Best-item Learning in Random Utility Models with Subset Choices

Aadirupa Saha
aadirupa@iisc.ac.in
Indian Institute of Science
Bengaluru, India

Aditya Gopalan
aditya@iisc.ac.in
Indian Institute of Science
Bengaluru, India

Abstract

We consider the problem of PAC learning the most valuable item from a pool of n items using sequential, adaptively chosen plays of subsets of k items, when, upon playing a subset, the learner receives relative feedback sampled according to a general Random Utility Model (RUM) with independent noise perturbations to the latent item utilities. We identify a new property of such a RUM, termed the minimum advantage, that helps in characterizing the complexity of separating pairs of items based on their relative win/loss empirical counts, and can be bounded as a function of the noise distribution alone. We give a learning algorithm for general RUMs, based on pairwise relative counts of items and hierarchical elimination, along with a new PAC sample complexity guarantee of $O(\frac{n}{c^2\epsilon^2}\log\frac{k}{\delta})$ rounds to identify an ϵ-optimal item with confidence $1-\delta$, when the worst case pairwise advantage in the RUM has sensitivity at least c to the parameter gaps of items. Fundamental lower bounds on PAC sample complexity show that this is near-optimal in terms of its dependence on n, k and c.

1 Introduction

Random utility models (RUMs) are a popular and well-established framework for studying behavioral choices by individuals and groups (Thurstone, 1927). In a RUM with finite alternatives or items, a distribution on the preferred alternative(s) is assumed to arise from a random utility drawn from a distribution for each item, followed by rank ordering the items according to their utilities.

Perhaps the most widely known RUM is the Plackett-Luce or multinomial logit model (Plackett, 1975; Luce, 2012) which results when each item's utility is sampled from an additive model with a Gumbel-distributed perturbation. It is unique in the sense of enjoying the property of independence of irrelevant attributes (IIA), which is often key in permitting efficient inference of Plackett-Luce models from data (Khetan and Oh, 2016). Other well-known RUMs include the probit model (Bliss, 1934) featuring random Gaussian perturbations to the intrinsic utilities, mixed logit, nested logit, etc.

A long line of work in statistics and machine learning focuses on estimating RUM properties from observed data (Soufiani et al., 2014; Zhao et al., 2018; Soufiani et al., 2013). Online learning or adaptive testing, on the other hand, has shown efficient ways of identifying the most attractive (i.e., highest utility) items in RUMs by learning from relative feedback from item pairs or more generally subsets (Szörényi et al., 2015; Saha and Gopalan, 2019; Jang et al., 2017). However, almost all existing work in this vein exclusively employs the Plackett-Luce model, arguably due to its very useful IIA property, and our understanding of learning performance in other, more general RUMs has been lacking. We take a step in this direction by framing the problem of sequentially learning the best item/items in general RUMs by adaptive testing of item subsets and observing relative RUM feedback. In the process, we uncover new structural properties in RUMs, including models with exponential, uniform, Gaussian (probit) utility distributions, and give algorithmic principles to exploit this structure, that permit provably sample-efficient online learning and allow us to go beyond Plackett-Luce.

Our contributions: We introduce a new property of a RUM, called the (pairwise) *advantage ratio*, which essentially measures the worst-case relative probabilities between an item pair across all possible contexts

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

(subsets) where they occur. We show that this ratio can be controlled (bounded below) as an affine function of the relative strengths of item pairs for RUMs based on several common centered utility distributions, e.g., exponential, Gumbel, uniform, Gamma, Weibull, normal, etc., even when the resulting RUM does not possess analytically favorable properties such as IIA.

We give an algorithm for sequentially and adaptively PAC (probably approximately correct) learning the best item from among a finite pool when, in each decision round, a subset of fixed size can be tested and top-m rank ordered feedback from the RUM can be observed. The algorithm is based on the idea of maintaining pairwise win/loss counts among items, hierarchically testing subsets and propagating the surviving winners – principles that have been shown to work optimally in the more structured Plackett-Luce RUM (Szörényi et al., 2015; Saha and Gopalan, 2019).

In terms of performance guarantees, we derive a PAC sample complexity bound for our algorithm: when working with a pool of n items in total with subsets of size-k chosen in each decision round, the algorithm terminates in $O(\frac{n}{c^2\epsilon^2}\log\frac{k}{\delta})$ rounds where c is a lower bound on the advantage ratio's sensitivity to intrinsic item utilities. This can in turn be shown to be a property of only the RUM's perturbation distribution, independent of the subset size k. A novel feature of the guarantee is that, unlike existing sample complexity results for sequential testing in the Plackett-Luce model, it does not rely on specific properties like IIA which are not present in general RUMs. We also extend the result to cover top-m rank ordered feedback, of which winner feedback ($m = 1$) is a special case. Finally, we show that the sample complexity of our algorithm is order-wise optimal across RUMs having a given advantage ratio sensitivity c, by arguing an information-theoretic lower bound on the sample complexity of any online learning algorithm.

Our results and techniques represent a conceptual advance in the problem of online learning in general RUMs, moving beyond the Plackett-Luce model for the first time to the best of our knowledge.

Related Work: For classical multiarmed bandits setting, there is a well studied literature on PAC-arm identification problem (Even-Dar et al., 2006; Audibert and Bubeck, 2010; Kalyanakrishnan et al., 2012; Karnin et al., 2013; Jamieson et al., 2014), where the learner gets to see a noisy draw of absolute reward feedback of an arm upon playing a single arm per round. On the contrary, learning to identify the best item(s) with only relative preference information (ordinal as opposed to cardinal feedback) has seen steady progress since the introduction of the dueling bandit framework (Zoghi et al., 2013) with pairs of items (size-2 subsets) that can be played, and subsequent work on generalisation to broader models both in terms of distributional parameters (Yue and Joachims, 2009; Gajane et al., 2015; Ailon et al., 2014; Zoghi et al., 2015) as well as combinatorial subset-wise plays (Mohajer et al., 2017; González et al., 2017; Saha and Gopalan, 2018b; Sui et al., 2017). There have been several 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) and the Mallows model (Busa-Fekete et al., 2014a)]. Recent work has studied PAC-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), or the true ranking of the items (Busa-Fekete et al., 2014b; Falahatgar et al., 2017). Some of the recent works also extended the PAC-learning objective with relative subsetwise preferences (Saha and Gopalan, 2018a; Chen et al., 2017, 2018; Saha and Gopalan, 2019; Ren et al., 2018).

However, none of the existing work considers strategies to learn efficiently in general RUMs with subset-wise preferences and to the best of our knowledge we are the first to address this general problem setup. In a different direction, there has been work on batch (non-adaptive) estimation in general RUMs, e.g., (Zhao et al., 2018; Soufiani et al., 2013); however, this does not consider the price of active learning and the associated exploration effort required as we study here. A related body of literature lies in dynamic assortment selection, where the goal is to offer a subset of items to customers in order to maximise expected revenue, which has been studied under different choice models, e.g. Multinomial-Logit (Talluri and Van Ryzin, 2004), Mallows and mixture of Mallows (Désir et al., 2016a), Markov chain-based choice models (Désir et al., 2016b), single transition model (Nip et al., 2017) etc., but again each of this work addresses a given and a very specific kind of choice model, and their objective is more suited to regret minimization type framework where playing every item comes with a associated cost.

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 Random Utility-based Discrete Choice Models

A discrete choice model specifies the relative preferences of two or more discrete alternatives in a given set. Random Utility Models (RUMs) are a widely-studied class of discrete choice models; they assume a (non-random) ground-truth utility score $\theta_i \in \mathbb{R}$ for each alternative $i \in [n]$, and assign a distribution $\mathcal{D}_i(\cdot|\theta_i)$ for scoring item i, where $\mathbf{E}[\mathcal{D}_i \mid \theta_i] = \theta_i$. 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. More formally, the probability that an item $i \in S$ emerges as the *winner* in set S is given by:

$$Pr(i|S) = Pr(X_i > X_j \ \ \forall j \in S \setminus \{i\}) \qquad (1)$$

In this paper, we assume that for each item $i \in [n]$, its random *utility score* X_i is of the form $X_i = \theta_i + \zeta_i$, where all the $\zeta_i \sim \mathcal{D}$ are 'noise' random variables drawn independently from a probability distribution $\mathcal{D}$.

A 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$(0, 1)$ distributions with location parameters 0 and scale parameter 1 (Azari et al., 2012), which results in score distributions $Pr(X_i \in [x, x + dx]) = e^{-(x-\theta_i)}e^{-e^{-(x-\theta_i)}}dx, \forall i \in [n]$. Moreover, it can be shown that the probability that an alternative i emerges as the winner in any set $S \ni i$ is simply proportional to its score parameter: $Pr(i|S) = \frac{e^{\theta_i}}{\sum_{j \in S} e^{\theta_j}}$.

Other families of discrete choice models can be obtained by imposing different probability distributions over the iid noise $\zeta_i \sim \mathcal{D}$; e.g.,

1. *Exponential* noise: $\mathcal{D}$ is the Exponential(λ) distribution $(\lambda > 0)$.
2. Noise from *Extreme value distributions*: $\mathcal{D}$ is the Extreme-value-distribution(μ, σ, ξ) $(\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R})$. Many well-known distributions fall in this class, e.g., *Frechet, Weibull, Gumbel*. For instance, when $\chi = 0$, this reduces to the *Gumbel*(μ, σ) distribution.
3. *Uniform* noise: $\mathcal{D}$ is the (continuous) Uniform(a, b) distribution $(a, b \in \mathbb{R}, b > a)$.
4. *Gaussian* or Frechet, Weibull, Gumbel noise: $\mathcal{D}$ is the Gaussian(μ, σ) distribution $(\mu \in \mathbb{R}, \sigma > 0)$.
5. *Gamma* noise: $\mathcal{D}$ is the Gamma(k, ξ) distribution (where $k, \xi > 0$).

Other distributions $\mathcal{D}$ can alternatively be used for modelling the noise distribution , depending on desired tail properties, domain-specific information, etc.

Finally, we denote a RUM choice model, comprised of an instance $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ (with its implicit dependence on the noise distribution $\mathcal{D}$) along with a playable subset size $k \leq n$, by RUM$(k, \boldsymbol{\theta})$.

3 Problem Setting

We consider the probably approximately correct (PAC) version of the sequential decision-making problem of finding the best item in a set of n items, by making only subset-wise comparisons.

Formally, the learner is given a finite set $[n]$ of $n > 2$ items or 'arms'[1] along with a playable subset size $k \leq n$. 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 RUM$(k, \boldsymbol{\theta})$ 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. For the purposes of analysis, we assume, without loss of generality[2], that $\theta_1 > \theta_i \, \forall i \in [n] \setminus \{1\}$ for ease of exposition[3]. We define a *best item* to be one with the highest score parameter: $i^* \in \underset{i \in [n]}{\text{argmax}} \ \theta_i = \{1\}$, under the assumptions above.

Remark 1. *Under the assumptions above, it follows that item* 1 *is the* Condorcet Winner *(Zoghi et al., 2014) for the underlying pairwise preference model induced by RUM*$(k, \boldsymbol{\theta})$.

3.1 Feedback models

We mean by 'feedback model' the information received (from the 'environment') once the learner plays a subset $S \subseteq [n]$ of k items. Similar to different types of feedback models introduced earlier in the context of the specific Plackett-Luce RUM (Saha and Gopalan, 2019), we consider the following feedback mechanisms:

- **Winner of the selected subset (WI:** The environment returns a single item $I \in S$, drawn

[1] terminology borrowed from multi-armed bandits

[2] under the assumption that the learner's decision rule does not contain any bias towards a specific item index

[3] The extension to the case where several items have the same highest parameter value is easily accomplished.

independently from the probability distribution $Pr(I = i|S) = Pr(X_i > X_j,\ \forall j \in S \setminus \{i\})\ \ \forall i \in S,\ S \subseteq [n]$.

- **Full ranking selected subset of items (FR):** The environment returns a full ranking $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_S$, drawn from the probability distribution $Pr(\boldsymbol{\sigma} = \sigma|S) = \prod_{i=1}^{|S|} Pr(X_{\sigma(i)} > X_{\sigma(j)},\ \forall j \in \{i+1, \dots |S|\}), \forall \boldsymbol{\sigma} \in \boldsymbol{\Sigma}_S$. In fact, this is equivalent to picking $\boldsymbol{\sigma}(1)$ according to the winner feedback from S, then picking $\boldsymbol{\sigma}(2)$ from $S \setminus \{\boldsymbol{\sigma}(1)\}$ following the same feedback model, and so on, until all elements from S are exhausted, or, in other words, successively sampling $|S|$ winners from S according to the RUM$(k, \boldsymbol{\theta})$ model, without replacement.

3.2 PAC Performance Objective: Correctness and Sample Complexity

For a RUM$(k, \boldsymbol{\theta})$ instance with $n \geq k$ arms, an arm $i \in [n]$ is said to be ϵ-optimal if $\theta_i > \theta_1 - \epsilon$. A sequential[4] learning algorithm that depends on feedback from an appropriate subset-wise feedback model is said to be (ϵ, δ)-PAC, for given constants $0 < \epsilon \leq \frac{1}{2}, 0 < \delta \leq 1$, if the following properties hold when it is run on any instance RUM$(k, \boldsymbol{\theta})$: (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 in RUM$(k, \boldsymbol{\theta})$ 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 when run on the instance RUM$(k, \boldsymbol{\theta})$.

4 Connecting Subsetwise preferences to Pairwise Scores

In this section, we introduce the key concept of Advantage ratio as a means to systematically relate subsetwise preference observations to pairwise scores in general RUMs.

Consider any set $S \subseteq [n]$, $|S| = k$, and recall that the probability of item i winning in S is $Pr(i|S) := Pr(X_i > X_j,\ \forall j \in [n] \setminus \{i\})$ for all $i \in S,\ S \subseteq [n]$. For any two items $i, j \in [n]$, let us denote $\Delta_{ij} = (\theta_i - \theta_j)$. Let us also denote by $f(\cdot), F(\cdot)$ and $\bar{F}(\cdot)$ the probability density function[5], cumulative distribution function and complementary cumulative distribution function of the noise distribution $\mathcal{D}$, respectively; thus, $F(x) = \int_{-\infty}^{x} f(x)dx$ for any $x \in \text{Support}(\mathcal{D})$ and $\bar{F}(x) = \int_{x}^{\infty} f(x)dx = 1 - F(x)$ for any $x \in \text{Support}(\mathcal{D})$.

We now introduce and analyse the *Advantage-Ratio* (Def. 1); we will see in Sec. 5.1 how this quantity helps us deriving an improved sample complexity guarantee for our (ϵ, δ)-PAC item identification problem.

Definition 1 (Advantage ratio and Minimum advantage ratio)**.** *Given any subsetwise preference model defined on n items, we define the advantage ratio of item i over item j within the subset $S \subseteq [n]$, $i, j \in S$ as Advantage-Ratio$(i, j, S) = \frac{Pr(i|S)}{Pr(j|S)}$.*

Moreover, given a playable subset size k, we define the minimum advantage ratio, Min-AR, of item-i over j, as the least advantage ratio of i over j across size-k subsets of $[n]$, i.e.,

$$\textit{Min-AR}(i, j) = \min_{S \subseteq [n], |S|=k, S \ni i,j} \frac{Pr(i|S)}{Pr(j|S)}. \tag{2}$$

The key intuition here is that when *Min-AR*(i, j) does not equal 1, it serves as a distinctive measure for identifying item i and j separately irrespective of the context S. We specifically build on this intuition later in Sec. 5.1 to propose a new algorithm (Alg. 1) which finds the (ϵ, δ)-PAC best item relying on the unique distinctive properly of the best-item $\theta_1 > \theta_j \forall j \in [n] \setminus \{1\}$ (as described in Sec. 3).

The following result shows a variational lower bound, in terms of the noise distribution, for the minimum advantage ratio in a RUM$(k, \boldsymbol{\theta})$ model with independent and identically distributed (iid) noise variables, that is often amenable to explicit calculation/bounding.

Lemma 2 (Variational lower bound for the advantage ratio)**.** *For any RUM$(k, \boldsymbol{\theta})$ based subsetwise preference model and any item pair (i, j),*[6]

$$\textit{Min-AR}(i, j) \geq \min_{z \in \mathbb{R}} \frac{Pr\big(X_i > \max(X_j, z)\big)}{Pr\big(X_j > \max(X_i, z)\big)}. \tag{3}$$

Moreover for RUM$(k, \boldsymbol{\theta})$ models one can show that for any triplet (i, j, S), $Pr\big(X_i > \max(X_j, z)\big) = F(z - \theta_j)\bar{F}(z - \theta_i) + \int_{z-\theta_j}^{\infty} \bar{F}(x - \Delta_{ij}) f(x)dx$, which further lower bounds Min-AR(i, j) by:

$$\min_{z \in \mathbb{R}} \frac{F(z - \theta_j)\bar{F}(z - \theta_i) + \int_{z-\theta_j}^{\infty} \bar{F}(x - \Delta_{ij}) f(x)dx}{F(z - \theta_i)\bar{F}(z - \theta_j) + \int_{z-\theta_i}^{\infty} \bar{F}(x + \Delta_{ij}) f(x)dx}.$$

The proof of the result appears in Appendix A.1. Fig. 1 shows a geometrical interpretation behind *Min-AR*(i, j), under the joint realization of the pair of values (ζ_i, ζ_j).

[4] We essentially mean a causal algorithm that makes present decisions using only past observed information at each time; the technical details for defining this precisely are omitted.

[5] We assume by default that all noise distributions have a density; the extension to more general noise distributions is left to future work.

[6] We assume $\frac{0}{0}$ to be ∞ in the right hand side of Eqn. 3.

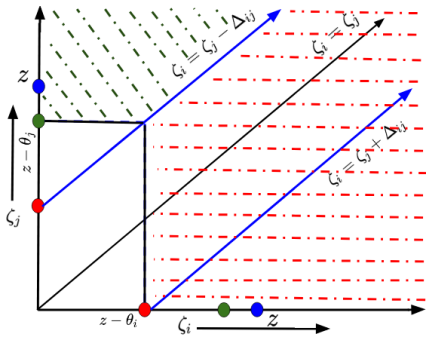

Figure 1: A two-dimensional geometrical interpretation for the quantity *Min-AR*(i,j). Let $z \in R$ be a random variable denoting the max score observed for the rest of the items, i.e. $\max_{a \in S\setminus\{i,j\}} X_a$. Let the blue, green and red dot respectively denote the position of z, $z-\theta_j$ and $z-\theta_i$. With $X_i = \theta_i + \zeta_i$, $\forall i \in [n]$, the green shaded region is where $X_j > \max(X_i, z)$, the red shaded region is where $X_i > \max(X_j, z)$ i.e. item i is the winner, and the white rectangle is where $\max(X_i, X_j) < z$ i.e. some other item wins. The shape of the green and red region varies as z moves on $\mathbb{R}$ (in the hindsight this basically covers the realizations of all z over all possible subsets S)–*Min-AR*(i,j) is attained at the particular z where the ratio of the mass of the red and green region is minimized (see Eqn. (3) for details).

Remark 2. *Suppose* $\bar{S} := \arg\min_{|S|=k, i,j \in S} \frac{Pr(i|S)}{Pr(j|S)}$. *It is sufficient to consider the domain of* z *in the right hand side of (3) to be just the set* $\max_{r \in \bar{S}\setminus\{i,j\}} \theta_r$ + *support*$(\mathcal{D})$, *as the proof of Lemma 2 brings out. However, for simplicity we use a smaller lower bound in Eqn. 3 and take* $z \in R$.

We next derive the *Min-AR*(i,j) values certain specific noise distributions:

Lemma 3 (Analysing *Min-AR* for specific noise models)**.** *Given a fixed item pair* (i,j) *such that* $\theta_i > \theta_j$, *the following bounds hold under the respective noise models in an iid RUM.*

1. *Exponential*(λ)*:* *Min-AR*$(i,j) \geq e^{\Delta_{ij}} > 1 + \Delta_{ij}$ *for Exponential noise with* $\lambda = 1$.
2. *Extreme value distribution*(μ, σ, χ)*:* *For Gumbel*(μ, σ) $(\chi = 0)$ *noise,* *Min-AR*$(i,j) = e^{\frac{\Delta_{ij}}{\sigma}} > 1 + \frac{\Delta_{ij}}{\sigma}$.
3. *Uniform*(a,b)*:* *Min-AR*$(i,j) \geq 1 + \frac{2\Delta_{ij}}{b-a}$ *for Uniform*(a,b) *noise* $(a, b \in \mathbb{R}, b > a,$ *and* $\Delta_{ij} < \frac{a}{2})$.
4. *Gamma*(k, ξ)*:* *Min-AR*$(i,j) \geq 1 + \Delta_{ij}$ *for Gamma*$(2,1)$ *noise.*
5. *Weibull*(λ, k)*:* *Min-AR*$(i,j) \geq e^{\lambda\Delta_{ij}} > 1 + \lambda\Delta_{ij}$ *for* $(k = 1)$.
6. *Normal* $\mathcal{N}(0,1)$*:* $\exists c > 0$ *such that, for* Δ_{ij} *small enough (in a neighborhood of* 0*),* *Min-AR*$(i,j) \geq 1 + c\Delta_{ij}$.

The proof appears in Appendix A.2.

5 An optimal algorithm for the winner feedback model

In this section, we propose an algorithm (*Sequential-Pairwise-Battle*, Algorithm 1) for the (ϵ, δ)-PAC objective with winner feedback. We then analyse its correctness and sample complexity guarantee (Theorem 4) for any noise distribution $\mathcal{D}$ (under a mild assumption of its being *Min-AR* bounded away from 1). Following this, we also prove a matching lower bound for the problem which shows that the sample complexity of Algorithm *Sequential-Pairwise-Battle* is unimprovable (up to a factor of $\log k$).

5.1 The *Sequential-Pairwise-Battle* algorithm

Our algorithm is based on the simple idea of dividing the set of n items into sub-groups of size k, querying each subgroup 'sufficiently enough', retaining thereafter only the empirically 'strongest item' of each sub-group, and recursing on the remaining set of items until only one item remains.

More specifically, it starts by partitioning the initial item pool into $G := \lceil \frac{n}{k} \rceil$ mutually exclusive and exhaustive sets $\mathcal{G}_1, \mathcal{G}_2, \cdots \mathcal{G}_G$ such that $\cup_{j=1}^{G} \mathcal{G}_j = S$ and $\mathcal{G}_j \cap \mathcal{G}_{j'} = \emptyset$, $\forall j, j' \in [G]\, |G_j| = k$, $\forall j \in [G-1]$. Each set $\mathcal{G}_g$, $g \in [G]$ is then queried for $t = O\Big(\frac{k}{\epsilon_\ell^2} \ln \frac{k}{\delta_\ell}\Big)$ rounds, and only the 'empirical winner' c_g of each group g is retained in a set S, rest are discarded. The algorithm next recurses the same procedure on the remaining set of surviving items, until a single item is left, which then is declared to be the (ϵ, δ) PAC-best item. Algorithm 1 presents the pseudocode in more detail.

Key idea: The primary novelty here is how the algorithm reasons about the 'strongest item' in each sub-group $\mathcal{G}_g$: It maintains the pairwise preferences of every item pair (i,j) in any sub-group $\mathcal{G}_g$ and simply chooses the item that beats the rest of the items in the sub-group with a positive advantage of greater than $\frac{1}{2}$ (alternatively, the item that wins maximum number of subset-wise plays). Our idea of maintaining pairwise preferences is motivated by a similar algorithm proposed in (Saha and Gopalan, 2019); however, their performance guarantee applies to only the very specific class of Plackett-Luce feedback models, whereas the novelty of our current analysis reveals the

Algorithm 1 *Sequential-Pairwise-Battle*(Seq-PB)

Input:
  Set of items: $[n]$, Subset size: $n \geq k > 1$
  Error bias: $\epsilon > 0$, Confidence parameter: $\delta > 0$
  Noise model ($\mathcal{D}$) dependent constant $c > 0$
Initialize:
  $S \leftarrow [n]$, $\epsilon_0 \leftarrow \frac{c\epsilon}{8}$, 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 $t := \lceil \frac{k}{2\epsilon_\ell^2} \ln \frac{k}{\delta_\ell} \rceil$ 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 $S \leftarrow S \cup S'$, $\epsilon_\ell \leftarrow \frac{c\epsilon}{2}$, $\delta_\ell \leftarrow \delta$
  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: The unique item left in S

power of maintaining pairwise-estimates for more general RUM$(k, \boldsymbol{\theta})$ subsetwise model (which includes the Plackett-Luce choice model as a special case). The pseudo code of *Sequential-Pairwise-Battle* is given in Alg. 1.

The following is our chief result; it proves correctness and a sample complexity bound for Algorithm 1.

Theorem 4 (*Sequential-Pairwise-Battle*: Correctness and Sample Complexity)**.** *Consider any iid subsetwise preference model RUM$(k, \boldsymbol{\theta})$ based on a noise distribution $\mathcal{D}$, and suppose that for any item pair i, j, we have Min-AR$(i, j) \geq 1 + \frac{4c\Delta_{ij}}{1-2c}$ for some $\mathcal{D}$-dependent constant $c > 0$. Then, Algorithm 1, with input constant $c > 0$, is an (ϵ, δ)-PAC algorithm with sample complexity $O(\frac{n}{c^2\epsilon^2} \log \frac{k}{\delta})$.*

The proof of the result appears in Appendix B.1.

Remark 3. *The linear dependence on the total number of items, n, is, in effect, indicates the price to pay for learning the n unknown model parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ which decide the subsetwise winning probabilities of the n items. Remarkably, however, the theorem shows that the PAC sample complexity of the (ϵ, δ)-best item identification problem, with only winner feedback information from k-size subsets, is independent of k (except some mild logarithmic dependencies). One may expect to see improved sample complexity as the number of items being simultaneously tested in each round is large ($k \geq 2$), but note that on the other side, the sample complexity could also worsen, since it is also harder for a good item to win and show itself in a few draws against a large population of $k-1$ other competitors – these effects roughly balance each other out, and the final sample complexity only depends on the total number of items n and the accuracy parameters (ϵ, δ).*

Note that Lemma 3 gives specific values of the noise-model $\mathcal{D}$ dependent constant $c > 0$, using which we can derive specific sample complexity bounds for certain noise models:

Corollary 5 (Model specific correctness and sample complexity guarantees)**.** *For the following representative noise distributions: Exponential(1), Gumbel(μ, σ) Gamma$(2, 1)$, Uniform(a, b), Weibull$(\lambda, 1)$, Standard normal or Normal$(0, 1)$, Seq-PB (Alg.1) finds an (ϵ, δ)-PAC item within sample complexity $O\big(\frac{n}{\epsilon^2} \ln \frac{k}{\delta}\big)$.*

Proof sketch. The proof follows from the general performance guarantee of Seq-PB (Thm. 4) and Lem. 3. More specifically from Lem. 3 it follows that the value of c for these specific distributions are constant, which concludes the claim. For completeness the distribution-specific values of c are given in Appendix B.2. □

5.2 Sample Complexity Lower Bound

In this section we derive a sample complexity lower bound for any (ϵ, δ)-PAC algorithm for any RUM$(k, \boldsymbol{\theta})$ model with *Min-AR*(i, j) strictly bounded away from 1 in terms of Δ_{ij}. Our formal claim goes as follows:

Theorem 6 (Sample Complexity Lower Bound for RUM$(k, \boldsymbol{\theta})$ model)**.** *Given $\epsilon \in (0, \frac{1}{4}]$, $\delta \in (0, 1]$, $c > 0$ and an (ϵ, δ)-PAC algorithm A with winner item feedback, there exists a RUM$(k, \boldsymbol{\theta})$ instance ν with Min-AR$(i, j) \geq 1 + 4c\Delta_{ij}$ for all $i, j \in [n]$, where the expected sample complexity of A on ν is at least $\Omega\big(\frac{n}{c^2\epsilon^2} \ln \frac{1}{2.4\delta}\big)$.*

The proof is given in Appendix B.3. It essentially involves a change of measure argument demonstrating a family of Plackett-Luce models (iid Gumbel noise), with the appropriate c value, that cannot easily be teased apart by any learning algorithm.

Comparing this result with the performance guarantee

of our proposed algorithm (Theorem 6) shows that the sample complexity of the algorithm is order-wise optimal (up to a $\log k$ factor). Moreover, this result also shows that the IIA (independence of irrelevant attributes) property of the Plackett-Luce choice model is not essential for exploiting pairwise preferences via rank breaking, as was claimed in (Saha and Gopalan, 2019). Indeed, except for the case of *Gumbel* noise, none of the RUM$(k, \boldsymbol{\theta})$ based models in Corollary 5 satisfies IIA, but they all respect the $O\Big(\frac{n}{\epsilon^2}\ln\frac{1}{\delta}\Big)$ (ϵ,δ)-PAC sample complexity guarantee.

Remark 4. *For constant $c = O(1)$, the fundamental sample complexity bound of Theorem 6 resembles that of PAC best arm identification in the standard multi-armed bandit (MAB) problem (Even-Dar et al., 2006). Recall that our problem objective is exactly same as MAB, however our feedback model is very different since in MAB, the learner gets to see the noisy rewards/scores (i.e. the exact values of X_i, which can be seen as a noisy feedback of the true reward/score θ_i of item-i), whereas here the learner only sees a k-wise relative preference feedback based on the underlying observed values of X_i, which is a more indirect way of giving feedback on the item scores, and thus intuitively our problem objective is at least as hard as that of MAB setup.*

6 Results for Top-m Ranking (TR) feedback model

We now address our (ϵ,δ)-PAC item identification problem for the case of more general, top-m rank ordered feedback for the RUM$(k, \boldsymbol{\theta})$ model, that generalises both the winner-item (WI) and full ranking (FR) feedback models.

Top-m ranking of items (TR-m): In this feedback setting, the environment is assumed to return a ranking of only m items from among S, i.e., the environment first draws a full ranking $\boldsymbol{\sigma}$ over S according to RUM$(k, \boldsymbol{\theta})$ 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 $Pr(\boldsymbol{\sigma} = \sigma|S) = \prod_{i=1}^{m} Pr(X_{\sigma(i)} > X_{\sigma(j)},\ \forall j \in \{i{+}1, \ldots m\})$, $\forall \boldsymbol{\sigma} \in \boldsymbol{\Sigma}_S^m$, where by $\boldsymbol{\Sigma}_S^m$ we denote the set of all possible m-length ranking of items in set S, it is easy to note that $|S| = \binom{k}{m}m!$. Thus, 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 $m = 1$. Note that the idea for top-m ranking feedback was introduced by (Saha and Gopalan, 2018a) but only for the specific Plackett Luce choice model.

6.1 Algorithm for top-m ranking feedback

In this section, we extend the algorithm proposed earlier (Alg. 1) to handle feedback from the general top-m ranking feedback model. We also show that we can achieve an $\frac{1}{m}$-factor improved sample complexity rate with top-m ranking feedback (Thm. 7). We finally give a fundamental sample complexity bound (Thm. 8), which shows the optimality of our proposed algorithm mSeq-PB up to logarithmic factors.

Main idea: Same as Seq-PB, the algorithm proposed in this section (Alg. 2) in principle follows the same sequential elimination based strategy to find the near-best item of the RUM$(k, \boldsymbol{\theta})$ model based on pairwise preferences. However, we use the idea of *rank breaking* (Soufiani et al., 2014; Saha and Gopalan, 2018a) to extract the pairwise preferences: 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, then 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: $\{(b \succ a), (b \succ c), (b \succ d), (a \succ c), (a \succ d), (c \succ d)\}$ etc.

As a whole, our new algorithm now again divides the set of n items into small groups of size k, say $\mathcal{G}_1, \ldots \mathcal{G}_G$, $G = \lceil \frac{n}{k} \rceil$, and play each sub-group some $t = O\Big(\frac{k}{m\epsilon^2}\ln\frac{1}{\delta}\Big)$ many rounds. Inside any fixed sub-group $\mathcal{G}_g$, after each round of play, it uses *Rank-Breaking* on the top-m ranking feedback $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_{\mathcal{G}_g}^m$, to extract out $\binom{m}{2} + (k-m)m$ many pairwise feedback, which is further used to estimate the empirical pairwise preferences $\hat{p}_{ij}$ for each pair of items $i, j \in \mathcal{G}_g$. Based on these pairwise estimates it then only retains the strongest item of $\mathcal{G}_g$ and recurse the same procedure on the set of surviving items, until just one item is left in the set. The complete algorithm is given in Alg. 2 (Appendix C.1).

Theorem 7 analyses the correctness and sample complexity bounds of mSeq-PB. Note that the sample complexity bound of mSeq-PB with top-m ranking (TR) feedback model is $\frac{1}{m}$-times that of the WI model (Thm. 4). This is justified since intuitively revealing a ranking on m items in a k-set provides about m many WI feedback per round, which essentially leads to the m-factor improvement in the sample complexity.

Theorem 7 (mSeq-PB(Alg. 2): Correctness and Sample Complexity)**.** *Consider any RUM$(k, \boldsymbol{\theta})$ subsetwise preference model based on noise distribution $\mathcal{D}$ and suppose for any item pair i, j, we have Min-AR$(i, j) \geq 1 + \frac{4c\Delta_{ij}}{1-2c}$ for some $\mathcal{D}$-dependent constant $c > 0$. Then mSeq-PB (Alg.2) with input constant*

$c > 0$ on top-m ranking feedback model is an (ϵ, δ)-PAC algorithm with sample complexity $O(\frac{n}{mc^2\epsilon^2} \log \frac{k}{\delta})$.

(Proof is given in Appendix C.2.) Similar to Cor. 5, for the top-m model again, we can derive specific sample complexity bounds for different noise distributions, e.g., *Exponential*, *Gumbel*, *Gaussian*, *Uniform*, *Gamma* etc., in this case as well.

6.2 Lower Bound: Top-m ranking feedback

In this section, we analyze the fundamental limit of sample complexity lower bound for any (ϵ, δ)-PAC algorithm for RUM$(k, \boldsymbol{\theta})$ model.

Theorem 8 (Sample Complexity Lower Bound for RUM$(k, \boldsymbol{\theta})$ model with TR-m feedback)**.** *Given $\epsilon \in (0, \frac{1}{4}]$ and $\delta \in (0, 1]$, and an (ϵ, δ)-PAC algorithm A with winner item feedback, there exists a RUM$(k, \boldsymbol{\theta})$ instance ν, in which for any pair $i, j \in [n]$ Min-AR$(i, j) \geq 1 + 4c\Delta_{ij}$, where the expected sample complexity of A on ν with top-m ranking feedback has to be at least $\Omega\left(\frac{n}{mc^2\epsilon^2} \ln \frac{1}{2.4\delta}\right)$ for A to be (ϵ, δ)-PAC.*

(The proof is given in Appendix C.3.) Similar to the case of winner feedback, comparing Theorem 7 with the above result shows that the sample complexity of mSeq-PB is orderwise optimal (up to logarithmic factors), for general case of top-m ranking feedback as well.

7 Experiments

To complement our theoretical guarantees, we carry out some empirical simulations, as detailed below.

RUM models. We use the following 4 different noise models: **1.** Gumbel$(0, 1)$, **2.** Normal$(0, 1)$, **3.** Uniform$(0, 1)$, **4.** Exponential(1).

Utility Scores. Towards modelling different RUM based choice models, we combine the above noise models with the following 4 different ground utility scores $(\boldsymbol{\theta})$: **1.** b1: $\theta_1 = 0.8$, $\theta_i = 0.6$, otherwise. **2.** g1: $\theta_1 = 0.8$, $\theta_i = 0.2$, otherwise. **3.** geo: $\theta_1 = 1, \frac{\theta_{i+1}}{\theta_i} = 0.9$, $\forall i \in [n]$. **4.** arith: $\theta_1 = 1, \theta_i - \theta_{i+1} = 0.01$, $\forall i \in [n]$. with respectively $n = 8, 16, 50$ and 100 items.

All reported performances are averaged across 50 runs. To the best of our knowledge no known algorithm address our problem setup for general RUM models, unfortunately we could not compare our method (Alg. 1 and 2) with any baseline. Given the above setup, we run two types the experiments to investigate:

Success probability $(1 - \delta)$ (i.e. rate of correctness) vs sample complexity. We set $k = \frac{n}{2}, m = \frac{k}{2}$, $\epsilon = \min_{ij} |\theta_i - \theta_j|$ (i.e. the minimum pairwise gap among the utility scores) for each different environment. As expected from Thm 4, 6—with higher sample complexity, the success probability $1 - \delta$ goes to 1 for each noise model showing that the algorithm is almost always correct with sufficient queries; similarly for lower number of observations the algorithms errors too often.

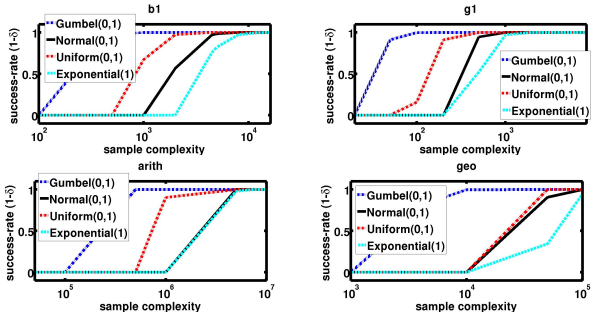

Figure 2: Success probability $(1 - \delta)$ vs sample complexity of Alg. 1 on different utility score-noise model combination

Sample complexity vs length of rank-ordered feedback (m). We run these expeiments on the *geo dataset*. Fig. 3 shows that the sample complexity seem to scale as $O(\frac{1}{m})$ while ϵ, δ is kept fixed to 0.1 (validating the claim from Thm. 7).

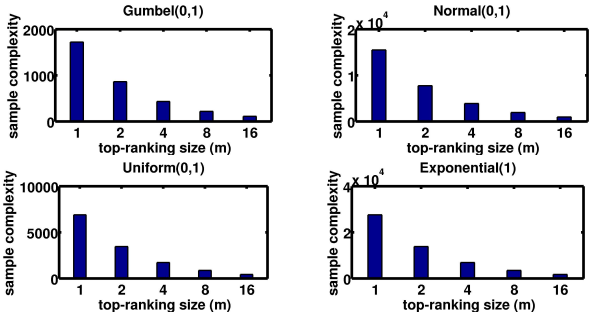

Figure 3: Sample complexity vs length of rank-ordered feedback (m) of Alg. 2 on *geo* utility score for different RUM models

8 Conclusion and Future Directions

We have identified a new principle to learn with general subset-size preference feedback in general iid RUMs – rank breaking followed by pairwise comparisons. This is by extending the concept of pairwise advantage from the Plackett-Luce (PL) choice model to more general RUMs, and by showing that the IIA property that PL models enjoy is not essential for optimal sample complexity. Several interesting directions exist for future investigation, e.g., considering correlated noise models (more general RUMs), explicitly modeling item features or attributes, other metrics like regret for online utility optimization, and relative preference learning in time-correlated Markov Decision Processes.

Acknowledgements. This work was supported by the Qualcomm Innovation Fellowship IND-417067, 2019, and by a grant from the Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science.

References

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

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p.

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

Bliss, C. I. (1934). The method of probits. *Science*.

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

Busa-Fekete, R., Szorenyi, B., Cheng, W., Weng, P., and Hüllermeier, E. (2013). Top-k selection based on adaptive sampling of noisy preferences. In *International Conference on Machine Learning*, pages 1094–1102.

Busa-Fekete, R., Szörényi, B., and Hüllermeier, E. (2014b). Pac rank elicitation through adaptive sampling of stochastic pairwise preferences. In *AAAI*, pages 1701–1707.

Chen, X., Gopi, S., Mao, J., and Schneider, J. (2017). 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.

Chen, X., Li, Y., and Mao, J. (2018). 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.

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

Désir, A., Goyal, V., Segev, D., and Ye, C. (2016b). Capacity constrained assortment optimization under the markov chain based choice model. *Operations Research*.

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

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

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

González, J., Dai, Z., Damianou, A., and Lawrence, N. D. (2017). Preferential Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1282–1291. JMLR. org.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil' ucb : An optimal exploration algorithm for multi-armed bandits. In Balcan, M. F., Feldman, V., and Szepesvari, C., editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 423–439. PMLR.

Jang, M., Kim, S., Suh, C., and Oh, S. (2017). Optimal sample complexity of m-wise data for top-k ranking. In *Advances in Neural Information Processing Systems*, pages 1685–1695.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246.

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

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

Luce, R. D. (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Mohajer, S., Suh, C., and Elmahdy, A. (2017). Active learning for top-k rank aggregation from noisy comparisons. In *International Conference on Machine Learning*, pages 2488–2497.

Nip, K., Wang, Z., and Wang, Z. (2017). Assortment optimization under a single transition model.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202.

Popescu, P. G., Dragomir, S., Slusanschi, E. I., and Stanasila, O. N. (2016). Bounds for Kullback-Leibler divergence. *Electronic Journal of Differential Equations*, 2016.

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

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

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

Saha, A. and Gopalan, A. (2019). PAC Battling Bandits in the Plackett-Luce Model. In *Algorithmic Learning Theory*, pages 700–737.

Soufiani, H. A., Diao, H., Lai, Z., and Parkes, D. C. (2013). Generalized random utility models with multiple types. In *Advances in Neural Information Processing Systems*, pages 73–81.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2014). Computing parametric ranking models via rank-breaking. In *ICML*, pages 360–368.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. (2017). Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*.

Szörényi, B., Busa-Fekete, R., Paul, A., and Hüllermeier, E. (2015). Online rank elicitation for plackett-luce: A dueling bandits approach. In *Advances in Neural Information Processing Systems*, pages 604–612.

Talluri, K. and Van Ryzin, G. (2004). Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4):273.

Urvoy, T., Clerot, F., Féraud, R., and Naamane, S. (2013). Generic exploration and k-armed voting bandits. In *International Conference on Machine Learning*, pages 91–99.

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

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

Zhao, Z., Villamil, T., and Xia, L. (2018). Learning mixtures of random utility models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Zoghi, M., Whiteson, S., Munos, R., and de Rijke, M. (2013). Relative upper confidence bound for the k-armed dueling bandit problem. *arXiv preprint arXiv:1312.3393*.

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