac{\theta_b}{\theta_a} > \frac{(1-2\epsilon_1)}{(1+2\epsilon_1)}$.

Similarly we have $\tilde{p}_{cb} > -\epsilon_2 \implies \frac{\theta_c}{\theta_b} > \frac{(1-2\epsilon_2)}{(1+2\epsilon_2)}$. Combining above we get

$$\frac{\theta_c}{\theta_a} > \frac{(1-2\epsilon_1)}{(1+2\epsilon_1)}\frac{(1-2\epsilon_2)}{(1+2\epsilon_2)} > \frac{1-2(\epsilon_1+\epsilon_2)+\epsilon_1\epsilon_2}{1+2(\epsilon_1+\epsilon_2)+\epsilon_1\epsilon_2} > \frac{1-2(\epsilon_1+\epsilon_2)}{1+2(\epsilon_1+\epsilon_2)}, \; \left[\text{since,}(\epsilon_1+\epsilon_2) < \frac{1}{2}\right]$$
$$\implies \tilde{p}_{ca} = \frac{\theta_c - \theta_a}{2(\theta_c + \theta_a)} > -(\epsilon_1 + \epsilon_2),$$

which concludes the proof. ■

We now analyze the required sample complexity of *Divide-and-Battle*. For clarity of notations we will denote the set S at iteration ℓ by S_ℓ. Note that at any iteration ℓ, any set $\mathcal{G}_g$ is played for exactly $t = \frac{k}{2\epsilon_\ell^2}\ln\frac{k}{\delta_\ell}$ many number of rounds. Also since the algorithm discards away exactly $k-1$ items from each set $\mathcal{G}_g$, hence the maximum number of iterations possible is $\lceil \ln_k n \rceil$. Now at any iteration ℓ, since $G = \left\lfloor \frac{|S_\ell|}{k} \right\rfloor < \frac{|S_\ell|}{k}$, the total sample complexity for iteration ℓ is at most $\frac{|S_\ell|}{k}t \le \frac{n}{2k^{\ell-1}\epsilon_\ell^2}\ln\frac{k}{\delta_\ell}$, as $|S_\ell| \le \frac{n}{k^\ell}$ for all $\ell \in [\lfloor \ln_k n \rfloor]$. Also note that for all but last iteration $\ell \in [\lfloor \ln_k n \rfloor]$, $\epsilon_\ell = \frac{\epsilon}{8}\left(\frac{3}{4}\right)^{\ell-1}$, and $\delta_\ell = \frac{\delta}{2^{\ell+1}}$. Moreover for the last iteration $\ell = \lceil \ln_k n \rceil$, the sample complexity is clearly $t = \frac{2k}{\epsilon^2}\ln\frac{2k}{\delta}$, as in this case $\epsilon_\ell = \frac{\epsilon}{2}$, and $\delta_\ell = \frac{\delta}{2}$, and $|S| = k$. Thus the total sample complexity of Algorithm 2 is given by

$$\begin{aligned}
\sum_{\ell=1}^{\lceil \ln_k n \rceil} \frac{|S_\ell|}{2\epsilon_\ell^2}\ln\frac{k}{\delta_\ell} &\le \sum_{\ell=1}^{\infty} \frac{n}{2k^\ell\left(\frac{\epsilon}{8}\left(\frac{3}{4}\right)^{\ell-1}\right)^2} k \ln\frac{k2^{\ell+1}}{\delta} + \frac{2k}{\epsilon^2}\ln\frac{2k}{\delta} \\
&\le \frac{64n}{2\epsilon^2}\sum_{\ell=1}^{\infty}\frac{16^{\ell-1}}{(9k)^{\ell-1}}\Big(\ln\frac{k}{\delta} + (\ell+1)\Big) + \frac{2k}{\epsilon^2}\ln\frac{2k}{\delta} \\
&\le \frac{32n}{\epsilon^2}\ln\frac{k}{\delta}\sum_{\ell=1}^{\infty}\frac{4^{\ell-1}}{(9k)^{\ell-1}}\Big(3\ell\Big) + \frac{2k}{\epsilon^2}\ln\frac{2k}{\delta} = O\left(\frac{n}{\epsilon^2}\ln\frac{k}{\delta}\right) \text{ [for any } k > 1].
\end{aligned}$$

Above proves the sample complexity bound of Theorem 6. We next prove the (ϵ, δ)-PAC property of *Divide-and-Battle*. The crucial observation lies in the fact that, at any iteration ℓ, for any set $\mathcal{G}_g$ ($g = 1, 2, \ldots, G$), the item c_g retained by the algorithm is likely to be not more than ϵ_ℓ-worse than the best item (the one with maximum score parameter θ) of the set $\mathcal{G}_g$, with probability at least $(1 - \delta_\ell)$. More precisely, we claim the following:

Lemma 7 *At any iteration ℓ, for any $\mathcal{G}_g$, if $i_g := \arg\max_{i \in \mathcal{G}_g} \theta_i$, then with probability at least $(1 - \delta_\ell)$, $p_{c_g i_g} > \frac{1}{2} - \epsilon_\ell$.*

Proof Let us define $\hat{p}_{ij} = \frac{w_i}{w_i + w_j}$, $\forall i, j \in \mathcal{G}_g, i \neq j$. Then clearly $\hat{p}_{c_g i_g} \geq \frac{1}{2}$, as c_g is the empirical winner in t rounds, i.e. $c_g \leftarrow \underset{i \in \mathcal{G}_g}{\arg\max}\, w_i$. Moreover c_g being the empirical winner of $\mathcal{G}_g$ we also have $w_{c_g} \geq \frac{t}{k}$, and thus $w_{c_\ell} + w_{r_\ell} \geq \frac{t}{k}$ as well. Let $n_{ij} := w_i + w_j$ denotes the number of pairwise comparisons of item i and j in t rounds, $i, j \in \mathcal{G}_g$. Clearly $0 \leq n_{ij} \leq t$. Then let us analyze the probability of a 'bad event' where c_g is indeed such that $p_{c_g i_g} < \frac{1}{2} - \epsilon_\ell$ but we have c_g beating i_g empirically:

$$
\begin{aligned}
& Pr\left(\left\{\hat{p}_{c_g i_g} \geq \frac{1}{2}\right\}\right) \\
& = Pr\left(\left\{\hat{p}_{c_g i_g} \geq \frac{1}{2}\right\} \cap \left\{n_{c_g i_g} \geq \frac{t}{k}\right\}\right) + \cancelto{0}{Pr\left\{n_{c_g i_g} < \frac{t}{k}\right\}} Pr\left(\left\{\hat{p}_{c_g i_g} \geq \frac{1}{2}\right\} \Big| \left\{n_{c_g i_g} < \frac{t}{k}\right\}\right) \\
& = Pr\left(\left\{\hat{p}_{c_g i_g} - \epsilon_\ell \geq \frac{1}{2} - \epsilon_\ell\right\} \cap \left\{n_{c_g i_g} \geq \frac{t}{k}\right\}\right) \\
& \leq Pr\left(\left\{\hat{p}_{c_g i_g} - p_{c_g i_g} \geq \epsilon_\ell\right\} \cap \left\{n_{c_g i_g} \geq \frac{t}{k}\right\}\right) \\
& \leq \exp\Big(-2\frac{t}{k}(\epsilon_\ell)^2\Big) = \frac{\delta_\ell}{k}.
\end{aligned}
$$

where the first inequality holds as $p_{c_g i_g} < \frac{1}{2} - \epsilon_\ell$, and the second inequality is by applying Lemma 1 with $\eta = \epsilon_\ell$ and $v = \frac{t}{k}$. Now taking union bound over all ϵ_ℓ-suboptimal elements i' of $\mathcal{G}_g$ (i.e. $p_{i' i_g} < \frac{1}{2} - \epsilon_\ell$), we get

$$
Pr\left(\left\{\exists i' \in \mathcal{G}_g \mid p_{i' i_g} < \frac{1}{2} - \epsilon_\ell, \text{and } c_g = i'\right\}\right) \leq \frac{\delta_\ell}{k}\left|\left\{\exists i' \in \mathcal{G}_g \mid p_{i' i_g} < \frac{1}{2} - \epsilon_\ell, \text{and } c_g = i'\right\}\right| \leq \delta_\ell,
$$

as $|\mathcal{G}_g| = k$, and the claim follows henceforth. ■

Remark 8 *For the last iteration $\ell = \lceil \ln_k n \rceil$, since $\epsilon_\ell = \frac{\epsilon}{2}$, and $\delta_\ell = \frac{\delta}{2}$, applying Lemma 7 on S, we get that $Pr\left(p_{r_* i_g} < \frac{1}{2} - \frac{\epsilon}{2}\right) \leq \frac{\delta}{2}$.*

Now for each iteration ℓ, let us define $g_\ell \in [G]$ to be the set that contains *best item* of the entire set S, i.e. $\arg\max_{i \in S} \theta_i \in \mathcal{G}_{g_\ell}$. Then applying Lemma 7, with probability at least $(1 - \delta_\ell)$, $\tilde{p}_{c_{g_\ell} i_{g_\ell}} > -\epsilon_\ell$. Then, for each iteration ℓ, applying Lemma 15 and Lemma 7 to $\mathcal{G}_{g_\ell}$, we finally get $\tilde{p}_{r_* 1} > -\left(\frac{\epsilon}{8} + \frac{\epsilon}{8}\left(\frac{3}{4}\right) + \cdots + \frac{\epsilon}{8}\left(\frac{3}{4}\right)^{\lfloor \ln_k n \rfloor}\right) + \frac{\epsilon}{2} \geq -\frac{\epsilon}{8}\left(\sum_{i=0}^{\infty}\left(\frac{3}{4}\right)^i\right) + \frac{\epsilon}{2} = \epsilon$. (Note that, for above analysis to go through, it is in fact sufficient to consider only the set of iterations $\{\ell \geq \ell_0 \mid \ell_0 = \min\{l \mid 1 \notin \mathcal{R}_l,\, l \geq 1\}\}$ because prior considering item 1, it does not matter even if the algorithm mistakes in any of the iteration $\ell < \ell_0$). Thus assuming the algorithm does not fail in any of the iteration ℓ, we finally have that $p_{r_* 1} > \frac{1}{2} - \epsilon$.

Finally since at each iteration ℓ, the algorithm fails with probability at most δ_ℓ, the total failure probability of the algorithm is at most $\Big(\frac{\delta}{4} + \frac{\delta}{8} + \cdots + \frac{\delta}{2^{\lceil \ln_k n \rceil}}\Big) + \frac{\delta}{2} \le \delta$. This concludes the correctness of the algorithm showing that it indeed satisfies the (ϵ, δ)-PAC objective. ■

B.4. Proof of Theorem 8

Theorem 8 (*Halving-Battle*: Correctness and Sample Complexity with WI) *Halving-Battle (Algorithm 3) is (ϵ, δ)-PAC with sample complexity $O\big(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\big)$.*

Proof For clarity of notation, we will denote the set of remaining items S at iteration ℓ by S_ℓ. We start by observing that at each iteration $\ell = 1, 2, \ldots$, the size of the set of remaining items $S_{\ell+1}$ gets halved compared to that of the previous iteration S_ℓ, since the algorithm discards away all the elements below the *median item* h_g, as follows from the definition of median. This implies that the maximum number of iterations possible is $\ell = \lceil \ln n \rceil$, after which $|S| = 1$ and the algorithm returns r_*.

We first analyze the sample complexity of the algorithm. Clearly each iteration ℓ uses a sample complexity of $t = \frac{k}{2\epsilon_\ell^2}\ln\frac{1}{\delta_\ell}$, and as argued before ℓ can be at most $\lceil \ln n \rceil$ which makes the total sample complexity of the algorithm:

$$\sum_{\ell=1}^{\lceil \ln n \rceil} \frac{|S_\ell|}{2\epsilon_\ell^2}\ln\frac{1}{\delta_\ell} \le \sum_{\ell=1}^{\infty} \frac{n}{2k^\ell\bigg(\frac{\epsilon}{4}\big(\frac{3}{4}\big)^{\ell-1}\bigg)^2} k \ln 4\Big(\frac{2^\ell}{\delta}\Big) \le \frac{16n}{2\epsilon^2}\sum_{\ell=1}^{\infty}\frac{16^{\ell-1}}{(9k)^{\ell-1}}\Big(\ln\frac{4}{\delta} + \ell\Big)$$

$$\le \frac{8n}{\epsilon^2}\ln\frac{4}{\delta}\sum_{\ell=1}^{\infty}\frac{16^{\ell-1}}{9k^{\ell-1}}\Big(2\ell\Big) = O\bigg(\frac{n}{\epsilon^2}\ln\frac{1}{\delta}\bigg) \text{ [for any} k > 1].$$

This ensures the sample complexity of Theorem 8 holds good.

We are now only left with verifying the (ϵ, δ)-PAC property of the algorithm where lies the main difference of the analysis of *Halving-Battle* from *Divide-and-Battle*. Consider any iteration $\ell \in [\lceil \ln n \rceil]$. The crucial observation is that, with high probability of at least $\big(1 - \delta_\ell\big)$ for any such ℓ, and any set $\mathcal{G}_g$ ($g = 1, 2, \ldots, G$), some ϵ_ℓ-approximation of the *'best-item'* (the one with the highest score parameter θ_i) of $\mathcal{G}_g$ must lie above the median in terms of the empirical win count w_i, and hence must be retained by the algorithm till the next iteration $\ell + 1$. We prove this formally below.

Our first claim starts by showing that for any set $\mathcal{G}_g$, the empirical win count estimate w_i of the best item $i_g := \arg\max_{i\in\mathcal{G}_g}\theta_i$ (i.e. the one with highest score parameter θ_i) can not be too small, as shown in Lemma 16:

Lemma 16 *Consider any particular set $\mathcal{G}_g$ at any iteration $\ell \in \lceil \ln n \rceil$. If $i_g := \arg\max_{i\in\mathcal{G}_g}\theta_i$, then with probability at least $\Big(1 - \frac{\delta_\ell}{4}\Big)$, the empirical win count $w_{i_g} > (1-\eta)\frac{t}{k}$, for any $\eta \in \big(\frac{3}{16}, 1\big]$.*

Proof The proof follows from an straightforward application of Chernoff-Hoeffding's inequality Boucheron et al. (2013). Note that the algorithm plays each set $\mathcal{G}_g$ for $t = \frac{k}{2\epsilon_\ell^2}\ln\frac{1}{\delta_\ell}$ number of times. Fix any iteration ℓ and a set $\mathcal{G}_g$, $g \in 1, 2, \ldots, G$. Suppose i_τ denotes the winner of τ-th play of $\mathcal{G}_g$, $\tau \in [t]$. Then clearly, for any item $i \in \mathcal{G}_g$, $w_i = \sum_{\tau=1}^{t}\mathbf{1}(i_\tau == i)$, where $\mathbf{1}(i_\tau == i)$ is a Bernoulli random variable with parameter $\frac{\theta_i}{\sum_{j \in \mathcal{G}_g} \theta_j}$, by definition of WI feedback model. Also for $i = i_g$, we have $Pr(\{i_\tau = i_g\}) = \frac{\theta_{i_g}}{\sum_{j \in \mathcal{G}_g} \theta_j} \ge \frac{1}{k}$, $\forall \tau \in [t]$, as follows from the definition $i_g := \arg\max_{i \in \mathcal{G}_g} \theta_i$. Hence $\mathbf{E}[w_{i_g}] = \sum_{\tau=1}^{t} \mathbf{E}[\mathbf{1}(i_\tau == i)] \ge \frac{t}{k}$. Now applying multiplicative Chernoff-Hoeffdings bound for w_{i_g}, we get that for any $\eta \in (\frac{3}{16}, 1]$,

$$
\begin{aligned}
Pr\Big(w_{i_g} \le (1-\eta)\mathbf{E}[w_{i_g}]\Big) &\le \exp\bigg(-\frac{\mathbf{E}[w_{i_g}]\eta^2}{2}\bigg) \le \exp\bigg(-\frac{t\eta^2}{2k}\bigg) \\
&\le \exp\bigg(-\frac{\eta^2}{\epsilon_\ell^2} \ln\bigg(\frac{4}{\delta_\ell}\bigg)\bigg) \le \exp\bigg(-\ln\bigg(\frac{4}{\delta_\ell}\bigg)\bigg) = \frac{\delta_\ell}{4},
\end{aligned}
$$

where the second last inequality holds as $\eta > \frac{3}{16}$ and $\epsilon_\ell \le \frac{3}{16}$, for any iteration $\ell \in \lceil \ln n \rceil$; in other words for any $\eta \ge \frac{1}{4}$, we have $\frac{\eta}{\epsilon_\ell} > 1$ which leads to the second last inequality, and the proof follows henceforth. ■

In particular, fixing $\eta = \frac{1}{2}$ in Lemma 16, we get that with probability at least $\big(1 - \frac{\delta_\ell}{4}\big)$, $w_{i_g} > (1 - \frac{1}{2})\mathbf{E}[w_{i_g}] > \frac{t}{2k}$. We now prove that for any set $\mathcal{G}_g$, given its best item i_g is selected as the winner for at least $\frac{t}{2k}$ times out of t plays of $\mathcal{G}_g$, the empirical estimate of $p_{i_g b}$, defined as $\hat{p}_{i_g b} = \frac{w_{i_g}}{w_{i_g} + w_b}$, for any suboptimal element $b \in \mathcal{G}_g$ (such that $p_{i_g b} > \frac{1}{2} + \epsilon$) can not be too misleading where empirical win count of b exceeds that of i_g, i.e. $w_b > w_{i_g}$. The formal claim is as follows:

Lemma 9 *At any iteration ℓ, for any set $\mathcal{G}_g$, let $i_g \leftarrow \underset{i \in \mathcal{G}_g}{\arg\max}\, \theta_i$, and consider any suboptimal item $b \in \mathcal{G}_g$ such that $p_{b i_g} < \frac{1}{2} - \epsilon_\ell$. Then with probability at least $\big(1 - \frac{\delta_\ell}{4}\big)$, the empirical win count of i_g lies above that of b, i.e. $w_{i_g} \ge w_b$ (equivalently $\hat{p}_{i_g b} = \frac{w_{i_g}}{w_{i_g} + w_b} \ge \frac{1}{2}$).*

Proof First note since $w_{i_g} \ge \frac{t}{2k}$, this implies $w_{i_g} + w_b \ge \frac{t}{2k}$ as well. Let us define $n_{ij} = w_i + w_j$ to be the number of pairwise comparisons of item i and j in t rounds, for any $i, j \in \mathcal{G}_g$, and $\hat{p}_{ij} = \frac{w_i}{w_i + w_j}$ to be the empirical estimate of pairwise probability of item i and j. Then,

$$
\begin{aligned}
Pr\Bigg(\bigg\{w_b \ge w_{i_g}\bigg\} \cap \bigg\{n_{i_g b} \ge \frac{t}{2k}\bigg\}\Bigg) &= Pr\Bigg(\bigg\{\hat{p}_{b i_g} \ge \frac{1}{2}\bigg\} \cap \bigg\{n_{i_g b} \ge \frac{t}{2k}\bigg\}\Bigg) \\
&= Pr\Bigg(\bigg\{\hat{p}_{b i_g} - \epsilon_\ell \ge \frac{1}{2} - \epsilon_\ell\bigg\} \cap \bigg\{n_{i_g b} \ge \frac{t}{2k}\bigg\}\Bigg) \\
&\le Pr\Bigg(\bigg\{\hat{p}_{b i_g} - p_{b i_g} \ge \epsilon_\ell\bigg\} \cap \bigg\{n_{i_g b} \ge \frac{t}{2k}\bigg\}\Bigg) \\
&\le \exp\bigg(-2\frac{t}{2k}\big(\epsilon_\ell\big)^2\bigg) \le \frac{\delta_\ell}{4},
\end{aligned}
$$

where the second last inequality holds since $p_{b i_g} < \frac{1}{2} - \epsilon_\ell$. The last inequality follows by applying Lemma 1 with $\eta = \epsilon_\ell$ and $v = \frac{t}{2k}$.

Using the results from Lemma 16 we further get that for any such suboptimal element $b \in \mathcal{C}_g$ with $p_{b i_g} < \frac{1}{2} - \epsilon_\ell$,

$$
\begin{aligned}
& Pr\big(\{w_b \ge w_{i_g}\}\big) \\
& = Pr\Bigg(\bigg\{w_b \ge w_{i_g}\bigg\} \cap \bigg\{w_{i_g} < \frac{t}{2k}\bigg\}\Bigg) + Pr\Bigg(\bigg\{w_b \ge w_{i_g}\bigg\} \cap \bigg\{w_{i_g} \ge \frac{t}{2k}\bigg\}\Bigg) \\
& \le Pr\big(\{w_{i_g} < \frac{t}{2k}\}\big) + Pr\Bigg(\bigg\{w_b \ge w_{i_g}\bigg\} \cap \bigg\{n_{i_g b} \ge \frac{t}{2k}\bigg\}\Bigg) \\
& \le \frac{\delta_\ell}{4} + \frac{\delta_\ell}{4} \ \Big[\text{Applying Lemma 16 with } \eta = \frac{1}{2}\Big] \le \frac{\delta_\ell}{2}
\end{aligned}
$$

■

Now for any particular $\mathcal{G}_g$, and for all suboptimal element $b \in \mathcal{G}_g$, let us define an indicator random variable $F_b := \mathbf{1}(w_b > w_{i_g})$. Note that by above claim we have $\mathbf{E}[F_b] = Pr(F_b) = Pr(w_b > w_{i_g}) \le \frac{\delta_\ell}{2}$. Moreover if $\mathcal{B} = \{b \in \mathcal{G}_g \mid p_{bi_g} < \frac{1}{2} - \epsilon\}$ denotes the set of all ϵ_ℓ-suboptimal elements of $\mathcal{G}_g$ (with respect to the best item i_g of $\mathcal{G}_g$), then clearly $|\mathcal{B}| < |\mathcal{G}_g|$, and thus we have $\mathbf{E}[\sum_{b\in\mathcal{B}} F_b] \le |\mathcal{G}_g|\frac{\delta_\ell}{2}$. Now using Markov's inequality Boucheron et al. (2013) we get:

$$
Pr\Bigg[\sum_{b\in\mathcal{B}} F_b \ge \frac{|\mathcal{G}_g|}{2}\Bigg] \le \frac{\mathbf{E}[\sum_{b\in\mathcal{B}} F_b]}{|\mathcal{G}_g|/2} \le \frac{|\mathcal{G}_g|\delta_\ell/2}{|\mathcal{G}_g|/2} = \delta_\ell.
$$

1 Above immediately implies that at any iteration ℓ, and for any set $\mathcal{G}_g$ in ℓ, more than $\frac{|\mathcal{G}_g|}{2}$ of the suboptimal elements of $\mathcal{G}_g$ can not beat the best item i_g in terms of empirical win count w_i. Thus there has to at least one non-suboptimal element $i' \in \mathcal{G}_g$ (i' could be i_g itself), i.e. $p_{i_g i'} < \frac{1}{2} - \epsilon$, and i' beats the median item h_g with $w_{i'} \ge w_{h_g}$. Hence i' would be retained by the algorithm in set S till the next iteration $\ell + 1$.

The above argument precisely shows that the best item of the set S at the beginning of iteration $\ell + 1$, can not be ϵ_ℓ worse than that of iteration ℓ, for any $\ell = [\lceil \ln n \rceil]$. More formally, if i_ℓ and $i_{\ell+1}$ respectively denote the best item of set S at the beginning of iteration ℓ and $\ell + 1$ respectively, i.e. $i_\ell := \arg\max_{i\in S_\ell} \theta_i$, and $i_{\ell+1} := \arg\max_{i\in S_{\ell+1}} \theta_i$, then by Lemma 9, with probability at least $(1 - \delta_\ell)$, $p_{i_{\ell+1} i_\ell} > \frac{1}{2} - \epsilon_\ell$. Note that, at the beginning $i_1 = 1$, which is the true best item (condorcet winner) $i^* = 1$ of $[n]$, as defined in Section 3. Now applying Lemma 7 and 15 for each iteration ℓ, we get that the final item r_* returned by the algorithm would satisfy $\tilde{p}_{r_* 1} > -\Big(\frac{\epsilon}{4} + \frac{\epsilon}{4}\Big(\frac{3}{4}\Big) + \cdots + \frac{\epsilon}{4}\big(\frac{3}{4}\big)^{\lfloor \ln n \rfloor}\Big) \ge -\frac{\epsilon}{4}\Big(\sum_{i=0}^{\infty}\big(\frac{3}{4}\big)^i\Big) = -\epsilon$. Thus assuming the algorithm does not fail in any of the iteration ℓ, we have that $p_{r_* 1} > \frac{1}{2} - \epsilon$.

Finally at each iteration ℓ, since the algorithm can fail with probability at most δ_ℓ, the total failure probability of the algorithm is at most $\Big(\frac{\delta}{2} + \frac{\delta}{4} + \cdots + \frac{\delta}{2^{\lceil \ln n \rceil}}\Big) \le \delta$. This concludes the proof as *Halving-Battle* indeed satisfies the (ϵ, δ)-PAC objective. ■

Appendix C. Appendix for Section 5

C.1. Proof of Theorem 10

Theorem 10 (Sample Complexity Lower Bound for TR) *Given $\epsilon \in (0, \frac{1}{\sqrt{8}}]$ and $\delta \in (0,1]$, and an (ϵ,δ)-PAC algorithm A with top-m ranking (TR) feedback ($2 \le m \le k$), there exists a PL instance ν such that the expected sample complexity of A on ν is at least $\Omega\bigg(\frac{n}{m\epsilon^2}\ln\frac{1}{2.4\delta}\bigg)$.*

Proof In this case too, we will use Lemma 3 to derive the desired lower bounds of Theorem 2 for BB-PL with **TR** feedback model.

Let us consider a bandit instance with the arm set containing all subsets of size k: $\mathcal{A} = \{S = (S(1), \ldots, S(k)) \subseteq [n] \mid S(i) < S(j), \forall i < j\}$. Let $\boldsymbol{\nu}^1$ be the true distribution associated with the bandit arms, given by the Plackett-Luce parameters:

$$\textbf{True Instance } (\boldsymbol{\nu}^1): \theta_j^1 = \theta\bigg(\frac{1}{2} - \epsilon\bigg), \forall j \in [n] \setminus \{1\}, \text{ and } \theta_1^1 = \theta\bigg(\frac{1}{2} + \epsilon\bigg),$$

for some $\theta \in \mathbb{R}_+,\ \epsilon > 0$. Now for every suboptimal item $a \in [n] \setminus \{1\}$, consider the modified instances $\boldsymbol{\nu}^a$ such that:

$$\textbf{Instance–a } (\boldsymbol{\nu}^a): \theta_j^a = \theta\bigg(\frac{1}{2} - \epsilon\bigg)^2, \forall j \in [n] \setminus \{a,1\},\ \theta_1^a = \theta\bigg(\frac{1}{4} - \epsilon^2\bigg), \text{ and } \theta_a^a = \theta\bigg(\frac{1}{2} + \epsilon\bigg)^2.$$

It is now interesting to note that how top-m ranking feedback affects the KL-divergence analysis, precisely the KL-divergence shoots up by a factor of m which in fact triggers an $\frac{1}{m}$ reduction in regret learning rate. Note that for top-m ranking feedback for any problem **Instance-a** (for any $a \in [n]$), each k-set $S \subseteq [n]$ is associated to $\binom{k}{m}(m!)$ number of possible outcomes, each representing one possible ranking of set of m items of S, say S_m. Also the probability of any permutation $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_{S_m}$ is given by $\boldsymbol{\nu}_S^a(\boldsymbol{\sigma}) = Pr_{\boldsymbol{\theta}^a}(\boldsymbol{\sigma}|S)$, where $Pr_{\boldsymbol{\theta}^a}(\boldsymbol{\sigma}|S)$ is as defined for top-m (TR-m) ranking feedback (as in Sec. 3.1). More formally, for any problem **Instance-a**, we have that:

$$\boldsymbol{\nu}_S^a(\boldsymbol{\sigma}) = \prod_{i=1}^{m} \frac{\theta_{\sigma(i)}^a}{\sum_{j=i}^{m} \theta_{\sigma(j)}^a + \sum_{j \in S \setminus \sigma(1:m)} \theta_{\sigma(j)}^a}, \quad \forall a \in [n],$$

The important thing now to note is that for any such top-m ranking of $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_S^m$, $KL(\boldsymbol{\nu}_S^1(\boldsymbol{\sigma}), \boldsymbol{\nu}_S^a(\boldsymbol{\sigma})) = 0$ for any set $S \not\ni a$. Hence while comparing the KL-divergence of instances $\boldsymbol{\theta}^1$ vs $\boldsymbol{\theta}^a$, we need to focus only on sets containing a. Applying *Chain-Rule* of KL-divergence, we now get

$$\begin{aligned} KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) = KL(\boldsymbol{\nu}_S^1(\sigma_1), \boldsymbol{\nu}_S^a(\sigma_1)) + KL(\boldsymbol{\nu}_S^1(\sigma_2 \mid \sigma_1), \boldsymbol{\nu}_S^a(\sigma_2 \mid \sigma_1)) + \cdots \\ + KL(\boldsymbol{\nu}_S^1(\sigma_m \mid \sigma(1:m-1)), \boldsymbol{\nu}_S^a(\sigma_m \mid \sigma(1:m-1))), \end{aligned} \tag{8}$$

where we abbreviate $\sigma(i)$ as σ_i and $KL(P(Y \mid X), Q(Y \mid X)) := \sum_x Pr\Big(X = x\Big)\big[KL(P(Y \mid X = x), Q(Y \mid X = x))\big]$ denotes the conditional KL-divergence. Moreover it is easy to note that for any $\sigma \in \Sigma_{S_m}$ such that $\sigma(i) = a$, we have $KL(\boldsymbol{\nu}_S^1(\sigma_{i+1} \mid \sigma(1:i)), \boldsymbol{\nu}_S^a(\sigma_{i+1} \mid \sigma(1:i))) := 0$, for all $i \in [m]$. We also denote the set of possible top-i rankings of set S, by Σ_{S_i}, for all $i \in [m]$. Now as derived in (3) in the proof of Theorem 2, we have

$$KL(\boldsymbol{\nu}_S^1(\sigma_1), \boldsymbol{\nu}_S^a(\sigma_1)) \le \frac{1}{k}\Big(R - \frac{1}{R}\Big)^2.$$

To bound the remaining terms of (8), note that for all $i \in [m-1]$

$$\begin{aligned}
&KL(\boldsymbol{\nu}_S^1(\sigma_{i+1} \mid \sigma(1:i)),\ \boldsymbol{\nu}_S^a(\sigma_{i+1} \mid \sigma(1:i))) \\
&\qquad = \sum_{\sigma' \in \Sigma_{S_i}} Pr_{\boldsymbol{\nu}^1}(\sigma') KL(\boldsymbol{\nu}_S^1(\sigma_{i+1} \mid \sigma(1:i)) = \sigma', \boldsymbol{\nu}_S^a(\sigma_{i+1} \mid \sigma(1:i)) = \sigma') \\
&\qquad = \sum_{\sigma' \in \Sigma_{S_i} \mid a \notin \sigma'} \left[\prod_{j=1}^{i} \left(\frac{\theta^1_{\sigma'_j}}{\theta^1_S - \sum_{j'=1}^{j-1} \theta^1_{\sigma'_{j'}}} \right) \right] \frac{1}{k-i}\Big(R - \frac{1}{R}\Big)^2 = \frac{1}{k}\Big(R - \frac{1}{R}\Big)^2
\end{aligned}$$

where $\theta_S^1 = \sum_{j' \in S} \theta^1_{j'}$. Thus applying above in (8) we get:

$$\begin{aligned}
&KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) = KL(\boldsymbol{\nu}_S^1(\sigma_1), \boldsymbol{\nu}_S^a(\sigma_1)) + \cdots + KL(\boldsymbol{\nu}_S^1(\sigma_m \mid \sigma(1:m-1)), \boldsymbol{\nu}_S^a(\sigma_m \mid \sigma(1:m-1))) \\
&\quad \le \frac{m}{k}\Big(R - \frac{1}{R}\Big)^2 \le \frac{m}{k} 256\epsilon^2 \ \left[\text{since} \left(R - \frac{1}{R}\right) = \frac{8\epsilon}{(1-4\epsilon^2)} \le 16\epsilon, \forall \epsilon \in [0, \frac{1}{\sqrt{8}}] \right] \qquad (9)
\end{aligned}$$

Eqn. (9) gives the main result to derive Theorem 10 as it shows an m-factor blow up in the KL-divergence terms owning to top-m ranking feedback.

Now, consider $\mathcal{E}_0 \in \mathcal{F}_\tau$ be an event such that the algorithm A returns the element $i = 1$, and let us analyse the left hand side of (2) for $\mathcal{E} = \mathcal{E}_0$. Clearly, A being an (ϵ, δ)-PAC algorithm, we have $Pr_{\boldsymbol{\nu}^1}(\mathcal{E}_0) > 1 - \delta$, and $Pr_{\boldsymbol{\nu}^a}(\mathcal{E}_0) < \delta$, for any suboptimal arm $a \in [n] \setminus \{1\}$. Then we have:

$$kl(Pr_{\boldsymbol{\nu}^1}(\mathcal{E}_0), Pr_{\boldsymbol{\nu}^a}(\mathcal{E}_0)) \ge kl(1-\delta, \delta) \ge \ln \frac{1}{2.4\delta} \qquad (10)$$

where the last inequality follows due to Equation (3) of Kaufmann et al. (2016).

Now applying (2) and (10) for each modified bandit **Instance-$\boldsymbol{\nu}^a$**, and summing over all suboptimal items $a \in [n] \setminus \{1\}$ we get,

$$\sum_{a=2}^{n} \sum_{\{S \in \mathcal{A} \mid a \in S\}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)] KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) \ge (n-1) \ln \frac{1}{2.4\delta}. \qquad (11)$$

Moreover, using (9), the term in the right hand side of (11) can be further upper bounded as:

$$\sum_{a=2}^{n} \sum_{\{S \in \mathcal{A} \mid a \in S\}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)] KL(\boldsymbol{\nu}_S^1, \boldsymbol{\nu}_S^a) \le \sum_{S \in \mathcal{A}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)] \sum_{\{a \in S \mid a \neq 1\}} \frac{m}{k} 256\epsilon^2$$

$$= \sum_{S \in \mathcal{A}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)]\big(k - \mathbf{1}(1 \in S)\big)\left(\frac{m}{k}256\epsilon^2\right) \le \sum_{S \in \mathcal{A}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)]m\Big(256\epsilon^2\Big) \tag{12}$$

Finally noting that $\tau_A = \sum_{S \in \mathcal{A}}[N_S(\tau_A)]$, combining (11) and (12), we get

$$m\Big(256\epsilon^2\Big)\mathbf{E}_{\boldsymbol{\nu}^1}[\tau_A] = \sum_{S \in \mathcal{A}} \mathbf{E}_{\boldsymbol{\nu}^1}[N_S(\tau_A)]m\Big(256\epsilon^2\Big) \ge (n-1)\ln\frac{1}{2.4\delta}.$$

Thus above construction shows the existence of a problem instance $\boldsymbol{\nu} = \boldsymbol{\nu}^1$, such that $\mathbf{E}_{\boldsymbol{\nu}^1}[\tau_A] = \Omega(\frac{n}{m\epsilon^2}\ln\frac{1}{2.4\delta})$, which concludes the proof. ■

C.2. Proof of Lemma 12

Lemma 12 (Rank-Breaking Update) *Consider any subset $S \subseteq [n]$ with $|S| = k$. Let S be played for t rounds of battle, and let $\boldsymbol{\sigma}_\tau \in \boldsymbol{\Sigma}_{S_m^\tau}$, $(S_m^\tau \subseteq S,\ |S_m^\tau| = m)$, denote the TR feedback at each round $\tau \in [t]$. For each item $i \in S$, let $q_i := \sum_{\tau=1}^{t}\mathbf{1}(i \in S_m^\tau)$ be the number of times i appears in the top-m ranked output in t rounds. Then, the most frequent item(s) in the top-m positions must appear at least $\frac{mt}{k}$ times, i.e. $\max_{i \in S} q_i \ge \frac{mt}{k}$.*

Proof Let us denote $\hat{i} := \arg\max_{i \in S} q_i$ to be the item (note that it need not be unique) that appears in the top-m set for maximum number of times in t rounds of battle. Note that, after the battle of any round $\tau \in [t]$, σ_τ chooses exactly m distinct items in the top-m set $S_m^\tau \subseteq S$. Thus t rounds of feedback places exactly mt items in the top-m slots, i.e. $\sum_{i \in S} q_i = mt$. Now at any round τ, since an item $i \in S$ can appear in S_m^τ at most once, and $\sum_{i \in S} q_i = mt$, item $\hat{i}$ must be selected for at least $\frac{mt}{k}$ many rounds in the top-m set implying that $q_{\hat{i}} \ge \frac{mt}{k}$ (as we have $|S| = k$). ■

C.3. Proof of Theorem 13

Theorem 13 (*Trace-the-Best*: Correctness and Sample Complexity with TR) *With top-m ranking (TR) feedback model, Trace-the-Best (Algorithm 5) is (ϵ, δ)-PAC with sample complexity $O(\frac{n}{m\epsilon^2}\log\frac{n}{\delta})$.*

Proof We start by analyzing the required sample complexity first. Note that the 'while loop' of Algorithm 5 always discards away $k-1$ items per iteration. Thus, n being the total number of items, the 'while loop' can be executed for at most $\lceil \frac{n}{k-1} \rceil$ many number of iterations. Clearly, the sample complexity of each iteration being $t = \frac{2k}{m\epsilon^2}\ln\frac{n}{2\delta}$, the total sample complexity of the algorithm becomes $\big(\lceil \frac{n}{k-1} \rceil\big)\frac{2k}{m\epsilon^2}\ln\frac{n}{2\delta} \le \big(\frac{n}{k-1} + 1\big)\frac{2k}{m\epsilon^2}\ln\frac{n}{2\delta} = \big(n + \frac{n}{k-1} + k\big)\frac{2}{m\epsilon^2}\ln\frac{n}{2\delta} = O(\frac{n}{m\epsilon^2}\ln\frac{n}{\delta})$.

We now proceed to prove the (ϵ, δ)-PAC correctness of the algorithm. As argued before, the 'while loop' of Algorithm 5 can run for maximum $\lceil \frac{n}{k-1} \rceil$ many number of iterations, say $\ell = 1, 2, \ldots, \lceil \frac{n}{k-1} \rceil$, and let us denote the corresponding set $\mathcal{A}$ of iteration ℓ as $\mathcal{A}_\ell$. Same as before, our idea is to retain the estimated best item as the *'running winner'* in r_ℓ and compare it with the *'empirical best item'* of $\mathcal{A}_\ell$ at every ℓ. We start by noting the following important property of item c_ℓ for any iteration ℓ:

Lemma 17 *Suppose $q_i := \sum_{\tau=1}^{t} \mathbf{1}(i \in \mathcal{A}_{\ell m}^{\tau})$ denotes the number of times item i appeared in the top-m ranking in t iterations, and let $B_\ell \subseteq \mathcal{A}_\ell$ is defined as $B_\ell := \{i \in \mathcal{A}_\ell \mid q_i = \max_{j \in \mathcal{A}_\ell} q_j\}$, that denotes the subset of items in $\mathcal{A}_\ell$ which are selected in the top m ranking for maximum number of times in t rounds of battle on set $\mathcal{A}_\ell$. Then $c_\ell \in B_\ell$.*

Proof We prove the claim by contradiction. Suppose, $c_\ell \notin B_\ell$ and consider any item $\hat{i} \in B_\ell$. Then by definition, $q_{\hat{i}} > q_{c_\ell}$. But in that case following our rank breaking update (see Algorithm 4) implies that $w_{\hat{i}c_\ell} > w_{c_\ell \hat{i}}$, since item $\hat{i}$ is ranked higher than item c_ℓ for at least $(q_{\hat{i}} - q_{c_\ell}) > 0$ many rounds of battle. Now consider any other item $j \in \mathcal{A}_\ell$. Note that j can belong to either of these two cases:

Case 1. ($j \notin B_\ell$) Following the same argument as above (i.e. for $\hat{i}$ vs c_ℓ), we again have $w_{\hat{i}j} > w_{j\hat{i}}$, whereas for c_ℓ vs j, either $w_{c_\ell j} > w_{jc_\ell}$, or $w_{c_\ell j} < w_{jc_\ell}$, both cases are plausible. Thus we get: $\mathbf{1}(w_{\hat{i}j} > w_{j\hat{i}}) = 1 \geq \mathbf{1}(w_{c_\ell j} > w_{jc_\ell})$.

Case 2. ($j \in B_\ell$) In this case since $j \in B_\ell$, again following the same argument as for $\hat{i}$ vs c_ℓ, we here have $w_{jc_\ell} > w_{c_\ell j}$; whereas for $\hat{i}$ vs j, either $w_{\hat{i}j} > w_{j\hat{i}}$, or $w_{\hat{i}j} < w_{j\hat{i}}$, both cases are plausible. Thus we get: $\mathbf{1}(w_{\hat{i}j} > w_{j\hat{i}}) \geq 0 = \mathbf{1}(w_{c_\ell j} > w_{jc_\ell})$.

Combining the results of Case 1 and 2 along with the fact that $w_{\hat{i}c_\ell} > w_{c_\ell \hat{i}}$, we get $\sum_{j \in \mathcal{A} \setminus \{\hat{i}\}} \mathbf{1}\big(w_{\hat{i}j} \geq w_{j\hat{i}}\big) > \sum_{j \in \mathcal{A} \setminus \{c_\ell\}} \mathbf{1}\big(w_{c_\ell j} \geq w_{jc_\ell}\big)$. But this violates the fact that c_ℓ is defined as $c_\ell := \underset{i \in \mathcal{A}_\ell}{\text{argmax}} \sum_{j \in \mathcal{A} \setminus \{i\}} \mathbf{1}\big(w_{ij} \geq w_{ji}\big)$ which leads to a contradiction. Then our initial assumption has to be wrong and $c_\ell \in B_\ell$, which concludes the proof. ■

The next crucial observation lies in noting that, the estimated best item r ('running winner') gets updated as per the following lemma:

Lemma 18 *At any iteration $\ell = 1, 2 \ldots, \left\lfloor \frac{n}{k-1} \right\rfloor$, for any set $\mathcal{A}_\ell$, nwith probability at least $(1 - \frac{\delta}{2n})$, Algorithm 1 retains $r_{\ell+1} \leftarrow r_\ell$ if $p_{c_\ell r_\ell} \leq \frac{1}{2}$, and set $r_{\ell+1} \leftarrow c_\ell$ if $p_{c_\ell r_\ell} \geq \frac{1}{2} + \epsilon$.*

Proof The main observation lies in proving that at any iteration ℓ, $w_{c_\ell r_\ell} + w_{r_\ell c_\ell} \geq \frac{mt}{k}$. We argue this as follows: Firstly note that by Lemma 12 and 17, $c_\ell \in \mathcal{B}_\ell$ (Lemma 17) and hence it must have appeared in top-m positions for at least $\frac{mt}{k}$ times (Lemma 12). But the rank breaking update ensures that every element in top-m position gets updated for exactly k times (it loses to all elements preceding it in the top-m ranking and wins over the rest). Define $n_{ij} = w_{ij} + w_{ji}$ to be the number of times item i and j are compared after rank-breaking, $i, j \in \mathcal{A}_\ell$. Clearly $0 \leq n_{ij} \leq tk$ and $n_{ij} = n_{ji}$. Now using above argument we have that $n_{c_\ell r_\ell} = w_{c_\ell r_\ell} + w_{r_\ell c_\ell} \geq \frac{mt}{k}$. We are now proof the claim using the following two case analyses:

Case 1. (If $p_{c_\ell r_\ell} \leq \frac{1}{2}$, *Trace-the-Best* retains $r_{\ell+1} \leftarrow r_\ell$): Note that *Trace-the-Best* replaces $r_{\ell+1}$ by c_ℓ only if $\hat{p}_{c_\ell, r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}$, but this happens with probability:

$$
\begin{aligned}
& Pr\left(\left\{\hat{p}_{c_\ell r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}\right\}\right) \\
& = Pr\left(\left\{\hat{p}_{c_\ell r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}\right\} \cap \left\{n_{c_\ell r_\ell} \geq \frac{mt}{k}\right\}\right) + \cancelto{0}{Pr\left\{n_{c_\ell r_\ell} < \frac{mt}{k}\right\}} Pr\left(\left\{\hat{p}_{c_\ell r_\ell} > \frac{1}{2} + \frac{\epsilon}{2}\right\} \Big| \left\{n_{c_\ell r_\ell} < \frac{mt}{k}\right\}\right)
\end{aligned}
$$

$$\le Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} - p_{c_\ell r_\ell} > \frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{mt}{k}\bigg\}\Bigg) \le \exp\Big(-2\frac{mt}{k}\Big(\frac{\epsilon}{2}\Big)^2\Big) = \frac{\delta}{2n},$$

where the first inequality follows as $p_{c_\ell r_\ell} \le \frac{1}{2}$, and the second inequality is simply by applying Lemma 1 with $\eta = \frac{\epsilon}{2}$ and $v = \frac{mt}{k}$. We now proceed to the second case:

Case 2. (If $p_{c_\ell r_\ell} \ge \frac{1}{2} + \epsilon$, *Trace-the-Best* sets $r_{\ell+1} \leftarrow c_\ell$): Again recall that *Trace-the-Best* retains $r_{\ell+1} \leftarrow r_\ell$ only if $\hat{p}_{c_\ell, r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}$. In this case, that happens with probability:

$$
\begin{aligned}
& Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}\bigg\}\Bigg) \\
& = Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{mt}{k}\bigg\}\Bigg) + \cancelto{0}{Pr\bigg\{n_{c_\ell r_\ell} < \frac{mt}{k}\bigg\}} Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \frac{\epsilon}{2}\bigg\} \bigg| \bigg\{n_{c_\ell r_\ell} < \frac{mt}{k}\bigg\}\Bigg) \\
& = Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} \le \frac{1}{2} + \epsilon - \frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{mt}{k}\bigg\}\Bigg) \\
& \le Pr\Bigg(\bigg\{\hat{p}_{c_\ell r_\ell} - p_{c_\ell r_\ell} \le -\frac{\epsilon}{2}\bigg\} \cap \bigg\{n_{c_\ell r_\ell} \ge \frac{mt}{k}\bigg\}\Bigg) \le \exp\Big(-2\frac{mt}{k}\Big(\frac{\epsilon}{2}\Big)^2\Big) = \frac{\delta}{2n},
\end{aligned}
$$

where the first inequality holds as $p_{c_\ell r_\ell} \ge \frac{1}{2} + \epsilon$, and the second one is simply by applying Lemma 1 with $\eta = \frac{\epsilon}{2}$ and $v = \frac{mt}{k}$. Combining the above two cases concludes the proof. ■

The rest of the proof follows exactly same as that of Theorem 4. We include the details for completeness. Given Algorithm 5 satisfies Lemma 18, and taking union bound over $(k-1)$ elements in $\mathcal{A}_\ell \setminus \{r_\ell\}$, we get that with probability at least $\bigg(1 - \frac{(k-1)\delta}{2n}\bigg)$,

$$p_{r_{\ell+1} r_\ell} \ge \frac{1}{2} \text{ and, } p_{r_{\ell+1} c_\ell} \ge \frac{1}{2} - \epsilon. \tag{13}$$

Above clearly suggests that for each iteration ℓ, the estimated 'best' item r_ℓ only gets improved as $p_{r_{\ell+1} r_\ell} \ge 0$. Let, ℓ_* denotes the specific iteration such that $1 \in \mathcal{A}_\ell$ for the first time, i.e. $\ell_* = \min\{\ell \mid 1 \in \mathcal{A}_\ell\}$. Clearly $\ell_* \le \lceil \frac{n}{k-1} \rceil$.

Now (13) suggests that with probability at least $(1 - \frac{(k-1)\delta}{2n})$, $p_{r_{\ell_*+1} 1} \ge \frac{1}{2} - \epsilon$. Moreover (13) also suggests that for all $\ell > \ell_*$, with probability at least $(1 - \frac{(k-1)\delta}{2n})$, $p_{r_{\ell+1} r_\ell} \ge \frac{1}{2}$, which implies for all $\ell > \ell_*$, $p_{r_{\ell+1} 1} \ge \frac{1}{2} - \epsilon$ as well.

Note that above holds due to the following transitivity property of the Plackett-Luce model: For any three items $i_1, i_2, i_3 \in [n]$, if $p_{i_1 i_2} \ge \frac{1}{2}$ and $p_{i_2 i_3} \ge \frac{1}{2}$, then we have $p_{i_1 i_3} \ge \frac{1}{2}$ as well. This argument finally leads to $p_{r_* 1} \ge \frac{1}{2} - \epsilon$. Since failure probability at each iteration ℓ is at most $\frac{(k-1)\delta}{2n}$, and Algorithm 5 runs for maximum $\lceil \frac{n}{k-1} \rceil$ many number of iterations, using union bound over ℓ, the total failure probability of the algorithm is at most $\lceil \frac{n}{k-1} \rceil \frac{(k-1)\delta}{2n} \le (\frac{n}{k-1} + 1)\frac{(k-1)\delta}{2n} = \delta\bigg(\frac{n+k-1}{2n}\bigg) \le \delta$ (since $k \le n$). This concludes the correctness of the algorithm showing that it indeed returns an ϵ-best element r_* such that $p_{r_* 1} \ge \frac{1}{2} - \epsilon$ with probability at least $1 - \delta$. ■

C.4. Proof of Theorem 14

Theorem 14 (*Divide-and-Battle*: Correctness and Sample Complexity with TR) *With top-m ranking (TR) feedback, Divide-and-Battle (Algorithm 6) is (ϵ,δ)-PAC with sample complexity $O(\frac{n}{m\epsilon^2}\log\frac{k}{\delta})$.*

Proof For the notational convenience we will use $\tilde p_{ij} = p_{ij} - \frac{1}{2}, \forall i,j \in [n]$.

We first analyze the required sample complexity of the algorithm. For clarity of notation, we will denote the set S at iteration ℓ by S_ℓ. Note that at any iteration ℓ, any set $\mathcal{G}_g$ is played for exactly $t = \frac{4k}{m\epsilon_\ell^2}\ln\frac{2k}{\delta_\ell}$ many number of times. Also since the algorithm discards away exactly $k-1$ items from each set $\mathcal{G}_g$, hence the maximum number of iterations possible is $\lceil \ln_k n\rceil$. Now at any iteration ℓ, since $G = \left\lfloor \frac{|S_\ell|}{k}\right\rfloor < \frac{|S_\ell|}{k}$, the total sample complexity for iteration ℓ is at most $\frac{|S_\ell|}{k}t \le \frac{4n}{mk^{\ell-1}\epsilon_\ell^2}\ln\frac{2k}{\delta_\ell}$, as $|S_\ell| \le \frac{n}{k^\ell}$ for all $\ell \in [\lfloor \ln_k n\rfloor]$. Also note that for all but last iteration $\ell \in [\lfloor \ln_k n\rfloor]$, $\epsilon_\ell = \frac{\epsilon}{8}\bigg(\frac{3}{4}\bigg)^{\ell-1}$, and $\delta_\ell = \frac{\delta}{2^{\ell+1}}$. Moreover for the last iteration $\ell = \lceil \ln_k n\rceil$, the sample complexity is clearly $t = \frac{4k}{m(\epsilon/2)^2}\ln\frac{4k}{\delta}$, as in this case $\epsilon_\ell = \frac{\epsilon}{2}$, and $\delta_\ell = \frac{\delta}{2}$, and $|S| = k$. Thus the total sample complexity of Algorithm 6 is given by

$$
\begin{aligned}
\sum_{\ell=1}^{\lceil \ln_k n\rceil}\frac{|S_\ell|}{m(\epsilon_\ell/2)^2}\ln\frac{2k}{\delta_\ell} &\le \sum_{\ell=1}^{\infty}\frac{4n}{mk^\ell\bigg(\frac{\epsilon}{8}\big(\frac{3}{4}\big)^{\ell-1}\bigg)^2}k\ln\frac{k2^{\ell+1}}{\delta} + \frac{16k}{m\epsilon^2}\ln\frac{4k}{\delta}\\
&\le \frac{256n}{m\epsilon^2}\sum_{\ell=1}^{\infty}\frac{16^{\ell-1}}{(9k)^{\ell-1}}\bigg(\ln\frac{k}{\delta}+(\ell+1)\bigg) + \frac{16k}{m\epsilon^2}\ln\frac{4k}{\delta}\\
&\le \frac{256n}{m\epsilon^2}\ln\frac{k}{\delta}\sum_{\ell=1}^{\infty}\frac{4^{\ell-1}}{(9k)^{\ell-1}}\bigg(3\ell\bigg) + \frac{16k}{m\epsilon^2}\ln\frac{4k}{\delta} = O\bigg(\frac{n}{m\epsilon^2}\ln\frac{k}{\delta}\bigg) \text{ [for any } k>1].
\end{aligned}
$$

Above proves the sample complexity bound of Theorem 14. We now proceed to prove the (ϵ,δ)-PAC correctness of the algorithm. We start by making the following observations:

Lemma 19 *Consider any particular set $\mathcal{G}_g$ at any iteration $\ell \in \lfloor\frac{n}{k}\rfloor$ and define $q_i := \sum_{\tau=1}^{t}\mathbf{1}(i \in \mathcal{G}_{gm}^\tau)$ as the number of times any item $i \in \mathcal{G}_g$ appears in the top-m rankings when items in the set $\mathcal{G}_g$ is made to battle for t rounds. Then if $i_g := \arg\max_{i\in\mathcal{G}_g}\theta_i$, then with probability at least $\bigg(1-\frac{\delta_\ell}{2k}\bigg)$, one can show that $q_{i_g} > (1-\eta)\frac{mt}{k}$, for any $\eta \in (\frac{3}{32\sqrt{2}}, 1]$.*

Proof Fix any iteration ℓ and a set $\mathcal{G}_g$, $g \in 1,2,\ldots,G$. Define $i^\tau := \mathbf{1}(i \in \mathcal{G}_{gm}^\tau)$ as the indicator variable if i^{th} element appeared in the top-m ranking at iteration $\tau \in [t]$. Recall the definition of TR feedback model (Sec. 3.1). Using this we get $\mathbf{E}[i_g^\tau] = Pr(\{i_g \in \mathcal{G}_{gm}^\tau\}) = Pr\big(\exists j \in [m] \mid \sigma(j) = i_g\big) = \sum_{j=1}^{m} Pr\big(\sigma(j) = i_g\big) = \sum_{j=0}^{m-1}\frac{1}{k-j} \ge \frac{m}{k}$, since $Pr(\{i_g|S\}) = \frac{\theta_{i_g}}{\sum_{j\in S}\theta_j} \ge \frac{1}{|S|}$ for any $S \subseteq [\mathcal{G}_g]$, as $i_g := \arg\max_{i\in\mathcal{G}_g}\theta_i$ is the best item of set $\mathcal{G}_g$. Hence $\mathbf{E}[q_{i_g}] = \sum_{\tau=1}^{t}\mathbf{E}[i_g^\tau] \ge \frac{mt}{k}$.

Now applying Chernoff-Hoeffdings bound for w_{i_g}, we get that for any $\eta \in (\frac{3}{32}, 1]$,

$$
Pr\Big(q_{i_g} \le (1-\eta)\mathbf{E}[q_{i_g}]\Big) \le \exp\Big(-\frac{\mathbf{E}[q_{i_g}]\eta^2}{2}\Big) \le \exp\Big(-\frac{mt\eta^2}{2k}\Big)
$$

$$= \exp\bigg(- \frac{2\eta^2}{\epsilon_\ell^2} \ln\bigg(\frac{2k}{\delta_\ell}\bigg)\bigg) = \exp\bigg(- \frac{(\sqrt{2}\eta)^2}{\epsilon_\ell^2} \ln\bigg(\frac{2k}{\delta_\ell}\bigg)\bigg)$$
$$\le \exp\bigg(- \ln\bigg(\frac{2k}{\delta_\ell}\bigg)\bigg) \le \frac{\delta_\ell}{2k},$$

where the second last inequality holds as $\eta \ge \frac{3}{32\sqrt{2}}$ and $\epsilon_\ell \le \frac{3}{32}$, for any iteration $\ell \in \lceil \ln n \rceil$; in other words for any $\eta \ge \frac{3}{32\sqrt{2}}$, we have $\frac{\sqrt{2}\eta}{\epsilon_\ell} \ge 1$ which leads to the second last inequality. Thus we finally derive that with probability at least $\Big(1 - \frac{\delta_\ell}{2k}\Big)$, one can show that $q_{i_g} > (1-\eta)\mathbf{E}[q_{i_g}] \ge (1-\eta)\frac{tm}{k}$, and the proof follows henceforth. ∎

In particular, fixing $\eta = \frac{1}{2}$ in Lemma 16, we get that with probability at least $\big(1 - \frac{\delta_\ell}{2}\big)$, $q_{i_g} > (1-\frac{1}{2})\mathbf{E}[w_{i_g}] > \frac{mt}{2k}$. Note that, for any round $\tau \in [t]$, whenever an item $i \in \mathcal{G}_g$ appears in the top-m set $\mathcal{G}_{gm}^\tau$, then the rank breaking update ensures that every element in the top-m set gets compared with rest of the $k-1$ elements of $\mathcal{G}_g$. Based on this observation, we now prove that for any set $\mathcal{G}_g$, its best item i_g is retained as the winner c_g with probability at least $\big(1 - \frac{\delta_\ell}{2}\big)$. More formally, first thing to observe is:

Lemma 20 *Consider any particular set $\mathcal{G}_g$ at any iteration $\ell \in \lfloor \frac{n}{k} \rfloor$. If $i_g \leftarrow \arg\max_{i \in \mathcal{G}_g} \theta_i$, then with probability at least $\Big(1 - \delta_\ell\Big)$, $\hat{p}_{i_g j} + \frac{\epsilon_\ell}{2} \ge \frac{1}{2}$ for all ϵ_ℓ-optimal item $\forall j \in \mathcal{G}_g$ such that $p_{i_g j} \in \big(\frac{1}{2}, \frac{1}{2} + \epsilon_\ell\big]$, and $\hat{p}_{i_g j} - \frac{\epsilon_\ell}{2} \ge \frac{1}{2}$ for all non ϵ_ℓ-optimal item $j \in \mathcal{G}_g$ such that $p_{i_g j} > \frac{1}{2} + \epsilon_\ell$.*

Proof With top-m ranking feedback, the crucial observation lies in the fact that at any round $\tau \in [t]$, whenever an item $i \in \mathcal{G}_g$ appears in the top-m set $\mathcal{G}_{gm}^\tau$, then the rank breaking update ensures that every element in the top-m set gets compared with each of the rest of the $k-1$ elements of $\mathcal{G}_g$ - it defeats to every element preceding item in $\sigma \in \Sigma_{\mathcal{G}_{gm}}$, and wins over the rest. Therefore defining $n_{ij} = w_{ij} + w_{ji}$ to be the number of times item i and j are compared after rank-breaking, $i, j \in \mathcal{G}_g$. Clearly $n_{ij} = n_{ji}$, and $0 \le n_{ij} \le tk$. Moreover, from Lemma 19 with $\eta = \frac{1}{2}$, we have that $n_{i_g j} \ge \frac{mt}{2k}$. Given the above arguments in place let us analyze the probability of a 'bad event' that indedd:

Case 1. j is ϵ_ℓ-optimal with respect to i_g, i.e. $p_{i_g j} \in \big(\frac{1}{2}, \frac{1}{2} + \epsilon_\ell\big]$. Then we have

$$Pr\Bigg(\bigg\{\hat{p}_{i_g j} + \frac{\epsilon_\ell}{2} < \frac{1}{2}\bigg\} \cap \bigg\{n_{i_g j} \ge \frac{mt}{2k}\bigg\}\Bigg) = Pr\Bigg(\bigg\{\hat{p}_{i_g j} < \frac{1}{2} - \frac{\epsilon_\ell}{2}\bigg\} \cap \bigg\{n_{i_g j} = \frac{mt}{2k}\bigg\}\Bigg)$$
$$\le Pr\Bigg(\bigg\{\hat{p}_{i_g j} - p_{i_g j} < -\frac{\epsilon_\ell}{2}\bigg\} \cap \bigg\{n_{i_g j} = \frac{mt}{2k}\bigg\}\Bigg)$$
$$\le \exp\bigg(-2\frac{mt}{2k}(\epsilon_\ell/2)^2\bigg) = \frac{\delta_\ell}{2k},$$

where the first inequality follows as $p_{i_g j} > \frac{1}{2}$, and the second inequality follows from Lemma 1 with $\eta = \frac{\epsilon_\ell}{2}$ and $v = \frac{mt}{2k}$.

Case 2. j is non ϵ_ℓ-optimal with respect to i_g, i.e. $p_{i_g j} > \frac{1}{2} + \epsilon_\ell$. Similar to before, we have

$$Pr\Bigg(\bigg\{\hat{p}_{i_g j} - \frac{\epsilon_\ell}{2} < \frac{1}{2}\bigg\} \cap \bigg\{n_{i_g j} \ge \frac{mt}{2k}\bigg\}\Bigg) = Pr\Bigg(\bigg\{\hat{p}_{i_g j} < \frac{1}{2} + \frac{\epsilon_\ell}{2}\bigg\} \cap \bigg\{n_{i_g j} = \frac{mt}{2k}\bigg\}\Bigg)$$

$$
\begin{aligned}
&\le Pr\bigg(\Big\{\hat p_{i_gj} - p_{i_gj} < -\frac{\epsilon_\ell}{2}\Big\} \cap \Big\{n_{i_gj} = \frac{mt}{2k}\Big\}\bigg)\\
&\le \exp\Big(-2\frac{mt}{2k}(\epsilon_\ell/2)^2\Big)\Big) = \frac{\delta_\ell}{2k},
\end{aligned}
$$

where the third last inequality follows since in this case $p_{i_gj} > \frac{1}{2} + \epsilon_\ell$, and the last inequality follows from Lemma 1 with $\eta = \frac{\epsilon_\ell}{2}$ and $v = \frac{mt}{2k}$.

Let us define the event $\mathcal{E} := \Big\{\exists j \in \mathcal{G}_g$ such that $\hat p_{i_gj} + \frac{\epsilon_\ell}{2} < \frac{1}{2}, p_{i_gj} \in \big(\frac{1}{2}, \frac{1}{2}+\epsilon_\ell\big]$ or $\hat p_{i_gj} - \frac{\epsilon_\ell}{2} < \frac{1}{2}, p_{i_gj} > \frac{1}{2} + \epsilon_\ell\Big\}$. Then by combining Case 1 and 2, we get

$$
\begin{aligned}
Pr\Big(\mathcal{E}\Big) &= Pr\bigg(\mathcal{E} \cap \Big\{n_{i_gj} \ge \frac{mt}{2k}\Big\}\bigg) + Pr\bigg(\mathcal{E} \cap \Big\{n_{i_gj} < \frac{mt}{2k}\Big\}\bigg)\\
&\le \sum_{j \in \mathcal{G}_g \text{ s.t. } p_{i_gj} \in \left(\frac{1}{2}, \frac{1}{2}+\epsilon_\ell\right]} Pr\bigg(\Big\{\hat p_{i_gj} + \frac{\epsilon_\ell}{2} < \frac{1}{2}\Big\} \cap \Big\{n_{i_gj} \ge \frac{mt}{2k}\Big\}\bigg)\\
&+ \sum_{j \in \mathcal{G}_g \text{ s.t. } p_{i_gj} > \frac{1}{2}+\epsilon_\ell} Pr\bigg(\Big\{\hat p_{i_gj} - \frac{\epsilon_\ell}{2} < \frac{1}{2}\Big\} \cap \Big\{n_{i_gj} \ge \frac{mt}{2k}\Big\}\bigg) + Pr\bigg(\Big\{n_{i_gj} < \frac{mt}{2k}\Big\}\bigg)\\
&\le \frac{(k-1)\delta_\ell}{2k} + \frac{\delta_\ell}{2k} \le \delta_\ell
\end{aligned}
$$

where the last inequality follows from the above two case analyses and Lemma 19.

■

Given Lemma 20 in place, let us now analyze with what probability the algorithm can select a non ϵ_ℓ-optimal item $j \in \mathcal{G}_g$ as c_g at any iteration $\ell \in \lceil \frac{n}{k} \rceil$. For any set $\mathcal{G}_g$ (or set S for the last iteration $\ell = \lceil \frac{n}{k} \rceil$), we define the set of non ϵ_ℓ-optimal element $\mathcal{O}_g = \{j \in \mathcal{G}_g \mid p_{i_gj} > \frac{1}{2} + \epsilon_\ell\}$, and recall the event $\mathcal{E} := \Big\{\exists j \in \mathcal{G}_g$ such that $\hat p_{i_gj} + \frac{\epsilon_\ell}{2} < \frac{1}{2}, p_{i_gj} \in \big(\frac{1}{2}, \frac{1}{2}+\epsilon_\ell\big]$ or $\hat p_{i_gj} - \frac{\epsilon_\ell}{2} < \frac{1}{2}, p_{i_gj} > \frac{1}{2} + \epsilon_\ell\Big\}$. Then we have

$$
\begin{aligned}
Pr(c_g \in \mathcal{O}_g) &\le Pr\bigg(\Big\{\exists j \in \mathcal{G}_g, \hat p_{i_gj} + \frac{\epsilon_\ell}{2} < \frac{1}{2}\Big\} \cup \Big\{\exists j \in \mathcal{O}_g, \hat p_{ji_g} + \frac{\epsilon_\ell}{2} \ge \frac{1}{2}\Big\}\bigg)\\
&\le Pr\bigg(\mathcal{E} \cup \Big\{\exists j \in \mathcal{O}_g, \hat p_{ji_g} + \frac{\epsilon_\ell}{2} \ge \frac{1}{2}\Big\}\bigg)\\
&= Pr\Big(\mathcal{E}\Big) + Pr\bigg(\Big\{\exists j \in \mathcal{O}_g, \hat p_{ji_g} + \frac{\epsilon_\ell}{2} \ge \frac{1}{2}\Big\} \cap \mathcal{E}^c\bigg)\\
&= Pr\Big(\mathcal{E}\Big) + Pr\bigg(\Big\{\exists j \in \mathcal{O}_g, \hat p_{ji_g} + \frac{\epsilon_\ell}{2} \ge \frac{1}{2}\Big\} \cap \mathcal{E}^c\bigg) \le \delta_\ell + 0 = \delta_\ell, \qquad (14)
\end{aligned}
$$

where the last inequality follows from Lemma 20, and the fact that $\hat{p}_{i_g j} - \frac{\epsilon_\ell}{2} \geq \frac{1}{2} \implies \hat{p}_{j i_g} + \frac{\epsilon_\ell}{2} < \frac{1}{2}$. The proof now follows combining all the above parts together.

More formally, for each iteration ℓ, let us define $g_\ell \in [G]$ to be the set that contains *best item* of the entire set S, i.e. $\arg\max_{i \in S} \theta_i \in \mathcal{G}_{g_\ell}$. Then from (14), with probability at least $(1 - \delta_\ell)$, $\tilde{p}_{c_{g_\ell} i_{g_\ell}} > -\epsilon_\ell$. Now for each iteration ℓ, recursively applying (14) and Lemma 15 to $\mathcal{G}_{g_\ell}$, we get that $\tilde{p}_{r_* 1} > -\Big(\frac{\epsilon}{8} + \frac{\epsilon}{8}\Big(\frac{3}{4}\Big) + \cdots + \frac{\epsilon}{8}\big(\frac{3}{4}\big)^{\lfloor \frac{n}{k} \rfloor}\Big) + \frac{\epsilon}{2} \geq -\frac{\epsilon}{8}\Big(\sum_{i=0}^{\infty}\big(\frac{3}{4}\big)^i\Big) + \frac{\epsilon}{2} = \epsilon$. (Note that, for above analysis to go through, it is in fact sufficient to consider only the set of iterations $\{\ell \geq \ell_0 \mid \ell_0 = \min\{l \mid 1 \notin \mathcal{R}_l,\, l \geq 1\}\}$ because prior considering item 1, it does not matter even if the algorithm mistakes in any of the iteration $\ell < \ell_0$). Thus assuming the algorithm does not fail in any of the iteration ℓ, we have that $p_{r_* 1} > \frac{1}{2} - \epsilon$.

Finally, since at each iteration ℓ, the algorithm fails with probability at most δ_ℓ, the total failure probability of the algorithm is at most $\Big(\frac{\delta}{4} + \frac{\delta}{8} + \cdots + \frac{\delta}{2^{\lceil \frac{n}{k} \rceil}}\Big) + \frac{\delta}{2} \leq \delta$. This concludes the correctness of the algorithm showing that it indeed returns an ϵ-best element r_* such that $p_{r_* 1} \geq \frac{1}{2} - \epsilon$ with probability at least $1 - \delta$. ■