Efficient and Accurate Top-K Recovery from Choice Data

Abstract

The intersection of learning to rank and choice modeling is an active area of research with applications in e-commerce, information retrieval and the social sciences. In some applications such as recommendation systems, the statistician is primarily interested in recovering the set of the top ranked items from a large pool of items as efficiently as possible using passively collected *discrete choice data*, i.e., the user picks one item from a set of multiple items. Motivated by this practical consideration, we propose *the choice-based Borda count algorithm* as a fast and accurate ranking algorithm for *top K-recovery* i.e., correctly identifying all of the top K items. We show that the choice-based Borda count algorithm has optimal sample complexity for top-K recovery under a broad class of *random utility models*. We prove that in the limit, the choice-based Borda count algorithm produces the same top-K estimate as the commonly used Maximum Likelihood Estimate method but the former's speed and simplicity brings considerable advantages in practice. Experiments on both synthetic and real datasets show that the counting algorithm is competitive with commonly used ranking algorithms in terms of accuracy while being several orders of magnitude faster.

1 INTRODUCTION

The research on discrete choice modeling and learning to rank has received a lot of interest in recent years thanks to the growing availability of discrete choice data generated by e-commerce platforms, search engines and the social sciences. In discrete choice data, when presented with a set of items, also referred to as *menu*, the user picks the most preferred item. Discrete choice data is an intermediate between pairwise comparison data and full ranking data. In many settings such as e-commerce and political surveys, a large quantity of passively collected data is in the form of discrete choice data, e.g., consumers choosing to buy a product when presented with a catalogue of items, voters picking a favorite candidate from a pool of candidates.

In this paper, we focus on the problem of learning to rank using choice data. Specifically, we are interested in the top-K recovery problem, i.e., identifying the set of the top K items out of a universe of n items, using *passively collected choice data*. This problem has many useful applications. For example, in e-commerce applications, marketers are interested in finding the set of the best items based on how consumers make purchasing decisions. In the social sciences, political scientists are interested in determining the most preferred candidates among a pool of candidates using survey questionnaires.

To ground our theoretical discussions, we posit that the choice data is generated according to a probabilistic choice model- when presented with a menu of items S, the user makes a non-deterministic decision, picking a single item i from S with some probability $p_{i|S}$. More specifically, we assume our choice model falls within the class of Random Utility Models with Independently and Identically Distributed noise (IID-RUMs), described in detail in section (2). IID-RUMs are an expressive and flexible framework that can be used to model pairwise comparison data, discrete choice data as well as full ranking data. For example, the Multinomial Logit (MNL) model is one of the most commonly used IID-RUMs to model discrete choice data [Train, 2009].

Our motivation: While expressive, random utility models also pose hard computational problems. For example, many models within the class of IID-RUMs with the few exceptions such as the MNL model do not admit analytical expression for the choice probabilities (while the pairwise comparison probabilities can be evaluated easily), limiting inference to MCMC-based algorithms. However, sampling-based algorithms can be time inefficient when running on

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

large choice datasets with many items and menus. Furthermore, most classical inference algorithms assume a parametric model generating the choice data. In practice, it is often hard to verify if the data comes from a specific parametric model. Therefore, developing efficient ranking algorithms that are robust to model misspecification is of timely interest.

Motivated by these considerations, we study the generalization of a simple yet powerful counting algorithm for ranking- Borda count - to the discrete choice setting. The Borda count algorithm itself has a long history, dating back to the 18th century and its analysis has been instantiated in various contexts such ranking from pairwise comparisons in Rajkumar and Agarwal [2014], Shah and Wainwright [2017]. Our work, however, is the first to study the theoretical guarantees of Borda count in the *discrete choice setting* under a broad class of discrete choice models.

Our contributions:

- In section (4) and section (5): We show that the choice-based Borda count algorithm needs $\theta(n \log n)$ samples in order to exactly recover all of the top K items using choice data. We further show that this sample complexity is optimal for a broad class of IID-RUMs. This hinges on a fundamental property shared by many IID-RUMs which we term *Borda consistency*.
- In section (6): We study the effect of the menu size m on the sample complexity for top K recovery. For the special case of the MNL model, which is a common assumption in the ranking literature, we present an asymptotic characterization of the optimal sample complexity for top K recovery in terms of m. This bound monotonically decreases, but at a decreasing rate, with m. This suggests that there is a benefit to increasing the menu size but such benefit comes with diminishing returns. To the best of our knowledge, this result is the first of its kind in the choice modeling and ranking literature.
- In section (7): We study the connections between the choice-based Borda count algorithm and two commonly used top-K recovery algorithms: Maximum Likelihood Estimate under MNL assumption (MNL-MLE) and Spectral Ranking [Negahban et al., 2017, Maystre and Grossglauser, 2015, Agarwal et al., 2018]. We prove that the choice-based Borda count algorithm and MNL-MLE produce the same top-K estimate in the limit of infinite data, even if the data has not been generated by an IID-RUM. On the other hand, Spectral Ranking does not in general give the same estimate as the choice-based Borda count algorithm/MNL-MLE even with infinite data.
- In section (8): We show through empirical experiments that the choice-based Borda count algorithm is competitive in terms of accuracy with both MNL-MLE and Spectral Ranking while being several orders of magnitude faster. This highlights the advantage of the choice-based Borda count algorithm in applications where the statistician is primarily interested in efficiently and accurately identifying the top items.

1.1 RELATED WORKS

Our work falls within the literature on learning to rank under Random Utility Models (RUMs). There has been a substantial amount of work on learning to rank under Random Utility Models and mixtures of Random Utility Models using *full ranking data* [Parkes et al., 2012, Azari Soufiani et al., 2013b,a, Soufiani et al., 2014, Zhao et al., 2016, Zhao and Xia, 2019]. Furthermore, most classical ranking methods assume that the data is generated by a well specified RUM. To the best of our knowledge, our paper is the first to propose a method for top-K ranking under a broad class of RUMs using *passively collected choice data alone*.

The related literature on ranking from pairwise comparisons is vast and we can only refer the interested reader to adjacent problems such as *active top-K recovery* from pairwise comparisons [Busa-Fekete et al., 2013, Agarwal et al., 2017, Mohajer et al., 2017, Falahatgar et al., 2017, 2018, Heckel et al., 2019]; top-K recovery from pairwise comparisons [Chen and Suh, 2015, Shah and Wainwright, 2017, Chen et al., 2019]; top-K recovery from *m-wise sorted data* (full rankings among some m items) [Jang et al., 2017, Chen et al., 2020].

Closest to our work is the analysis of Borda count by Shah and Wainwright [2017] who showed that it is optimal for top-K recovery from *pairwise comparisons*. Our work complements theirs by showing that the choice-based Borda count is optimal even in the *general choice setting*. To this end, we obtain in section (4) sample complexity upper and lower bounds that are *both more general and refined* than those given by Shah and Wainwright [2017]. We also study in section (6) the effect of the menu size on the sample complexity. To the best of our knowledge, our paper presents the first asymptotic characterization of the sample complexity for ranking from choice data in terms of the menu size under the very commonly used MNL model. Operating on *m-wise sorted data*, Jang et al. [2017] showed that the optimal sample complexity for top-K recovery under the Plackett Luce model[1] scales with $O(\frac{1}{m})$. Our results complement theirs by showing that the sample complexity for top-Krecovery from *discrete choice data* scales as $O(1 + \frac{1}{m})$. Furthermore, the choice-based Borda count algorithm is different from the Spectral-MLE algorithm studied there which is specialized to the Plackett-Luce model.

[1]Within the ranking literature, Plackett-Luce (PL) is a class of distributions over permutations, induced by the IID-RUM with standard Gumbel noise.

2 NOTATIONS AND PROBLEM FORMULATION

Let there be n items in the universe. Each item i has a *deterministic and hidden* utility, also referred to as partworth, U_i for $i = 1, \ldots, n$. Let us assume the non-degenerate case where no two items have identical partworths. Without loss of generality, we also assume that $U_{\max} = U_1 > U_2 > \ldots > U_n = U_{\min} > 0$. Let $\mathcal{S}_K^* = \{1, \ldots, K\}$ denote the set of K items with the highest parthworths.

Items are presented to the consumer in a set S, also referred to as menu, of size at least 2. When S is presented to the consumer, the perceived utility of each item $i \in S$ is the sum of its parthworth and a random noise term: $X_i = U_i + \epsilon_i$ where the ϵ_i's are independently and identically distributed according to an *unknown* universal noise distribution D. The consumer then picks the item i with the highest perceived utility among all the items in S. Such a choice model is referred to as a random utility model with independent and identically distributed noise (IID-RUM). In short, a choice model ρ within the class of IID-RUMs is parametrized by a set of partworths $\{U_1, \ldots U_n\}$ and noise distribution D.

As an overload of notation, we will also use $\rho(i|S)$ to denote the probability that a consumer picks item i from menu S under choice model ρ. By definition, $\rho(i|S) = \mathbb{P}(X_i > X_k \, \forall k \in S \backslash \{i\})$. For simplicity, we consider a fixed menu size m. However, our analysis can be easily extended to account for a mixture of menu sizes.

A choice sample is a (menu, item) tuple (S, y) where the consumer chooses item y from menu S. A choice dataset is a set of choice samples. A top-K recovery algorithm takes in a choice dataset and returns an estimate of the top K items, $\hat{\mathcal{S}}_K$. The goal is to exactly recover the top K items and the performance metric of interest is the 0-1 loss: $L_{01}(\hat{\mathcal{S}}_K, \mathcal{S}_K^*) = \mathbf{1}[\hat{\mathcal{S}}_K = \mathcal{S}_K^*]$ [2]. We emphasize that as opposed to the *top-K ranking* problem, the objective of the top-K recovery prolem is to accurately identify the set of the top K items, while allowing for mis-ranking among these items.

3 THE CHOICE-BASED BORDA COUNT ALGORITHM

As discussed previously, the general counting approach referred to as Borda count has a long history and has been instantiated in various contexts such as ranking from pairwise comparisons. Here, we instantiate the Borda count approach to the more general discrete choice setting. This is shown in Algorithm (1).

As would be expected, the algorithm essentially tallies the number of observed 'wins' by each item and finally ranking the items by their number of wins, returning the top K items. As with other versions of the Borda count approach, the algorithm is simple and easy to implement; and very efficient in practice. This makes the choice-based Borda count algorithm appropriate in settings where the statistician is primarily interested in efficiently and accurately recovering the set of the top items from a large pool of choice data.

Algorithm 1 The choice-based Borda count algorithm

Input: Choice dataset $\mathcal{B} = \{(S_l, y_l)\}_{l=1}^N$
Output: Top-K estimate $\hat{\mathcal{S}}_K$

1: For each item $i = 1, \ldots, n$
2: Compute the number of times i gets chosen:
3: $\hat{W}_i := \sum_{l=1}^N \mathbf{1}[y_l = i]$
4: Return the set of K items corresponding to the highest $\hat{W}_i$'s. Ties are broken arbitrarily.

4 SAMPLE COMPLEXITY BOUND

In this section, we present the sample complexity of the choice-based Borda count algorithm for top-K recovery. We first formalize our sampling model in section (4.1). In section (4.2), we characterize the class of IID-RUMs under which the choice-based Borda count can successfully identify all of the top K items via a theoretical quantity we term the *generalized Borda score*. The main theorems on the sample complexity of the choice-based Borda count algorithm are presented in section (4.3).

4.1 THE SAMPLING MODEL

Let $\mathcal{C}^{(m)}$ be the set of *all menus* of size $m \geq 2$ (i.e., $|C^{(m)}| = \binom{n}{m}$). Additionally, let $\mathcal{C}_i^{(m)}$ be the set of all menus of size m containing item i (i.e., $|\mathcal{C}_i^{(m)}| = \binom{n-1}{m-1}$). We consider a multiple-round uniform sampling model with R rounds of sampling in total. In each round $r = 1, \ldots, R$, each menu $S \in \mathcal{C}^{(m)}$ is independently offered with probability $p > 0$. Let $\hat{\mathcal{C}}^{(m,r)}$ denote the set of menus of size m that are offered in round r. If offered menu S, the user responds with a random choice $y_S^{(r)}$ where

$$\mathbb{P}(y_S^{(r)} = i) = \rho(i|S) \, .$$

It is easy to check that we have, in expectation, $pR\binom{n}{m}$ samples over R rounds.

As a practical example, this sampling procedure can be used to design online political surveys. Suppose that there are R voters willing to take part in answering survey questionnaires to determine support for n political candidates. Fix a ballot size m. For each voter, each ballot of size m is independently presented to that user with probability p. For each ballot, the voter picks one favourite candidate.

[2] $\mathbf{1}$ is the indicator function and the equality is with respect to set equality.

4.2 THE GENERALIZED BORDA SCORE

For each item i, define the following theoretical quantity, which we term the *generalized Borda score*:

$$\tau_i^{(m)} = \frac{1}{\binom{n-1}{m-1}} \cdot \sum_{S \in \mathcal{C}_i^{(m)}} \rho(i|S) \,.$$

Intuitively, the generalized Borda score is the expected probability that an item i is chosen from a menu S where S is uniformly sampled from $C_i^{(m)}$. Note that $\tau_i^{(m)} \in [0, 1]$ for all $i \in [n]$. The generalized Borda score is interesting to us because *for a large class of IID-RUMs*, the order among the items with respect to the generalized Borda scores is the same as that with respect to the partworths. Therefore, it suffices to rank the items by their generalized Borda scores to recover the items with the highest partworths. Formally, we can characterize this class of IID-RUMs using a property we term *Borda consistency*.

Definition 1. *An IID-RUM ρ satisfies Borda consistency if for any two items i, j and any menu size $m \geq 2$:*

$$\tau_i^{(m)} > \tau_j^{(m)} \Leftrightarrow U_i > U_j \,.$$

The follow lemma establishes that many commonly used IID-RUMs such as the MNL (Gumbel distributed noise) and the Probit (Normal distributed noise) model satisfy Borda consistency.

Lemma 4.1. *All IID-RUMs whose noise distribution has absolutely continuous density function and support on the real line satisfy Borda consistency.*

In the supplementary materials, we will show that Borda consistency is satisfied by an even broader class of IID-RUMs that include other commonly used models such as the IID-RUM with exponentially distributed noise. Intuitively, this stems from the property enjoyed by many IID-RUMs: for any two items i, j where $U_i > U_j$, $\rho(i|S) > \rho(j|S)\ \forall S \in \mathcal{C}^{(m)} : i, j \in S$; and $\rho(i|S \cup \{i\}) > \rho(j|S \cup \{j\})\ \forall S \in \mathcal{C}^{(m-1)} : i, j \notin S$. To the best of our knowledge, this fundamental property that holds across a very broad class of IID-RUMs has not been previously decribed in the literature and may be useful to future works exploring the intersection of ranking and choice modeling.

4.3 EXACT TOP-K RECOVERY

Having established Borda consistency as a property enjoyed by many IID-RUMs, we will now present the finite sample guarantees of the choice-based Borda count algorithm for top-K recovery that holds for all choice models in this broad class of IID-RUMs. To show that the choice-based Borda count algorithm accurately identifies all of the top K items with high probability, it suffices to bound the probability that the algorithm mistakenly ranks an item $j \notin \mathcal{S}_K^*$ higher than another item $i \in \mathcal{S}_K^*$. Specifically, we want to bound the following probabilities:

$$\mathbb{P}(\hat{W}_j > \hat{W}_i) \quad \forall i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*$$

where $\hat{W}_i$ is defined in Algorithm (1). Considering this, the fundamental hardness of top-K ranking lies in distinguishing between the K-th and $K+1$-th best item, and therefore depends on the gap between their generalized Borda scores:

$$\Delta_K^{(m)} = \tau_K^{(m)} - \tau_{K+1}^{(m)} \,.$$

The smaller this gap, the more data the algorithm requires in order to correctly separate between the top K and the bottom $n - K$ items. Building on this intuition and generalizing to any pair of items (i, j) where $\tau_i > \tau_j$, we obtain the following upper bound on $\mathbb{P}(\hat{W}_j > \hat{W}_i)$:

Lemma 4.2. *Consider an IID-RUM that satisfies Borda consistency per definition (1). Assume input choice data with menu size m is generated according to the sampling model described in section (4.1). For any two items i and j where $\tau_i^{(m)} > \tau_j^{(m)}$, the choice-based Borda count algorithm satisfies:*

$$\mathbb{P}(\hat{W}_j > \hat{W}_i) \leq \exp\left(\frac{-3pR\binom{n}{m} m (\tau_i^{(m)} - \tau_j^{(m)})^2}{8n(\tau_i^{(m)} + \tau_j^{(m)})} \right) \,.$$

The proof of Lemma (4.2) uses a standard concentration inequality argument based on Bernstein's inequality (cf. Theorem 2.8.4 Vershynin [2018]). The lemma itself states that, for each pair $i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*$, if $pR\binom{n}{m} \geq \frac{8n \log n (\tau_i^{(m)} + \tau_j^{(m)})}{m(\tau_i^{(m)} - \tau_j^{(m)})^2}$, then $\mathbb{P}(\hat{W}_j > \hat{W}_i) = O(\frac{1}{n^3})$. We also have the following lemma which presents an upper bound on the item-dependent term $\frac{\tau_i^{(m)} + \tau_j^{(m)}}{(\tau_i^{(m)} - \tau_j^{(m)})^2}$.

Lemma 4.3. *Consider an IID-RUM that satisfies Borda consistency per definition (1). For any K, we have:*

$$\frac{\tau_K^{(m)} + \tau_{K+1}^{(m)}}{{\Delta_K^{(m)}}^2} = \max_{i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*} \left\{ \frac{\tau_i^{(m)} + \tau_j^{(m)}}{(\tau_i^{(m)} - \tau_j^{(m)})^2} \right\} .$$

By combining the two lemmas above and applying union bound over all pairs $i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*$, we obtain the following sample complexity bound for exact top-K recovery:

Theorem 4.4. *Assume the conditions of lemma (4.2). Given sufficiently large p, R such that $pR\binom{n}{m} \geq \frac{8n \log n}{m {\Delta_K^{(m)}}^2} \cdot (\Delta_K^{(m)} + 2\tau_{K+1}^{(m)})$, the choice-based Borda count algorithm correctly identifies all of the top K items with probability at least $1 - O(\frac{K}{n^2})$.*

The reader may also recognize that $\Delta_K^{(m)} + 2\tau_{K+1}^{(m)}$ is simply $\tau_K^{(m)} + \tau_{K+1}^{(m)}$. The former presentation is, however, useful in highlighting the main quantities that will also reappear in our matching lower bound. In summary, the choice-based Borda count algorithm has the following sample complexity for exact top-K recovery:

$$O\Big(\frac{n\log n}{m\Delta_K^{(m)}} \cdot (1 + \frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}})\Big).$$

This shows that overall, we only need $O(n \log n)$ examples to recover the top K items from choice data with high accuracy. Our upper bound (and matching lower bound to be shown) can be seen as both *generalization and refinement* of theorem 1 of Shah and Wainwright [2017]. Under the pairwise comparison setting ($m = 2$), we can simply upper bound $\tau_{K+1}^{(m)} \leq 1$ and recover the (optimal) sample complexity $O\big(\frac{n\log n}{\Delta_K^{(2)\,2}}\big)$ of Borda count obtained by Shah and Wainwright [2017]. The analysis approach there, however, is insufficient to produce an optimal sample complexity bound in the discrete choice setting. Note also that there can be combinatorially many realizations of the data in the discrete choice setting as $|\mathcal{C}^{(m)}| = \binom{n}{m}$. Our proof therefore requires considerably more effort. Our bound also shows that the sample complexity depends not only on the gap $\Delta_K^{(m)}$ between the K-th and $K+1$-th item, but also the relative 'strength' of the $K+1$-th item, as captured by the $\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}$ term.

In general, the factors $\tau_{K+1}^{(m)}$ and $\Delta_K^{(m)}$ don't admit closed form expressions because both are sums of $\binom{n-1}{m-1}$ terms. The reader may also recognize that these parameters also depend on the menu size m, the partworth parameters and the noise distribution. In the next section, we will show a *matching lower bound* in terms of the same parameters, establishing the optimality of the choice-based Borda count algorithm, and discuss why the exact relation between $\Delta_K^{(m)}$, $\tau_{K+1}^{(m)}$ and the model parameters remains elusive.

Often in practice, we may tolerate some error for top-K ranking by allowing the algorithm to misidentify, up to a threshold, some number of items. This is known as *approximate top-K recovery*. We include detailed discussions of this problem in the supplementary materials and show that the choice-based Borda count algorithm also has *optimal sample complexity* for approximate top-K recovery under the broad class of IID-RUMs that satisfy Borda consistency.

5 INFORMATION-THEORETIC LOWER BOUND

In this section, we will show that the choice-based Borda count algorithm enjoys optimal sample complexity by furnishing a matching lower bound.

To show a lower bound, we will construct a special subclass of the MNL family where any estimator requires $\Omega(n \log n)$ examples in order to exactly recover the top K items. We defer detailed descriptions of this model to the supplementary materials while stating the main results as follows:

Theorem 5.1. *Consider the sampling model described in subsection (4.1). There exists a class of MNL models such that for $n \geq 20$, if $pR\binom{n}{m} \leq \frac{n\log n}{8} \cdot \frac{\tau_{K+1}^{(m)} + \Delta_K^{(m)}}{m\Delta_K^{(m)\,2}}$ then any estimator fails to correctly identify all of the top K items with probability at least $\frac{1}{12}$.*

The proof of Theorem (5.1) first reduces the problem of exact top-K recovery to a multiple hypothesis testing problem and then applies Fano's lemma [Cover, 1999]. Each hypothesis in the testing problem corresponds to an MNL model. Within each model, the set of the top K items always includes items $1, \ldots, K-1$. However, the index of remaining item in the top-K set is different for each model (i.e., there are $n - K + 1$ different models). We make all of the top K items have the same partworths while the bottom $n - K$ items have the same (and lower) partworths. The key challenge then is to obtain a tight upper bound on the KL divergence between any two hypothesis models. In summary, Theorem (5.1) implies the following *minimum* sample complexity for any algorithm for top-K recovery:

$$\Omega\Big(\frac{n\log n}{m\Delta_K^{(m)}} \cdot \big(1 + \frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}\big)\Big)$$

Comparing with the bound in theorem (4.4), one can see that the sample complexity of Borda Count is optimal in terms of both m, n as well as the model dependent parameters $\Delta_K^{(m)}$ and $\tau_{K+1}^{(m)}$.

6 THE ROLE OF THE MENU SIZE m

The effect of the menu size on the performance of top-K recovery algorithms is an aspect of both theoretical and practical importance. In real life applications, the menu size could range from 2 to hundreds of items. One may suspect that increasing the menu size means the data carries more information per data point, and thereby reduces the sample complexity for top-K recovery. However, to the best of our knowledge, such a relationship has not been theoretically established in the literature on choice modeling, even for the very commonly used MNL model.

As seen in the matching lower and upper bound for the sample complexity of top-K recovery, the menu size enters in complex ways through the factors $\Delta_K^{(m)}$ and $\tau_{K+1}^{(m)}$. Both factors can vary in subtle ways with m, depending on the underlying choice model. Even for the class of MNL models which admit closed form choice probabilities, these factors don't seem to have a closed form expression as each of them

is a sum of $\binom{n-1}{m-1}$ terms. To bypass the difficulty of exactly evaluating $\Delta_K^{(m)}$ and $\tau_{K+1}^{(m)}$, we characterize the asymptotic dependency of $\frac{1}{m\Delta_K^{(m)}}$ and $\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}$ on m *under the MNL class of models* and show that both of these factors monotonically decrease with m but at a *decreasing rate*. This implies that while there is an advantage to using choice data of larger menu sizes, there is a diminishing return to increasing the menu size.

Theorem 6.1. *For any MNL model and a fixed K*

$$\frac{1}{m\Delta_K^{(m)}} = \theta\left(\frac{1}{e^{U_K} - e^{U_{K+1}}} \cdot \left(1 + \frac{1}{m-1}\right)\right);$$

$$\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}} = \theta\left(\frac{e^{U_{K+1}}}{e^{U_K} - e^{U_{K+1}}} \cdot \left(1 + \frac{1}{m-1}\right)\right).$$

It can be seen in both $\frac{1}{m\Delta_K^{(m)}}$ and $\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}$ that the term which depends on m, $1 + \frac{1}{m-1}$, montonically decreases with m but at a diminishing rate. Combining the above theorem and the matching sample complexity bounds obtained earlier, one can see that the optimal sample complexity for top-K recovery from choice data scales as $\theta(1 + \frac{1}{m})$.

Outside of the MNL family of models, we are not aware of any IID-RUM that admits a closed form expression for the choice probabilities. However, suppose that we know all of the partworths and the noise distribution, we can still approximate the choice probabilities via Monte Carlo sampling. Given these (approximated) choice probabilities, one can then evaluate $\frac{1}{m\Delta_K^{(m)}}$ and $\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}$. As an example, Figure (1) shows how these quantities vary with m under a randomly generated MNL and Probit model (IID-RUM with standard normal noise) [Train, 2009] with $n = 15$, $K = 3$. The partworths were independently generated from a zero-mean normal distribution which is also a commonly chosen prior in the literature [Parkes et al., 2012, Train, 2009]. The curves for $\frac{1}{m\Delta_K^{(m)}}$ and $\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}$ decrease at a rate approximately similar to those of the MNL model as stated in Theorem (6.1).

Ranking from choice data under MNL model assumptions remains an active area of research [Agarwal et al., 2018, 2020] and to the best of our knowledge, our work presents *the first asymptotic characterization* of the optimal sample complexity for top-K recovery in terms of the menu size m under this often used class of choice models.

7 CONNECTIONS TO COMMONLY USED RANKING ALGORITHMS

In this section, we establish close connections among choice-based Borda count, the method of maximum likelihood estimate under MNL assumptions (MNL-MLE) [Train, 2009] and spectral ranking [Negahban et al., 2017, Maystre and Grossglauser, 2015, Agarwal et al., 2018] which will explain many experimental results we present in later sections.

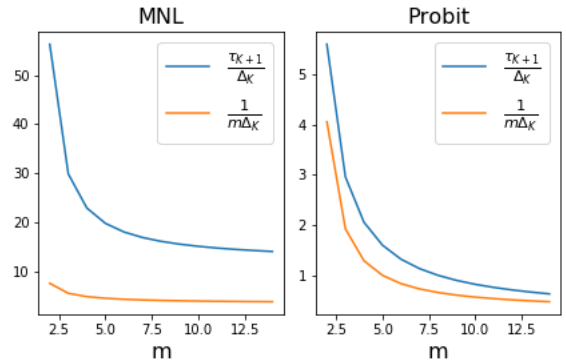

Figure 1: $\frac{1}{m\Delta_K^{(m)}}$ and $\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}$ decrease with larger m under a randomly generated MNL and Probit model. This suggests that there is an advantage, albeit with dimnishing return, to using larger menu sizes.

Firstly, one can prove that choice-based Borda count and MNL-MLE are 'equivalent' top-K recovery algorithms in the limit of infinite data. This connection is formalized as follows:

Theorem 7.1. *Consider the sampling model described in section (4.1), for any $p > 0$, in the limit as $R \to \infty$, MNL-MLE and choice-based Borda count will produce the same top-K estimate. Moreover, this holds even if the data does not come from the MNL model or any IID-RUM.*

A similar observation was made in Rajkumar and Agarwal [2014]: under the *pairwise comparison* setting the Borda count algorithm and MNL-MLE are both consistent for full ranking under a class of pairwise comparison models that is strictly more general than the BTL model [3]. Our results generalize the relation between the two algorithms to the choice setting and show that in fact the Borda count algorithm and MNL-MLE produce the same estimate in the limit of infinite data under any choice models. This connection between the two methods is reflected in our experiments where the performance of the choice-based Borda count algorithm is almost identical to that of MLE, when the sample size is large. While performing similarly to MNL-MLE, the choice-based Borda count algorithm is several orders of magnitude faster thanks to its simplicity. This suggests that if the statistician is mostly concerned with recovering a small number of top items, the choice-based Borda count algorithm should be seriously considered

[3] The BTL model is the instantiation of the MNL to the pairwise comparison setting

due to its speed, simplicity and guaranteed optimal sample complexity.

The above result also means that MNL-MLE is a consistent top-K ranking algorithm under the broad class of IID-RUMs, since the choice-based Borda count algorithm is consistent in recovering the top K items. This shows that MNL-MLE may be used for ranking applications even when the data does not satisfy the MNL assumption. Consistency of MLE under model misspecification is an underexplored question and we leave the careful characterization of the sample complexity of MNL-MLE when the data comes from a non-MNL distribution as a subject of future studies.

On the other hand, Spectral Ranking does not in general produce the same top-K estimate as MNL-MLE/choice-based Borda count. However, when the underlying choice model falls within a broad class of IID-RUMs which include many commonly used choice models such as the MNL and Probit model, all three algorithms produce the same estimate given infinite data.

Theorem 7.2. *Consider the sampling model described in subsection (4.1). Assume that the underlying choice model generating the data is in the class of IID-RUMs whose noise distribution has absolutely continuous density function with support on the real line. For any $p > 0$, in the limit as $R \to \infty$, then spectral ranking, MNL-MLE and choice-based Borda count produce the same top-K estimate.*

On the other hand, there exists a choice model where in the limit as $R \to \infty$, the spectral ranking algorithm produces a different top-K estimate from MNL-MLE/Borda count.

8 EXPERIMENTS

In this section, we present experiment results on both synthetic and real datasets. The main performance metric is top-K accuracy. More specifically, we measure top-K accuracy as the frequency at which the respective algorithm correctly identifies *all* of the true top K items, *over 100 trials*.

8.1 SYNTHETIC DATA

We verify, via synthetic experiments, the efficacy of the choice-based Borda count algorithm and the effect of the menu size m on its performance. Let there be $n = 50$ items in the universe. We experiment with 3 different noise distributions: standard Normal noise (Probit), standard Gumbel noise (MNL) and standard Exponential noise. We vary the menu size $m = 2, 4, 6, 8$ and $K = 1, 3, 5$. Figure (2) shows top-K accuracy against the sample size. In all experiments, choice-based Borda count successfully identifies the top K items with high probability given sufficiently large sample size. Furthermore, using larger menu sizes improves the performance of Borda Count. However, it can be seen that there is a diminishing return in performance gains from using larger menu sizes, agreeing with our theoretical analysis in section (5).

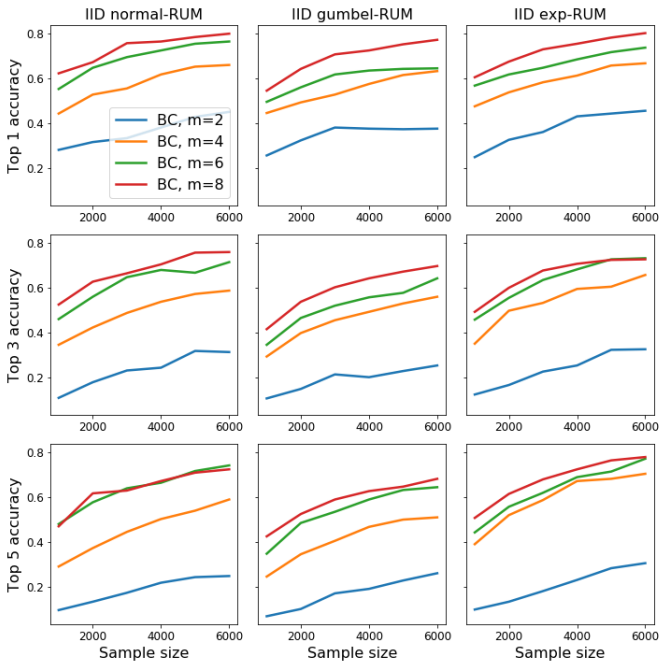

Figure 2: **Synthetic data:** Exact top K accuracy of choice-based Borda count against sample size for different menu sizes 2 (blue), 4 (orange), 6 (green), 8 (red) with $K = 1, 3, 5$ and $n = 50$. Increasing the menu size improves performance but with a diminishing return.

8.2 REAL DATA

Baseline algorithms: We compare choice-based Borda count against Accelerated Spectral Ranking (ASR) [Agarwal et al., 2018] and Maximum Likelihood Estimate under MNL assumptions [Train, 2009] in terms of top K accuracy. We implement MLE using Scipy's L-BFGS optimizer [Virtanen et al., 2020].

Data description: We follow standard procedures commonly used in previous works such as Rajkumar and Agarwal [2014], Agarwal et al. [2020]. Operating on full ranking datasets, we can estimate the choice probabilities for any menu S, i.e., choice probability $\rho(i|S)$ is the proportion of rankings that ranks item i highest among all them items in S. Given these probabilities, we can simulate the sampling model as described in subsection (4.1). Our datasets include SUSHI [Kamishima, 2003], APA election dataset [Diaconis, 1989], 3 Irish election datsets and F1 race dataset included in the library PrefLib [Mattei and Walsh, 2013]. Notably, the induced pairwise choice probabilities of these datasets all satisfy stochastic transitivity. Therefore, there exists a universal ordering of the items which we can use as a true global ranking over the items. Due to space constraint, we can only present a few representative experimental findings

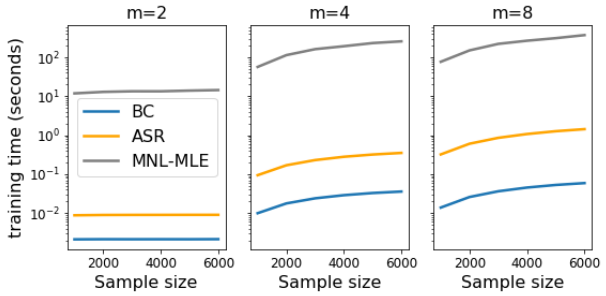

Figure 3: **F1 dataset** ($n = 22$)**:** Average training time (seconds) against sample size for $m = 2, 4, 8$. Choice-based Borda count (blue) is several orders of magnitude faster than its competitors.

and leave additional results with detailed descriptions of data processing in the supplementary materials.

Speed advantage: Across all experiments, choice-based Borda count is several orders of magnitude faster than ASR and MLE. This difference is especially pronounced in datasets with more items such as the F1 dataset, as shown in Figure (3).

Competitive accuracy: Figure (4) show the performance of the algorithms under the Irish-Meath dataset and Figure (5) shows the results for the Irish-West dataset. Our theoretical analysis in section (7) is reflected in our experimental findings: the performance of MNL-MLE and the Borda count algorithm are very similar given sufficiently large sample size. Spectral Ranking, on the other hand, may perform better or worse than MLE/Borda count depending on the dataset and the choice of m and K. For many combinations of m and K, we observe that the choice-based Borda count algorithm accurately recovers the top K items and is highly competitive with MNL-MLE and Spectral Ranking. Notably, in most datasets, for smaller K and large m, ithe Borda count algorithm has considerable advantages thanks to its accuracy and faster running time. In practice, this means choice-based Borda count is appropriate for applications where the statistician is interested in quickly determining a single (or a few) top candidate(s) from a large amount of data such as aggregating political surveys.

9 CONCLUSION

Ranking under Random Utility Models is a promising area of research with many practical applications. Our work shows how an efficient algorithm can perform very well under a broad family of RUMs. That being said, the class of IID-RUMs constitutes only a subset of models within the class of general RUMs. Beyond IID-RUMs, not much is known in terms of efficient inference and ranking algorithms. In the future, we hope to see more ranking methods developed for more expressive RUMs which have non-identical noise distributions or dependent noise distributions.

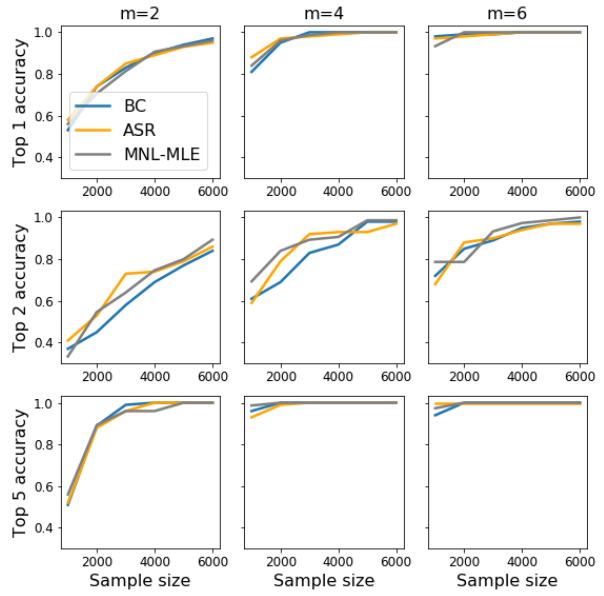

Figure 4: **Irish-Meath dataset** ($n = 14$)**:** Exact top-K accuracy against sample size. choice-based Borda count (blue) is competitive with baseline algorithms. Using larger menu sizes generally improves the performance of the algorithms.

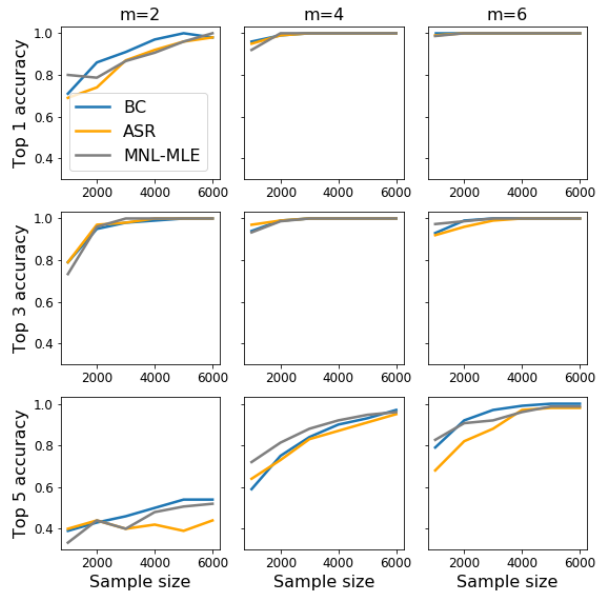

Figure 5: **Irish West dataset** ($n = 9$)**:** choice-based Borda count (blue) performs very similarly to MNL-MLE (gray) while the performance of ASR (orange) may diverge from MNL-MLE/choice-based Borda count.

References

Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pages 39–75. PMLR, 2017.

Arpit Agarwal, Prathamesh Patil, and Shivani Agarwal. Accelerated spectral ranking. In *International Conference on Machine Learning*, pages 70–79. PMLR, 2018.

Arpit Agarwal, Nicholas Johnson, and Shivani Agarwal. Choice bandits. In *NeurIPS*, 2020.

Hossein Azari Soufiani, William Ziwei Chen, David C Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. 2013a.

Hossein Azari Soufiani, Hansheng Diao, Zhenyu Lai, and David C Parkes. Generalized random utility models with multiple types. *Advances in Neural In*, 2013b.

Gordon M Becker, Morris H DeGroot, and Jacob Marschak. Stochastic models of choice behavior. *Behavioral science*, 8(1):41–55, 1963.

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. PMLR, 2013.

Wenjing Chen, Ruida Zhou, Chao Tian, and Cong Shen. On top-k selection from m-wise partial rankings via borda counting. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2759–2764. IEEE, 2020.

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

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.

Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Light rums. In *International Conference on Machine Learning*, pages 1888–1897. PMLR, 2021.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Persi Diaconis. A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*, pages 949–979, 1989.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. Maxing and ranking with few assumptions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 7063–7073, 2017.

Moein Falahatgar, Ayush Jain, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. The limits of maxing, ranking, and preference learning. In *International Conference on Machine Learning*, pages 1427–1436. PMLR, 2018.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126, 2019.

Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Optimal sample complexity of m-wise data for top-k ranking. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1685–1695, 2017.

Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 583–588, 2003.

Vladimir Iosifovich Levenshtein. Upper-bound estimates for fixed-weight codes. *Problemy Peredachi Informatsii*, 7(4):3–12, 1971.

Nicholas Mattei and Toby Walsh. Preflib: A library for preferences http://www. preflib. org. In *International Conference on Algorithmic Decision Theory*, pages 259–270. Springer, 2013.

Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. In *Advances in neural information processing systems*, pages 172–180, 2015.

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. PMLR, 2017.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

David C Parkes, Houssein Azari Soufiani, and Lirong Xia. Random utility theory for social choice. In *Proceeedings of the 25th Annual Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2012.

Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, pages 118–126. PMLR, 2014.

Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

Hossein Azari Soufiani, David Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *International Conference on Machine Learning*, pages 360–368. PMLR, 2014.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Zhibing Zhao and Lirong Xia. Learning mixtures of plackett-luce models from structured partial orders. *arXiv preprint arXiv:1910.11721*, 2019.

Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models. In *International Conference on Machine Learning*, pages 2906–2914. PMLR, 2016.

Supplemental Materials: Efficient and Accurate Top-K Recovery from Choice Data

Table of Contents

10 Proofs **11**

10.1 A closer look at IID-RUMs . . . 11

10.2 Exact top-K ranking . . . 14

10.3 Borda Count: Approximate top-K ranking . . . 17

10.4 Information theoretic lower bound . . . 18

10.5 Connection among MNL-MLE, Generalized Borda Count and Spectral Ranking . . . 30

11 Additional experiments **32**

11.1 Real datasets . . . 33

10 PROOFS

In this section, we will prove all the theoretical results of the paper. The first subsection (10.1), however, details some of the important properties of IID-RUMs that are useful in proving the main theorems and lemmas. Many of these results can be of independent interest such as the *Borda consistency* property which is shared by a very broad class of IID-RUMs. The reader may also skip to subsection (10.2) for the proofs of the main results stated earlier in the paper.

10.1 A CLOSER LOOK AT IID-RUMS

All IID-RUMs are in a family of choice models known as Fechnerian models Becker et al. [1963].

Definition 2. *A choice model ρ is a Fechnerian model if there exists a function $F : \mathbb{R} \times \mathbb{R} \to [0, 1]$ that is increasing in the first argument and decreasing in the second argument such that:*

$$\rho(i|\{i, j\}) = F(U_i, U_j)$$

It is well known that Fechnerian models satisfy strongly stochastic transitivity (SST). That is, if $\rho(i|\{i, j\}) > \frac{1}{2}$ then $\rho(i|\{i, k\}) \geq \rho(j|\{j, k\})$ for all $k \neq i, j$. A weaker notion of stochastic transitivity is weak stochastic transitivity (WST): if $\rho(i|\{i, j\}) > \frac{1}{2}, \rho(j|\{j, k\}) > \frac{1}{2}$ then $\rho(i|\{i, k\}) > \frac{1}{2}$. The following lemma establishes that all IID-RUMs are Fechnerian models.

Lemma 10.1. *All IID-RUMs are Fechnerian models.*

Proof. The proof is simply showing the existence of the function F in the definition of Fechnerian models. Consider any two items i and j:

$$\begin{aligned}\rho(i|\{i, j\}) &= \mathbb{P}(X_i > X_j) = \mathbb{P}(U_i + \epsilon_i > U_j + \epsilon_j) \\ &= \mathbb{P}(\epsilon_j - \epsilon_i < U_i - U_j) \\ &= F_{ij}(U_i - U_j)\end{aligned}$$

where F_{ij} is the CDF of the random variable $\epsilon_j - \epsilon_i$. Note that all the ϵ_i's are identically distributed and thus $F = F_{ij}$ for all pairs (i, j). This completes the proof. ■

As the class of SST models are a special subclass of WST models, there exists a universal ordering among the items, induced by pairwise preference. Under IID-RUMs, this ordering the same as that of the partworth parameters. It's also notable that while IID-RUMs satisfy SST, RUMs with independent but not identically distributed noise (Independent RUMs) don't in general. In fact, Independent RUMs may not even satisfy WST. Next, we introduce a property of random utility models which we term the *order preserving* property.

Definition 3. *A choice model ρ satisfies the order preserving property if:*

$$\rho(i|\{i,j\}) > \frac{1}{2} \Rightarrow \rho(i|S) > \rho(j|S) \, \forall \, S : i,j \in S$$

The following lemma states that a broad class of IID-RUMs which include the MNL and Probit model satisfy the order preserving property.

Lemma 10.2. *All IID-RUMs whose noise distribution has absolutely continuous density function and support on the real line satisfy the order preserving property.*

Proof. Consider a pair of item i, j such that $U_i > U_j$. The proof for the case when $S = \{i, j\}$ is trivial. We immediately have $\rho(i|\{i,j\}) > \frac{1}{2}$. Now consider any choice set S containing both i and j where $|S| \geq 3$. Let f_{ϵ_i} denote the density function of the item-specific noise distribution of item i. Of course, under IID-RUM, $f_{\epsilon_i} = f_{\epsilon_j} = f$ for all $i \neq j$. We have

$$\begin{aligned}
\rho(i|S) &= \mathbb{P}(X_i > X_k, \forall k \in S \backslash \{i\}) \\
&= \mathbb{P}(U_i + \epsilon_i > U_k + \epsilon_k, \forall k \in S \backslash \{i\}) \\
&= \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon_i, \forall k \in S \backslash \{i\}) \\
&= \int_\epsilon \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon_i, \forall k \in S \backslash \{i\}, \epsilon_i = \epsilon) d\epsilon \\
&= \int_\epsilon \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon_i, \forall k \in S \backslash \{i\} | \epsilon_i = \epsilon) f_{\epsilon_i}(\epsilon) d\epsilon \\
&= \int_\epsilon \Big(\prod_{k \in S \backslash \{i\}} \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon) \Big) f(\epsilon) d\epsilon
\end{aligned}$$

In the last equality, we have used the independence assumption of IID-RUMs. Further breaking down the product gives:

$$\begin{aligned}
&\int_\epsilon \Big(\prod_{k \in S \backslash \{i\}} \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon) \Big) f(\epsilon) d\epsilon \\
&= \int_\epsilon \Big(\prod_{k \in S \backslash \{i,j\}} \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon) \Big) \mathbb{P}(\epsilon_j < U_i - U_j + \epsilon) f(\epsilon) d\epsilon
\end{aligned} \tag{1}$$

By our assumption of identicallity and that the noise distribution has continuous density with support on the real line, $\mathbb{P}(\epsilon_k < U_i - U_k + \epsilon) < \mathbb{P}(\epsilon_k < U_j - U_k + \epsilon)$ for all $k \in S \backslash \{i,j\}$. On the other hand, for a random variable ϵ' distributed identically to the ϵ_i's:

$$\mathbb{P}(\epsilon_j < U_i - U_j + \epsilon) > \mathbb{P}(\epsilon' < U_j - U_i + \epsilon) \quad \forall \epsilon$$

Putting these two inequalities back into (1):

$$\begin{aligned}
\rho(i|S) &= \int_\epsilon \Big(\prod_{k \in S \backslash \{i,j\}} \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon) \Big) \mathbb{P}(\epsilon_j < U_i - U_j + \epsilon) f(\epsilon) d\epsilon \\
&> \int_\epsilon \Big(\prod_{k \in S \backslash \{i,j\}} \mathbb{P}(\epsilon_k < U_j - U_k + \epsilon) \Big) \mathbb{P}(\epsilon' < U_j - U_i + \epsilon) f(\epsilon) d\epsilon \\
&= \rho(j|S)
\end{aligned}$$

This completes the proof.■

While the class of IID-RUMs whose noise distribution has continuous density function and support on the real line includes many commonly used IID-RUMs such as the MNL and Probit model, there are other IID-RUMs outside this class that also satisfy the order preserving property. An example is IID-Exponential RUM.

Lemma 10.3. *IID-Exponential RUM satisfies the order preserving property.*

Proof. Since the pairwise case ($m = 2$) is trivial, we will focus on proving lemma (10.3) for $m \geq 3$. Consider a menu $S \in \mathcal{C}^{(m)} : i, j \in S$, let $S' = S \backslash \{i, j\}$. For this proof, let us consider a 'copy' of i and call it i'. That is, $X_{i'}$ is distributed identically to X_i. We have:

$$\begin{aligned}
\rho(i|S) &= \mathbb{P}(X_i > \max\{\{X_k\}_{k \in S'} \cup \{j\}\}) \\
&= \int_0^\infty \mathbb{P}(X_k < U_i + \epsilon \, \forall k \in S', X_j < U_i + \epsilon) \cdot f(\epsilon) d\epsilon \\
&= \int_0^\infty \mathbb{P}(X_k < U_i + \epsilon \, \forall k \in S') \cdot \mathbb{P}(X_j < U_i + \epsilon) \cdot f(\epsilon) d\epsilon \\
&> \int_0^\infty \mathbb{P}(X_k < U_i + \epsilon \, \forall k \in S') \cdot \mathbb{P}(X_{i'} < U_i + \epsilon) \cdot f(\epsilon) d\epsilon \\
&> \int_0^\infty \mathbb{P}(X_k < U_j + \epsilon \, \forall k \in S') \cdot \mathbb{P}(X_{i'} < U_j + \epsilon) \cdot f(\epsilon) d\epsilon \\
&= \int_0^\infty \mathbb{P}(X_k < U_j + \epsilon \, \forall k \in S', X_{i'} < U_j + \epsilon) \cdot f(\epsilon) d\epsilon \\
&= \mathbb{P}(X_j > \max\{\{X_k\}_{k \in S'} \cup \{i'\}\}) \\
&= \rho(j|S)
\end{aligned}$$

The first inequality holds because for any $\epsilon > U_j - U_i$, $\mathbb{P}(X_j < U_i + \epsilon) > \mathbb{P}(X_{i'} < U_i + \epsilon)$. The second inequality holds because for any $\epsilon > \max\{0, \max\{U_k - U_j\}_{k \in S' \cup \{j\}}\}$, $\mathbb{P}(X_k < U_i + \epsilon) > \mathbb{P}(X_k < U_j + \epsilon)$ for all $k \in S'$ and $\mathbb{P}(X_{i'} < U_j + \epsilon) > \mathbb{P}(X_{i'} < U_j + \epsilon)$. This completes the proof. ■

Recall the Borda Consistency property introduced in subsection (4.2). One can show that a broad class of IID-RUMs which contains many commonly used RUMs such as the MNL and the Probit model satisfies Borda Consistency, the key property that ensures correctness of Generalized Borda Count.

Theorem 10.4. *Any IID-RUM ρ whose noise distribution has continuous density function and support on the real line satisfies Borda consistency.*

Proof. Note that $|\mathcal{C}_i^{(m)}| = |\mathcal{C}_j^{(j)}| = \binom{n-1}{m-1}$ for any i, j and m. The proof for $m = 2$ is trivial so we will focus on $m \geq 3$. By definition,

$$\begin{aligned}
\tau_i^{(m)} &= \frac{1}{|\mathcal{C}_i^{(m)}|} \sum_{S \in \mathcal{C}_i^{(m)}} \rho(i|S) \\
&= \frac{1}{|\mathcal{C}_i^{(m)}|} \Bigg(\sum_{S : |S| = m, i, j \in S} \rho(i|S) + \sum_{S : |S| = m, i \in S, j \notin S} \rho(i|S) \Bigg)
\end{aligned}$$

By the ordering preserving property of IID-RUMs, for any $S : i, j \in S$, $p(\rho|S) > \rho(j|S)$. To prove that $\tau_i^{(m)} > \tau_j^{(m)}$, we only need to prove that for any $S' : |S'| = m - 1; i, j \notin S'$:

$$\rho(i|S' \cup \{i\}) > \rho(j|S' \cup \{j\})$$

Let f_ϵ denote the density function of the noise distribution, we have:

$$\begin{aligned}
\rho(i|S' \cup \{i\}) &= \mathbb{P}(X_i > X_k \forall k \in S') \\
&= \mathbb{P}(U_i + \epsilon_i > U_k + \epsilon_k \forall k \in S') \\
&= \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon_i \forall k \in S') \\
&= \int_{-\infty}^{\infty} \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon_i, \epsilon_i = \epsilon) \\
&= \int_{-\infty}^{\infty} \mathbb{P}(\epsilon_k < U_i - U_k + \epsilon) f(\epsilon) d\epsilon \\
&> \int_{-\infty}^{\infty} \mathbb{P}(\epsilon_k < U_j - U_k + \epsilon) f(\epsilon) d\epsilon \\
&= \rho(j|S' \cup \{j\})
\end{aligned}$$

This completes the proof.■

We also have a more general theorem that characterizes the class of IID-RUMs, beyond those whose noise distribution has continuous density function and support on the real line, that satisfy Borda consistency

Theorem 10.5. *Consider an IID-RUM ρ. Fix a menu size m. If, for any pair of item i, j where $U_i > U_j$:*

- *There exists a menu $S \in \mathcal{C}^{(m)} : i, j \in S$ such that $\rho(i|S) \neq \rho(j|S)$, or*
- *$m \geq 3$ and there exists a menu $S' \in \mathcal{C}^{(m-1)} : i, j \notin S$ such that $\rho(i|S \cup \{i\}) \neq \rho(j|S \cup \{j\})$,*

then ρ satisfies Borda consistency.

Proof. Consider an IID-RUM ρ and a fixed menu size m. For any pair of items i, j where $U_i > U_j$, one can easily check that:

$$\rho(i|S) \geq \rho(j|S) \, \forall S \in \mathcal{C}^{(m)} : i, j \in S$$

and

$$\rho(i|S' \cup \{i\}) \geq \rho(j|S' \cup \{j\}) \, \forall S' \in \mathcal{C}^{(m-1)} : i, j \notin S'$$

If there exists a menu $S \in \mathcal{C}^{(m)} : i, j \in S$ where $\rho(i|S) \neq \rho(j|S)$, then $\rho(i|S) > \rho(j|S)$. Similarly, if there exists a menu $S' \in \mathcal{C}^{(m-1)} : i, j \notin S$ where $\rho(i|S \cup \{i\}) \neq \rho(j|S \cup \{j\})$ then $\rho(i|S \cup \{i\}) > \rho(j|S \cup \{j\})$. If either of these cases hold, we have:

$$\begin{aligned}
\tau_i^{(m)} &= \frac{1}{|\mathcal{C}_i^{(m)}|} \sum_{S \in \mathcal{C}_i^{(m)}} \rho(i|S) \\
&= \frac{1}{|\mathcal{C}_i^{(m)}|} \left(\sum_{S:|S|=m, i,j \in S} \rho(i|S) + \sum_{S:|S|=m, i \in S, j \notin S} \rho(i|S) \right) \\
&> \frac{1}{|\mathcal{C}_j^{(m)}|} \left(\sum_{S:|S|=m, i,j \in S} \rho(j|S) + \sum_{S:|S|=m, j \in S, i \notin S} \rho(j|S) \right) = \tau_j^{(m)}
\end{aligned}$$

This completes the proof.■

10.2 SAMPLE COMPLEXITY FOR EXACT TOP-K RECOVERY

In this section we prove all theorems stated in section (4) as well as the sample complexity for approximate top-K ranking of the choice based Borda count algorithm.

Lemma 4.2. *Consider an IID-RUM that satisfies Borda consistency per definition (1). Assume input choice data with menu size m is generated according to the sampling model described in section (4.1). For any two items i and j where $\tau_i^{(m)} > \tau_j^{(m)}$, the choice-based Borda count algorithm satisfies:*

$$\mathbb{P}(\hat{W}_j > \hat{W}_i) \leq \exp\left(\frac{-3pR\binom{n}{m} m (\tau_i^{(m)} - \tau_j^{(m)})^2}{8n(\tau_i^{(m)} + \tau_j^{(m)})} \right).$$

Proof. Let $\mathcal{E}$ denote the event $\{i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*\}$. We wish to prove that if $\tau_i - \tau_j > \Delta_K^{(m)}$, then with sufficiently large p and R, $\hat{W}_i$ is smaller than $\hat{W}_j$ with very small probability. The main external concentration inequality used in this proof is Bernstein's inequality (cf. Theorem 2.8.4 Vershynin [2018]). We start our proof by expanding on the probability of

misranking i and j:

$$
\begin{aligned}
&\mathbb{P}\Big(\hat{W}_j > \hat{W}_i \,|\, \mathcal{E}\Big) \\
&= \mathbb{P}\Big(\sum_{r=1}^{R}\sum_{S\in\mathcal{C}_j^{(m)}} \underbrace{\mathbf{1}[y_S^{(r)}=j]}_{X_S^{j,r}} - \sum_{r=1}^{R}\sum_{S\in\mathcal{C}_i^{(m)}} \underbrace{\mathbf{1}[y_S^{(r)}=i]}_{X_S^{i,r}} > 0 \,|\, \mathcal{E}\Big) \\
&= \mathbb{P}\Big(\sum_{r=1}^{R}\sum_{S\in\mathcal{C}_j^{(m)}} X_S^{j,r} - \mathbb{E}[X_S^{j,r}] + \mathbb{E}[X_S^{j,r}] - \sum_{r=1}^{R}\sum_{S\in\mathcal{C}_i^{(m)}} X_S^{i,r} - \mathbb{E}[X_S^{i,r}] + \mathbb{E}[X_S^{i,r}] > 0 \,|\, \mathcal{E}\Big) \\
&= \mathbb{P}\Big(\sum_{r=1}^{R}\sum_{S\in\mathcal{C}_j^{(m)}} X_S^{j,r} - \mathbb{E}[X_S^{j,r}] - \sum_{r=1}^{R}\sum_{S\in\mathcal{C}_i^{(m)}} X_S^{i,r} - \mathbb{E}[X_S^{i,r}] > \sum_{r=1}^{R}\sum_{S\in\mathcal{C}_i^{(m)}} \mathbb{E}[X_S^{i,r}] - \sum_{r=1}^{R}\sum_{S\in\mathcal{C}_j^{(m)}} \mathbb{E}[X_S^{j,r}] \,|\, \mathcal{E}\Big) \\
&= \mathbb{P}\Big(\sum_{r=1}^{R}\sum_{S\in\mathcal{C}_j^{(m)}} X_S^{j,r} - \mathbb{E}[X_S^{j,r}] - \sum_{r=1}^{R}\sum_{S\in\mathcal{C}_i^{(m)}} X_S^{i,r} - \mathbb{E}[X_S^{i,r}] > pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})\Big)
\end{aligned}
\tag{2}
$$

In the last equality, we have used the fact that $\sum_{r=1}^{R}\sum_{S\in\mathcal{C}_i^{(m)}} \mathbb{E}[X_S^{i,r}] - \sum_{r=1}^{R}\sum_{S\in\mathcal{C}_j^{(m)}} \mathbb{E}[X_S^{j,r}] = pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)}) \geq pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})$. Now, we will expand the two terms on the LHS of the inequality as:

$$
\begin{aligned}
&\mathbb{P}\Big(\hat{W}_j > \hat{W}_j \,|\, \mathcal{E}\Big) \\
&= \mathbb{P}\Big(\sum_{S\in\mathcal{C}_j^{(m)}:i\notin S}\sum_{r=1}^{R}\big[X_S^{j,r} - \mathbb{E}[X_S^{j,r}]\big] - \sum_{S\in\mathcal{C}_i^{(m)}:j\notin S}\sum_{r=1}^{R}\big[X_S^{i,r} - \mathbb{E}[X_S^{i,r}]\big] \\
&+ \sum_{S\in\mathcal{C}^{(m)}:i,j\in S}\sum_{r=1}^{R}\big[(X_S^{j,r} - X_S^{i,r}) - (\mathbb{E}[X_S^{j,r} - X_S^{i,r}])\big] > pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})\Big)
\end{aligned}
\tag{3}
$$

Note that with the above decomposition, the terms $\{X_S^{j,r}\}_{S\in\mathcal{C}_j^{(m)}:i\notin S}, \{X_S^{i,r}\}_{S\in\mathcal{C}_i^{(m)}:j\notin S}, \{X_S^{j,r} - X_S^{i,r}\}_{S\in\mathcal{C}^{(m)}:i,j\in S}$ are all mutually independent. To apply Bernstein's inequality, we need to evaluate (or upper bound) the following:

$$
\sum_{S\in\mathcal{C}_j^{(m)}:i\notin S}\sum_{r=1}^{R}\mathbb{E}[(X_S^{j,r})^2] + \sum_{S\in\mathcal{C}_i^{(m)}:j\notin S}\sum_{r=1}^{R}\mathbb{E}[(X_S^{i,r})^2] + \sum_{S\in\mathcal{C}^{(m)}:i,j\in S}\sum_{r=1}^{R}\mathbb{E}[(X_S^{j,r} - X_S^{i,r})^2]
$$

We can easily see that for $S \in \mathcal{C}_j^{(m)} : i \notin S$:

$$
X_S^{j,r} = \begin{cases} 1 & \text{with probability } p\cdot\rho(j|S) \\ 0 & \text{with probability } 1 - p\cdot\rho(j|S) \end{cases}
$$

and for $S \in \mathcal{C}^{(m)} : i, j \in S$:

$$
X_S^{j,r} - X_S^{i,r} = \begin{cases} 1 & \text{with probability } p\cdot\rho(j|S) \\ -1 & \text{with probability } p\cdot\rho(i|S) \\ 0 & \text{with probability } 1 - p\cdot\rho(j|S) - p\cdot\rho(i|S) \end{cases}
$$

As such $\mathbb{E}[(X_S^{j,r})^2] = p \cdot \rho(j|S)$ and $\mathbb{E}[(X_S^{j,r} - X_S^{i,r})^2] = p \cdot (\rho(j|S) + \rho(i|S))$. We have:

$$
\begin{aligned}
&\sum_{S \in \mathcal{C}_j^{(m)}: i \notin S} \sum_{r=1}^{R} \mathbb{E}[(X_S^{j,r})^2] + \sum_{S \in \mathcal{C}_i^{(m)}: j \notin S} \sum_{r=1}^{R} \mathbb{E}[(X_S^{i,\ r})^2] + \sum_{S \in \mathcal{C}^{(m)}: i,j \in S} \sum_{r=1}^{R} \mathbb{E}[(X_S^{j,r} - X_S^{i,r})^2] \\
&= \sum_{S \in \mathcal{C}_j^{(m)}: i \notin S} \sum_{r=1}^{R} p \cdot \rho(j|S) + \sum_{S \in \mathcal{C}_i^{(m)}: j \notin S} \sum_{r=1}^{R} p \cdot \rho(i|S) + \sum_{S \in \mathcal{C}^{(m)}: i,j \in S} \sum_{r=1}^{R} p \cdot (\rho(j|S) + \rho(i|S)) \\
&= \sum_{S \in \mathcal{C}_j^{(m)}} \sum_{r=1}^{R} p \cdot \rho(j|S) + \sum_{S \in \mathcal{C}_i^{(m)}} \sum_{r=1}^{R} p \cdot \rho(i|S) \\
&= pR\binom{n-1}{m-1}\tau_i^{(m)} + pR\binom{n-1}{m-1}\tau_j^{(m)} \\
&= pR\binom{n-1}{m-1}(\tau_i^{(m)} + \tau_j^{(m)})
\end{aligned}
$$

Now, applying Bernstein's inequality to (3) directly yields:

$$
\begin{aligned}
&\mathbb{P}\left(\hat{W}_j > \hat{W}_j \,|\, \mathcal{E}\right) \\
&\leq \exp\left(-\frac{\left(pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})\right)^2}{2(pR\binom{n-1}{m-1}(\tau_i^{(m)} + \tau_j^{(m)}) + \frac{pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})}{3})}\right) \\
&= \exp\left(-\frac{pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})^2}{2(\tau_i^{(m)} + \tau_j^{(m)}) + \frac{2(\tau_i^{(m)} - \tau_j^{(m)})}{3}}\right) \\
&\leq \exp\left(-\frac{pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})^2}{8/3(\tau_i^{(m)} + \tau_j^{(m)})}\right)
\end{aligned}
$$

This completes the proof. ■

Next we present the proof of Lemma (4.2).

Lemma 4.3. *Consider an IID-RUM that satisfies Borda consistency per definition (1). Fix a K, we have:*

$$
\frac{\tau_K^{(m)} + \tau_{K+1}^{(m)}}{{\Delta_K^{(m)}}^2} = \max_{i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*} \left\{ \frac{\tau_i^{(m)} + \tau_j^{(m)}}{(\tau_i^{(m)} - \tau_j^{(m)})^2} \right\}.
$$

Proof. Firstly, it is easy to see that, fixing an $i \in \mathcal{S}_K^*$, for all $j \notin \mathcal{S}_K^*$:

$$
\frac{\tau_i^{(m)} + \tau_{K+1}^{(m)}}{(\tau_i^{(m)} - \tau_{K+1}^{(m)})^2} \geq \frac{\tau_i^{(m)} + \tau_j^{(m)}}{(\tau_i^{(m)} - \tau_j^{(m)})^2}
$$

It remains to prove that for any $i \in \mathcal{S}_K^*$:

$$
\frac{\tau_K^{(m)} + \tau_{K+1}^{(m)}}{(\tau_K^{(m)} - \tau_{K+1}^{(m)})^2} \geq \frac{\tau_i^{(m)} + \tau_{K+1}^{(m)}}{(\tau_i^{(m)} - \tau_{K+1}^{(m)})^2}
$$

To declutter the notation, we'll remove the superscript m. Let $\Delta_{iK} = \tau_i - \tau_K$ for some $i \in \mathcal{S}_K^*$. For any $i \in \mathcal{S}_K^* \backslash \{K\}$, we

have:

$$\begin{aligned}
\frac{\tau_K + \tau_{K+1}}{\Delta_K^2} &\geq \frac{\tau_i + \tau_{K+1}}{(\tau_i - \tau_{K+1})^2} \\
\Leftrightarrow \frac{\tau_K + \tau_{K+1}}{\Delta_K^2} &\geq \frac{\tau_K + \Delta_{iK} + \tau_{K+1}}{(\Delta_{iK} + \Delta_K)^2} \\
\Leftrightarrow \frac{\tau_K + \tau_{K+1}}{\Delta_K^2} &\geq \frac{\tau_K + \Delta_{iK} + \tau_{K+1}}{\Delta_{iK}^2 + \Delta_K^2 + 2\Delta_{iK}\Delta_K} \\
\Leftrightarrow \tau_K\Delta_{iK}^2 + \tau_K\Delta_K^2 + 2\tau_K\Delta_{iK}\Delta_K + \tau_{K+1}\Delta_{iK}^2 &+ \\
\tau_{K+1}\Delta_K^2 + 2\tau_{K+1}\Delta_{iK}\Delta_K &\geq \tau_K\Delta_K^2 + \tau_{K+1}\Delta_K^2 + \Delta_{iK}\Delta_K^2 \\
\Leftrightarrow \tau_K\Delta_{iK}^2 + 2\tau_K\Delta_{iK}\Delta_K + \tau_{K+1}\Delta_{iK}^2 + 2\tau_{K+1}\Delta_{iK}\Delta_K &\geq \Delta_{iK}\Delta_K^2
\end{aligned}$$

In deriving the last statement, the first two terms on the RHS get canceled out. Note that $2\tau_K\Delta_{iK}\Delta_K \geq 2\Delta_K\Delta_{iK}\Delta_K \geq \Delta_{iK}\Delta_K^2$. This completes the proof. ■

Theorem 4.4. *Assume the conditions of lemma (4.2). Given sufficiently large p, R such that $pR\binom{n}{m} \geq \frac{8n\log n}{m\Delta_K^{(m)2}} \cdot (\Delta_K^{(m)} + 2\tau_{K+1}^{(m)})$, the choice-based Borda count algorithm correctly identifies all of the top K items with probability at least $1 - O(\frac{K}{n^2})$.*

Proof. The exponential bound in Lemma (4.2) holds simultaneously for all pairs $i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*$ given sufficiently large p, R such that:

$$pR\binom{n-1}{m-1} \geq \max_{i\in\mathcal{S}_K^*, j\notin\mathcal{S}_K^*} \frac{8\log n}{(\tau_i^{(m)} - \tau_j^{(m)})^2} \cdot (\tau_i^{(m)} + \tau_j^{(m)}) \,.$$

From Lemma (4.3), we have:

$$\max_{i\in\mathcal{S}_K^*, j\notin\mathcal{S}_K^*} \frac{\tau_i^{(m)} + \tau_j^{(m)}}{(\tau_i^{(m)} - \tau_j^{(m)})^2} = \frac{\tau_K^{(m)} + \tau_{K+1}^{(m)}}{{\Delta_K^{(m)}}^2} \,.$$

This means that if

$$pR\binom{n-1}{m-1} \geq 8\log n \cdot \frac{\tau_K^{(m)} + \tau_{K+1}^{(m)}}{{\Delta_K^{(m)}}^2}$$

then

$$\mathbb{P}(\hat{W}_i < \hat{W}_j) \leq \frac{1}{n^3} \,.$$

Applying union bound over all pairs i, j such that $i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*$ gives

$$\hat{W}_i > \hat{W}_j \;\forall i \in \mathcal{S}_K^*, j \notin \mathcal{S}_K^*$$

with probability at least $1 - O(\frac{K}{n^2})$. ■

10.3 SAMPLE COMPLEXITY FOR APPROXIMATE TOP-K RECOVERY

In practice, some error in the top K estimate could be tolerable. This is known as approximate top-K ranking. The metric of interest in approximate top-K ranking is the edit distance between the estimate $\hat{\mathcal{S}}_K$ and the true top K items $\mathcal{S}_K^*$:

$$D_{01}(\hat{\mathcal{S}}_K, \mathcal{S}_K^*) = K - |\hat{\mathcal{S}}_K \cup \mathcal{S}_K| \,.$$

One can see that correctly separating the top $K - h$ items and the bottom $n - K - h$ items, i.e., unable to identify at most h of the top K items, is sufficient to guarantee that the approximate loss is bounded by h. Considering this, the fundamental quantity that determines the hardness of approximate top-K ranking is the gap between the generalized Borda Score of the $K - h$ item and that of the $K + h + 1$ item. For convenience, let us denote such quantity as follows:

$$\Delta_{K,h}^{(m)} = \tau_{K-h-1}^{(m)} - \tau_{K+h+1}^{(m)}$$

Clearly, this quantity generalizes $\Delta_K^{(m)}$ as $\Delta_{K,0}^{(m)} = \Delta_K^{(m)}$. We are interested in bounding the probability that, given an error threshold h, Generalized Borda Count fails to output an estimate with loss bounded by h.

$$\mathbb{P}(D_{01}(\hat{\mathcal{S}}_K, \mathcal{S}_K^*) \leq h)$$

Building on lemma (4.2), we obtain the following sample complexity of Generalized Borda Count for approximate top-K ranking:

Theorem 10.6. *Assume the conditions of lemma (4.2). Fix an error threshold h. Given sufficiently large p, R such that $pR\binom{n}{m} \geq \frac{8n\log n}{m\Delta_{K,h}^{(m)}} \cdot (1 + \frac{\tau_{K+h+1}^{(m)}}{\Delta_{K,h}^{(m)}})$, the choice based Borda count algorithm outputs a set $\hat{\mathcal{S}}_K$ that satisfies:*

$$D_H(\mathcal{S}_K^*, \hat{\mathcal{S}}_K) \leq h$$

with probability at least $1 - O(\frac{K}{n^2})$.

Proof. Applying lemma (4.2), we have for every pair i, j where $i \in \mathcal{S}_{K-h}^*$, $j \notin \mathcal{S}_{K+h}^*$:

$$\mathbb{P}(\hat{W}_j > \hat{W}_i) \leq \exp\left(-\frac{3pR\binom{n-1}{m-1}(\tau_i^{(m)} - \tau_j^{(m)})^2}{8(\tau_i^{(m)} + \tau_j^{(m)})}\right)$$

Following the same argument as in the proof of theorem (4.4), one can show that:

$$\frac{\tau_{K-h} + \tau_{K+h+1}}{\Delta_{K,h}^2} = \max_{i \in \mathcal{S}_{K-h}^*, j \notin \mathcal{S}_{K+h}^*} \frac{\tau_i^{(m)} + \tau_j^{(m)}}{(\tau_i^{(m)} - \tau_j^{(m)})^2}$$

The rest of the proof follows by showing that given the condition on p, R in the theorem statement, the probability that Generalized Borda Count wrongly flips the relative order of any two items $i \in \mathcal{S}_{K-h}^*, j \notin \mathcal{S}_{K+h}^*$ is upper bounded by $O(\frac{1}{n^3})$. Applying union bound over all such pairs of i, j completes the proof. ■

In short, Generalized Borda Count has the following sample complexity for approximate top-K ranking

$$O\left(\frac{n\log n}{m\Delta_{K,h}^{(m)}} \cdot (1 + \frac{\tau_{K+h+1}^{(m)}}{\Delta_{K,h}^{(m)}}\right).$$

10.4 LOWER BOUND ON THE SAMPLE COMPLEXITY OF TOP-K RECOVERY

10.4.1 Preliminaries and notations

We first restate a version of Fano's inequality Cover [1999] which will be useful to the construction of our lower bound:

Lemma 10.7 (Fano's inequality)**.** *Consider a set of L distributions $\{\mathbb{P}^1, \ldots, \mathbb{P}^L\}$. Suppose that we observe a random variable (or a set of random variables) Y that was generated by first picking an index $A \in \{1, \ldots, L\}$ uniformly at random and then $Y \sim \mathbb{P}^A$. Fano's inequality states that any hypothesis test ϕ for this problem has an error probability lower bounded as:*

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{\max_{a,b \in [L], a \neq b} \mathbb{D}_{KL}(\mathbb{P}^a(Y) \| \mathbb{P}^b(Y)) + \log 2}{\log L}$$

To obtain a lower bound in the form of theorem 5.1, we need to construct an IID-RUM such that any method requires $\Omega(n\log n)$ examples in order to accurately recover the top K items. For this purpose, we look for a model within the Multinomial Logit family. It is well known that the MNL model is an instance of IID-RUM where the noise distribution is the standard Gumbel distribution. While the MNL model has a random utility characterization, it is often more convenient to describe the MNL model in terms of weighted probabilities. Specifically, an MNL model over n items can be parameterized

by a set of n positive real numbers $w_1, \ldots, w_n$ called weights. There is also a well defined relation between the weighted probabilities representation and the utility partworth representations. Specifically, the resulting choice probability is defined to be:

$$\rho(i|S) = \frac{w_i}{\sum_{k \in S} w_k} = \frac{e^{U_i}}{\sum_{k \in S} e^{U_k}}$$

Now consider a special sub-class of MNL models that is defined by 3 parameters $(v, \delta, \mathcal{S}_K^*)$ for some $0 < \delta < v$ and a set $\mathcal{S}_K^*$ of size K. The item-specific weights are then defined as follows:

$$w_i = \begin{cases} v + \delta & \text{if } i \in \mathcal{S}_K^* \\ v & \text{otherwise} \end{cases}$$

That is, all the top items have the same weights and all the bottom items have the same (but smaller) weight. For any $A \in [K, \ldots, n]$ let $\mathcal{M}^A$ be a special MNL model parametrized by $(v, \delta, \{1, \ldots, K-1, A\})$ as described above. One can see that Fano's inequality can be used to lower bound the error probability of any hypothesis test for the identity of A. That is, to lower bound the probability that any estimator fails to correctly identify the K-th item.

Recall our uniform sampling model described in subsection (4.1), let us use $\{Y_S^r\}_{S \in \mathcal{C}^{(m)}, r=1:R}$ to denote the observed choice data where:

$$Y_S^r = \begin{cases} y & \text{for some } y \in S \text{ if menu } S \text{ is offered in round } r \\ 0 & \text{if menu } S \text{ is not offered in round } r \end{cases}$$

As a short hand, denote ρ^a as the choice rule corresponding to $\mathcal{M}^a$ and ρ^b corresponding to $\mathcal{M}^b$. Let $\mathbb{P}^a(\{Y_S^r\})$ denote the likelihood of choice data $\{Y_S^r\}$ under choice model $\mathcal{M}^a$ (define analogously for $\mathcal{M}^b$).

10.4.2 Exact top-K recovery

The following pair of lemma and theorem establish the lower bound for exact top-K recovery via showing an upper bound on the KL divergence between any two models $\mathcal{M}^a, \mathcal{M}^b$ and then applying Fano's inequality.

Lemma 10.8. *Assuming the same sampling model as in theorem (4.2), we have, for any $a, b \in [K, \ldots, n]$:*

$$\mathbb{D}_{KL}(\mathbb{P}^a(\{Y_S^r\}) \| \mathbb{P}^b(\{Y_S^r\})) \leq 4pR \binom{n-1}{m-1} \frac{{\Delta_K^{(m)}}^2}{\tau_K^{(m)}}$$

Theorem 5.1. *Consider the sampling model described in subsection (4.1). There exists a choice model $\mathcal{M}^*$ within the class of Multinomial Logit Models (MNLs) such that for $n \geq 20$, if $pR\binom{n}{m} \leq \frac{n \log n}{8} \cdot \frac{\tau_K^{(m)}}{m {\Delta_K^{(m)}}^2}$ then any statistical estimator fails to correctly identify all of the top K items with probability at least $\frac{1}{12}$.*

Once lemma (10.8) has been established, we can prove theorem (5.1) as follows. For any $n \geq 20$, if ${\Delta_K^{(m)}}^2 \leq \frac{\tau_K^{(m)} \cdot \log n}{8pR\binom{n-1}{m-1}}$, Fano's inequality gives us a lower bound on the error of any estimator for finding the top K items under $\mathcal{M}(v, \delta, \mathcal{S}_K^*)$:

$$\mathbb{P}(\hat{\mathcal{S}}_K \neq \mathcal{S}_K^*) \geq 1 - \frac{0.5 \log n + \log 2}{\log(n - K + 1)} \geq 1 - \frac{0.5 \log n + \log 2}{\log(n/2 + 1)} \geq 0.086\,.$$

The rest of the proof of theorem (5.1) involves simple algebra, noting that $\binom{n}{m} = \binom{n-1}{m-1} \cdot \frac{n}{m}$. We now move on to the proof of lemma (10.8).

Proof. (Proof of lemma (10.8)). Let $\hat{\mathbb{C}}^{(m,r)}$ denote the set of menus of size m being picked in round r under the sampling

model described in subsection (4.1). Let us first decompose $\mathbb{D}_{\text{KL}}(\mathbb{P}^a(\{Y_S^r\})\|\mathbb{P}^b(\{Y_S^r\}))$ into a more manageable form:

$$
\begin{aligned}
&\mathbb{D}_{\text{KL}}(\mathbb{P}^a(\{Y_S^r\})\|\mathbb{P}^b(\{Y_S^r\})) \\
&[\text{thanks to mutual independence among } \{Y_S^r\}] \\
&= \sum_{r=1}^{R} \sum_{S\in\mathcal{C}^{(m)}} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^r)\|\mathbb{P}^b(Y_S^r)) \\
&= \sum_{r=1}^{R} \sum_{S\in\mathcal{C}^{(m)}} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^r|S\in\mathcal{C}_r^{(m,r)})\mathbb{P}^a(S\in\hat{\mathcal{C}}^{(m)})\|\mathbb{P}^b(Y_S^r|S\in\hat{\mathcal{C}}^{(m,r)})\mathbb{P}^a(S\in\hat{\mathcal{C}}^{(m)})) \\
&= \sum_{r=1}^{R} \sum_{S\in\mathcal{C}^{(m)}} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^r|S\in\hat{\mathcal{C}}^{(m,r)})\mathbb{P}^a(S\in\hat{\mathcal{C}}^{(m,r)})\|\mathbb{P}^b(Y_S^r|S\in\hat{\mathcal{C}}_r^{(m,r)})\mathbb{P}^b(S\in\hat{\mathcal{C}}^{(m,r)})) \\
&[\text{Using the fact that } \mathbb{P}^a(S\in\hat{\mathcal{C}}^{(m,r)}) = \mathbb{P}^b(S\in\hat{\mathcal{C}}^{(m,r)}) = p] \\
&= pR \sum_{S\in\mathcal{C}^{(m)}} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^1|S\in\hat{\mathcal{C}}^{(m,1)})\|\mathbb{P}^b(Y_S^1|S\in\hat{\mathcal{C}}^{(m,1)}))
\end{aligned}
$$

To handle the above sum, partition $\mathcal{C}^{(m)}$ into 4 subsets $\{S\in\mathcal{C}^{(m)} : a,b\notin S\}, \{S\in\mathcal{C}^{(m)} : a\in S, b\notin S\}, \{S\in\mathcal{C}^{(m)} : a\notin S, b\in S\}, \{S\in\mathcal{C}^{(m)} : a,b\in S\}$. It is easy to see that for $S : a,b\notin S$, the corresponding KL divergence terms are all 0. Hence, we only have to work with other 3 subsets. Before going into the details, let us define some shorthand notations: let $w^a(S)$ denote the sum of the weights of the items in menu S under $\mathcal{M}^a$ and $S+a$ denote set union.

Subset 1: $\{S\in\mathcal{C}^{(m)} : a,b\in S\}$

$$
\begin{aligned}
&\sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^1|S\in\hat{\mathcal{C}}^{(m,1)})\|\mathbb{P}^b(Y_S^1|S\in\hat{\mathcal{C}}^{(m,1)})) \\
&[\text{Note that } \rho^a(x|S) = \rho^b(x|S) \quad \forall x\neq a,b \text{ and } w^a(S) = w^b(S) \text{ for all } S : a,b\in S] \\
&= \sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \rho^a(a|S)\log\frac{\rho^a(a|S)}{\rho^b(a|S)} + \rho^a(b|S)\log\frac{\rho^a(b|S)}{\rho^b(b|S)} \\
&= \sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{v+\delta}{w^a(S)}\log\frac{v+\delta}{v} + \frac{v}{w^a(S)}\log\frac{v}{v+\delta} \\
&= \sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{\delta}{w^a(S)}\log\frac{v+\delta}{v}
\end{aligned}
$$

Subset 2: $\{S\in\mathcal{C}^{(m)} : a\in S, b\notin S\}$. Firstly, it is easy to verify the following useful identities: $w^a(S'+a) = w^b(S'+b), w^a(S'+b) = w^b(S'+a)\forall S' : a,b\notin S'$. Furthermore, $w^a(x) = w^b(x)\forall x\neq a,b$. Now we have

$$
\begin{aligned}
&\sum_{S\in\mathcal{C}^{(m)}:a\in S,b\notin S} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^1|S\in\hat{\mathcal{C}}^{(m,1)})\|\mathbb{P}^b(Y_S^1|S\in\hat{\mathcal{C}}^{(m,1)})) \\
&= \sum_{S'\in\mathcal{C}^{(m-1)}:a,b\notin S'} \left[\left(\sum_{x\in S'} \rho^a(x|S'+a)\log\frac{\rho^a(x|S'+a)}{\rho^b(x|S'+a)}\right) + \rho^a(a|S'+a)\log\frac{\rho^a(a|S'+a)}{\rho^b(a|S'+a)}\right] \\
&= \sum_{S'\in\mathcal{C}^{(m-1)}:a,b\notin S'} \left[\left(\sum_{x\in S'} \frac{w^a(x)}{w^a(S'+a)}\log\frac{w^b(S'+a)}{w^a(S'+a)}\right) + \frac{v+\delta}{w^a(S'+a)}\log\frac{v+\delta}{v}\cdot\frac{w^b(S'+a)}{w^a(S'+a)}\right] \\
&= \sum_{S'\in\mathcal{C}^{(m-1)}:a,b\notin S'} \left[\left(\sum_{x\in S'} \frac{w^a(x)}{w^a(S'+a)}\log\frac{w^a(S'+b)}{w^a(S'+a)}\right) + \frac{v+\delta}{w^a(S'+a)}\log\frac{v+\delta}{v}\cdot\frac{w^a(S'+b)}{w^a(S'+a)}\right]
\end{aligned}
$$

Subset 3: $\{S \in \mathcal{C}^{(m)} : a \in S, b \notin S\}$. Following a very similar procedure to the 2nd subset, we obtain:

$$\sum_{S \in \mathcal{C}^{(m)}: a \notin S, b \in S} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}) \| \mathbb{P}^b(Y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}))$$
$$= \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S'} \left[\left(\sum_{x \in S'} \frac{w^a(x)}{w^a(S'+b)} \log \frac{w^a(S'+a)}{w^a(S'+b)} \right) + \frac{v}{w^a(S'+b)} \log \frac{v}{v+\delta} \cdot \frac{w^a(S'+a)}{w^a(S'+b)} \right]$$

Focusing on the terms from the 2nd and 3rd subset and grouping them in an intelligent way gives us:

$$\sum_{S \in \mathcal{C}^{(m)}: a \notin S, b \in S} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}) \| \mathbb{P}^b(Y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}))$$
$$+ \sum_{S \in \mathcal{C}^{(m)}: a \in S, b \notin S} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(Y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}) \| \mathbb{P}^b(Y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}))$$
$$= \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S} \left[\left(\sum_{x \in S'} \frac{w^a(x)}{w^a(S'+a)} \log \frac{w^a(S'+b)}{w^a(S'+a)} \right) + \frac{v+\delta}{w^a(S'+a)} \log \frac{v+\delta}{v} \cdot \frac{w^a(S'+b)}{w^a(S'+a)} \right]$$
$$+ \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S} \left[\left(\sum_{x \in S'} \frac{w^a(x)}{w^a(S'+b)} \log \frac{w^a(S'+a)}{w^a(S'+b)} \right) + \frac{v}{w^a(S'+b)} \log \frac{v}{v+\delta} \cdot \frac{w^a(S'+a)}{w^a(S'+b)} \right]$$
$$= \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S} \left[\left(\frac{v+\delta}{w^a(S'+a)} - \frac{v}{w^a(S'+b)} \right) \log \frac{v+\delta}{v} \right]$$
$$+ \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S} \log \frac{w^a(S'+a)}{w^a(S'+b)} \cdot \left[\sum_{x \in S'} \left(\frac{w^a(x)}{w^a(S'+b)} - \frac{w^a(x)}{w^a(S'+a)} \right) + \frac{v}{w^a(S'+b)} - \frac{v+\delta}{w^a(S'+a)} \right]$$
$$= \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S} \left[\left(\frac{v+\delta}{w^a(S'+a)} - \frac{v}{w^a(S'+b)} \right) \log \frac{v+\delta}{v} \right]$$

The last equality comes from recognizing that $\sum_{x \in S'} \frac{w^a(x)}{w^a(S'+b)} + \frac{v}{w^a(S'+b)} = \frac{w^a(S'+b)}{w^a(S'+b)} = 1$ and similarly $\sum_{x \in S'} \frac{w^a(x)}{w^a(S'+a)} + \frac{v+\delta}{w^a(S'+a)} = \frac{w^a(S'+a)}{w^a(S'+a)} = 1$ so they cancel out and the second term becomes 0. We now have the following much more compact identity:

$$\mathbb{D}_{\text{KL}}(\mathbb{P}^a(\{Y_S^r\}) \| \mathbb{P}^b(\{Y_S^r\}))$$
$$= pR \cdot \log \frac{v+\delta}{v} \cdot \left(\sum_{S \in \mathcal{C}^{(m)}: a, b \in S} \frac{\delta}{w^a(S)} + \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S} \left[\frac{v+\delta}{w^a(S'+a)} - \frac{v}{w^a(S'+b)} \right] \right)$$

To make a connection between this and $\Delta_K^{(m)}$, recognize that under $\mathcal{M}^a$, $\Delta_K^{(m)} = \tau_a^{(m)} - \tau_b^{(m)}$. We therefore have:

$$\begin{aligned}
\Delta_K^{(m)} &= \tau_a^{(m)} - \tau_b^{(m)} \\
&= \frac{1}{\binom{n-1}{m-1}} \cdot \left(\sum_{S \in \mathcal{C}^{(m)}: a \in S} \rho^a(a|S) - \sum_{S \in \mathcal{C}^{(m)}: b \in S} \rho^a(b|S) \right) \\
&= \frac{1}{\binom{n-1}{m-1}} \cdot \left(\sum_{S \in \mathcal{C}^{(m)}: a \in S} \frac{v+\delta}{w^a(S)} - \sum_{S \in \mathcal{C}^{(m)}: b \in S} \frac{v}{w^a(S)} \right) \\
&= \frac{1}{\binom{n-1}{m-1}} \cdot \left(\sum_{S \in \mathcal{C}^{(m)}: a, b \in S} \frac{v}{w^a(S)} + \sum_{S \in \mathcal{C}^{(m)}: a \in S, b \notin S} \frac{v+\delta}{w^a(S)} - \sum_{S \in \mathcal{C}^{(m)}: a \notin S, b \in S} \frac{v}{w^a(S)} \right) \\
&= \frac{1}{\binom{n-1}{m-1}} \cdot \left(\sum_{S \in \mathcal{C}^{(m)}: a, b \in S} \frac{\delta}{w^a(S)} + \sum_{S' \in \mathcal{C}^{(m-1)}: a, b \notin S'} \left[\frac{v+\delta}{w^a(S'+a)} - \frac{v}{w^a(S'+b)} \right] \right)
\end{aligned}$$

We thus have:

$$\mathbb{D}_{\text{KL}}(\mathbb{P}^a(\{Y_S^r\}) \| \mathbb{P}^b(\{Y_S^r\})) = pR \log \frac{v+\delta}{v} \binom{n-1}{m-1} \Delta_K^{(m)} \tag{4}$$

We will now bound $\log \frac{v+\delta}{v}$ in terms of $\frac{\Delta_K^{(m)}}{\tau_K^{(m)}}$

$$\begin{aligned}
\frac{\Delta_K^{(m)}}{\tau_K^{(m)}} &= \frac{\sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{w^a(a)-w^a(b)}{w^a(S)+2v+\delta} + \sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{w^a(a)}{w^a(S)+v+\delta} - \frac{w^a(b)}{w^a(S)+v}}{\sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{w^a(a)}{w^a(S)+2v+\delta} + \sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{w^a(a)}{w^a(S)+v+\delta}} \\
&= \frac{\sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{\delta}{w^a(S)+2v+\delta} + \sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{\delta(w^a(S))}{(w^a(S)+v+\delta)(w^a(S)+v)}}{\sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{v+\delta}{w^a(S)+2v+\delta} + \sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{v+\delta}{w^a(S)+v+\delta}} \\
&= \frac{\sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{1}{w^a(S)+2v+\delta} + \sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{(w^a(S))}{(w^a(S)+v+\delta)(w^a(S)+v)}}{\sum_{S\in\mathcal{C}^{(m)}:a,b\in S} \frac{1}{w^a(S)+2v+\delta} + \sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{1}{w^a(S)+v+\delta}} \cdot \frac{\delta}{v+\delta} \\
&\geq \frac{\sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{1}{w^a(S)+v+\delta} \cdot \frac{w^a(S)}{w^a(S)+v}}{\sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{1}{w^a(S)+v+\delta}} \cdot \frac{\delta}{v+\delta} \\
&\geq \frac{\sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{1}{w^a(S)+v+\delta} \cdot \frac{(m-1)v}{(m-1)v+v}}{\sum_{S\in\mathcal{C}^{(m-1)}:a,b\notin S} \frac{1}{w^a(S)+v+\delta}} \cdot \frac{\delta}{v+\delta} \\
&= \frac{(m-1)v}{mv} \cdot \frac{\delta}{v+\delta} \\
&= \frac{m-1}{m}\frac{\delta}{v+\delta} \geq \frac{1}{2}\frac{\delta}{v+\delta}
\end{aligned}$$

We can now bound $\log \frac{v+\delta}{v}$ as:

$$\log \frac{v+\delta}{v} \leq \frac{\delta}{v} \leq \frac{2\delta}{v+\delta} \leq \frac{4\Delta_K^{(m)}}{\tau_K^{(m)}}$$

Combining with equation (4) completes the proof. ■

10.4.3 Approximate top-K recovery

To obtain a lower bound on the sample complexity for approximate top-K recovery, we require a more nuanced construction than that in the proof of theorem (5.1). At a high level, we need to construct a multiset of K-size subsets $\{\mathcal{S}^1, \ldots, \mathcal{S}^L\}$. Each K-size subset $\mathcal{S}^l$ corresponds to an MNL model whose top K items are exactly $\mathcal{S}^l$. We need to carefully design this multiset such that the pairwise edit distance between any two subsets is larger than $2h$. This automatically ensures that any top-K estimate is close in edit distance to at most one of the L subsets. At the same time, we also want the distance between the distributions induced by any two models to be small in KL divergence sense. This would then allow us to invoke Fano's lemma to lower bound the probability that any statistical estimator outputs a top-K estimate set with small edit distance error. Lastly, we also want L to be sufficiently large in order to obtain a good lower bound using Fano's lemma.

We first present a reworded version of lemma 9 of Shah and Wainwright [2017]. The original lemma, in turn, is based on a result due to Levenshtein [1971] for fixed weighted binary codes.

Lemma 10.9. *[Shah and Wainwright, 2017, lemma 9] Consider the regime where $h < \frac{2}{3}\min\{K, \sqrt{n}, n-K\}$. For sufficiently large n, there exists a multiset $\{s^1, \ldots, s^L\}$ with cardinality $L \geq \exp\left(\frac{9}{20}h\log n\right)$ and $s^l \subseteq [\frac{n}{2}, n]$ for $l \in [L]$ such that:*

$$|s^l| = \frac{3h}{2}\ \forall l \in [L] \quad \textit{and} \quad D_{01}(s^j, s^l) = 2h+1 \textit{ for all } j \neq l \in [L]$$

By the above lemma, there exists a multiset of cardinality $\exp\left(\frac{9}{20}h\log n\right)$, consisting of sets of size exactly $\frac{3h}{2}$, with elements from $[\frac{n}{2}, n]$. For s^l, let us construct a MNL model as follows: let $u = \{1, \ldots, K - \frac{3h}{2}\}$ and $\mathcal{S}_K^l = u \cup s^l$ be the set of the top K items for model $l \in [L]$. Note that this construction is valid since $h < \frac{2K}{3}$. Following closely the description of the special MNL model used in subsection (10.4.2), we assign v as weight for all items in $[n]\backslash\mathcal{S}_K^l$, $v+\delta/2$ for all items in u and $v+\delta$ for all items in s_l for some $v > \delta > 0$. In other words, v^l is the set of the top $\frac{3h}{2}$ items for the l-th model while u is the set composing of the $\frac{3h}{2}+1$-th, $\ldots$, K-th best item for model l.

We emphasize that by design, $D_{01}(\mathcal{S}_K^l, \mathcal{S}_K^j) > 2h$ for any $l \neq j \in [L]$. This also means that for any estimate $\hat{\mathcal{S}}_K$, there exists at most one index $a \in [L]$ such that $D_{01}(\hat{\mathcal{S}}_K, \mathcal{S}_K^a) \leq h$. In other words, outputting an estimate with low edit distance error is equivalent to exactly identifying the set of the top K items. All that remains is to prove that it is hard to distinguish between any two models as their distributions over observed choice data have small KL divergence. Such a result is summarized in the lemma below:

Lemma 10.10. *Consider the construction described above, for any two models indexed by $a, b \in [L]$ with $a \neq b$. Let $\mathbb{P}^a$ and $\mathbb{P}^b$ be the distributions parametrized by $\mathcal{S}_K^a$ and $\mathcal{S}_K^b$ as described above, respectively. Under the sampling model described in subsection (4.1), we have:*

$$\mathbb{D}_{KL}(\mathbb{P}^a(\{Y\}) \| \mathbb{P}^b(\{Y\})) \leq 12hpR\binom{n-1}{m-1}\frac{{\Delta_{K,h}^{(m)}}^2}{\tau_{K-h}^{(m)}}$$

Before proving lemma (10.10), let us state and prove the sample complexity lower bound for approximate top-K ranking, the proof of which directly makes use of the lemma.

Theorem 10.11. *Consider the sampling model described in subsection (4.1). There exists a choice model $\mathcal{M}^*$ within the class of Multinomial Logit Models (MNLs) such that for $n \geq 20$, if $pR\binom{n}{m} \leq \frac{1}{96}\frac{n \log n}{m\Delta_{K,h}^{(m)}} \cdot \frac{\tau_{K-h}^{(m)}}{\Delta_{K,h}^{(m)}}$ then any statistical estimator outputs an estimate $\hat{\mathcal{S}}_K$ where:*

$$D_{01}(\hat{\mathcal{S}}_K, \mathcal{S}_K^*) > h$$

with probability at least $\frac{1}{5}$.

Proof. Consider the regime when p, R are small enough such that:

$$pR\binom{n-1}{m-1} \leq \frac{\log n}{96{\Delta_{K,h}^{(m)}}^2} \cdot \tau_{K-h}^{(m)} = \frac{\log n}{96{\Delta_{K,h}^{(m)}}^2} \cdot (\tau_{K-h}^{(m)} + \Delta_{K,h}^{(m)})$$

and the multiset construction as described above. Invoking Fano's inequality, we have the probability that any statistical estimator failing to output an estimate $\hat{\mathcal{S}}_K$ such that $D_{01}(\hat{\mathcal{S}}_K, \mathcal{S}_K^*) \leq h$, is lower bounded as:

$$1 - \frac{\frac{1}{8}h\log n + \log 2}{\frac{9}{20}h\log n} \geq \frac{1}{5} \quad \text{for } n \geq 20$$

This finishes the proof. ■

We will now proceed to prove lemma (10.10)

Proof. Following the same notation and argument as in the proof of lemma (10.8), we have:

$$\mathbb{D}_{\text{KL}}(\mathbb{P}^a(\{Y\}) \| \mathbb{P}^b(\{Y\})) = pR \sum_{S \in \mathcal{C}^{(m)}} \mathbb{D}_{\text{KL}}(\mathbb{P}^a(y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}) \| \mathbb{P}^b(y_S^1 | S \in \hat{\mathcal{C}}^{(m,1)}))$$

Similarly to the proof of lemma (10.8) let us define some shorthand notations: let $w^a(S)$ denote the sum of the weights of the items in menu S and $w^a(i)$ the weight of item i under the special MNL model whose top K items is $\mathcal{S}_K^a$ (similarly defined for $b \neq a$). Expanding on the sum of the KL divergence terms and rearranging the summation order gives us:

$$
\begin{aligned}
&\frac{1}{pR}\mathbb{D}_{\mathrm{KL}}(\mathbb{P}^a(\{Y\})\|\mathbb{P}^b(\{Y\}))\\
&= \sum_{i\in\mathcal{S}_K^a\setminus\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^a(i)\cdot w^b(S)}{w^b(i)\cdot w^a(S)} + \sum_{i\in\mathcal{S}_K^b\setminus\mathcal{S}_K^a}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^a(i)\cdot w^b(S)}{w^b(i)\cdot w^a(S)}\\
&+ \sum_{i\in\mathcal{S}_K^a\cap\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^a(i)w^b(S)}{w^b(i)w^a(S)} + \sum_{i\in[n]\setminus(\mathcal{S}_K^a\cup\mathcal{S}_K^b)}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^a(i)w^b(S)}{w^b(i)w^a(S)}\\
&= \sum_{i\in\mathcal{S}_K^a\setminus\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{v+\delta}{w^a(S)}\log\frac{(v+\delta)\cdot w^b(S)}{v\cdot w^a(S)} + \sum_{i\in\mathcal{S}_K^b\setminus\mathcal{S}_K^a}\sum_{S\in C_i^{(m)}}\frac{v}{w^a(S)}\log\frac{v\cdot w^b(S)}{(v+\delta)\cdot w^a(S)}\\
&+ \sum_{i\in\mathcal{S}_K^a\cap\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)} + \sum_{i\in[n]\setminus(\mathcal{S}_K^a\cup\mathcal{S}_K^b)}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}\\
&= \sum_{i\in\mathcal{S}_K^a\setminus\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{v+\delta}{w^a(S)}\log\frac{v+\delta}{v} + \sum_{i\in\mathcal{S}_K^b\setminus\mathcal{S}_K^a}\sum_{S\in C_i^{(m)}}\frac{v}{w^a(S)}\log\frac{v}{v+\delta}\\
&+ \sum_{i\in\mathcal{S}_K^a\setminus\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{v+\delta}{w^a(S)}\log\frac{w^b(S)}{w^a(S)} + \sum_{i\in\mathcal{S}_K^b\setminus\mathcal{S}_K^a}\sum_{S\in C_i^{(m)}}\frac{v}{w^a(S)}\log\frac{w^b(S)}{w^a(S)} \qquad (*)\\
&+ \sum_{i\in\mathcal{S}_K^a\cap\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)} + \sum_{i\in[n]\setminus(\mathcal{S}_K^a\cup\mathcal{S}_K^b)}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}
\end{aligned}
$$

In the second equality, we have made use of the fact that for any item i in $\mathcal{S}_K^a \cup \mathcal{S}_K^b$ and $\mathcal{S}_K^a \cap \mathcal{S}_K^b$, $w^a(i) = w^b(i)$. Focusing on the first two summations:

$$
\begin{aligned}
&\sum_{i\in\mathcal{S}_K^a\setminus\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{v+\delta}{w^a(S)}\log\frac{v+\delta}{v} + \sum_{i\in\mathcal{S}_K^b\setminus\mathcal{S}_K^a}\sum_{S\in C_i^{(m)}}\frac{v}{w^a(S)}\log\frac{v}{v+\delta}\\
&= \left(\sum_{i\in\mathcal{S}_K^a\setminus\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{v+\delta}{w^a(S)} - \sum_{i\in\mathcal{S}_K^b\setminus\mathcal{S}_K^a}\sum_{S\in C_i^{(m)}}\frac{v}{w^a(S)}\right)\log\frac{v+\delta}{v}\\
&= (2h+1)\binom{n-1}{m-1}\Delta_{K,h}^{(m)}\log\frac{\delta+v}{v} < 3h\binom{n-1}{m-1}\Delta_{K,h}^{(m)}\log\frac{\delta+v}{v}
\end{aligned}
$$

The last equality comes from recognizing that $|\mathcal{S}_K^a\setminus\mathcal{S}_K^b| = |\mathcal{S}_K^b\setminus\mathcal{S}_K^a| = D_{01}(\mathcal{S}_K^a, \mathcal{S}_K^b) = 2h+1$. We now simplify the

remaining four summation terms of $(*)$:

$$
\begin{aligned}
&\sum_{i\in\mathcal{S}_K^a\setminus\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{v+\delta}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}+\sum_{i\in\mathcal{S}_K^b\setminus\mathcal{S}_K^a}\sum_{S\in C_i^{(m)}}\frac{v}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}\\
&+\sum_{i\in\mathcal{S}_K^a\cap\mathcal{S}_K^b}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}+\sum_{i\in[n]\setminus(\mathcal{S}_K^a\cup\mathcal{S}_K^b)}\sum_{S\in C_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}\\
&=\sum_{i\in[n]}\sum_{S\in\mathcal{C}_i^{(m)}}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}\\
&=\sum_{S\in\mathcal{C}^{(m)}}\sum_{i\in S}\frac{w^a(i)}{w^a(S)}\log\frac{w^b(S)}{w^a(S)}\\
&=\sum_{S\in\mathcal{C}^{(m)}}\log\frac{w^b(S)}{w^a(S)}\\
&=\log\frac{\prod_{S\in\mathcal{C}^{(m)}}w^b(S)}{\prod_{S\in\mathcal{C}^{(m)}}w^a(S)}\\
&=\log 1=0
\end{aligned}
$$

The last equality comes from recognizing that the two models a and b only differ by the the identities of the items while the weights are the same. Therefore, the two products in the log term are identical. We thus obtain:

$$\mathbb{D}_{\mathrm{KL}}(\mathbb{P}^a(\{Y\})\|\mathbb{P}^b(\{Y\}))<pR3h\binom{n-1}{m-1}\Delta_{K,h}^{(m)}\log\frac{\delta+v}{v}\qquad(**)$$

Following the same argument as in the proof of lemma (10.8), we can bound:

$$\log\frac{\delta+v}{v}\le 4\frac{\Delta_{K,h}^{(m)}}{\tau_{K-h}^{(m)}}$$

Substituting this into $(**)$ completes the proof. ■

In short, the (matching) lower bound on the sample complexity for approximate top-K ranking is:

$$\Omega\left(\frac{n\log n}{m\Delta_{K,h}^{(m)}}\cdot\left(1+\frac{\tau_{K+h+1}^{(m)}}{\Delta_{K,h}^{(m)}}\right)\right)$$

10.5 EFFECT OF THE MENU SIZE ON THE SAMPLE COMPLEXITY

Theorem 6.1. *Consider an MNL model with $n\ge 2$ items and fix a K, we have:*

$$\frac{1}{m\Delta_K^{(m)}}=\theta\left(\frac{1}{e^{U_K}-e^{U_{K+1}}}\cdot\left(1+\frac{1}{m-1}\right)^2\right);$$

$$\frac{\tau_{K+1}^{(m)}}{\Delta_K^{(m)}}=\theta\left(\frac{e^{U_{K+1}}}{e^{U_K}-e^{U_{K+1}}}\cdot\left(1+\frac{1}{m-1}\right)\right).$$

Proof. **Part 1**. We will prove that there exists a positive valued function $l(m)$ such that $m\Delta_K^{(m)}<l(m)$ and that $l(m)$ monotonically increases with m, and at diminishing rate. Once we have obtained $l(m)$, we can set $f_1(m)=\frac{1}{l(m)}$. In fact, we will prove a slightly more general result concerning any pair of items i,j where $w_i:=e^{U_i}>e^{U_j}=:w_j$. Let $\Delta_{ij}^{(m)}=\tau_i^{(m)}-\tau_j^{(m)}$. Additionally, let $w_n:=e^{U_n}=e^{U_{\min}}$ and $w_1:=e^{U_1}=e^{U_{\max}}$ be two model-dependent constants. We have:

$$
\begin{aligned}
m\Delta_{ij}^{(m)} &= \frac{m}{\binom{n-1}{m-1}} \cdot \Bigg(\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_i}{w_i + w_j + \sum_{k \in S} w_k} - \frac{w_j}{w_i + w_j + \sum_{k \in S} w_k} \\
&+ \sum_{S' \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_i}{w_i + \sum_{k \in S} w_k} - \frac{w_j}{w_j + \sum_{k \in S} w_k} \Bigg) \\
&= \frac{m(w_i - w_j)}{\binom{n-1}{m-1}} \Bigg(\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{w_i + w_j + \sum_{k \in S} w_k} + \sum_{S' \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{1}{w_j + \sum_{k \in S} w_k} \cdot \frac{\sum_{k \in S} w_k}{w_i + \sum_{k \in S} w_k} \Bigg) \\
&\leq \frac{m(w_i - w_j)}{\binom{n-1}{m-1}} \Bigg(\binom{n-2}{m-2} \cdot \frac{1}{w_i + w_j + (m-2)w_n} + \binom{n-2}{m-1} \frac{1}{w_j + (m-1)w_n} \cdot \frac{(m-1)w_1}{w_i + (m-1)w_1} \Bigg) \\
&= \frac{m(w_i - w_j)}{n-1} \Bigg((m-1) \cdot \frac{1}{w_i + w_j + (m-2)w_n} + (n-m) \frac{1}{w_j + (m-1)w_n} \cdot \frac{(m-1)w_1}{w_i + (m-1)w_1} \Bigg) \\
&= \frac{(w_i - w_j)}{n-1} \Bigg(n \frac{m}{[w_j + (m-1)w_n]} \frac{(m-1)w_1}{[w_i + (m-1)w_1]} \\
&+ m(m-1) \cdot \left[\frac{1}{w_i + w_j + (m-2)w_n} - \frac{m w_1}{[w_i + (m-1)w_1][w_j + (m-1)w_n]} \right] \Bigg) \\
&= \frac{(w_i - w_j)}{n-1} \Bigg(n \frac{m}{[m w_n + w_j - w_n]} \frac{(m-1)w_1}{[w_i + (m-1)w_1]} \\
&+ m(m-1) \cdot \frac{(w_1 - w_i)(w_n - w_j)}{[(m-2)w_n + w_i + w_j][(m-1)w_n + w_j][(m-1)w_1 + w_i]} \Bigg) \\
&= \frac{(w_i - w_j)}{n-1} \Bigg(n \frac{m}{[m w_n + w_j - w_n]} \frac{(m-1)w_1}{[w_i + (m-1)w_1]} \\
&+ \underbrace{\frac{(w_1 - w_i)(w_n - w_j)}{(m-2)w_n + w_i + w_j} \cdot \frac{m}{m w_n + w_j - w_n} \cdot \frac{(m-1)}{(m-1)w_1 + w_i]}}_{\leq 0} \Bigg) \\
&\leq (w_i - w_j) \cdot \frac{n}{n-1} \cdot \frac{m}{[m w_n + w_j - w_n]} \frac{(m-1)w_1}{[w_i + (m-1)w_1]} \\
&= (w_i - w_j) \cdot \frac{n}{n-1} \cdot (1 - \frac{w_j/w_n - 1}{m - 1 + w_j/w_n}) \cdot (1 - \frac{w_i/w_1}{m - 1 + w_i/w_1})
\end{aligned}
$$

It is easy to see that all of the terms containing m are positive and *increase* with m. Additionally, it is easy to show that all of these terms increase with *decreasing* rate with respect to m. Though complicated, the above lower bound can somewhat be simplified using big-O notations as follows:

$$
m\Delta_{ij}^{(m)} = \Omega\Bigg((w_i - w_j) \cdot (1 - \frac{1}{m})^2 \Bigg)
$$

By modifying the above argument, we can also show a variational lower bound on $m\Delta_{ij}^{(m)}$.

$$
\begin{aligned}
m\Delta_{ij}^{(m)} &= \frac{m}{\binom{n-1}{m-1}} \cdot \Bigg(\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_i}{w_i + w_j + \sum_{k \in S} w_k} - \frac{w_j}{w_i + w_j + \sum_{k \in S} w_k} \\
&+ \sum_{S' \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_i}{w_i + \sum_{k \in S} w_k} - \frac{w_j}{w_j + \sum_{k \in S} w_k} \Bigg) \\
&= \frac{m}{\binom{n-1}{m-1}} \cdot \Bigg(\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_i - w_j}{w_i + w_j + \sum_{k \in S} w_k} + \sum_{S' \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_i}{w_i + \sum_{k \in S} w_k} - \frac{w_j}{w_j + \sum_{k \in S} w_k} \Bigg) \\
&= \frac{m}{\binom{n-1}{m-1}} \cdot \Bigg(\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_i - w_j}{w_i + w_j + \sum_{k \in S} w_k} + \sum_{S' \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{(w_i - w_j)(\sum_{k \in S} w_k)}{(w_i + \sum_{k \in S} w_k)(w_j + \sum_{k \in S} w_k)} \Bigg) \\
&= \frac{m(w_i - w_j)}{\binom{n-1}{m-1}} \Bigg(\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{w_i + w_j + \sum_{k \in S} w_k} + \sum_{S' \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{1}{w_i + \sum_{k \in S} w_k} \cdot \frac{\sum_{k \in S} w_k}{w_j + \sum_{k \in S} w_k} \Bigg) \\
&\text{[Using the fact that } \frac{\sum_{k \in S} w_k}{w_j + \sum_{k \in S} w_k} \geq \frac{\sum_{k \in S} w_n}{w_j + \sum_{k \in S} w_n} \; \forall S : j \in S] \\
&\geq \frac{m(w_i - w_j)}{\binom{n-1}{m-1}} \Bigg(\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{w_i + w_j + \sum_{k \in S} w_k} + \sum_{S' \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{1}{w_i + \sum_{k \in S} w_k} \cdot \frac{(m-1)w_n}{w_j + (m-1)w_n} \Bigg) \\
&\geq \frac{m(w_i - w_j)}{\binom{n-1}{m-1}} \Bigg(\binom{n-2}{m-2} \frac{1}{w_i + w_j + (m-2)\bar{w}} + \binom{n-2}{m-1} \frac{1}{w_i + (m-1)\bar{w}} \cdot \frac{(m-1)w_n}{w_j + (m-1)w_n} \Bigg)
\end{aligned}
$$

where $\bar{w} = \frac{1}{n-2} \sum_{k \neq i,j} w_k$. The last inequality comes from applying Jensen's inequality. More precisely, one could treat the sum $\sum_{k \in S} w_k$ as a random variable and the summation as the unnormalized 'expectation' over the uniform distribution of the menus. As $\frac{1}{x}$ is a convex function in x for positive x, Jensen's inequality applies. Furthermore, without loss of generality, one can assume that $\bar{w} = 1$ (via scaling of the weights). Consequently, $w_n \leq 1$. Continuing with the expansion gives us:

$$
\begin{aligned}
m\Delta_{ij}^{(m)} &\geq \frac{m(w_i - w_j)}{\binom{n-1}{m-1}}\left(\binom{n-2}{m-2}\frac{1}{w_i + w_j + (m-2)\bar{w}} + \binom{n-2}{m-1}\frac{1}{w_i + (m-1)\bar{w}} \cdot \frac{(m-1)w_n}{w_j + (m-1)w_n}\right) \\
&= \frac{m(w_i - w_j)}{n-1}\left(\frac{m-1}{w_i + w_j + (m-2)} + \frac{n-m}{w_i + (m-1)} \cdot \frac{(m-1)w_n}{w_j + (m-1)w_n}\right) \\
&= \frac{(w_i - w_j)}{n-1}\left(\frac{m(m-1)}{w_i + w_j + (m-2)} + \frac{m(n-m)}{w_i + (m-1)} \cdot \frac{(m-1)w_n}{w_j + (m-1)w_n}\right) \\
&\geq \frac{(w_i - w_j)}{n-1}\left(\frac{m(m-1)}{w_i + w_j + (m-2)} + \frac{(m-1)(n-m)}{w_i + (m-1)} \cdot \frac{(m-1)w_n}{w_j + (m-1)w_n}\right) \\
&\geq \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j} + \frac{m(m-1)}{w_i + w_j + (m-2)} - \frac{m(m-1)(m-1)w_n}{(w_i + m - 1)(w_j + (m-1)w_n)}\right) \\
&\geq \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j}\right. \\
&\left. + m(m-1)\left(\frac{1}{w_i + w_j + (m-2)} - \frac{(m-1)w_n}{(w_i + m - 1)(w_j + (m-1)w_n)}\right)\right) \\
&\text{[Expanding on the difference term]} \\
&\geq \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j}\right. \\
&\left. + m(m-1)\left(\frac{1}{w_i + w_j + (m-1)} - \frac{(m-1)w_n}{(w_i + m - 1)(w_j + (m-1)w_n)}\right)\right) \\
&= \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j}\right. \\
&+ m(m-1) \cdot (w_i w_j + w_i(m-1)w_n + m w_j + m(m-1)w_n - w_j - (m-1)w_n - w_i(m-1)w_n - w_j(m-1)w_n - \\
&\left. m(m-1)w_n + 1(m-1)w_n))/\big((w_i + w_j + (m-1))(w_i + m - 1)(w_j + (m-1)w_n)\big)\right) \\
&= \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j}\right. \\
&\left. + m(m-1) \cdot \frac{m w_j + w_i w_j - w_j - w_j(m-1)w_n}{(w_i + w_j + (m-1))(w_i + m - 1)(w_j + (m-1)w_n)}\right) \\
&= \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j}\right. \\
&\left. + m(m-1)w_j \cdot \frac{m + w_i - 1 - (m-1)w_n}{(w_i + w_j + (m-1))(w_i + m - 1)(w_j + (m-1)w_n)}\right) \\
&= \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j} + \frac{m(m-1)w_j((m-1)(1-w_n) + w_i)}{(w_i + w_j + (m-1))(w_i + m - 1)(w_j + (m-1)w_n)}\right) \\
&\geq \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j} + \frac{m(m-1)w_j((m-1)(1-w_n) + w_i)}{(w_i + w_j + m)(w_i + m - 1)(w_j + (m-1)w_n)}\right) \\
&= \frac{(w_i - w_j)}{n-1}\left(n \cdot \frac{m-1}{m-1+w_i} \cdot \frac{(m-1)w_n}{(m-1)w_n + w_j} + \underbrace{w_j \cdot \frac{m}{w_i + w_j + m} \cdot \frac{m-1}{w_j + (m-1)w_n} \cdot \frac{w_i + (m-1)(1-w_n)}{w_i + (m-1)}}_{\geq 0}\right) \\
&\geq \frac{(w_i - w_j)}{n-1}\left(n \cdot (1 - \frac{w_i}{m-1+w_i}) \cdot (1 - \frac{w_j}{(m-1)w_n + w_j})\right) \\
&= (w_i - w_j) \cdot \frac{n}{n-1} \cdot (1 - \frac{w_i}{m-1+w_i}) \cdot (1 - \frac{w_j/w_n}{m-1+w_j/w_n})
\end{aligned}
$$

We thus have $m\Delta_{ij} = O((w_i - w_j) \cdot 1 - \frac{1}{m})^2)$. Combining with the upper bound shown earlier, we get:

$$m\Delta_{ij} = \theta\bigg((w_i - w_j) \cdot (1 - \frac{1}{m})^2\bigg)$$

Part 2. We will prove that there exists a positive valued function $l(m)$ such that $\frac{\Delta_{ij}^{(m)}}{\tau_j^{(m)}} \geq l(m)$ and that $l(m_2) > l(m_1)$ for $m_2 > m_1 \geq 2$. Consider for any two items i, j such that $w_i > w_j$. By definition:

$$\begin{aligned}
\frac{\Delta_{ij}^{(m)}}{\tau_j^{(m)}} &= \frac{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_i - w_j}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_i}{w_i + \sum_{k \in S} w_k} - \frac{w_j}{w_j + \sum_{k \in S} w_k}}{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_j}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_j}{w_j + \sum_{k \in S} w_k}} \\
&= \frac{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_i - w_j}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{(w_i - w_j)(\sum_{k \in S} w_k)}{(w_i + \sum_{k \in S} w_k)(w_j + \sum_{k \in S} w_k)}}{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{w_j}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_j}{w_j + \sum_{k \in S} w_k}} \\
&= \frac{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{\sum_{k \in S} w_k}{(w_i + \sum_{k \in S} w_k)(w_j + \sum_{k \in S} w_k)}}{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{1}{w_j + \sum_{k \in S} w_k}} \cdot \frac{w_i - w_j}{w_j} \\
&= (1 - \frac{\sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_j}{(w_i + \sum_{k \in S} w_k)(w_j + \sum_{k \in S} w_k)}}{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{1}{w_j + \sum_{k \in S} w_k}}) \cdot \frac{w_i - w_j}{w_j}
\end{aligned}$$

The goal is to upper (and lower) bound the quantity

$$\frac{\sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_j}{(w_i + \sum_{k \in S} w_k)(w_j + \sum_{k \in S} w_k)}}{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{1}{w_j + \sum_{k \in S} w_k}}$$

For a lower bound, we can show that:

$$\begin{aligned}
&\frac{\sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{w_j}{(w_i + \sum_{k \in S} w_k)(w_j + \sum_{k \in S} w_k)}}{\sum_{S \in \mathcal{C}_i^{(m)} \cap \mathcal{C}_j^{(m)}} \frac{1}{\sum_{k \in S} w_k} + \sum_{S \in \mathcal{C}^{(m-1)}: i,j \notin S} \frac{1}{w_j + \sum_{k \in S} w_k}} \\
&\geq \frac{\binom{n-2}{m-1} \frac{1}{(m-1)w_1 + w_i} \frac{w_j}{w_j + (m-1)w_1}}{\binom{n-2}{m-2} \cdot \frac{1}{(m-2)w_n + w_i + w_j} + \binom{n-2}{m-1} \cdot \frac{1}{(m-1)w_n + w_j}} \\
&= \frac{(n-m) \cdot \frac{1}{(m-1)w_1 + w_i} \frac{w_j}{w_j + (m-1)w_1}}{(m-1) \cdot \frac{1}{(m-2)w_n + w_i + w_j} + (n-m) \cdot \frac{1}{(m-1)w_n + w_j}} \\
&[\text{Noting that } (m-2)w_n + w_i + w_j = (m-1)w_n + w_j + (w_i - w_n) \geq (m-1)w_n + w_j] \\
&\geq \frac{(n-m) \cdot \frac{1}{(m-1)w_1 + w_i} \frac{w_j}{w_j + (m-1)w_1}}{(m-1) \cdot \frac{1}{(m-1)w_n + w_j} + (n-m) \cdot \frac{1}{(m-1)w_n + w_j}} \\
&= \frac{n-m}{n-1} \cdot \frac{1}{(m-1)w_1 + w_i} \cdot \frac{(m-1)w_n + w_j}{(m-1)w_1 + w_j} \\
&\geq \frac{n-m}{n-1} \cdot \frac{1/w_1}{m - 1 + w_i/w_1} \cdot \frac{w_n}{w_1}
\end{aligned}$$

For an upper bound, we can show that:

$$
\begin{aligned}
&\frac{\sum_{S\in\mathcal{C}^{(m-1)}:i,j\notin S}\frac{w_j}{(w_i+\sum_{k\in S}w_k)(w_j+\sum_{k\in S}w_k)}}{\sum_{S\in\mathcal{C}_i^{(m)}\cap\mathcal{C}_j^{(m)}}\frac{1}{\sum_{k\in S}w_k}+\sum_{S\in\mathcal{C}^{(m-1)}:i,j\notin S}\frac{1}{w_j+\sum_{k\in S}w_k}}\\
&\leq\frac{\binom{n-2}{m-1}\frac{1}{(m-1)w_n+w_i}\frac{w_j}{w_j+(m-1)w_n}}{\binom{n-2}{m-2}\cdot\frac{1}{(m-2)w_1+w_i+w_j}+\binom{n-2}{m-1}\cdot\frac{1}{(m-1)w_1+w_j}}\\
&=\frac{(n-m)\cdot\frac{1}{(m-1)w_n+w_i}\frac{w_j}{w_j+(m-1)w_n}}{(m-1)\cdot\frac{1}{(m-2)w_1+w_i+w_j}+(n-m)\cdot\frac{1}{(m-1)w_1+w_j}}\\
&[\text{Noting that } (m-2)w_1+w_i+w_j=(m-1)w_1+w_j+(w_i-w_1)\leq(m-1)w_1+w_j]\\
&\leq\frac{(n-m)\cdot\frac{1}{(m-1)w_n+w_i}\frac{w_j}{w_j+(m-1)w_n}}{(m-1)\cdot\frac{1}{(m-1)w_1+w_j}+(n-m)\cdot\frac{1}{(m-1)w_1+w_j}}\\
&=\frac{n-m}{n-1}\cdot\frac{1/w_n}{m-1+w_i/w_n}\cdot\frac{w_1}{w_n}
\end{aligned}
$$

In both cases, it is straight forward to check that both of the upper and lower bounds decrease with increasing m and at a decreasing rate. Furthermore, they have the same asympotic dependency on m. Putting both bounds together, we have:

$$
\frac{\Delta_{ij}^{(m)}}{\tau_j^{(m)}}=\theta\left((1-\frac{n-m}{n-1}\cdot\frac{1}{m})\cdot\frac{w_i-w_j}{w_j}\right)=\theta\left(\frac{n}{n-1}\cdot(1-\frac{1}{m})\cdot\frac{w_i-w_j}{w_j}\right)
$$

Simplifying the above expression by removing the dependency on n gives:

$$
\frac{\Delta_{ij}^{(m)}}{\tau_j^{(m)}}=\theta\left((1-\frac{1}{m})\cdot\frac{w_i-w_j}{w_j}\right)
$$

This completes the proof. ■

10.6 CONNECTION AMONG MNL-MLE, GENERALIZED BORDA COUNT AND SPECTRAL RANKING

In this section, we will prove the theorems about the connections among the three algorithms. Here, we provide more general results pertaining to ranking output (as opposed to just top K estimate).

Theorem 7.1. *Consider the sampling model described in subsection (4.1), for any $p>0$, in the limit as $R\to\infty$, MNL-MLE and choice-based Borda count will produce the same top K estimate. Moreover, this holds even if the data does not come from the MNL model or any IID-RUM.*

Proof. **Notation:** As a shorthand notation, we will use $p(i|S)$ denote the probability of item i being chosen from menu S. Note that this probability does not necessarily follow any parametric RUM. Let ord denote the ordering function whose input is a set of real numbers and whose output is the full ordering of the indices of those numbers in increasing order. Under our sampling model, in the limit as $R\to\infty$, all the possible menus in $\mathcal{C}^{(m)}$ are observed given that $p>0$. Furthermore, given infinite data, the observed probability becomes exact. Consider the log likelihood function:

$$
\mathcal{L}(\mathbf{U})=\sum_{i=1}^{n}\left(\sum_{S\in\mathcal{C}_i^{(m)}}p(i|S)\cdot\log\frac{e^{U_i}}{\sum_{j\in S}e^{U_j}}\right)
$$

The derivative of the log likelihood with respect to individual partworth parameter is given as:

$$
\begin{aligned}
\nabla_{U_i}\mathcal{L}(\mathbf{U})&=\sum_{S\in\mathcal{C}_i^{(m)}}\left(-\sum_{j\in S,j\neq i}p(j|S)\frac{e^{U_i}}{\sum_{k\in S}e^{U_k}}-p(i|S)\frac{e^{U_i}}{\sum_{k\in S}e^{U_k}}+p(i|S)\right)\\
&=\sum_{S\in\mathcal{C}_i^{(m)}}\left(p(i|S)-\frac{e^{U_i}}{\sum_{k\in S}e^{U_k}}\right)
\end{aligned}
$$

As the log likelihood function is concave in U, the MLE estimate $\hat{\mathbf{U}}$ must satisfy:

$$\sum_{S \in \mathcal{C}_i^{(m)}} \frac{e^{\hat{U}_i}}{\sum_{j \in S} e^{\hat{U}_j}} = \sum_{S \in \mathcal{C}_i^{(m)}} p(i|S) \quad \forall i$$

On the other hand it can easily be shown that for any two items $i \neq j$,

$$\hat{U}_i > \hat{U}_j \Rightarrow \sum_{S \in \mathcal{C}_i^{(m)}} \frac{e^{\hat{U}_i}}{\sum_{k \in S} e^{\hat{U}_k}} > \sum_{S \in \mathcal{C}_j^{(m)}} \frac{e^{\hat{U}_j}}{\sum_{k \in S} e^{\hat{U}_k}}$$

Hence,

$$\text{ord}(\{\hat{U}_i\}_{i=1}^n) = \text{ord}(\{\sum_{S \in \mathcal{C}_i^{(m)}} \frac{e^{\hat{U}_i}}{\sum_{k \in S} e^{\hat{U}_k}}\}_{i=1}^n) = \text{ord}(\{\sum_{S \in \mathcal{C}_i^{(m)}} p(i|S)\}_{i=1}^n)$$

Observing that the ordering given by the Borda Count algorithm is consistent with the ordering induced by $\{\sum_{S \in \mathcal{C}_i^{(m)}} p(i|S)\}_{i=1}^n$ completes the proof. ■

Theorem 7.2. *Consider the sampling model described in subsection (4.1). Assume that the underlying choice model generating the data is in the class of IID-RUMs whose noise distribution has absolutely continuous density function with support on the real line. For any $p > 0$, in the limit as $R \to \infty$, then spectral ranking, MNL-MLE and choice-based Borda count produce the same top K estimate.*

On the other hand, there exists a choice model where in the limit as $R \to \infty$, the spectral ranking algorithm produces a different top K estimate from MNL-MLE/Borda count.

Proof. Following the argument as in the proof of theorem (7.1), given $p > 0$ as $R \to \infty$, all possible menus of size m are observed and all choice probabilities are exact. For the first part of the theorem it suffices to prove that under IID-RUMs, the Spectral Ranking algorithm is consistent in recovering the true ordering among the items, as this will be the same ordering as given by Generalized Borda Count/MLE.

The Spectral Ranking algorithm ranks the items by the stationary distribution of a Markov Chain constructed using choice data. For analysis purpose, we follow the construction due to Negahban et al. [2017], Maystre and Grossglauser [2015]. Fix a menu size $m \geq 2$ and consider the following Markov Chain where for any two items i, j:

$$\mathbf{M}_{ij} = \begin{cases} \frac{1}{\binom{n-1}{m-1}} \sum_{S \in \mathcal{C}^{(m)}: i,j \in S} \rho(j|S) & \text{if } j \neq i \\ 1 - \sum_{k \neq i} \mathbf{M}_{ik} & \text{if } j = i \end{cases}$$

Consider any two items i, j such that $U_i > U_j$. Under IID-RUMs assumptions in the theorem statement, we have:

$$\rho(i|S) > \rho(j|S) \quad \forall S \in \mathcal{C}^{(m)} : i, j \in S$$

and

$$\rho(i|S' \cup \{i\}) > \rho(j|S' \cup \{j\}) \quad \forall S' \in \mathcal{C}^{(m-1)} : i, j \notin S$$

It is easy to show that:

$$\mathbf{M}_{ki} > \mathbf{M}_{kj} \quad \forall k \neq i, j$$

and

$$\mathbf{M}_{ii} > \mathbf{M}_{jj}$$

Let π denote the stationary distribution of the Markov Chain constructed above. By definition, the stationary distribution of the Markov Chain satisfies:

$$\pi_i = \sum_{k \in [n]} \pi_k \cdot \mathbf{M}_{ki} > \sum_{k \in [n]} \pi_k \cdot \mathbf{M}_{kj} = \pi_j$$

For the second half of the theorem, let us consider the pairwise comparison setting ($m = 2$). We will construct a pairwise choice model such that MNL-MLE/Generalized Borda Count give a different ordering among the items from Spectral

Ranking algorithm in the limit as $R \to \infty$. Consider an universe of 4 items with the following pairwise choice probability. Note that $P_{ij} = P(j|\{i,j\})$:

$$P = \begin{bmatrix} 0.5 & 0.6 & 0.55 & 0.55 \\ 0.2 & 0.5 & 0.85 & 0.60 \\ 0.45 & 0.40 & 0.5 & 0.95 \\ 0.45 & 0.45 & 0.15 & 0.5 \end{bmatrix}$$

It is easy to check that $\tau_4 > \tau_3 > \tau_2 > \tau_1$. In the limit of infinite data, MLE/Borda Count will output the ordering $4, 3, 2, 1$ (best item first). However, the Spectral Ranking algorithm will give the ordering $4, 2, 3, 1$. This completes the proof.■

11 ADDITIONAL EXPERIMENTS

11.1 DATASET DESCRIPTIONS

Table (1) shows the characteristics of the datasets used in our experiments: number of items, number of rankings available and whether the data contains partial rankings.

Dataset	n	Num rankings	Partial ranking?
APA	5	6000	No
Sushi	11	5000	No
Irish-North	12	$44k$	Yes
Irish-West	9	$30k$	Yes
Irish-Meath	14	$64k$	Yes
F1	22	18	Yes

Table 1: Characteristics of the datasets used in our experiments.

Data preparation: Given menu size m, to estimate choice probability from rankings and partial rankings, we first enumerate all the possible menus of size m. For each menu S, and for each $i \in S$, we count the number of rankings where i is ranked ahead of all the other items in S, a 'win' for i. For a partial ranking where there are l items ranked as 'equal', we count $\frac{1}{l}$-'win' for each item. Choice probability is obtained by normalizing the number of wins by the number of rankings.

Experiments: For each independent trial, we first generate all the data as specified by the maximum sample size. We then feed increasing portion of this data to the algorithm and check if the algorithm correctly identifies *all* of the top K items. For all experiments, we keep $R = 100$ and adjust p to obtain the appropriate expected sample size.

11.2 ADDITIONAL EXPERIMENTAL RESULTS

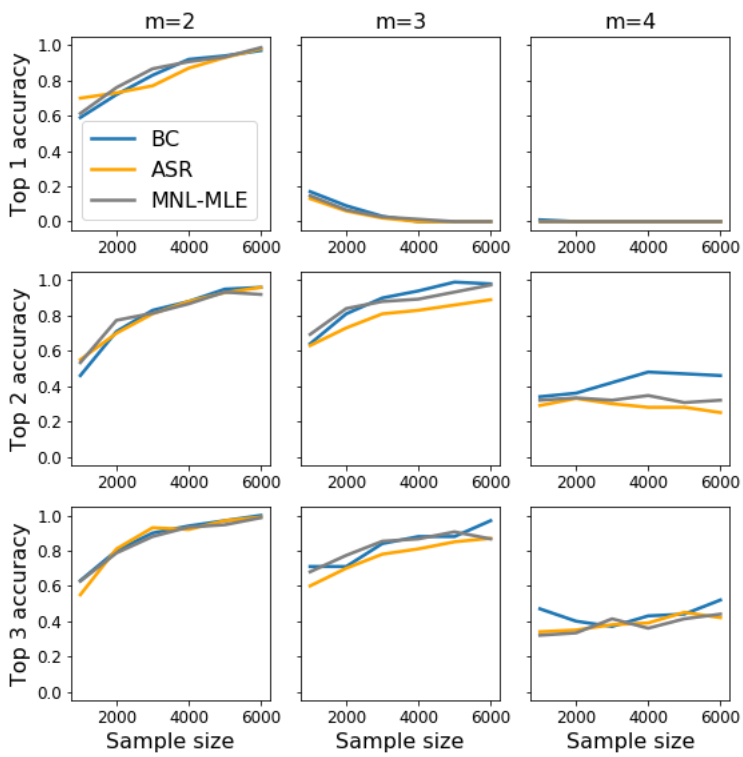

Figure 6: **APA election dataset:** Exact top K accuracy against sample size

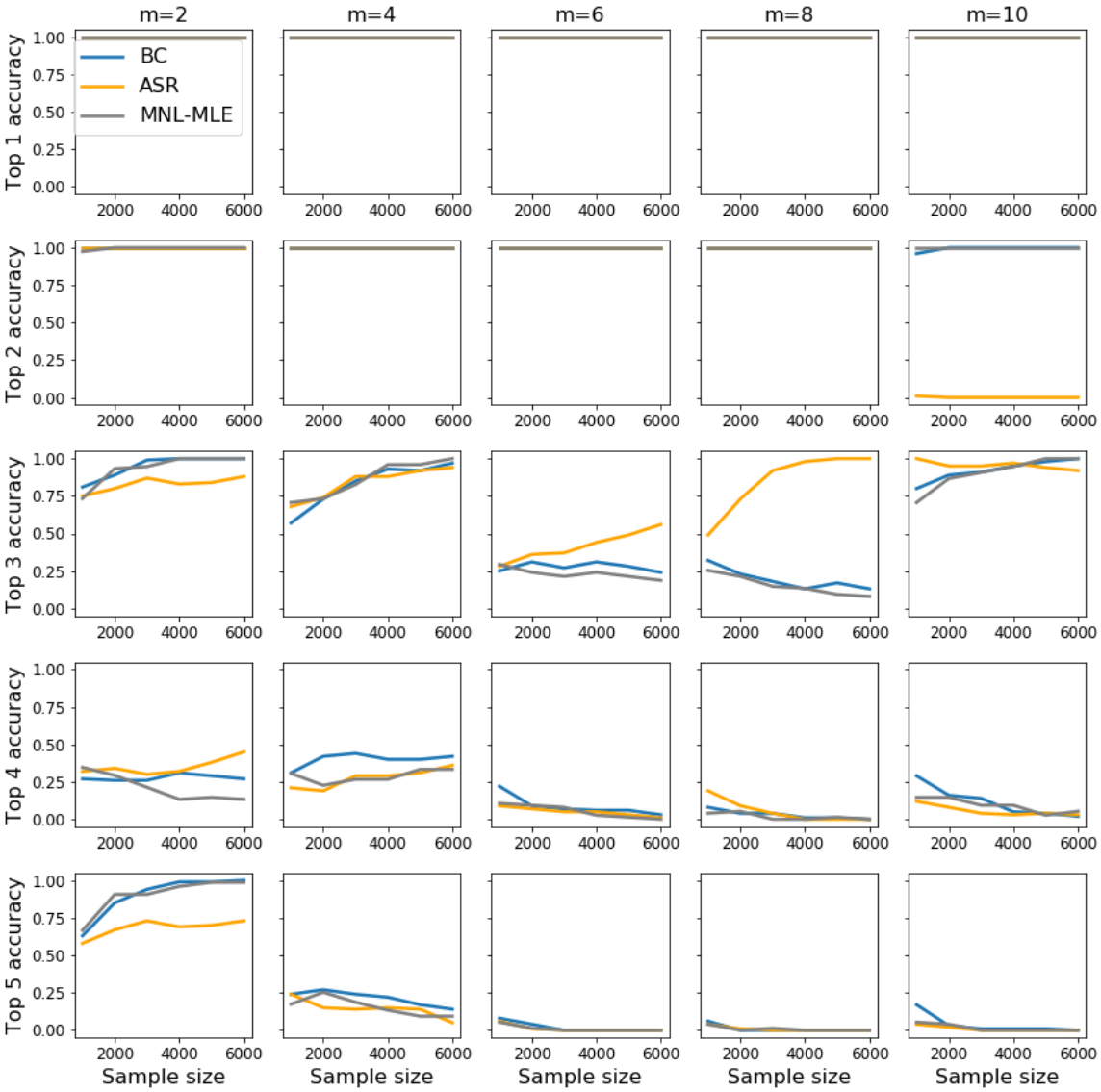

Figure 7: **SUSHI dataset:** Exact top K accuracy against sample size

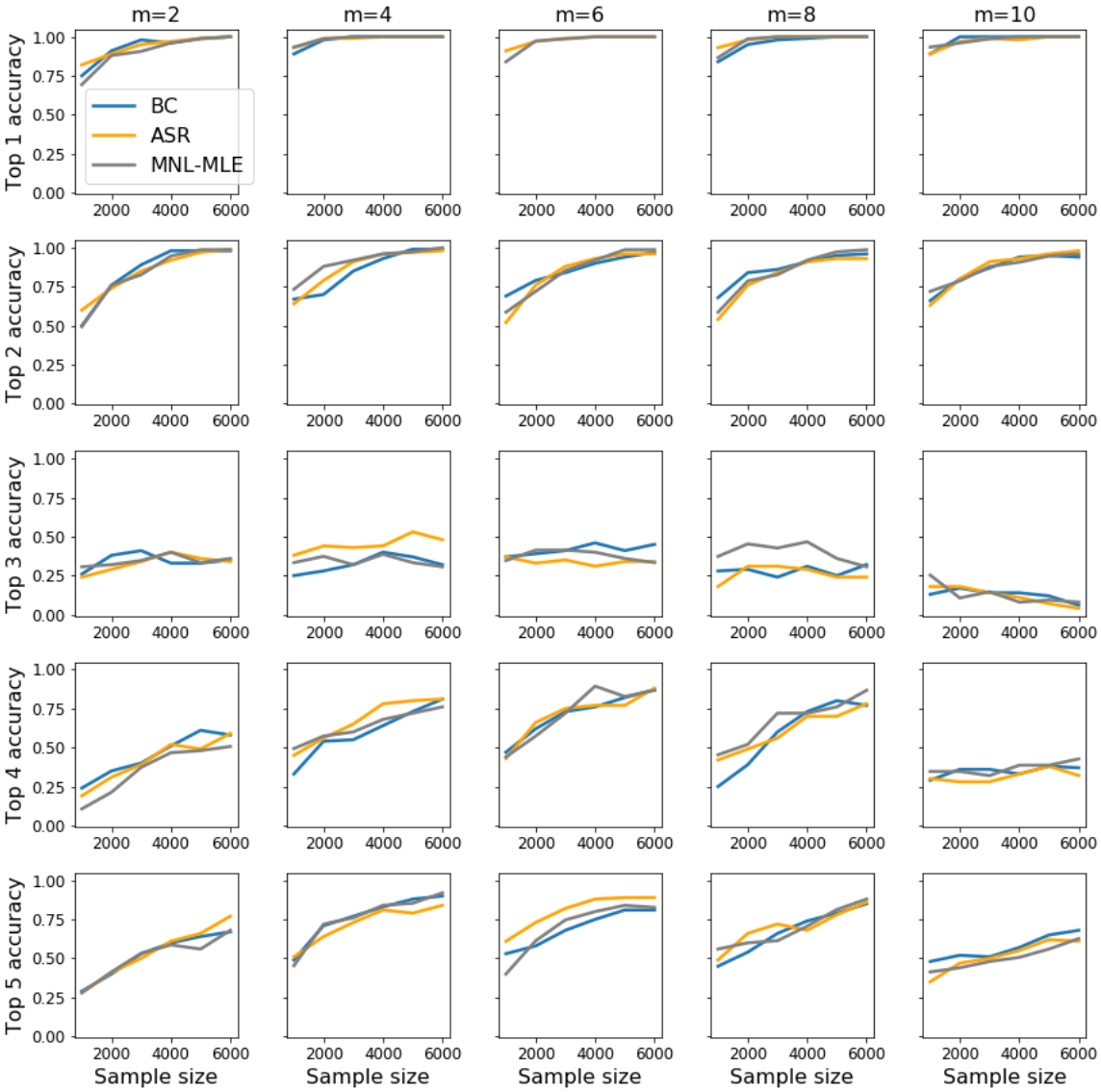

Figure 8: **Irish North dataset:** Exact top K accuracy against sample size

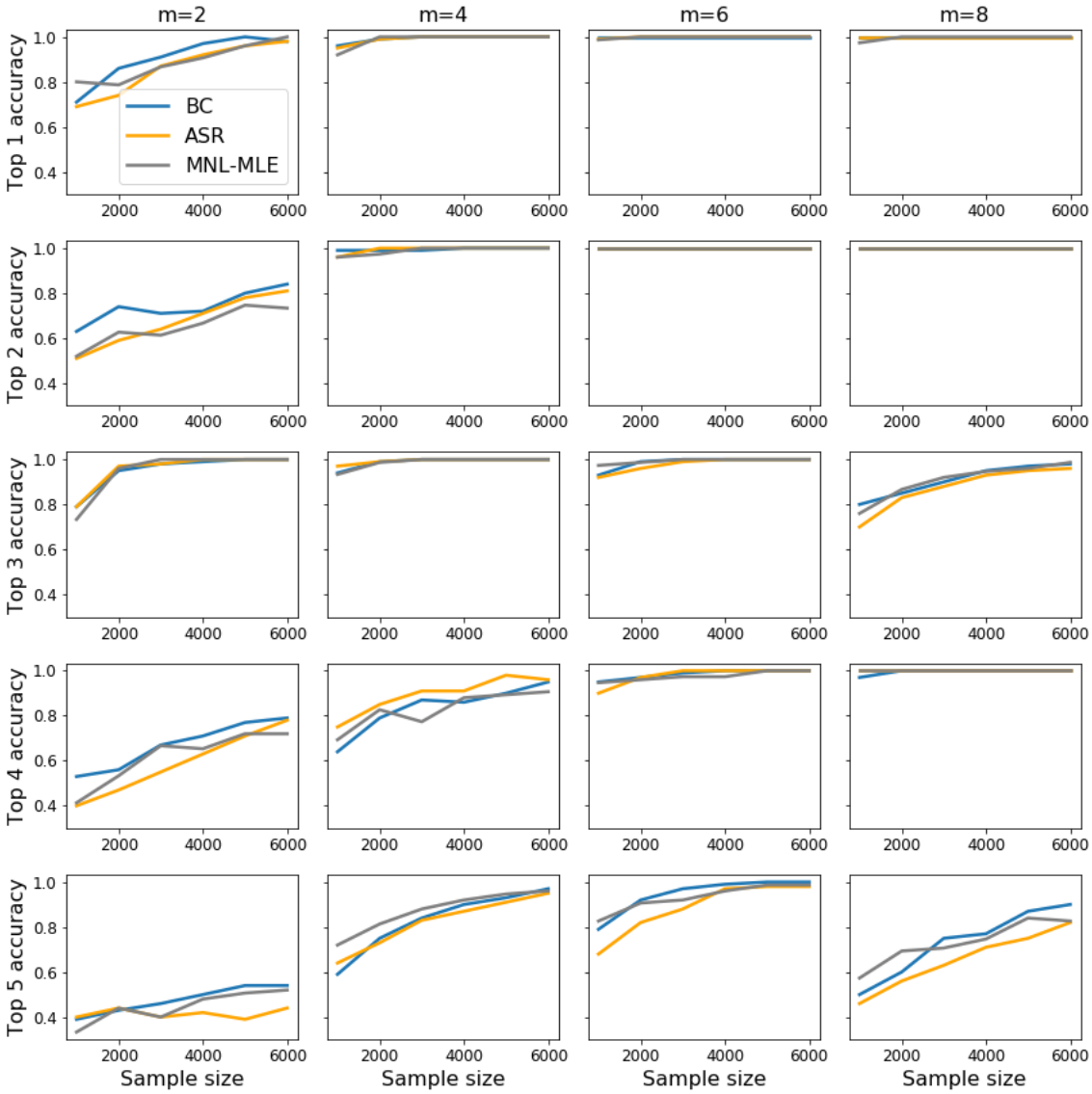

Figure 9: **Irish West dataset:** Exact top K accuracy against sample sizes

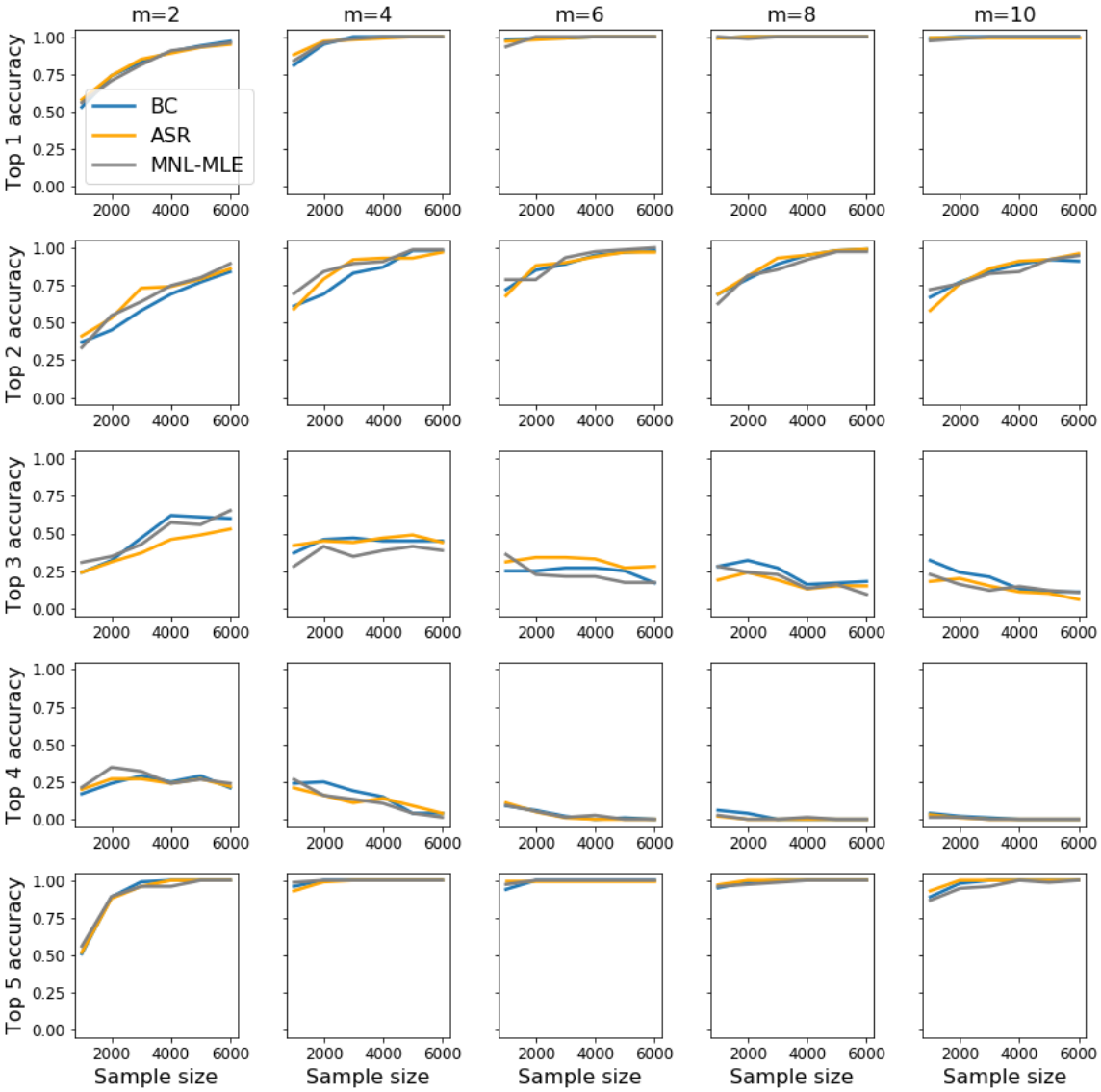

Figure 10: **Irish Meath dataset:** Exact top K accuracy against sample size

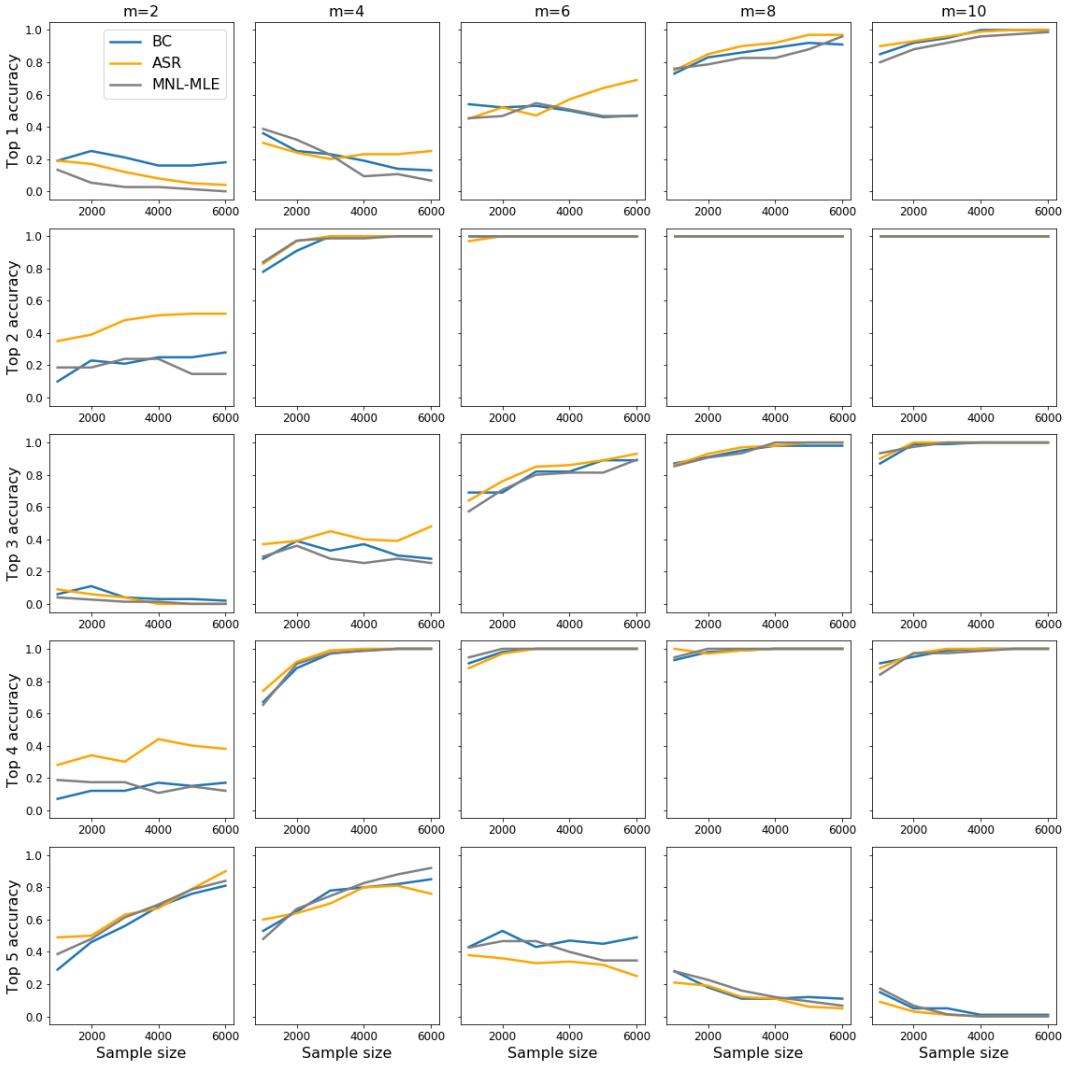

Figure 11: **F1 dataset:** Exact top K accuracy against sample size.