Journal of Machine Learning Research 17 (2016) 1-54 Submitted 4/16; Published 10/16

Data-driven Rank Breaking for Efficient Rank Aggregation

Ashish Khetan KHETAN2@ILLINOIS.EDU
Sewoong Oh SWOH@ILLINOIS.EDU
Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Editor: Benjamin Recht

Abstract

Rank aggregation systems collect ordinal preferences from individuals to produce a global ranking that represents the social preference. Rank-breaking is a common practice to reduce the computational complexity of learning the global ranking. The individual preferences are broken into pairwise comparisons and applied to efficient algorithms tailored for independent paired comparisons. However, due to the ignored dependencies in the data, naive rank-breaking approaches can result in inconsistent estimates. The key idea to produce accurate and consistent estimates is to treat the pairwise comparisons unequally, depending on the topology of the collected data. In this paper, we provide the optimal rank-breaking estimator, which not only achieves consistency but also achieves the best error bound. This allows us to characterize the fundamental tradeoff between accuracy and complexity. Further, the analysis identifies how the accuracy depends on the spectral gap of a corresponding comparison graph.

Keywords: Rank aggregation, Plackett-Luce model, Sample complexity

1. Introduction

In several applications such as electing officials, choosing policies, or making recommendations, we are given partial preferences from individuals over a set of alternatives, with the goal of producing a global ranking that represents the collective preference of the population or the society. This process is referred to as *rank aggregation*. One popular approach is *learning to rank*. Economists have modeled each individual as a rational being maximizing his/her perceived utility. Parametric probabilistic models, known collectively as Random Utility Models (RUMs), have been proposed to model such individual choices and preferences (McFadden, 1980). This allows one to infer the global ranking by learning the inherent utility from individuals' revealed preferences, which are noisy manifestations of the underlying true utility of the alternatives.

Traditionally, learning to rank has been studied under the following data collection scenarios: pairwise comparisons, best-out-of-k comparisons, and k-way comparisons. *Pairwise comparisons* are commonly studied in the classical context of sports matches as well as more recent applications in crowdsourcing, where each worker is presented with a pair of choices and asked to choose the more favorable one. *Best-out-of-k comparisons* data sets are commonly available from purchase history of customers. Typically, a set of k alternatives are offered among which one is chosen or purchased by each customer. This has been widely studied in operations research in the context of modeling customer choices for revenue management and assortment optimization. The *k-way comparisons* are assumed in traditional rank aggregation scenarios, where each person reveals his/her preference as

a ranked list over a set of k items. In some real-world elections, voters provide ranked preferences over the whole set of candidates (Lundell, 2007). We refer to these three types of ordinal data collection scenarios as 'traditional' throughout this paper.

For such traditional data sets, there are several computationally efficient inference algorithms for finding the Maximum Likelihood (ML) estimates that provably achieve the minimax optimal performance (Negahban et al., 2012; Shah et al., 2015a; Hajek et al., 2014). However, modern data sets can be unstructured. Individual's revealed ordinal preferences can be implicit, such as movie ratings, time spent on the news articles, and whether the user finished watching the movie or not. In crowdsourcing, it has also been observed that humans are more efficient at performing batch comparisons (Gomes et al., 2011), as opposed to providing the full ranking or choosing the top item. This calls for more flexible approaches for rank aggregation that can take such diverse forms of ordinal data into account. For such non-traditional data sets, finding the ML estimate can become significantly more challenging, requiring run-time exponential in the problem parameters.

To avoid such a computational bottleneck, a common heuristic is to resort to *rank-breaking*. The collected ordinal data is first transformed into a bag of pairwise comparisons, ignoring the dependencies that were present in the original data. This is then processed via existing inference algorithms tailored for *independent* pairwise comparisons, hoping that the dependency present in the input data does not lead to inconsistency in estimation. This idea is one of the main motivations for numerous approaches specializing in learning to rank from pairwise comparisons, e.g., (Ford Jr., 1957; Negahban et al., 2014; Azari Soufiani et al., 2013). However, such a heuristic of full rank-breaking defined explicitly in (1), where all pairwise comparisons are weighted and treated equally ignoring their dependencies, has been recently shown to introduce inconsistency (Azari Soufiani et al., 2014).

The key idea to produce accurate and consistent estimates is to treat the pairwise comparisons unequally, depending on the topology of the collected data. A fundamental question of interest to practitioners is how to choose the weight of each pairwise comparison in order to achieve not only consistency but also the best accuracy, among those consistent estimators using rank-breaking. We study how the accuracy of resulting estimate depends on the topology of the data and the weights on the pairwise comparisons. This provides a guideline for the optimal choice of the weights, driven by the topology of the data, that leads to accurate estimates.

Problem formulation. The premise in the current race to collect more data on user activities is that, a hidden true preference manifests in the user's activities and choices. Such data can be explicit, as in ratings, ranked lists, pairwise comparisons, and like/dislike buttons. Others are more implicit, such as purchase history and viewing times. While more data in general allows for a more accurate inference, the heterogeneity of user activities makes it difficult to infer the underlying preferences directly. Further, each user reveals her preference on only a few contents.

Traditional collaborative filtering fails to capture the diversity of modern data sets. The sparsity and heterogeneity of the data renders typical similarity measures ineffective in the nearest-neighbor methods. Consequently, simple measures of similarity prevail in practice, as in Amazon's "people who bought ... also bought ..." scheme. Score-based methods require translating heterogeneous data into numeric scores, which is a priori a difficult task. Even if explicit ratings are observed, those are often unreliable and the scale of such ratings vary from user to user.

We propose aggregating ordinal data based on users' revealed preferences that are expressed in the form of *partial orderings* (notice that our use of the term is slightly different from its original

use in revealed preference theory). We interpret user activities as manifestation of the hidden preferences according to discrete choice models (in particular the Plackett-Luce model defined in (1)). This provides a more reliable, scale-free, and widely applicable representation of the heterogeneous data as partial orderings, as well as a probabilistic interpretation of how preferences manifest. In full generality, the data collected from each individual can be represented by a *partially ordered set (poset)*. Assuming consistency in a user's revealed preferences, any ordered relations can be seamlessly translated into a poset, represented as a Hasse diagram by a directed acyclic graph (DAG). The DAG below represents ordered relations $a > \{b, d\}$, $b > c$, $\{c, d\} > e$, and $e > f$. For example, this could have been translated from two sources: a five star rating on a and a three star ratings on b, c, d, a two star rating on e, and a one star rating on f; and the item b being purchased after reviewing c as well.

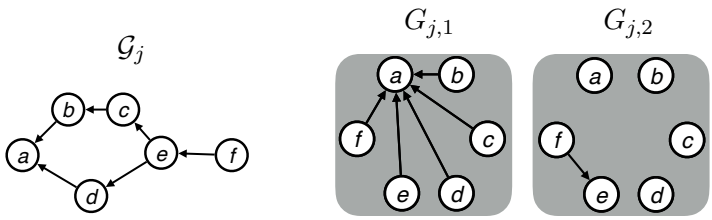

Figure 1: A DAG representation of consistent partial ordering of a user j, also called a Hasse diagram (left). A set of rank-breaking graphs extracted from the Hasse diagram for the separator item a and e, respectively (right).

There are n users or agents, and each agent j provides his/her ordinal evaluation on a subset S_j of d items or alternatives. We refer to $S_j \subset \{1, 2, \ldots, d\}$ as *offerings* provided to j, and use $\kappa_j = |S_j|$ to denote the size of the offerings. We assume that the partial ordering over the offerings is a manifestation of her preferences as per a popular choice model known as Plackett-Luce (PL) model. As we explain in detail below, the PL model produces total orderings (rather than partial ones). The data collector queries each user for a partial ranking in the form of a poset over S_j. For example, the data collector can ask for the top item, unordered subset of three next preferred items, the fifth item, and the least preferred item. In this case, an example of such poset could be $a < \{b, c, d\} < e < f$, which could have been generated from a total ordering produced by the PL model and taking the corresponding partial ordering from the total ordering. Notice that we fix the topology of the DAG first and ask the user to fill in the node identities corresponding to her total ordering as (randomly) generated by the PL model. Hence, the structure of the poset is considered deterministic, and only the identity of the nodes in the poset is considered random. Alternatively, one could consider a different scenario where the topology of the poset is also random and depends on the outcome of the preference, which is out-side the scope of this paper and provides an interesting future research direction.

The PL model is a special case of *random utility models*, defined as follows (Walker and Ben-Akiva, 2002; Azari Soufiani et al., 2012). Each item i has a real-valued latent utility θ_i. When presented with a set of items, a user's reveled preference is a partial ordering according to noisy manifestation of the utilities, i.e. i.i.d. noise added to the true utility θ_i's. The PL model is a special case where the noise follows the standard Gumbel distribution, and is one of the most popular model in social choice theory (McFadden, 1973; McFadden and Train, 2000). PL has several important

properties, making this model realistic in various domains, including marketing (Guadagni and Little, 1983), transportation (McFadden, 1980; Ben-Akiva and Lerman, 1985), biology (Sham and Curtis, 1995), and natural language processing (Mikolov et al., 2013). Precisely, each user j, when presented with a set S_j of items, draws a noisy utility of each item i according to

$$u_i \quad = \quad \theta_i + Z_i \ ,$$

where Z_i's follow the independent standard Gumbel distribution. Then we observe the ranking resulting from sorting the items as per noisy observed utilities u_j's. Alternatively, the PL model is also equivalent to the following random process. For a set of alternatives S_j, a ranking $\sigma_j : [|S|] \to S$ is generated in two steps: (1) independently assign each item $i \in S_j$ an unobserved value X_i, exponentially distributed with mean $e^{-\theta_i}$; (2) select a ranking σ_j so that $X_{\sigma_j(1)} \leq X_{\sigma_j(2)} \leq \cdots \leq X_{\sigma_j(|S_j|)}$.

The PL model (i) satisfies Luce's 'independence of irrelevant alternatives' in social choice theory (Ray, 1973), and has a simple characterization as sequential (random) choices as explained below; and (ii) has a maximum likelihood estimator (MLE) which is a convex program in θ in the traditional scenarios of pairwise, best-out-of-k and k-way comparisons. Let $\mathbb{P}(a > \{b, c, d\})$ denote the probability a was chosen as the best alternative among the set $\{a, b, c, d\}$. Then, the probability that a user reveals a linear order $(a > b > c > d)$ is equivalent as making sequential choice from the top to bottom:

$$\begin{aligned} \mathbb{P}(a > b > c > d) \quad &= \quad \mathbb{P}(a > \{b, c, d\}) \ \mathbb{P}(b > \{c, d\}) \ \mathbb{P}(c > d) \\ &= \quad \frac{e^{\theta_a}}{(e^{\theta_a} + e^{\theta_b} + e^{\theta_c} + e^{\theta_d})} \ \frac{e^{\theta_b}}{(e^{\theta_b} + e^{\theta_c} + e^{\theta_d})} \ \frac{e^{\theta_c}}{(e^{\theta_c} + e^{\theta_d})} \ . \end{aligned}$$

We use the notation $(a > b)$ to denote the event that a is preferred over b. In general, for user j presented with offerings S_j, the probability that the revealed preference is a total ordering σ_j is $\mathbb{P}(\sigma_j) = \prod_{i \in \{1, \ldots, \kappa_j - 1\}} (e^{\theta_{\sigma^{-1}(i)}}) / (\sum_{i'=i}^{\kappa_j} e^{\theta_{\sigma^{-1}(i')}})$. We consider the true utility $\theta^* \in \Omega_b$, where we define Ω_b as

$$\Omega_b \quad \equiv \quad \left\{ \theta \in \mathbb{R}^d \,\middle|\, \sum_{i \in [d]} \theta_i = 0 \, , \, |\theta_i| \leq b \text{ for all } i \in [d] \right\} .$$

Note that by definition, the PL model is invariant under shifting the utility θ_i's. Hence, the centering ensures uniqueness of the parameters for each PL model. The bound b on the dynamic range is not a restriction, but is written explicitly to capture the dependence of the accuracy in our main results.

We have n users each providing a partial ordering of a set of offerings S_j according to the PL model. Let $\mathcal{G}_j$ denote both the DAG representing the partial ordering from user j's preferences. With a slight abuse of notations, we also let $\mathcal{G}_j$ denote the set of rankings that are consistent with this DAG. For general partial orderings, the probability of observing $\mathcal{G}_j$ is the sum of all total orderings that is consistent with the observation, i.e. $\mathbb{P}(\mathcal{G}_j) = \sum_{\sigma \in \mathcal{G}_j} \mathbb{P}(\sigma)$. The goal is to efficiently learn the true utility $\theta^* \in \Omega_b$, from the n sampled partial orderings. One popular approach is to compute the maximum likelihood estimate (MLE) by solving the following optimization:

$$\underset{\theta \in \Omega_b}{\text{maximize}} \qquad \sum_{j=1}^{n} \log \mathbb{P}(\mathcal{G}_j) \ .$$

This optimization is a simple convex optimization, in particular a logit regression, when the structure of the data $\{\mathcal{G}_j\}_{j\in[n]}$ is traditional. This is one of the reasons the PL model is attractive. However, for general posets, this can be computationally challenging. Consider an example of position-p ranking, where each user provides which item is at p-th position in his/her ranking. Each term in the log-likelihood for this data involves summation over $O((p-1)!)$ rankings, which takes $O(n\,(p-1)!)$ operations to evaluate the objective function. Since p can be as large as d, such a computational blow-up renders MLE approach impractical. A common remedy is to resort to rank-breaking, which might result in inconsistent estimates.

Rank-breaking. Rank-breaking refers to the idea of extracting a set of pairwise comparisons from the observed partial orderings and applying estimators tailored for paired comparisons treating each piece of comparisons as independent. Both the choice of which paired comparisons to extract and the choice of parameters in the estimator, which we call *weights*, turns out to be crucial as we will show. Inappropriate selection of the paired comparisons can lead to inconsistent estimators as proved in Azari Soufiani et al. (2014), and the standard choice of the parameters can lead to a significantly suboptimal performance.

A naive rank-breaking that is widely used in practice is to apply rank-breaking to all possible pairwise relations that one can read from the partial ordering and weighing them equally. We refer to this practice as *full rank-breaking*. In the example in Figure 1, full rank-breaking first extracts the bag of comparisons $\mathcal{C} = \{(a>b),(a>c),(a>d),(a>e),(a>f),\ldots,(e>f)\}$ with 13 paired comparison outcomes, and apply the maximum likelihood estimator treating each paired outcome as independent. Precisely, the *full rank-breaking estimator* solves the convex optimization of

$$\widehat{\theta} \;\in\; \arg\max_{\theta\in\Omega_b} \sum_{(i>i')\in\mathcal{C}} \left(\theta_i - \log\left(e^{\theta_i}+e^{\theta_{i'}}\right)\right). \tag{1}$$

There are several efficient implementation tailored for this problem (Ford Jr., 1957; Hunter, 2004; Negahban et al., 2012; Maystre and Grossglauser, 2015a), and under the traditional scenarios, these approaches provably achieve the minimax optimal rate (Hajek et al., 2014; Shah et al., 2015a). For general non-traditional data sets, there is a significant gain in computational complexity. In the case of position-p ranking, where each of the n users report his/her p-th ranking item among κ items, the computational complexity reduces from $O(n\,(p-1)!)$ for the MLE in (1) to $O(n\,p\,(\kappa-p))$ for the full rank-breaking estimator in (1). However, this gain comes at the cost of accuracy. It is known that the full-rank breaking estimator is inconsistent (Azari Soufiani et al., 2014); the error is strictly bounded away from zero even with infinite samples.

Perhaps surprisingly, Azari Soufiani et al. (2014) recently characterized the entire set of consistent rank-breaking estimators. Instead of using the bag of paired comparisons, the sufficient information for consistent rank-breaking is a set of rank-breaking graphs defined as follows.

Recall that a user j provides his/her preference as a poset represented by a DAG $\mathcal{G}_j$. Consistent rank-breaking first identifies all *separators* in the DAG. A node in the DAG is a separator if one can partition the rest of the nodes into two parts. A partition A_{top} which is the set of items that are preferred over the separator item, and a partition A_{bottom} which is the set of items that are less preferred than the separator item. One caveat is that we allow A_{top} to be empty, but A_{bottom} must have at least one item. In the example in Figure 1, there are two separators: the item a and the item e. Using these separators, one can extract the following partial ordering from the original poset: $(a > \{b,c,d\} > e > f)$. The items a and e separate the set of offerings into partitions, hence

the name separator. We use ℓ_j to denote the number of separators in the poset $\mathcal{G}_j$ from user j. We let $p_{j,a}$ denote the ranked position of the a-th separator in the poset $\mathcal{G}_j$, and we sort the positions such that $p_{j,1} < p_{j,2} < \ldots < p_{j,\ell_j}$. The set of separators is denoted by $\mathcal{P}_j = \{p_{j,1}, p_{j,2}, \cdots, p_{j,\ell_j}\}$. For example, since the separator a is ranked at position 1 and e is at the 5-th position, $\ell_j = 2$, $p_{j,1} = 1$, and $p_{j,2} = 5$. Note that f is not a separator (whereas a is) since corresponding A_{bottom} is empty.

Conveniently, we represent this extracted partial ordering using a set of DAGs, which are called *rank-breaking graphs*. We generate one rank-breaking graph per separator. A rank breaking graph $G_{j,a} = (S_j, E_{j,a})$ for user j and the a-th separator is defined as a directed graph over the set of offerings S_j, where we add an edge from a node that is less preferred than the a-th separator to the separator, i.e. $E_{j,a} = \{(i, i') \,|\, i' \text{ is the } a\text{-th separator, and } \sigma_j^{-1}(i) > p_{j,a}\}$. Note that by the definition of the separator, $E_{j,a}$ is a non-empty set. An example of rank-breaking graphs are shown in Figure 1.

This rank-breaking graphs were introduced in Azari Soufiani et al. (2013), where it was shown that the pairwise ordinal relations that is represented by edges in the rank-breaking graphs are sufficient information for using any estimation based on the idea of rank-breaking. Precisely, on the converse side, it was proved in Azari Soufiani et al. (2014) that any pairwise outcomes that is not present in the rank-breaking graphs $G_{j,a}$'s lead to inconsistency for a general θ^*. On the achievability side, it was proved that all pairwise outcomes that are present in the rank-breaking graphs give a consistent estimator, as long as all the paired comparisons in each $G_{j,a}$ are weighted equally.

It should be noted that rank-breaking graphs are defined slightly differently in Azari Soufiani et al. (2013). Specifically, Azari Soufiani et al. (2013) introduced a different notion of rank-breaking graph, where the vertices represent positions in total ordering. An edge between two vertices i_1 and i_2 denotes that the pairwise comparison between items ranked at position i_1 and i_2 is included in the estimator. Given such observation from the PL model, Azari Soufiani et al. (2013) and Azari Soufiani et al. (2014) prove that a rank-breaking graph is consistent if and only if it satisfies the following property. If a vertex i_1 is connected to any vertex i_2, where $i_2 > i_1$, then i_1 must be connected to all the vertices i_3 such that $i_3 > i_1$. Although the specific definitions of rank-breaking graphs are different from our setting, the mathematical analysis of Azari Soufiani et al. (2013) still holds when interpreted appropriately. Specifically, we consider only those rank-breaking that are consistent under the conditions given in Azari Soufiani et al. (2013). In our rank-breaking graph $G_{j,a}$, a separator node is connected to all the other item nodes that are ranked below it (numerically higher positions).

In the algorithm described in (33), we satisfy this sufficient condition for consistency by restricting to a class of convex optimizations that use the same weight $\lambda_{j,a}$ for all $(\kappa - p_{j,a})$ paired comparisons in the objective function, as opposed to allowing more general weights that defer from a pair to another pair in a rank-breaking graph $G_{j,a}$.

Algorithm. Consistent rank-breaking first identifies separators in the collected posets $\{\mathcal{G}_j\}_{j\in[n]}$ and transform them into rank-breaking graphs $\{G_{j,a}\}_{j\in[n],a\in[\ell_j]}$ as explained above. These rank-breaking graphs are input to the MLE for paired comparisons, assuming all directed edges in the rank-breaking graphs are independent outcome of pairwise comparisons. Precisely, the *consistent*

rank-breaking estimator solves the convex optimization of maximizing the paired log likelihoods

$$\mathcal{L}_{\mathrm{RB}}(\theta) \;=\; \sum_{j=1}^{n}\sum_{a=1}^{\ell_j} \lambda_{j,a} \Big\{ \sum_{(i,i')\in E_{j,a}} \Big(\theta_{i'} - \log\Big(e^{\theta_i} + e^{\theta_{i'}}\Big)\Big)\Big\}\,, \tag{2}$$

where $E_{j,a}$'s are defined as above via separators and different choices of the non-negative weights $\lambda_{j,a}$'s are possible and the performance depends on such choices. Each weight $\lambda_{j,a}$ determine how much we want to weigh the contribution of a corresponding rank-breaking graph $G_{j,a}$. We define the *consistent rank-breaking estimate* $\widehat{\theta}$ as the optimal solution of the convex program:

$$\widehat{\theta} \;\in\; \arg\max_{\theta\in\Omega_b}\; \mathcal{L}_{\mathrm{RB}}(\theta)\,. \tag{3}$$

By changing how we weigh each rank-breaking graph (by choosing the $\lambda_{j,a}$'s), the convex program (3) spans the entire set of consistent rank-breaking estimators, as characterized in Azari Soufiani et al. (2014). However, only asymptotic consistency was known, which holds independent of the choice of the weights $\lambda_{j,a}$'s. Naturally, a uniform choice of $\lambda_{j,a} = \lambda$ was proposed in (Azari Soufiani et al., 2014).

Note that this can be efficiently solved, since this is a simple convex optimization, in particular a logit regression, with only $O(\sum_{j=1}^{n} \ell_j\,\kappa_j)$ terms. For a special case of position-p breaking, the $O(n\,(p-1)!)$ complexity of evaluating the objective function for the MLE is now significantly reduced to $O(n\,(\kappa - p))$ by rank-breaking. Given this potential exponential gain in efficiency, a natural question of interest is "what is the price we pay in the accuracy?". We provide a sharp analysis of the performance of rank-breaking estimators in the finite sample regime, that quantifies the price of rank-breaking. Similarly, for a practitioner, a core problem of interest is how to choose the weights in the optimization in order to achieve the best accuracy. Our analysis provides a data-driven guideline for choosing the optimal weights.

Contributions. In this paper, we provide an upper bound on the error achieved by the rank-breaking estimator of (3) for any choice of the weights in Theorem 8. This explicitly shows how the error depends on the choice of the weights, and provides a guideline for choosing the optimal weights $\lambda_{j,a}$'s in a data-driven manner. We provide the explicit formula for the optimal choice of the weights and provide the the error bound in Theorem 2. The analysis shows the explicit dependence of the error in the problem dimension d and the number of users n that matches the numerical experiments.

If we are designing surveys and can choose which subset of items to offer to each user and also can decide which type of ordinal data we can collect, then we want to design such surveys in a way to maximize the accuracy for a given number of questions asked. Our analysis provides how the accuracy depends on the topology of the collected data, and provides a guidance when we do have some control over which questions to ask and which data to collect. One should maximize the spectral gap of corresponding comparison graph. Further, for some canonical scenarios, we quantify the price of rank-breaking by comparing the error bound of the proposed data-driven rank-breaking with the lower bound on the MLE, which can have a significantly larger computational cost (Theorem 4).

Notations. Following is a summary of all the notations defined above. We use d to denote the total number of items and index each item by $i \in \{1, 2, \ldots, d\}$. $\theta \in \Omega_b$ denotes vector of utilities

associated with each item. θ^* represents true utility and $\widehat{\theta}$ denotes the estimated utility. We use n to denote the number of users/agents and index each user by $j \in \{1, 2, \ldots, n\}$. $S_j \subseteq \{1, \ldots, d\}$ refer to the offerings provided to the j-th user and we use $\kappa_j = |S_j|$ to denote the size of the offerings. $\mathcal{G}_j$ denote the DAG (Hasse diagram) representing the partial ordering from user j's preferences. $\mathcal{P}_j = \{p_{j,1}, p_{j,2}, \cdots, p_{j,\ell_j}\}$ denotes the set of separators in the DAG $\mathcal{G}_j$, where $p_{j,1}, \cdots, p_{j,\ell_j}$ are the positions of the separators, and ℓ_j is the number of separators. $G_{j,a} = (S_j, E_{j,a})$ denote the rank-breaking graph for the a-th separator extracted from the partial ordering $\mathcal{G}_j$ of user j.

For any positive integer N, let $[N] = \{1, \cdots, N\}$. For a ranking σ over S, i.e., σ is a mapping from $[|S|]$ to S, let σ^{-1} denote the inverse mapping.For a vector x, let $\|x\|_2$ denote the standard l_2 norm. Let $\mathbf{1}$ denote the all-ones vector and $\mathbf{0}$ denote the all-zeros vector with the appropriate dimension. Let $\mathcal{S}^d$ denote the set of $d \times d$ symmetric matrices with real-valued entries. For $X \in \mathcal{S}^d$, let $\lambda_1(X) \leq \lambda_2(X) \leq \cdots \leq \lambda_d(X)$ denote its eigenvalues sorted in increasing order. Let $\mathrm{Tr}(X) = \sum_{i=1}^d \lambda_i(X)$ denote its trace and $\|X\| = \max\{|\lambda_1(X)|, |\lambda_d(X)|\}$ denote its spectral norm. For two matrices $X, Y \in \mathcal{S}^d$, we write $X \succeq Y$ if $X - Y$ is positive semi-definite, i.e., $\lambda_1(X - Y) \geq 0$. Let e_i denote a unit vector in $\mathbb{R}^d$ along the i-th direction.

2. Comparisons Graph and the Graph Laplacian

In the analysis of the convex program (3), we show that, with high probability, the objective function is strictly concave with $\lambda_2(H(\theta)) \leq -C_b\, \gamma\, \lambda_2(L) < 0$ (Lemma 11) for all $\theta \in \Omega_b$ and the gradient is bounded by $\|\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)\|_2 \leq C_b' \sqrt{\log d \sum_{j \in [n]} \ell_j}$ (Lemma 10). Shortly, we will define γ and $\lambda_2(L)$, which captures the dependence on the topology of the data, and C_b' and C_b are constants that only depend on b. Putting these together, we will show that there exists a $\theta \in \Omega_b$ such that

$$\|\widehat{\theta} - \theta^*\|_2 \;\leq\; \frac{2\|\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)\|_2}{-\lambda_2(H(\theta))} \;\leq\; C_b'' \frac{\sqrt{\log d \sum_{j \in [n]} \ell_j}}{\gamma\, \lambda_2(L)}\,.$$

Here $\lambda_2(H(\theta))$ denotes the second largest eigenvalue of a negative semi-definite Hessian matrix $H(\theta)$ of the objective function. The reason the second largest eigenvalue shows up is because the top eigenvector is always the all-ones vector which by the definition of Ω_b is infeasible. The accuracy depends on the topology of the collected data via the comparison graph of given data.

Definition 1. *(Comparison graph $\mathcal{H}$). We define a graph $\mathcal{H}([d], E)$ where each alternative corresponds to a node, and we put an edge (i, i') if there exists an agent j whose offerings is a set S_j such that $i, i' \in S_j$. Each edge $(i, i') \in E$ has a weight $A_{ii'}$ defined as*

$$A_{ii'} \;=\; \sum_{j \in [n]: i, i' \in S_j} \frac{\ell_j}{\kappa_j(\kappa_j - 1)}\,,$$

where $\kappa_j = |S_j|$ is the size of each sampled set and ℓ_j is the number of separators in S_j defined by rank-breaking in Section 1.

Define a diagonal matrix $D = \mathrm{diag}(A\mathbf{1})$, and the corresponding graph Laplacian $L = D - A$, such that

$$L \;=\; \sum_{j=1}^{n} \frac{\ell_j}{\kappa_j(\kappa_j - 1)} \sum_{i < i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top. \tag{4}$$

Let $0 = \lambda_1(L) \leq \lambda_2(L) \leq \cdots \leq \lambda_d(L)$ denote the (sorted) eigenvalues of L. Of special interest is $\lambda_2(L)$, also called the spectral gap, which measured how well-connected the graph is. Intuitively, one can expect better accuracy when the spectral gap is larger, as evidenced in previous learning to rank results in simpler settings (Negahban et al., 2014; Shah et al., 2015a; Hajek et al., 2014). This is made precise in (4), and in the main result of Theorem 2, we appropriately rescale the spectral gap and use $\alpha \in [0, 1]$ defined as

$$\alpha \equiv \frac{\lambda_2(L)(d-1)}{\mathrm{Tr}(L)} = \frac{\lambda_2(L)(d-1)}{\sum_{j=1}^{n} \ell_j} . \tag{5}$$

The accuracy also depends on the topology via the maximum weighted degree defined as $D_{\max} \equiv \max_{i\in[d]} D_{ii} = \max_{i\in[d]} \{\sum_{j:i\in S_j} \ell_j/\kappa_j\}$. Note that the average weighted degree is $\sum_i D_{ii}/d = \mathrm{Tr}(L)/d$, and we rescale it by $D_{\max}$ such that

$$\beta \equiv \frac{\mathrm{Tr}(L)}{dD_{\max}} = \frac{\sum_{j=1}^{n} \ell_j}{dD_{\max}} . \tag{6}$$

We will show that the performance of rank breaking estimator depends on the topology of the graph through these two parameters. The larger the spectral gap α the smaller error we get with the same effective sample size. The degree imbalance $\beta \in [0, 1]$ determines how many samples are required for the analysis to hold. We need smaller number of samples if the weighted degrees are balanced, which happens if β is large (close to one).

The following quantity also determines the convexity of the objective function.

$$\gamma \equiv \min_{j\in[n]} \left\{ \left(1 - \frac{p_{j,\ell_j}}{\kappa_j} \right)^{\lceil 2e^{2b} \rceil - 2} \right\} . \tag{7}$$

Note that γ is between zero and one, and a larger value is desired as the objective function becomes more concave and a better accuracy follows. When we are collecting data where the size of the offerings κ_j's are increasing with d but the position of the separators are close to the top, such that $\kappa_j = \omega(d)$ and $p_{j,\ell_j} = O(1)$, then for $b = O(1)$ the above quantity γ can be made arbitrarily close to one, for large enough problem size d. On the other hand, when p_{j,ℓ_j} is close to κ_j, the accuracy can degrade significantly as stronger alternatives might have small chance of showing up in the rank breaking. The value of γ is quite sensitive to b. The reason we have such a inferior dependence on b is because we wanted to give a universal bound on the Hessian that is simple. It is not difficult to get a tighter bound with a larger value of γ, but will inevitably depend on the structure of the data in a complicated fashion.

To ensure that the (second) largest eigenvalue of the Hessian is small enough, we need enough samples. This is captured by η defined as

$$\eta \equiv \max_{j\in[n]} \{\eta_j\} , \qquad \text{where} \qquad \eta_j = \frac{\kappa_j}{\max\{\ell_j, \kappa_j - p_{j,\ell_j}\}} . \tag{8}$$

Note that $1 < \eta_j \leq \kappa_j/\ell_j$. A smaller value of η is desired as we require smaller number of samples, as shown in Theorem 2. This happens, for instance, when all separators are at the top, such that $p_{j,\ell_j} = \ell_j$ and $\eta_j = \kappa_j/(\kappa_j - \ell_j)$, which is close to one for large κ_j. On the other hand, when all separators are at the bottom of the list, then η can be as large as κ_j.

We discuss the role of the topology of data captures by these parameters in Section 4.

3. Main Results

We present the main theoretical results accompanied by corresponding numerical simulations in this section.

3.1 Upper Bound on the Achievable Error

We present the main result that provides an upper bound on the resulting error and explicitly shows the dependence on the topology of the data. As explained in Section 1, we assume that each user provides a partial ranking according to his/her position of the separators. Precisely, we assume the set of offerings S_j, the number of separators ℓ_j, and their respective positions $\mathcal{P}_j = \{p_{j,1}, \ldots, p_{j,\ell_j}\}$ are predetermined. Each user draws the ranking of items from the PL model, and provides the partial ranking according to the separators of the form of $\{a > \{b, c, d\} > e > f\}$ in the example in the Figure 1.

Theorem 2. *Suppose there are n users, d items parametrized by $\theta^* \in \Omega_b$, each user j is presented with a set of offerings $S_j \subseteq [d]$, and provides a partial ordering under the PL model. When the effective sample size $\sum_{j=1}^n \ell_j$ is large enough such that*

$$\sum_{j=1}^{n} \ell_j \;\; \geq \;\; \frac{2^{11} e^{18b} \eta \log(\ell_{\max} + 2)^2}{\alpha^2 \gamma^2 \beta} d \log d \,, \tag{9}$$

where $b \equiv \max_i |\theta_i^*|$ is the dynamic range, $\ell_{\max} \equiv \max_{j \in [n]} \ell_j$, α is the (rescaled) spectral gap defined in (5)*, β is the (rescaled) maximum degree defined in* (6)*, γ and η are defined in Eqs.* (7) *and* (8)*, then the* rank-breaking estimator *in* (3) *with the choice of*

$$\lambda_{j,a} \;\; = \;\; \frac{1}{\kappa_j - p_{j,a}} \,, \tag{10}$$

for all $a \in [\ell_j]$ and $j \in [n]$ achieves

$$\frac{1}{\sqrt{d}} \big\| \widehat{\theta} - \theta^* \big\|_2 \;\; \leq \;\; \frac{4\sqrt{2} e^{4b} (1 + e^{2b})^2}{\alpha \gamma} \sqrt{\frac{d \log d}{\sum_{j=1}^n \ell_j}} \,, \tag{11}$$

with probability at least $1 - 3e^3 d^{-3}$.

Consider an ideal case where the spectral gap is large such that α is a strictly positive constant and the dynamic range b is finite and $\max_{j \in [n]} p_{j,\ell_j}/\kappa_j = C$ for some constant $C < 1$ such that γ is also a constant independent of the problem size d. Then the upper bound in (11) implies that we need the effective sample size to scale as $O(d \log d)$, which is only a logarithmic factor larger than the number of parameters to be estimated. Such a logarithmic gap is also unavoidable and due to the fact that we require high probability bounds, where we want the tail probability to decrease at least polynomially in d. We discuss the role of the topology of the data in Section 4.

The upper bound follows from an analysis of the convex program similar to those in (Negahban et al., 2012; Hajek et al., 2014; Shah et al., 2015a). However, unlike the traditional data collection scenarios, the main technical challenge is in analyzing the probability that a particular pair of items appear in the rank-breaking. We provide a proof in Section 8.1.

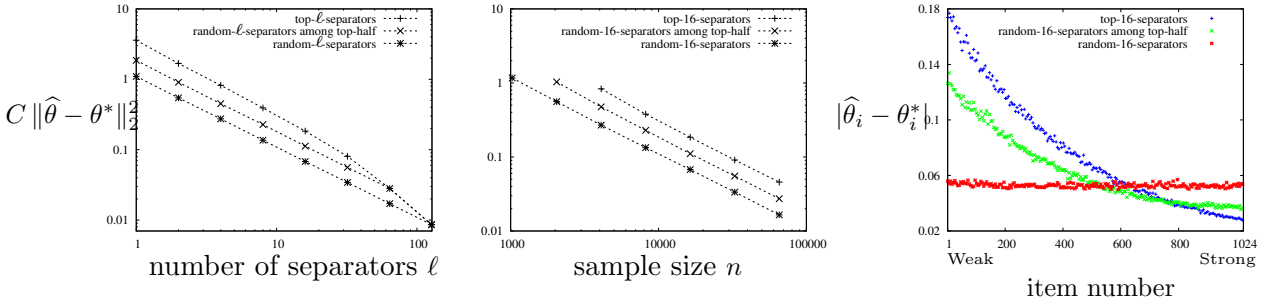

Figure 2: Simulation confirms $\|\theta^* - \widehat{\theta}\|_2^2 \propto 1/(\ell\, n)$, and smaller error is achieved for separators that are well spread out.

In Figure 2 , we verify the scaling of the resulting error via numerical simulations. We fix $d = 1024$ and $\kappa_j = \kappa = 128$, and vary the number of separators $\ell_j = \ell$ for fixed $n = 128000$ (left), and vary the number of samples n for fixed $\ell_j = \ell = 16$ (middle). Each point is average over 100 instances. The plot confirms that the mean squared error scales as $1/(\ell\, n)$. Each sample is a partial ranking from a set of κ alternatives chosen uniformly at random, where the partial ranking is from a PL model with weights θ^* chosen i.i.d. uniformly over $[-b, b]$ with $b = 2$. To investigate the role of the position of the separators, we compare three scenarios. The *top-ℓ-separators* choose the top ℓ positions for separators, the *random-ℓ-separators among top-half* choose ℓ positions uniformly random from the top half, and the *random-ℓ-separators* choose the positions uniformly at random. We observe that when the positions of the separators are well spread out among the κ offerings, which happens for *random-ℓ-separators*, we get better accuracy.

The figure on the right provides an insight into this trend for $\ell = 16$ and $n = 16000$. The absolute error $|\theta_i^* - \widehat{\theta}_i|$ is roughly same for each item $i \in [d]$ when breaking positions are chosen uniformly at random between 1 to $\kappa - 1$ whereas it is significantly higher for weak preference score items when breaking positions are restricted between 1 to $\kappa/2$ or are top-ℓ. This is due to the fact that the probability of each item being ranked at different positions is different, and in particular probability of the low preference score items being ranked in top-ℓ is very small. The third figure is averaged over 1000 instances. Normalization constant C is n/d^2 and $10^3\ell/d^2$ for the first and second figures respectively. For the first figure n is chosen relatively large such that $n\ell$ is large enough even for $\ell = 1$.

3.2 The Price of Rank Breaking for the Special Case of Position-p Ranking

Rank-breaking achieves computational efficiency at the cost of estimation accuracy. In this section, we quantify this tradeoff for a canonical example of position-p ranking, where each sample provides the following information: an unordered set of $p-1$ items that are ranked high, one item that is ranked at the p-th position, and the rest of $\kappa_j - p$ items that are ranked on the bottom. An example of a sample with position-4 ranking six items $\{a, b, c, d, e, f\}$ might be a partial ranking of $(\{a, b, d\} > \{e\} > \{c, f\})$. Since each sample has only one separator for $2 < p$, Theorem 2 simplifies to the following Corollary.

Corollary 3. *Under the hypotheses of Theorem 2, there exist positive constants C and c that only depend on b such that if $n \geq C(\eta d \log d)/(\alpha^2\gamma^2\beta)$ then*

$$\frac{1}{\sqrt{d}}\left\|\widehat{\theta} - \theta^*\right\|_2 \;\leq\; \frac{c}{\alpha\gamma}\sqrt{\frac{d\,\log d}{n}}\,. \tag{12}$$

Note that the error only depends on the position p through γ and η, and is not sensitive. To quantify the price of rank-breaking, we compare this result to a fundamental lower bound on the minimax rate in Theorem 4. We can compute a sharp lower bound on the minimax rate, using the Cramér-Rao bound, and a proof is provided in Section 8.3.

Theorem 4. *Let $\mathcal{U}$ denote the set of all unbiased estimators of θ^* and suppose $b > 0$, then*

$$\inf_{\widehat{\theta}\in\mathcal{U}}\sup_{\theta^*\in\Omega_b}\mathbb{E}[\|\widehat{\theta}-\theta^*\|^2] \;\geq\; \frac{1}{2p\log(\kappa_{\max})^2}\sum_{i=2}^{d}\frac{1}{\lambda_i(L)} \;\geq\; \frac{1}{2p\log(\kappa_{\max})^2}\frac{(d-1)^2}{n}\,,$$

where $\kappa_{\max} = \max_{j\in[n]}|S_j|$ and the second inequality follows from the Jensen's inequality.

Note that the second inequality is tight up to a constant factor, when the graph is an expander with a large spectral gap. For expanders, α in the bound (12) is also a strictly positive constant. This suggests that rank-breaking gains in computational efficiency by a super-exponential factor of $(p-1)!$, at the price of increased error by a factor of p, ignoring poly-logarithmic factors.

3.3 Tighter Analysis for the Special Case of Top-ℓ Separators Scenario

The main result in Theorem 2 is general in the sense that it applies to any partial ranking data that is represented by positions of the separators. However, the bound can be quite loose, especially when γ is small, i.e. p_{j,ℓ_j} is close to κ_j. For some special cases, we can tighten the analysis to get a sharper bound. One caveat is that we use a slightly sub-optimal choice of parameters $\lambda_{j,a} = 1/\kappa_j$ instead of $1/(\kappa_j - a)$, to simplify the analysis and still get the order optimal error bound we want. Concretely, we consider a special case of top-ℓ separators scenario, where each agent gives a ranked list of her most preferred ℓ_j alternatives among κ_j offered set of items. Precisely, the locations of the separators are $(p_{j,1}, p_{j,2}, \ldots, p_{j,\ell_j}) = (1, 2, \ldots, \ell_j)$.

Theorem 5. *Under the PL model, n partial orderings are sampled over d items parametrized by $\theta^* \in \Omega_b$, where the j-th sample is a ranked list of the top-ℓ_j items among the κ_j items offered to the agent. If*

$$\sum_{j=1}^{n}\ell_j \;\geq\; \frac{2^{12}e^{6b}}{\beta\alpha^2}d\log d\,, \tag{13}$$

where $b \equiv \max_{i,i'}|\theta^*_i - \theta^*_{i'}|$ and α, β are defined in (5) *and* (6)*, then the* rank-breaking estimator *in* (3) *with the choice of $\lambda_{j,a} = 1/\kappa_j$ for all $a \in [\ell_j]$ and $j \in [n]$ achieves*

$$\frac{1}{\sqrt{d}}\left\|\widehat{\theta} - \theta^*\right\|_2 \;\leq\; \frac{16(1+e^{2b})^2}{\alpha}\sqrt{\frac{d\,\log d}{\sum_{j=1}^{n}\ell_j}}\,, \tag{14}$$

with probability at least $1 - 3e^3 d^{-3}$.

A proof is provided in Section 8.4. In comparison to the general bound in Theorem 2, this is tighter since there is no dependence in γ or η. This gain is significant when, for example, p_{j,ℓ_j} is close to κ_j. As an extreme example, if all agents are offered the entire set of alternatives and are asked to rank all of them, such that $\kappa_j = d$ and $\ell_j = d-1$ for all $j \in [n]$, then the generic bound in (11) is loose by a factor of $(e^{4b}/2\sqrt{2})d^{\lceil 2e^{2b}\rceil - 2}$, compared to the above bound.

In the top-ℓ separators scenario, the data set consists of the ranking among top-ℓ_j items of the set S_j, i.e., $[\sigma_j(1), \sigma_j(2), \cdots, \sigma_j(\ell_j)]$. The corresponding log-likelihood of the PL model is

$$\mathcal{L}(\theta) = \sum_{j=1}^{n} \sum_{m=1}^{\ell_j} \left[\theta_{\sigma_j(m)} - \log\Big(\exp(\theta_{\sigma_j(m)}) + \exp(\theta_{\sigma_j(m+1)}) + \cdots + \exp(\theta_{\sigma_j(\kappa_j)}) \Big) \right], \tag{15}$$

where $\sigma_j(a)$ is the alternative ranked at the a-th position by agent j. The Maximum Likelihood Estimator (MLE) for this *traditional* data set is efficient. Hence, there is no computational gain in rank-breaking. Consequently, there is no loss in accuracy either, when we use the optimal weights proposed in the above theorem. Figure 3 illustrates that the MLE and the data-driven rank-breaking estimator achieve performance that is identical, and improve over naive rank-breaking that uses uniform weights. We also compare performance of Generalized Method-of-Moments (GMM) proposed by Azari Soufiani et al. (2013) with our algorithm. In addition, we show that performance of GMM can be improved by optimally weighing pairwise comparisons with $\lambda_{j,a}$. MSE of GMM in both the cases, uniform weights and optimal weights, is larger than our rank-breaking estimator. However, GMM is on average about four times faster than our algorithm. We choose $\lambda_{j,a} = 1/(\kappa_j - a)$ in the simulations, as opposed to the $1/\kappa_j$ assumed in the above theorem. This settles the question raised in Hajek et al. (2014) on whether it is possible to achieve optimal accuracy using rank-breaking under the top-ℓ separators scenario. Analytically, it was proved in (Hajek et al., 2014) that under the top-ℓ separators scenario, naive rank-breaking with uniform weights achieves the same error bound as the MLE, up to a constant factor. However, we show that this constant factor gap is not a weakness of the analyses, but the choice of the weights. Theorem 5 provides a guideline for choosing the optimal weights, and the numerical simulation results in Figure 3 show that there is in fact no gap in practice, if we use the optimal weights. We use the same settings as that of the first figure of Figure 2 for the figure below.

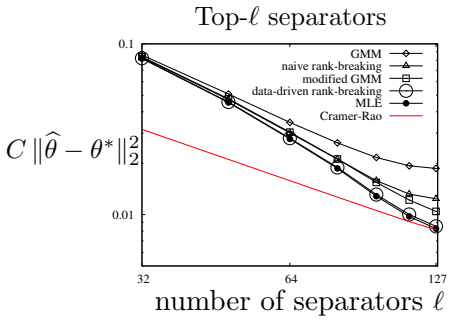

Figure 3: The proposed data-driven rank-breaking achieves performance identical to the MLE, and improves over naive rank-breaking with uniform weights.

To prove the order-optimality of the rank-breaking approach up to a constant factor, we can compare the upper bound to a Cramér-Rao lower bound on any unbiased estimators, in the following theorem. A proof is provided in Section 8.5.

Theorem 6. *Consider ranking $\{\sigma_j(i)\}_{i\in[\ell_j]}$ revealed for the set of items S_j, for $j \in [n]$. Let $\mathcal{U}$ denote the set of all unbiased estimators of $\theta^* \in \Omega_b$. If $b > 0$, then*

$$\inf_{\widehat{\theta}\in\mathcal{U}} \sup_{\theta^*\in\Omega_b} \mathbb{E}[\|\widehat{\theta} - \theta^*\|^2] \;\geq\; \left(1 - \frac{1}{\ell_{\max}} \sum_{i=1}^{\ell_{\max}} \frac{1}{\kappa_{\max} - i + 1}\right)^{-1} \sum_{i=2}^{d} \frac{1}{\lambda_i(L)} \;\geq\; \frac{(d-1)^2}{\sum_{j=1}^{n} \ell_j}\,, \tag{16}$$

where $\ell_{\max} = \max_{j\in[n]} \ell_j$ and $\kappa_{\max} = \max_{j\in[n]} \kappa_j$. The second inequality follows from the Jensen's inequality.

Consider a case when the comparison graph is an expander such that α is a strictly positive constant, and $b = O(1)$ is also finite. Then, the Cramér-Rao lower bound show that the upper bound in (14) is optimal up to a logarithmic factor.

3.4 Optimality of the Choice of the Weights

We propose the optimal choice of the weights $\lambda_{j,a}$'s in Theorem 2. In this section, we show numerical simulations results comparing the proposed approach to other naive choices of the weights under various scenarios. We fix $d = 1024$ items and the underlying preference vector θ^* is uniformly distributed over $[-b, b]$ for $b = 2$. We generate n rankings over sets S_j of size κ for $j \in [n]$ according to the PL model with parameter θ^*. The comparison sets S_j's are chosen independently and uniformly at random from $[d]$.

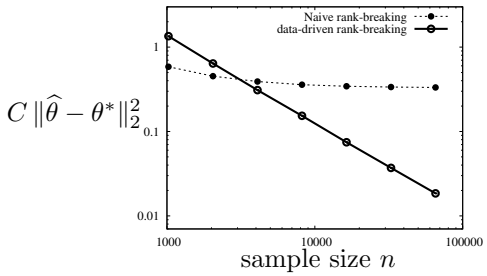

Figure 4: Data-driven rank-breaking is consistent, while a random rank-breaking results in inconsistency.

Figure 4 illustrates that a naive choice of rank-breakings can result in inconsistency. We create partial orderings data set by fixing $\kappa = 128$ and select $\ell = 8$ random positions in $\{1, \ldots, 127\}$. Each data set consists of partial orderings with separators at those 8 random positions, over 128 randomly chosen subset of items. We vary the sample size n and plot the resulting mean squared error for the two approaches. The data-driven rank-breaking, which uses the optimal choice of the weights, achieves error scaling as $1/n$ as predicted by Theorem 2, which implies consistency. For fair comparisons, we feed the same number of pairwise orderings to a naive rank-breaking estimator. This estimator uses randomly chosen pairwise orderings with uniform weights, and is

generally inconsistent. However, when sample size is small, inconsistent estimators can achieve smaller variance leading to smaller error. Normalization constant C is $10^3\ell/d^2$, and each point is averaged over 100 trials. We use the minorization-maximization algorithm from Hunter (2004) for computing the estimates from the rank-breakings.

Even if we use the consistent rank-breakings first proposed in Azari Soufiani et al. (2014), there is ambiguity in the choice of the weights. We next study how much we gain by using the proposed optimal choice of the weights. The optimal choice, $\lambda_{j,a} = 1/(\kappa_j - p_{j,a})$, depends on two parameters: the size of the offerings κ_j and the position of the separators $p_{j,a}$. To distinguish the effect of these two parameters, we first experiment with fixed $\kappa_j = \kappa$ and illustrate the gain of the optimal choice of $\lambda_{j,a}$'s.

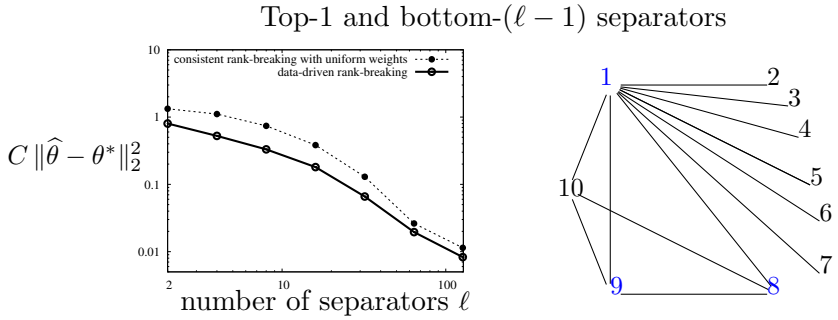

Figure 5: There is a constant factor gain of choosing optimal $\lambda_{j,a}$'s when the size of offerings are fixed, i.e. $\kappa_j = \kappa$ (left). We choose a particular set of separators where one separators is at position one and the rest are at the bottom. An example for $\ell = 3$ and $\kappa = 10$ is shown, where the separators are indicated by blue (right).

Figure 5 illustrates that the optimal choice of the weights improves over consistent rank-breaking with uniform weights by a constant factor. We fix $\kappa = 128$ and $n = 128000$. As illustrated by a figure on the right, the position of the separators are chosen such that there is one separator at position one, and the rest of $\ell - 1$ separators are at the bottom. Precisely, $(p_{j,1}, p_{j,2}, p_{j,3}, \ldots, p_{j,\ell}) = (1, 128 - \ell + 1, 128 - \ell + 2, \ldots, 127)$. We consider this scenario to emphasize the gain of optimal weights. Observe that the MSE does not decrease at a rate of $1/\ell$ in this case. The parameter γ which appears in the bound of Theorem 2 is very small when the breaking positions $p_{j,a}$ are of the order κ_j as is the case here, when ℓ is small. Normalization constant C is n/d^2.

The gain of optimal weights is significant when the size of S_j's are highly heterogeneous. Figure 6 compares performance of the proposed algorithm, for the optimal choice and uniform choice of weights $\lambda_{j,a}$ when the comparison sets S_j's are of different sizes. We consider the case when n_1 agents provide their top-ℓ_1 choices over the sets of size κ_1, and n_2 agents provide their top-1 choice over the sets of size κ_2. We take $n_1 = 1024$, $\ell_1 = 8$, and $n_2 = 10n_1\ell_1$. Figure 6 shows MSE for the two choice of weights, when we fix $\kappa_1 = 128$, and vary κ_2 from 2 to 128. As predicted from our bounds, when optimal choice of $\lambda_{j,a}$ is used MSE is not sensitive to sample set sizes κ_2. The error decays at the rate proportional to the inverse of the effective sample size, which is $n_1\ell_1 + n_2\ell_2 = 11n_1\ell_1$. However, with $\lambda_{j,a} = 1$ when $\kappa_2 = 2$, the MSE is roughly 10 times worse. Which reflects that the effective sample size is approximately $n_1\ell_1$, i.e. pairwise comparisons coming

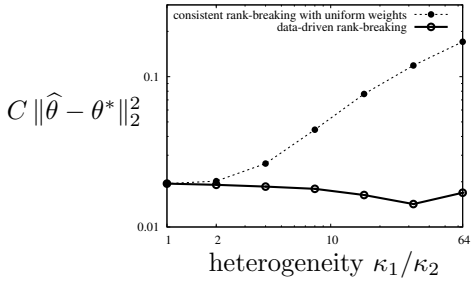

Figure 6: The gain of choosing optimal $\lambda_{j,a}$'s is significant when κ_j's are highly heterogeneous.

from small set size do not contribute without proper normalization. This gap in MSE corroborates bounds of Theorem 8. Normalization constant C is $10^3/d^2$.

4. The Role of the Topology of the Data

We study the role of topology of the data that provides a guideline for designing the collection of data when we do have some control, as in recommendation systems, designing surveys, and crowdsourcing. The core optimization problem of interest to the designer of such a system is to achieve the best accuracy while minimizing the number of questions.

4.1 The Role of the Graph Laplacian

Using the same number of samples, comparison graphs with larger spectral gap achieve better accuracy, compared to those with smaller spectral gaps. To illustrate how graph topology effects the accuracy, we reproduce known spectral properties of canonical graphs, and numerically compare the performance of data-driven rank-breaking for several graph topologies. We follow the examples and experimental setup from Shah et al. (2015a) for a similar result with pairwise comparisons. Spectral properties of graphs have been a topic of wide interest for decades. We consider a scenario where we fix the size of offerings as $\kappa_j = \kappa = O(1)$ and each agent provides partial ranking with ℓ separators, positions of which are chosen uniformly at random. The resulting spectral gap α of different choices of the set S_j's are provided below. The total number edges in the comparisons graph (counting hyper-edges as multiple edges) is defined as $|E| \equiv \binom{\kappa}{2} n$.

- Complete graph: when $|E|$ is larger than $\binom{d}{2}$, we can design the comparison graph to be a complete graph over d nodes. The weight $A_{ii'}$ on each edge is $n\,\ell/(d(d-1))$, which is the effective number of samples divided by twice the number of edges. Resulting spectral gap is one, which is the maximum possible value. Hence, complete graph is optimal for rank aggregation.

- Sparse random graph: when we have limited resources we might not be able to afford a dense graph. When $|E|$ is of order $o(d^2)$, we have a sparse graph. Consider a scenario where each set S_j is chosen uniformly at random. To ensure connectivity, we need $n = \Omega(\log d)$. Following standard spectral analysis of random graphs, we have $\alpha = \Theta(1)$. Hence, sparse random graphs are near-optimal for rank-aggregation.

- Chain graph: we consider a chain of sets of size κ overlapping only by one item. For example, $S_1 = \{1, \ldots, \kappa\}$ and $S_2 = \{\kappa, \kappa+1, \ldots, 2\kappa-1\}$, etc. We choose n to be a multiple of $\tau \equiv (d-1)/(\kappa-1)$ and offer each set n/τ times. The resulting graph is a chain of size κ cliques, and standard spectral analysis shows that $\alpha = \Theta(1/d^2)$. Hence, a chain graph is strictly sub-optimal for rank aggregation.

- Star-like graph: We choose one item to be the center, and every offer set consists of this center node and a set of $\kappa - 1$ other nodes chosen uniformly at random without replacement. For example, center node $= \{1\}$, $S_1 = \{1, 2, \ldots, \kappa\}$ and $S_2 = \{1, \kappa+1, \kappa+2, \ldots, 2\kappa-1\}$, etc. n is chosen in the way similar to that of the Chain graph. Standard spectral analysis shows that $\alpha = \Theta(1)$ and star-like graphs are near-optimal for rank-aggregation.

- Barbell-like graph: We select an offering $S = \{S', i, j\}$, $|S'| = \kappa - 2$ uniformly at random and divide rest of the items into two groups V_1 and V_2. We offer set S $n\kappa/d$ times. For each offering of set S, we offer $d/\kappa - 1$ sets chosen uniformly at random from the two groups $\{V_1, i\}$ and $\{V_2, j\}$. The resulting graph is a barbell-like graph, and standard spectral analysis shows that $\alpha = \Theta(1/d^2)$. Hence, a chain graph is strictly sub-optimal for rank aggregation.

Figure 7 illustrates how graph topology effects the accuracy. When θ^* is chosen uniformly at random, the accuracy does not change with d (left), and the accuracy is better for those graphs with larger spectral gap. However, for a certain worst-case θ^*, the error increases with d for the chain graph and the barbell-like graph, as predicted by the above analysis of the spectral gap. We use $\ell = 4$, $\kappa = 17$ and vary d from 129 to 2049. κ is kept small to make the resulting graphs more like the above discussed graphs. Figure on left shows accuracy when θ^* is chosen i.i.d. uniformly over $[-b, b]$ with $b = 2$. Error in this case is roughly same for each of the graph topologies with chain graph being the worst. However, when θ^* is chosen carefully error for chain graph and barbell-like graph increases with d as shown in the figure right. We chose θ^* such that all the items of a set have same weight, either $\theta_i = 0$ or $\theta_i = b$ for chain graph and barbell-like graph. We divide all the sets equally between the two types for chain graph. For barbell-like graph, we keep the two types of sets on the two different sides of the connector set and equally divide items of the connector set into two types. Number of samples n is $100(d-1)/(\kappa-1)$ and each point is averaged over 100 instances. Normalization constant C is $n\ell/d^2$.

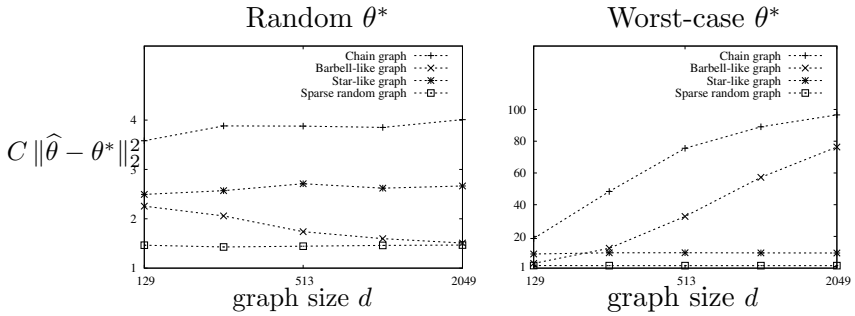

Figure 7: For randomly chosen θ^* the error does not change with d (left). However, for particular worst-case θ^* the error increases with d for the Chain graph and the Barbell-like graph as predicted by the analysis of the spectral gap (right).

4.2 The Role of the Position of the Separators

As predicted by theorem 2, rank-breaking fails when γ is small, i.e. the position of the separators are very close to the bottom. An extreme example is the bottom-ℓ separators scenario, where each person is offered κ randomly chosen alternatives, and is asked to give a ranked list of bottom ℓ alternatives. In other words, the ℓ separators are placed at $(p_{j,1}, \ldots, p_{j,\ell}) = (\kappa_j - \ell, \ldots, \kappa - 1)$. In this case, $\gamma \simeq 0$ and the error bound is large. This is not a weakness of the analysis. In fact we observe large errors under this scenario. The reason is that many alternatives that have large weights θ_i's will rarely be even compared once, making any reasonable estimation infeasible.

Figure 8 illustrates this scenario. We choose $\ell = 8$, $\kappa = 128$, and $d = 1024$. The other settings are same as that of the first figure of Figure 2. The left figure plots the magnitude of the estimation error for each item. For about 200 strong items among 1024, we do not even get a single comparison, hence we omit any estimation error. It clearly shows the trend: we get good estimates for about 400 items in the bottom, and we get large errors for the rest. Consequently, even if we only take those items that have at least one comparison into account, we still get large errors. This is shown in the figure right. The error barely decays with the sample size. However, if we focus on the error for the bottom 400 items, we get good error rate decaying inversely with the sample size. Normalization constant C in the second figure is $10^2\, x\, d/\ell$ and $10^2(400)d/\ell$ for the first and second lines respectively, where x is the number of items that appeared in rank-breaking at least once. We solve convex program (3) for θ restricted to the items that appear in rank-breaking at least once. The second figure of Figure 8 is averaged over 1000 instances.

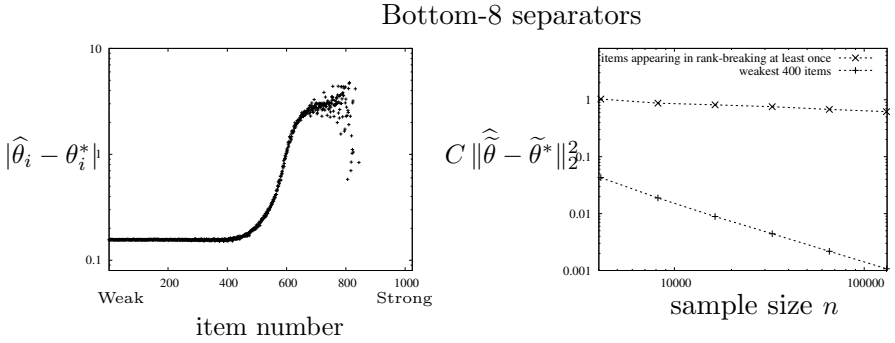

Figure 8: Under the bottom-ℓ separators scenario, accuracy is good only for the bottom 400 items (left). As predicted by Theorem 7, the mean squared error on the bottom 400 items scale as $1/n$, where as the overall mean squared error does not decay (right).

We make this observation precise in the following theorem. Applying rank-breaking to only to those weakest $\widetilde{d}$ items, we prove an upper bound on the achieved error rate that depends on the choice of the $\widetilde{d}$. Without loss of generality, we suppose the items are sorted such that $\theta^*_1 \leq \theta^*_2 \leq \cdots \leq \theta^*_d$. For a choice of $\widetilde{d} = \ell d/(2\kappa)$, we denote the weakest $\widetilde{d}$ items by $\widetilde{\theta}^* \in \mathbb{R}^{\widetilde{d}}$ such that $\widetilde{\theta}^*_i = \theta^*_i - (1/\widetilde{d}) \sum_{i'=1}^{\widetilde{d}} \theta^*_{i'}$, for $i \in [\widetilde{d}]$. Since $\theta^* \in \Omega_b$, $\widetilde{\theta}^* \in [-2b, 2b]^{\widetilde{d}}$. The space of all possible preference vectors for $[\widetilde{d}]$ items is given by $\widetilde{\Omega} = \{\widetilde{\theta} \in \mathbb{R}^{\widetilde{d}} : \sum_{i=1}^{\widetilde{d}} \widetilde{\theta}_i = 0\}$ and $\widetilde{\Omega}_{2b} = \widetilde{\Omega} \cap [-2b, 2b]^{\widetilde{d}}$.

Although the analysis can be easily generalized, to simplify notations, we fix $\kappa_j = \kappa$ and $\ell_j = \ell$ and assume that the comparison sets S_j, $|S_j| = \kappa$, are chosen uniformly at random from the set of

d items for all $j \in [n]$. The rank-breaking log likelihood function $\mathcal{L}_{\mathrm{RB}}(\widetilde{\theta})$ for the set of items $[\widetilde{d}]$ is given by

$$\mathcal{L}_{\mathrm{RB}}(\widetilde{\theta}) \;=\; \sum_{j=1}^{n}\sum_{a=1}^{\ell_j} \lambda_{j,a} \Big\{ \sum_{(i,i')\in E_{j,a}} \mathbb{I}_{\left\{i,i'\in[\widetilde{d}]\right\}} \Big(\theta_{i'} - \log\Big(e^{\theta_i}+e^{\theta_{i'}}\Big)\Big)\Big\}\,. \tag{17}$$

We analyze the rank-breaking estimator

$$\widehat{\widetilde{\theta}} \;\equiv\; \max_{\widetilde{\theta}\in\widetilde{\Omega}_{2b}} \mathcal{L}_{\mathrm{RB}}(\widetilde{\theta})\,. \tag{18}$$

We further simplify notations by fixing $\lambda_{j,a}=1$, since from Equation (24), we know that the error increases by at most a factor of 4 due to this sub-optimal choice of the weights, under the special scenario studied in this theorem.

Theorem 7. *Under the bottom-ℓ separators scenario and the PL model, S_j's are chosen uniformly at random of size κ and n partial orderings are sampled over d items parametrized by $\theta^* \in \Omega_b$. For $\widetilde{d} = \ell d/(2\kappa)$ and any $\ell \geq 4$, if the effective sample size is large enough such that*

$$n\ell \;\geq\; \left(\frac{2^{14}e^{8b}}{\chi^2}\frac{\kappa^3}{\ell^3}\right) d\log d\,, \tag{19}$$

where

$$\chi \;\equiv\; \frac{1}{4}\left(1-\exp\left(-\frac{2}{9(\kappa-2)}\right)\right), \tag{20}$$

then the rank-breaking *estimator in* (18) *achieves*

$$\frac{1}{\sqrt{\widetilde{d}}}\big\|\widehat{\widetilde{\theta}}-\widetilde{\theta}^*\big\|_2 \;\leq\; \frac{128(1+e^{4b})^2}{\chi}\frac{\kappa^{3/2}}{\ell^{3/2}}\sqrt{\frac{d\log d}{n\ell}}\,, \tag{21}$$

with probability at least $1-3e^3d^{-3}$.

Consider a scenario where $\kappa = O(1)$ and $\ell = \Theta(\kappa)$. Then, χ is a strictly positive constant, and also κ/ℓ is s finite constant. It follows that rank-breaking requires the effective sample size $n\ell = O(d\log d/\varepsilon^2)$ in order to achieve arbitrarily small error of $\varepsilon > 0$, on the weakest $\widetilde{d} = \ell\, d/(2\kappa)$ items.

5. Real-World Data Sets

On real-world data sets on sushi preferences (Kamishima, 2003), we show that the data-driven rank-breaking improves over Generalized Method-of-Moments (GMM) proposed by Azari Soufiani et al. (2013). This is a widely used data set for rank aggregation, for instance in Azari Soufiani et al. (2013, 2012); Maystre and Grossglauser (2015b); Le Van et al. (2015); Lu and Boutilier (2011a,b). The data set consists of complete rankings over 10 types of sushi from $n = 5000$ individuals. Below, we follow the experimental scenarios of the GMM approach in Azari Soufiani et al. (2013) for fair comparisons.

To validate our approach, we first take the estimated PL weights of the 10 types of sushi, using Hunter (2004) implementation of the ML estimator, over the entire input data of 5000 complete rankings. We take thus created output as the ground truth θ^*. To create partial rankings and compare the performance of the data-driven rank-breaking to the state-of-the-art GMM approach in Figure 9, we first fix $\ell = 6$ and vary n to simulate top-ℓ-separators scenario by removing the known ordering among bottom $10 - \ell$ alternatives for each sample in the data set (left). We next fix $n = 1000$ and vary ℓ and simulate top-ℓ-separators scenarios (right). Each point is averaged over 1000 instances. The mean squared error is plotted for both algorithms.

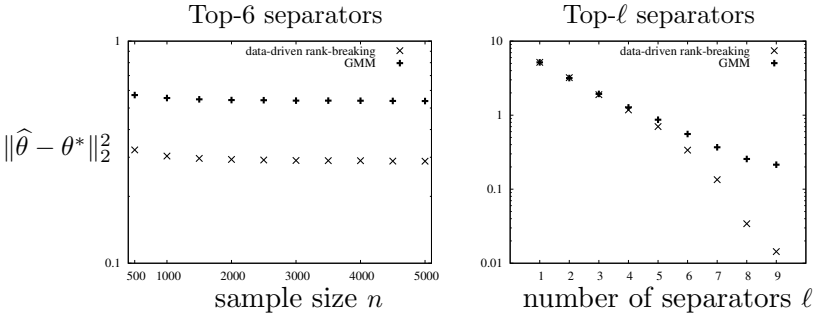

Figure 9: The data-driven rank-breaking achieves smaller error compared to the state-of-the-art GMM approach.

Figure 10 illustrates the Kendall rank correlation of the rankings estimated by the two algorithms and the ground truth. Larger value indicates that the estimate is closer to the ground truth, and the data-driven rank-breaking outperforms the state-of-the-art GMM approach.

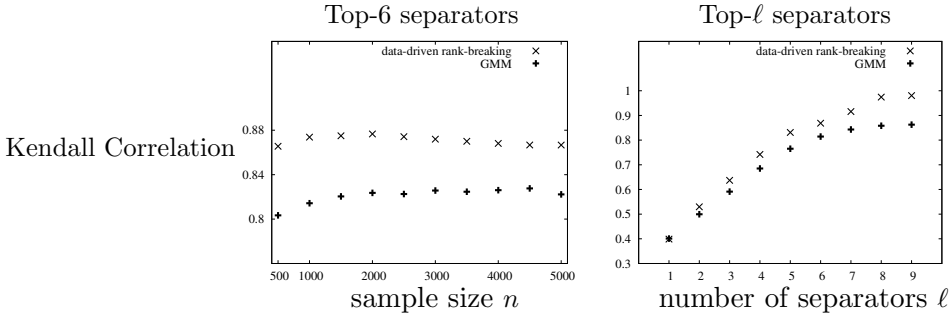

Figure 10: The data-driven rank-breaking achieves larger Kendall rank correlation compared to the state-of-the-art GMM approach.

To validate whether PL model is the right model to explain the sushi data set, we compare the data-driven rank-breaking, MLE for the PL model, GMM for the PL model, Borda count and Spearman's footrule optimal aggregation. We measure the Kendall rank correlation between the estimates and the samples and show the result in Table 1. In particular, if $\sigma_1, \sigma_2, \cdots, \sigma_n$ denote sample rankings and $\widehat{\sigma}$ denote the aggregated ranking then the correlation value is $(1/n)\sum_{i=1}^{n}\left(1 - \frac{4\mathcal{K}(\widehat{\sigma},\sigma_i)}{\kappa(\kappa-1)}\right)$, where $\mathcal{K}(\sigma_1,\sigma_2) = \sum_{i<j\in[\kappa]} \mathbb{I}_{\{(\sigma_1^{-1}(i)-\sigma_1^{-1}(j))(\sigma_2^{-1}(i)-\sigma_2^{-1}(j))<0\}}$. The results are reported

for different number of samples n and different values of ℓ under the top-ℓ separators scenarios. When $\ell = 9$, we are using all the complete rankings, and all algorithms are efficient. When $\ell < 9$, we have partial orderings, and Spearman's footrule optimal aggregation is NP-hard. We instead use scaled footrule aggregation (SFO) given in Dwork et al. (2001). Most approaches achieve similar performance, except for the Spearman's footrule. The proposed data-driven rank-breaking achieves a slightly worse correlation compared to other approaches. However, note that none of the algorithms are necessarily maximizing the Kendall correlation, and are not expected to be particularly good in this metric.

	MLE under PL	data-driven RB	GMM	Borda count	Spearman's footrule
$n = 500$, $\ell = 9$	0.306	0.291	0.315	0.315	0.159
$n = 5000$, $\ell = 9$	0.309	0.309	0.315	0.315	0.079
$n = 5000$, $\ell = 2$	0.199	0.199	0.201	0.200	0.113
$n = 5000$, $\ell = 5$	0.217	0.200	0.217	0.295	0.152

Table 1: Kendall rank correlation on sushi data set.

We compare our algorithm with the GMM algorithm on two other real-world data-sets as well. We use jester data set (Goldberg et al., 2001) that consists of over 4.1 million continuous ratings between -10 to $+10$ of 100 jokes from $48,483$ users. The average number of jokes rated by an user is 72.6 with minimum and maximum being 36 and 100 respectively. We convert continuous ratings into ordinal rankings. This data-set has been used by Miyahara and Pazzani (2000); Polat and Du (2005); Cortes et al. (2007); Lebanon and Mao (2007) for rank aggregation and collaborative filtering.

Similar to the settings of sushi data experiments, we take the estimated PL weights of the 100 jokes over all the rankings as ground truth. Figure 11 shows comparative performance of the data-driven rank-breaking and the GMM for the two scenarios. We first fix $\ell = 10$ and vary n to simulate random-10 separators scenario (left). We next take all the rankings $n = 73421$ and vary ℓ to simulate random-ℓ separators scenario (rights). Since sets have different sizes, while varying ℓ we use full breaking if the setsize is smaller than ℓ. Each point is averaged over 100 instances. The mean squared error is plotted for both algorithms.

We perform similar experiments on American Psychological Association (APA) data-set (Diaconis, 1989). The APA elects a president each year by asking each member to rank order a slate of five candidates. The data-set represents full rankings given by 5738 members of the association in 1980's election. The mean squared error is plotted for both algorithms under the settings similar to that of jester data-set.

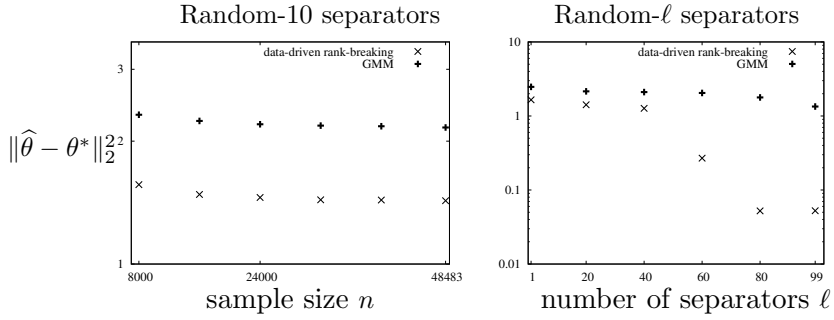

Figure 11: jester data set: The data-driven rank-breaking achieves smaller error compared to the state-of-the-art GMM approach.

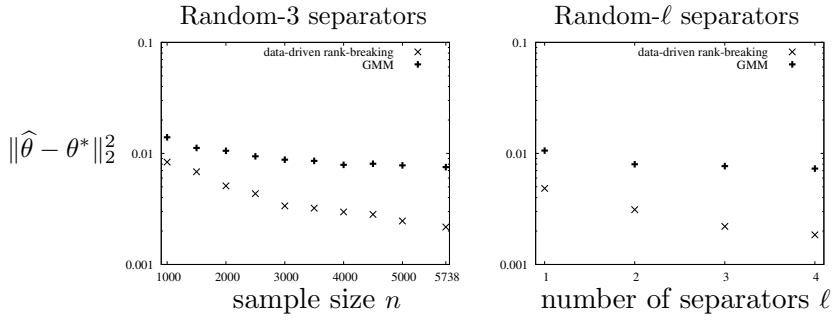

Figure 12: APA data set: The data-driven rank-breaking achieves smaller error compared to the state-of-the-art GMM approach.

6. Related Work

Initially motivated by elections and voting, rank aggregation has been a topic of mathematical interest dating back to Condorcet and Borda (De Condorcet, 1785; de Borda, 1781). Using probabilistic models to infer preferences has been popularized in operations research community for applications such as assortment optimization and revenue management. The PL model studied in this paper is a special case of MultiNomial Logit (MNL) models commonly used in discrete choice modeling, which has a long history in operations research (McFadden, 1980). Efficient inference algorithms has been proposed to either find the MLE efficiently or approximately, such as the iterative approaches in Ford Jr. (1957); Dykstra (1960), minorization-maximization approach in Hunter (2004), and Markov chain approaches in Negahban et al. (2012); Maystre and Grossglauser (2015a). These approaches are shown to achieve minimax optimal error rate in the traditional comparisons scenarios. Under the pairwise comparisons scenario, Negahban et al. (2012) provided Rank Centrality that provably achieves minimax optimal error rate for randomly chosen pairs, which was later generalized to arbitrary pairwise comparisons in Negahban et al. (2014). The analysis shows the explicit dependence on the topology of data shows that the spectral gap of comparisons graph similar to the one presented in this paper. This analysis was generalized to k-way comparisons in Hajek et al. (2014) and generalized to best-out-of-k comparisons with sharper bounds in Shah et al. (2015a). In an effort to give a guarantee for exact recovery of the top-ℓ items in the ranking, Chen

and Suh (2015) proposed a new algorithm based on Rank Centrality that provides a tighter error bound for L_∞ norm, as opposed to the existing L_2 error bounds. Another interesting direction in learning to rank is non-parametric learning from paired comparisons, initiated in several recent papers such as Duchi et al. (2010); Rajkumar and Agarwal (2014); Shah et al. (2015b); Shah and Wainwright (2015).

More recently, a more general problem of learning *personal* preferences from ordinal data has been studied (Yi et al., 2013; Lu and Boutilier, 2011b; Ding et al., 2015). The MNL model provides a natural generalization of the PL model to this problem. When users are classified into a small number of groups with same preferences, mixed MNL model can be learned from data as studied in Ammar et al. (2014); Oh and Shah (2014); Wu et al. (2015). A more general scenario is when each user has his/her individual preferences, but inherently represented by a lower dimensional feature. This problem was first posed as an inference problem in Lu and Negahban (2014) where convex relaxation of nuclear norm minimization was proposed with provably optimal guarantees. This was later generalized to k-way comparisons in Oh et al. (2015). A similar approach was studied with a different guarantees and assumptions in Park et al. (2015). Our algorithm and ideas of rank-breaking can be directly applied to this collaborative ranking under MNL, with the same guarantees for consistency in the asymptotic regime where sample size grows to infinity. However, the analysis techniques for MNL rely on stronger assumptions on how the data is collected, and especially on the independence of the samples. It is not immediate how the analysis techniques developed in this paper can be applied to learn MNL.

In an orthogonal direction, new discrete choice models with sparse structures has been proposed recently in Farias et al. (2009) and optimization algorithms for revenue management has been proposed Farias et al. (2013). In a similar direction, new discrete choice models based on Markov chains has been introduced in Blanchet et al. (2013), and corresponding revenue management algorithms has been studied in Feldman and Topaloglu (2014). However, typically these models are analyzed in the asymptotic regime with infinite samples, with the exception of Ammar and Shah (2011). A non-parametric choice models for pairwise comparisons also have been studied in Rajkumar and Agarwal (2014); Shah et al. (2015b). This provides an interesting opportunities to studying learning to rank for these new choice models.

We consider a fixed design setting, where inference is separate from data collection. There is a parallel line of research which focuses on adaptive ranking, mainly based on pairwise comparisons. When performing sorting from noisy pairwise comparisons, Braverman and Mossel (2009) proposed efficient approaches and provided performance guarantees. Following this work, there has been recent advances in adaptive ranking Ailon (2011); Jamieson and Nowak (2011); Maystre and Grossglauser (2015b).

7. Discussion

We study the problem of learning the PL model from ordinal data. Under the traditional data collection scenarios, several efficient algorithms find the maximum likelihood estimates and at the same time provably achieve minimax optimal performance. However, for some non-traditional scenarios, computational complexity of finding the maximum likelihood estimate can scale super-exponentially in the problem size. We provide the first finite-sample analysis of computationally efficient estimators known as rank-breaking estimators. This provides guidelines for choosing the

weights in the estimator to achieve optimal performance, and also explicitly shows how the accuracy depends on the topology of the data.

This paper provides the first analytical result in the sample complexity of rank-breaking estimators, and quantifies the price we pay in accuracy for the computational gain. In general, more complex higher-order rank-breaking can also be considered, where instead of breaking a partial ordering into a collection of paired comparisons, we break it into a collection of higher-order comparisons. The resulting higher-order rank-breakings will enable us to traverse the whole spectrum of computational complexity between the pairwise rank-breaking and the MLE. We believe this paper opens an interesting new direction towards understanding the whole spectrum of such approaches. However, analyzing the Hessian of the corresponding objective function is significantly more involved and requires new technical innovations.

8. Proofs

8.1 Proof of Theorem 2

We prove a more general result for an arbitrary choice of the parameter $\lambda_{j,a} > 0$ for all $j \in [n]$ and $a \in [\ell_j]$. The following theorem proves the (near)-optimality of the choice of $\lambda_{j,a}$'s proposed in (10), and implies the corresponding error bound as a corollary.

Theorem 8. *Under the hypotheses of Theorem 2 and any $\lambda_{j,a}$'s, the rank-breaking estimator achieves*

$$\frac{1}{\sqrt{d}}\big\|\,\widehat{\theta}-\theta^*\,\big\|_2 \;\;\leq\;\; \frac{4\sqrt{2}e^{4b}(1+e^{2b})^2\sqrt{d\log d}}{\alpha\,\gamma}\,\frac{\sqrt{\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\big(\lambda_{j,a}\big)^2\big(\kappa_j-p_{j,a}\big)\big(\kappa_j-p_{j,a}+1\big)}}{\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}(\kappa_j-p_{j,a})}\,, \tag{22}$$

with probability at least $1-3e^3d^{-3}$, if

$$\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}(\kappa_j-p_{j,a}) \;\;\geq\;\; 2^6e^{18b}\frac{\eta\delta}{\alpha^2\beta\gamma^2\tau}d\log d\,, \tag{23}$$

where γ, η, τ, δ, α, β, are now functions of $\lambda_{j,a}$'s and defined in (7), (8), (25), (27) *and* (30).

We first claim that $\lambda_{j,a}=1/(\kappa_j-p_{j,a}+1)$ is the optimal choice for minimizing the above upper bound on the error. From Cauchy-Schwartz inequality and the fact that all terms are non-negative, we have that

$$\frac{\sqrt{\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\big(\lambda_{j,a}\big)^2(\kappa_j-p_{j,a})(\kappa_j-p_{j,a}+1)}}{\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}(\kappa_j-p_{j,a})} \;\;\geq\;\; \frac{1}{\sqrt{\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\frac{(\kappa_j-p_{j,a})}{(\kappa_j-p_{j,a}+1)}}}\,, \tag{24}$$

where $\lambda_{j,a}=1/(\kappa_j-p_{j,a}+1)$ achieves the universal lower bound on the right-hand side with an equality. Since $\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\frac{(\kappa_j-p_{j,a})}{(\kappa_j-p_{j,a}+1)}\geq\sum_{j=1}^{n}\ell_j$, substituting this into (22) gives the desired error bound in (11). Although we have identified the optimal choice of $\lambda_{j,a}$'s, we choose a slightly different value of $\lambda=1/(\kappa_j-p_{j,a})$ for the analysis. This achieves the same desired error bound in (11), and significantly simplifies the notations of the sufficient conditions.

We first define all the parameters in the above theorem for general $\lambda_{j,a}$. With a slight abuse of notations, we use the same notations for $\mathcal{H}$, L, α and β for both the general $\lambda_{j,a}$'s and also the specific choice of $\lambda_{j,a} = 1/(\kappa_j - p_{j,a})$. It should be clear from the context what we mean in each case. Define

$$\tau \equiv \min_{j\in[n]} \tau_j \,, \quad \text{where } \tau_j \equiv \frac{\sum_{a=1}^{\ell_j} \lambda_{j,a}(\kappa_j - p_{j,a})}{\ell_j} \tag{25}$$

$$\delta_{j,1} \equiv \left\{ \max_{a\in[\ell_j]} \Big\{ \lambda_{j,a}(\kappa_j - p_{j,a}) \Big\} + \sum_{a=1}^{\ell_j} \lambda_{j,a} \right\} \,, \text{ and } \quad \delta_{j,2} \equiv \sum_{a=1}^{\ell_j} \lambda_{j,a} \tag{26}$$

$$\delta \equiv \max_{j\in[n]} \left\{ 4\delta_{j,1}^2 + \frac{2\big(\delta_{j,1}\delta_{j,2} + \delta_{j,2}^2\big)\kappa_j}{\eta_j \ell_j} \right\} \,. \tag{27}$$

Note that $\delta \geq \delta_{j,1}^2 \geq \max_a \lambda_{j,a}^2 (\kappa_j - p_{j,a})^2 \geq \tau^2$, and for the choice of $\lambda_{j,a} = 1/(\kappa_j - p_{j,a})$ it simplifies as $\tau = \tau_j = 1$. We next define a comparison graph $\mathcal{H}$ for general $\lambda_{j,a}$, which recovers the proposed comparison graph for the optimal choice of $\lambda_{j,a}$'s

Definition 9. *(Comparison graph $\mathcal{H}$). Each item $i \in [d]$ corresponds to a vertex i. For any pair of vertices i, i', there is a weighted edge between them if there exists a set S_j such that $i, i' \in S_j$; the weight equals $\sum_{j: i,i' \in S_j} \frac{\tau_j \ell_j}{\kappa_j(\kappa_j - 1)}$.*

Let A denote the weighted adjacency matrix, and let $D = \mathrm{diag}(A\mathbf{1})$. Define,

$$D_{\max} \equiv \max_{i\in[d]} D_{ii} = \max_{i\in[d]} \left\{ \sum_{j: i\in S_j} \frac{\tau_j \ell_j}{\kappa_j} \right\} \geq \tau_{\min} \max_{i\in[d]} \left\{ \sum_{j: i\in S_j} \frac{\ell_j}{\kappa_j} \right\}. \tag{28}$$

Define graph Laplacian L as $L = D - A$, i.e.,

$$L = \sum_{j=1}^{n} \frac{\tau_j \ell_j}{\kappa_j(\kappa_j - 1)} \sum_{i<i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top. \tag{29}$$

Let $0 = \lambda_1(L) \leq \lambda_2(L) \leq \cdots \leq \lambda_d(L)$ denote the sorted eigenvalues of L. Note that $\mathrm{Tr}(L) = \sum_{i=1}^{d} \sum_{j: i\in S_j} \tau_j \ell_j / \kappa_j = \sum_{j=1}^{n} \tau_j \ell_j$. Define α and β such that

$$\alpha \equiv \frac{\lambda_2(L)(d-1)}{\mathrm{Tr}(L)} = \frac{\lambda_2(L)(d-1)}{\sum_{j=1}^{n} \tau_j \ell_j} \text{ and } \beta \equiv \frac{\mathrm{Tr}(L)}{d D_{\max}} = \frac{\sum_{j=1}^{n} \tau_j \ell_j}{d D_{\max}} \,. \tag{30}$$

For the proposed choice of $\lambda_{j,a} = 1/(\kappa_j - p_{j,a})$, we have $\tau_j = 1$ and the definitions of $\mathcal{H}$, L, α, and β reduce to those defined in Definition 1. We are left to prove an upper bound, $\delta \leq 32(\log(\ell_{\max} + 2))^2$, which implies the sufficient condition in (9) and finishes the proof of Theorem 2. We have,

$$\begin{aligned} \delta_{j,1} = \max_{a\in[\ell_j]} \Big\{ \lambda_{j,a}(\kappa_j - p_{j,a}) \Big\} + \sum_{a=1}^{\ell_j} \lambda_{j,a} &= 1 + \sum_{a=1}^{\ell_j} \frac{1}{\kappa_j - p_{j,a}} \\ &\leq 1 + \sum_{a=1}^{\ell_j} \frac{1}{a} \\ &\leq 2\log(\ell_j + 2)\,, \end{aligned} \tag{31}$$

where in the first inequality follows from taking the worst case for the positions, i.e. $p_{j,a} = \kappa_j - \ell_j + a - 1$ Using the fact that for any integer x, $\sum_{a=0}^{\ell-1} 1/(x+a) \le \log((x+\ell-1)/(x-1))$, we also have

$$\begin{aligned}
\frac{\delta_{j,2}\kappa_j}{\eta_j \ell_j} &\le \sum_{a=1}^{\ell_j} \frac{1}{\kappa_j - p_{j,a}} \frac{\max\{\ell_j, \kappa_j - p_{j,\ell_j}\}}{\ell_j} \\
&\le \min\Big\{ \log(\ell_j+2)\,, \log\Big(\frac{\kappa_j - p_{j,\ell_j} + \ell_j - 1}{\kappa_j - p_{j,\ell_j} - 1}\Big)\Big\} \frac{\max\{\ell_j, \kappa_j - p_{j,\ell_j}\}}{\ell_j} \\
&\le \frac{\log(\ell_j+2)\ell_j}{\max\{\ell_j, \kappa_j - p_{j,\ell_j} - 1\}} \frac{\max\{\ell_j, \kappa_j - p_{j,\ell_j}\}}{\ell_j} \\
&\le 2\log(\ell_j+2)\,, \qquad (32)
\end{aligned}$$

where the first inequality follows from the definition of η_j, Equation (8). From (31), (32), and the fact that $\delta_{j,2} \le \log(\ell_j+2)$, we have

$$\delta = \max_{j\in[n]} \left\{ 4\delta_{j,1}^2 + \frac{2\big(\delta_{j,1}\delta_{j,2} + \delta_{j,2}^2\big)\kappa_j}{\eta_j \ell_j} \right\} \le 28(\log(\ell_{\max}+2))^2\,.$$

8.2 Proof of Theorem 8

We first introduce two key technical lemmas. In the following lemma we show that $\mathbb{E}_{\theta^*}[\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)] = 0$ and provide a bound on the deviation of $\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)$ from its mean. The expectation $\mathbb{E}_{\theta^*}[\cdot]$ is with respect to the randomness in the samples drawn according to θ^*. The log likelihood Equation (2) can be rewritten as

$$\mathcal{L}_{\mathrm{RB}}(\theta) = \sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\sum_{i<i'\in S_j} \mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}} \lambda_{j,a}\Big(\theta_i \mathbb{I}_{\left\{\sigma_j^{-1}(i)<\sigma_j^{-1}(i')\right\}} + \theta_{i'}\mathbb{I}_{\left\{\sigma_j^{-1}(i)>\sigma_j^{-1}(i')\right\}} - \log\left(e^{\theta_i} + e^{\theta_{i'}}\right)\Big)\,. \qquad (33)$$

We use $(i,i') \in G_{j,a}$ to mean either (i,i') or (i',i) belong to $E_{j,a}$. Taking the first-order partial derivative of $\mathcal{L}_{\mathrm{RB}}(\theta)$, we get

$$\nabla_i \mathcal{L}_{\mathrm{RB}}(\theta^*) = \sum_{j:i\in S_j}\sum_{a=1}^{\ell_j}\sum_{\substack{i'\in S_j \\ i'\neq i}} \lambda_{j,a}\, \mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}} \left(\mathbb{I}_{\left\{\sigma_j^{-1}(i)<\sigma_j^{-1}(i')\right\}} - \frac{\exp(\theta_i^*)}{\exp(\theta_i^*)+\exp(\theta_{i'}^*)} \right). \qquad (34)$$

Lemma 10. *Under the hypotheses of Theorem 2, with probability at least* $1 - 2e^3 d^{-3}$,

$$\big\|\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)\big\|_2 \le \sqrt{6\log d \sum_{j=1}^{n}\sum_{a=1}^{\ell_j} \big(\lambda_{j,a}\big)^2 \big(\kappa_j - p_{j,a}\big)\big(\kappa_j - p_{j,a}+1\big)}\,.$$

The Hessian matrix $H(\theta) \in \mathcal{S}^d$ with $H_{ii'}(\theta) = \frac{\partial^2 \mathcal{L}_{\mathrm{RB}}(\theta)}{\partial\theta_i \partial\theta_{i'}}$ is given by

$$H(\theta) = -\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\sum_{i<i'\in S_j} \mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}} \lambda_{j,a} \left((e_i - e_{i'})(e_i - e_{i'})^\top \frac{\exp(\theta_i+\theta_{i'})}{[\exp(\theta_i)+\exp(\theta_{i'})]^2} \right). \qquad (35)$$

It follows from the definition that $-H(\theta)$ is positive semi-definite for any $\theta \in \mathbb{R}^d$. The smallest eigenvalue of $-H(\theta)$ is equal to zero and the corresponding eigenvector is all-ones vector. The following lemma lower bounds its second smallest eigenvalue $\lambda_2(-H(\theta))$.

Lemma 11. *Under the hypotheses of Theorem 2, if*

$$\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}(\kappa_j - p_{j,a}) \geq 2^6 e^{18b}\frac{\eta\delta}{\alpha^2\beta\gamma^2\tau}d\log d \tag{36}$$

then with probability at least $1 - d^{-3}$, the following holds for any $\theta \in \Omega_b$:

$$\lambda_2(-H(\theta)) \;\geq\; \frac{e^{-4b}}{(1+e^{2b})^2}\frac{\alpha\gamma}{d-1}\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}(\kappa_j - p_{j,a})\,. \tag{37}$$

Define $\Delta = \widehat{\theta} - \theta^*$. It follows from the definition that Δ is orthogonal to the all-ones vector. By the definition of $\hat{\theta}$ as the optimal solution of the optimization (3), we know that $\mathcal{L}_{\mathrm{RB}}(\widehat{\theta}) \geq \mathcal{L}_{\mathrm{RB}}(\theta^*)$ and thus

$$\mathcal{L}_{\mathrm{RB}}(\widehat{\theta}) - \mathcal{L}_{\mathrm{RB}}(\theta^*) - \langle\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*), \Delta\rangle \;\geq\; -\langle\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*), \Delta\rangle \;\geq\; -\|\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)\|_2\|\Delta\|_2, \tag{38}$$

where the last inequality follows from the Cauchy-Schwartz inequality. By the mean value theorem, there exists a $\theta = a\widehat{\theta} + (1-a)\theta^*$ for some $a \in [0,1]$ such that $\theta \in \Omega_b$ and

$$\mathcal{L}_{\mathrm{RB}}(\widehat{\theta}) - \mathcal{L}_{\mathrm{RB}}(\theta^*) - \langle\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*), \Delta\rangle \;=\; \frac{1}{2}\Delta^\top H(\theta)\Delta \leq -\frac{1}{2}\lambda_2(-H(\theta))\|\Delta\|_2^2, \tag{39}$$

where the last inequality holds because the Hessian matrix $-H(\theta)$ is positive semi-definite with $H(\theta)\mathbf{1} = \mathbf{0}$ and $\Delta^\top\mathbf{1} = 0$. Combining (38) and (39),

$$\|\Delta\|_2 \;\leq\; \frac{2\|\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)\|_2}{\lambda_2(-H(\theta))}. \tag{40}$$

Note that $\theta \in \Omega_b$ by definition. Theorem 8 follows by combining Equation (40) with Lemma 10 and Lemma 11.

8.2.1 Proof of Lemma 10

The idea of the proof is to view $\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)$ as the final value of a discrete time vector-valued martingale with values in $\mathbb{R}^d$. Define $\nabla\mathcal{L}_{G_{j,a}}(\theta^*)$ as the gradient vector arising out of each rank-breaking graph $\{G_{j,a}\}_{j\in[n],a\in[\ell_j]}$ that is

$$\nabla_i\mathcal{L}_{G_{j,a}}(\theta^*) \equiv \sum_{\substack{i'\in S_j\\ i'\neq i}}\lambda_{j,a}\,\mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}}\left(\mathbb{I}_{\left\{\sigma_j^{-1}(i)<\sigma_j^{-1}(i')\right\}} - \frac{\exp(\theta_i^*)}{\exp(\theta_i^*)+\exp(\theta_{i'}^*)}\right). \tag{41}$$

Consider $\nabla\mathcal{L}_{G_{j,a}}(\theta^*)$ as the incremental random vector in a martingale of $\sum_{j=1}\ell_j$ time steps. Lemma 12 shows that the expectation of each incremental vector is zero. Observe that the conditioning event $\{i'' \in S \,:\, \sigma^{-1}(i'') < p_{j,a}\}$ given in (43) is equivalent to conditioning on the history

$\{G_{j,a'}\}_{a'<a}$. Therefore, using the assumption that the rankings $\{\sigma_j\}_{j\in[n]}$ are mutually independent, we have that the conditional expectation of $\nabla\mathcal{L}_{G_{j,a}}(\theta^*)$ conditioned on $\{G_{j',a''}\}_{j'<j,a''\in[\ell_{j'}]}$ is zero. Further, the conditional expectation of $\nabla\mathcal{L}_{G_{j,a}}(\theta^*)$ is zero even when conditioned on the rank breaking due to previous separators $\{G_{j,a'}\}_{a'<a}$ that are ranked higher (i.e. $a'<a$), which follows from the next lemma.

Lemma 12. *For a position-p rank breaking graph G_p, defined over a set of items S, where $p\in[|S|-1]$,*

$$\mathbb{P}\Big[\sigma^{-1}(i)<\sigma^{-1}(i')\;\Big|\;(i,i')\in G_p\Big]\;=\;\frac{\exp(\theta^*_i)}{\exp(\theta^*_i)+\exp(\theta^*_{i'})}\,, \tag{42}$$

for all $i,i'\in S$ and also

$$\mathbb{P}\Big[\sigma^{-1}(i)<\sigma^{-1}(i')\;\Big|\;(i,i')\in G_p \text{ and } \{i''\in S:\sigma^{-1}(i'')<p\}\Big]\;=\;\frac{\exp(\theta^*_i)}{\exp(\theta^*_i)+\exp(\theta^*_{i'})}\,. \tag{43}$$

This is one of the key technical lemmas since it implies that the proposed rank-breaking is consistent, i.e. $\mathbb{E}_{\theta^*}[\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)]=0$. Throughout the proof of Theorem 2, this is the only place where the assumption on the proposed (consistent) rank-breaking is used. According to a companion theorem in Azari Soufiani et al. (2014, Theorem 2), it also follows that any rank-breaking that is not union of position-p rank-breakings results in inconsistency, i.e. $\mathbb{E}_{\theta^*}[\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)]\neq 0$. We claim that for each rank-breaking graph $G_{j,a}$, $\|\nabla\mathcal{L}_{G_{j,a}}(\theta^*)\|_2^2\le(\lambda_{j,a})^2(\kappa_j-p_{j,a})(\kappa_j-p_{j,a}+1)$. By Lemma 13 which is a generalization of the vector version of the Azuma-Hoeffding inequality found in (Hayes, 2005, Theorem 1.8), we have

$$\mathbb{P}\big[\big\|\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)\big\|_2\ge\delta\big]\;\;\le\;\;2e^3\exp\left(\frac{-\delta^2}{2\sum_{j=1}^n\sum_{a=1}^{\ell_j}\big(\lambda_{j,a}\big)^2\big(\kappa_j-p_{j,a}\big)\big(\kappa_j-p_{j,a}+1\big)}\right),$$

which implies the result.

Lemma 13. *Let $(X_1,X_2,\cdots,X_n)$ be real-valued martingale taking values in $\mathbb{R}^d$ such that $X_0=0$ and for every $1\le i\le n$, $\|X_i-X_{i-1}\|_2\le c_i$, for some non-negative constant c_i. Then for every $\delta>0$,*

$$\mathbb{P}[\|X_n\|_2\ge\delta]\;\;\le\;\;2e^3e^{-\frac{\delta^2}{2\sum_{i=1}^n c_i^2}}\,. \tag{44}$$

It follows from the upper bound on $\|\nabla\mathcal{L}_{G_{j,a}}(\theta^*)\|_2^2\le c_i^2$ with $c_i^2=\lambda^2\big((k_j-p_{j,a})^2+(k_j-p_{j,a})\big)$. In the expression (41), $\nabla\mathcal{L}_{G_{j,a}}(\theta^*)$ has one entry at $p_{j,a}$-th position that is compared to $(k_j-p_{j,a})$ other items and $(k_j-p_{j,a})$ entries that is compared only once, giving the bound

$$\|\nabla\mathcal{L}_{G_{j,a}}(\theta^*)\|_2^2\;\;\le\;\;\lambda_{j,a}^2(k_j-p_{j,a})^2+\lambda_{j,a}^2(k_j-p_{j,a})\,.$$

8.2.2 Proof of Lemma 12

Define event $E\equiv\{(i,i')\in G_p\}$. Observe that

$$E=\Big\{\Big(\mathbb{I}_{\{(\sigma^{-1}(i)=p\}}+\mathbb{I}_{\{\sigma^{-1}(i'))=p\}}=1\Big)\wedge\Big(\sigma^{-1}(i),\sigma^{-1}(i')\ge p\Big)\Big\}\,.$$

Consider any set $\Omega \subset S \setminus \{i, i'\}$ such that $|\Omega| = p-1$. Let M denote an event that items of the set Ω are ranked in top-$(p-1)$ positions in a particular order. It is easy to verify the following:

$$\begin{aligned}
\mathbb{P}\Big[\sigma^{-1}(i) < \sigma^{-1}(i')\Big| E, M\Big] &= \frac{\mathbb{P}\Big[\big(\sigma^{-1}(i) < \sigma^{-1}(i')\big), E, M\Big]}{\mathbb{P}\Big[E, M\Big]} \\
&= \frac{\mathbb{P}\Big[\big(\sigma^{-1}(i) = p\big), M\Big]}{\mathbb{P}\Big[\big(\sigma^{-1}(i) = p\big), M\Big] + \mathbb{P}\Big[\big(\sigma^{-1}(i') = p\big), M\Big]} \\
&= \frac{\exp(\theta^*_i)}{\exp(\theta^*_i) + \exp(\theta^*_{i'})} = \mathbb{P}\Big[\sigma^{-1}(i) < \sigma^{-1}(i')\Big] .
\end{aligned}$$

Since M is any particular ordering of the set Ω and Ω is any subset of $S\setminus\{i,i'\}$ such that $|\Omega| = p-1$, conditioned on event E probabilities of all the possible events M over all the possible choices of set Ω sum to 1.

8.2.3 Proof of Lemma 13

It follows exactly along the lines of proof of Theorem 1.8 in (Hayes, 2005).

8.2.4 Proof of Lemma 11

The Hessian $H(\theta)$ is given in (35). For all $j \in [n]$, define $M^{(j)} \in \mathcal{S}^d$ as

$$M^{(j)} \equiv \sum_{a=1}^{\ell_j} \lambda_{j,a} \sum_{i<i' \in S_j} \mathbb{I}_{\big\{(i,i') \in G_{j,a}\big\}} (e_i - e_{i'})(e_i - e_{i'})^\top, \tag{45}$$

and let $M \equiv \sum_{j=1}^n M^{(j)}$. Observe that M is positive semi-definite and the smallest eigenvalue of M is zero with the corresponding eigenvector given by the all-ones vector. If $|\theta_i| \le b$, for all $i \in [d]$, $\frac{\exp(\theta_i+\theta_{i'})}{[\exp(\theta_i)+\exp(\theta_{i'})]^2} \ge \frac{e^{2b}}{(1+e^{2b})^2}$. Recall the definition of $H(\theta)$ from Equation (35). It follows that $-H(\theta) \succeq \frac{e^{2b}}{(1+e^{2b})^2} M$ for $\theta \in \Omega_b$. Since, $-H(\theta)$ and M are symmetric matrices, from Weyl's inequality we have, $\lambda_2(-H(\theta)) \ge \frac{e^{2b}}{(1+e^{2b})^2}\lambda_2(M)$. Again from Weyl's inequality, it follows that

$$\lambda_2(M) \ge \lambda_2(\mathbb{E}[M]) - \|M - \mathbb{E}[M]\| , \tag{46}$$

where $\|\cdot\|$ denotes the spectral norm. We will show in (51) that $\lambda_2(\mathbb{E}[M]) \ge 2\gamma e^{-6b}(\alpha/(d-1))\sum_{j=1}^n \tau_j \ell_j$, and in (63) that $\|M - \mathbb{E}[M]\| \le 8e^{3b}\sqrt{\frac{\eta\delta\log d}{\beta\tau d}\sum_{j=1}^n \tau_j\ell_j}$.

$$\lambda_2(M) \ge \frac{2e^{-6b}\alpha\gamma}{d-1}\sum_{j=1}^n \tau_j\ell_j - 8e^{3b}\sqrt{\frac{\eta\delta\log d}{\beta\tau d}\sum_{j=1}^n \tau_j\ell_j} \ge \frac{e^{-6b}\alpha\gamma}{d-1}\sum_{j=1}^n \tau_j\ell_j , \tag{47}$$

where the last inequality follows from the assumption that $\sum_{j=1}^n \tau_j\ell_j \ge 2^6 e^{18b}\frac{\eta\delta}{\alpha^2\beta\gamma^2\tau} d\log d$. This proves the desired claim.

To prove the lower bound on $\lambda_2(\mathbb{E}[M])$, notice that

$$\mathbb{E}[M] \;=\; \sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}\sum_{i<i'\in S_j}\mathbb{P}\Big[(i,i')\in G_{j,a}\Big|(i,i'\in S_j)\Big](e_i-e_{i'})(e_i-e_{i'})^\top\,. \tag{48}$$

The following lemma provides a lower bound on $\mathbb{P}[(i,i')\in G_{j,a}|(i,i'\in S_j)]$.

Lemma 14. *Consider a ranking σ over a set $S\subseteq[d]$ such that $|S|=\kappa$. For any two items $i,i'\in S$, $\theta\in\Omega_b$, and $1\le\ell\le\kappa-1$,*

$$\mathbb{P}_\theta\Big[\sigma^{-1}(i)=\ell,\sigma^{-1}(i')>\ell\Big] \;\ge\; \frac{e^{-6b}(\kappa-\ell)}{\kappa(\kappa-1)}\left(1-\frac{\ell}{\kappa}\right)^{\alpha_{i,i',\ell,\theta}-2}\,, \tag{49}$$

where the probability $\mathbb{P}_\theta$ is with respect to the sampled ranking resulting from PL weights $\theta\in\Omega_b$, and $\alpha_{i,i',\ell,\theta}$ is defined as $1\le\alpha_{i,i',\ell,\theta}=\lceil\widetilde{\alpha}_{i,i',\ell,\theta}\rceil$, and $\widetilde{\alpha}_{i,i',\ell,\theta}$ is,

$$\widetilde{\alpha}_{i,i',\ell,\theta} \;\equiv\; \max_{\ell'\in[\ell]}\max_{\substack{\Omega\subseteq S\setminus\{i,i'\}\\:|\Omega|=\kappa-\ell'}}\left\{\frac{\exp(\theta_i)+\exp(\theta_{i'})}{\big(\sum_{j\in\Omega}\exp(\theta_j)\big)/|\Omega|}\right\}. \tag{50}$$

Note that we do not need $\max_{\ell'\in[\ell]}$ in the above equation as the expression achieves its maxima at $\ell'=\ell$, but we keep the definition to avoid any confusion. In the worst case, $2e^{-2b}\le\widetilde{\alpha}_{i,i',\ell,\theta}\le 2e^{2b}$. Therefore, using definition of rank breaking graph $G_{j,a}$, and Equations (48) and (49) we have,

$$\begin{aligned}
\mathbb{E}[M] \;&\succeq\; \gamma e^{-6b}\sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}\frac{2(\kappa_j-p_{j,a})}{\kappa_j(\kappa_j-1)}\sum_{i<i'\in S_j}(e_i-e_{i'})(e_i-e_{i'})^\top\\
&\succeq\; 2\gamma e^{-6b}\sum_{j=1}^{n}\frac{1}{\kappa_j(\kappa_j-1)}\sum_{a=1}^{\ell_j}\lambda_{j,a}(\kappa_j-p_{j,a})\sum_{i<i'\in S_j}(e_i-e_{i'})(e_i-e_{i'})^\top\\
&=\; 2\gamma e^{-6b}L,
\end{aligned} \tag{51}$$

where we used $\gamma\le(1-p_{j,\ell_j}/\kappa_j)^{\alpha_1-2}$ which follows for the definition in (7). (51) follows from the definition of Laplacian L, defined for the comparison graph $\mathcal{H}$ in Definition 9. Using $\lambda_2(L)=(\alpha/(d-1))\sum_{j=1}^n\tau_j\ell_j$ from (30), we get the desired bound $\lambda_2(\mathbb{E}[M])\ge 2\gamma e^{-6b}(\alpha/(d-1))\sum_{j=1}^n\tau_j\ell_j$.

Next we need to upper bound $\|\sum_{j=1}^n\mathbb{E}[(M^j)^2]\|$ to bound the deviation of M from its expectation. To this end, we prove an upper bound on $\mathbb{P}[\sigma_j^{-1}(i)=p_{j,a}\mid i\in S_j]$ in the following lemma.

Lemma 15. *Under the hypotheses of Lemma 14,*

$$\mathbb{P}_\theta\Big[\sigma^{-1}(i)=\ell\Big] \;\le\; \frac{e^{6b}}{\kappa}\left(1-\frac{\ell}{\kappa+\alpha_{i,\ell,\theta}}\right)^{\alpha_{i,\ell,\theta}-1} \;\le\; \frac{e^{6b}}{\kappa-\ell}\,, \tag{52}$$

where $0\le\alpha_{i,\ell,\theta}=\lfloor\widetilde{\alpha}_{i,\ell,\theta}\rfloor$, and $\widetilde{\alpha}_{i,\ell,\theta}$ is,

$$\widetilde{\alpha}_{i,\ell,\theta} \;\equiv\; \min_{\ell'\in[\ell]}\min_{\substack{\Omega\in S\setminus\{i\}\\:|\Omega|=\kappa-\ell'+1}}\left\{\frac{\exp(\theta_i)}{\big(\sum_{j\in\Omega}\exp(\theta_j)\big)/|\Omega|}\right\}. \tag{53}$$

In the worst case, $e^{-2b}\le\widetilde{\alpha}_{i,\ell,\theta}\le e^{2b}$. Note that $\alpha_{i,\ell,\theta}=0$ gives the worst upper bound.

Therefore using Equation (52), for all $i \in [d]$, we have,

$$\mathbb{P}\Big[\sigma_j^{-1}(i) \in \mathcal{P}_j\Big] \leq \min\left\{1, \frac{e^{6b}\ell_j}{\kappa_j - p_{j,\ell_j}}\right\} \;\leq\; \frac{e^{6b}\ell_j}{\max\{\ell_j, \kappa_j - p_{j,\ell_j}\}} \leq \frac{e^{6b}\eta\ell_j}{\kappa_j}\,, \tag{54}$$

where we used η defined in Equation (8). Define a diagonal matrix $D^{(j)} \in \mathcal{S}^d$ and a matrix $A^{(j)} \in \mathcal{S}^d$,

$$A^{(j)}_{ii'} \;\equiv\; \mathbb{I}_{\left\{i,i' \in S_j\right\}} \sum_{a=1}^{\ell_j} \lambda_{j,a} \mathbb{I}_{\left\{(i,i') \in G_{j,a}\right\}}\,, \text{ for all } \; i, i' \in [d]\,, \tag{55}$$

and $D^{(j)}_{ii} = \sum_{i' \neq i} A^{(j)}_{ii'}$. Observe that $M^{(j)} = D^{(j)} - A^{(j)}$. For all $i \in [d]$, we have,

$$\begin{aligned}
D^{(j)}_{ii} \;&=\; \mathbb{I}_{\left\{i \in S_j\right\}} \sum_{i'=1}^{\kappa_j} \mathbb{I}_{\left\{\sigma_j^{-1}(i) = i'\right\}} \sum_{a=1}^{\ell_j} \lambda_{j,a} \mathrm{deg}_{G_{j,a}}(\sigma_j^{-1}(i')) \\
&\leq\; \mathbb{I}_{\left\{i \in S_j\right\}} \left\{ \mathbb{I}_{\left\{\sigma_j^{-1}(i) \in \mathcal{P}_j\right\}} \left(\max_{a \in [\ell_j]} \left\{ \lambda_{j,a}(\kappa_j - p_{j,a}) \right\} + \sum_{a=1}^{\ell_j} \lambda_{j,a} \right) + \mathbb{I}_{\left\{\sigma_j^{-1}(i) \notin \mathcal{P}_j\right\}} \left(\sum_{a=1}^{\ell_j} \lambda_{j,a} \right) \right\} \\
&=\; \mathbb{I}_{\left\{i \in S_j\right\}} \left\{ \mathbb{I}_{\left\{\sigma_j^{-1}(i) \in \mathcal{P}_j\right\}} \delta_{j,1} \;+\; \mathbb{I}_{\left\{\sigma_j^{-1}(i) \notin \mathcal{P}_j\right\}} \delta_{j,2} \right\},
\end{aligned} \tag{56}$$

where the last equality follows from the definition of $\delta_{j,1}$ and $\delta_{j,2}$ in Equation (26). Note that $\max_{i \in [d]}\{D_{ii}\} = \delta_{j,1}$. Using (54) and (56), we have,

$$\mathbb{E}\Big[D^{(j)}_{ii}\Big] \;\leq\; \mathbb{I}_{\left\{i \in S_j\right\}} \left\{ \frac{e^{6b}\eta\ell_j}{\kappa_j} \left(\delta_{j,1} + \frac{\delta_{j,2}\kappa_j}{\eta\ell_j} \right) \right\}. \tag{57}$$

Similarly we have,

$$\mathbb{E}\Big[\big(D^{(j)}_{ii}\big)^2\Big] \;\leq\; \mathbb{I}_{\left\{i \in S_j\right\}} \left\{ \frac{e^{6b}\eta\ell_j}{\kappa_j} \left(\delta_{j,1}^2 + \frac{\delta_{j,2}^2\kappa_j}{\eta\ell_j} \right) \right\} \tag{58}$$

For all $i \in [d]$, we have,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i'=1}^{d} \left(\big(A^{(j)}\big)^2\right)_{ii'}\right] \;&\leq\; \mathbb{E}\left[\left(\sum_{i'=1}^{d} A^{(j)}_{ii'}\right) \max_{i \in [d]} \left\{ \sum_{i'=1}^{d} A^{(j)}_{ii'} \right\}\right] \\
&\leq\; \mathbb{E}\Big[D^{(j)}_{ii}\delta_{j,1}\Big] \\
&\leq\; \mathbb{I}_{\left\{i \in S_j\right\}} \left\{ \frac{e^{6b}\eta\ell_j}{\kappa_j} \left(\delta_{j,1}^2 + \frac{\delta_{j,1}\delta_{j,2}\kappa_j}{\eta\ell_j} \right) \right\}.
\end{aligned} \tag{59}$$

Using (58) and (59), we have, for all $i \in [d]$,

$$\begin{aligned}
&\sum_{i'=1}^{d} \left| \mathbb{E}\Big[\big((M^{(j)})^2 \big)_{ii'} \Big] \right| \\
&= \sum_{i'=1}^{d} \left| \mathbb{E}\Big[\big((D^{(j)})^2 \big)_{ii'} \Big] - \mathbb{E}\Big[\big(D^{(j)} A^{(j)} \big)_{ii'} \Big] - \mathbb{E}\Big[\big(A^{(j)} D^{(j)} \big)_{ii'} \Big] + \mathbb{E}\Big[\big((A^{(j)})^2 \big)_{ii'} \Big] \right| \\
&\leq 2\mathbb{E}\Big[\big(D^{(j)}_{ii} \big)^2 \Big] + \sum_{i'=1}^{d} \left(\mathbb{E}\Big[\delta_{j,1} \big(A^{(j)} \big)_{ii'} \Big] + \mathbb{E}\Big[\big((A^{(j)})^2 \big)_{ii'} \Big] \right) \\
&\leq \mathbb{I}_{\left\{ i \in S_j \right\}} \left\{ \frac{e^{6b} \eta \ell_j}{\kappa_j} \left(4\delta_{j,1}^2 + \frac{2\big(\delta_{j,1}\delta_{j,2} + \delta_{j,2}^2 \big)\kappa_j}{\eta \ell_j} \right) \right\} \\
&= \mathbb{I}_{\left\{ i \in S_j \right\}} \left\{ \frac{e^{6b} \delta \eta \ell_j}{\kappa_j} \right\}, \qquad (60)
\end{aligned}$$

where the last equality follows from the definition of δ, Equation (27).

To bound $\| \sum_{j=1}^{n} \mathbb{E}[(M^{(j)})^2] \|$, we use the fact that for $J \in \mathbb{R}^{d \times d}$, $\|J\| \leq \max_{i \in [d]} \sum_{i'=1}^{d} |J_{ii'}|$. Therefore, we have

$$\begin{aligned}
\left\| \sum_{j=1}^{n} \mathbb{E}\Big[(M^{(j)})^2 \Big] \right\| &\leq e^{6b} \delta \eta \max_{i \in [d]} \left\{ \sum_{j : i \in S_j} \frac{\ell_j}{\kappa_j} \right\} \\
&= \frac{e^{6b} \eta \delta}{\tau} D_{\max} \qquad (61) \\
&= \frac{e^{6b} \eta \delta}{\beta \tau d} \sum_{j=1}^{n} \tau_j \ell_j \ , \qquad (62)
\end{aligned}$$

where (61) follows from the definition of $D_{\max}$ in Equation(28) and (62) follows from the definition of β in (30). Observe that from Equation (56), $\|M^{(j)}\| \leq 2\delta_{j,1} \leq 2\sqrt{\delta}$. Applying matrix Bernstein inequality, we have,

$$\mathbb{P}\Big[\big\| M - \mathbb{E}[M] \big\| \geq t \Big] \leq d \, \exp\left(\frac{-t^2/2}{\frac{e^{6b}\eta\delta}{\beta\tau d} \sum_{j=1}^{n} \tau_j \ell_j + 4\sqrt{\delta} t/3} \right).$$

Therefore, with probability at least $1 - d^{-3}$, we have,

$$\big\| M - \mathbb{E}[M] \big\| \leq 4e^{3b} \sqrt{\frac{\eta\delta \log d}{\beta\tau d} \sum_{j=1}^{n} \tau_j \ell_j} + \frac{64\sqrt{\delta} \log d}{3} \leq 8e^{3b} \sqrt{\frac{\eta\delta \log d}{\beta\tau d} \sum_{j=1}^{n} \tau_j \ell_j} \ , \qquad (63)$$

where the second inequality uses $\sum_{j=1}^{n} \tau_j \ell_j \geq 2^6 (\beta\tau/\eta) d \log d$ which follows from the assumption that $\sum_{j=1}^{n} \tau_j \ell_j \geq 2^6 e^{18b} \frac{\eta\delta}{\tau\gamma^2\alpha^2\beta} d \log d$ and the fact that $\alpha, \beta \leq 1$, $\gamma \leq 1$, $\eta \geq 1$, and $\delta > \tau^2$.

8.2.5 Proof of Lemma 14

Since providing a lower bound on $\mathbb{P}_\theta\big[\sigma^{-1}(i)=\ell,\sigma^{-1}(i')>\ell\big]$ for arbitrary θ is challenging, we construct a new set of parameters $\{\widetilde{\theta}_j\}_{j\in[d]}$ from the original θ. These new parameters are constructed such that it is both easy to compute the probability and also provides a lower bound on the original distribution. We denote the sum of the weights by $W \equiv \sum_{j\in S}\exp(\theta_j)$. We define a new set of parameters $\{\widetilde{\theta}_j\}_{j\in S}$:

$$\widetilde{\theta}_j \;=\; \begin{cases} \log(\widetilde{\alpha}_{i,i',\ell,\theta}/2) & \text{for } j=i \text{ or } i' \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{64}$$

Similarly define $\widetilde{W} \equiv \sum_{j\in S}\exp(\widetilde{\theta}_j) = \kappa - 2 + \widetilde{\alpha}_{i,i',\ell,\theta}$. We have,

$$\begin{aligned}
&\mathbb{P}_\theta\Big[\sigma^{-1}(i)=\ell,\sigma^{-1}(i')>\ell\Big] \\
=\;& \sum_{\substack{j_1\in S\\ j_1\neq i,i'}}\left(\frac{\exp(\theta_{j_1})}{W}\sum_{\substack{j_2\in S\\ j_2\neq i,i',j_1}}\left(\frac{\exp(\theta_{j_2})}{W-\exp(\theta_{j_1})}\cdots\right.\right. \\
&\left.\left.\left(\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,i',\\ j_1,\cdots,j_{\ell-2}}}\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-2}}\exp(\theta_k)}\frac{\exp(\theta_i)}{W-\sum_{k=j_1}^{j_{\ell-1}}\exp(\theta_k)}\right)\cdots\right)\right) \\
=\;& \frac{\exp(\theta_i)}{W}\sum_{\substack{j_1\in S\\ j_1\neq i,i'}}\left(\frac{\exp(\theta_{j_1})}{W-\exp(\theta_{j_1})}\sum_{\substack{j_2\in S\\ j_2\neq i,i',j_1}}\left(\frac{\exp(\theta_{j_2})}{W-\exp(\theta_{j_1})-\exp(\theta_{j_2})}\cdots\right.\right. \\
&\left.\left.\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,i',\\ j_1,\cdots,j_{\ell-2}}}\left(\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-1}}\exp(\theta_k)}\right)\cdots\right)\right)
\end{aligned} \tag{65}$$

Consider the last summation term in the above equation and let $\Omega_\ell = S \setminus \{i, i', j_1, \ldots, j_{\ell-2}\}$. Observe that, $|\Omega_\ell| = \kappa - \ell$ and from equation (50), $\frac{\exp(\theta_i)+\exp(\theta_{i'})}{\sum_{j\in\Omega_\ell}\exp(\theta_j)} \le \frac{\widetilde{\alpha}_{i,i',\ell,\theta}}{\kappa-\ell}$. We have,

$$
\begin{aligned}
& \sum_{j_{\ell-1}\in\Omega_\ell} \frac{\exp(\theta_{j_{\ell-1}})}{W - \sum_{k=j_1}^{j_{\ell-1}} \exp(\theta_k)} \\
= & \sum_{j_{\ell-1}\in\Omega_\ell} \frac{\exp(\theta_{j_{\ell-1}})}{W - \sum_{k=j_1}^{j_{\ell-2}} \exp(\theta_k) - \exp(\theta_{j_{\ell-1}})} \\
\ge & \frac{\sum_{j_{\ell-1}\in\Omega_\ell} \exp(\theta_{j_{\ell-1}})}{W - \sum_{k=j_1}^{j_{\ell-2}} \exp(\theta_k) - \big(\sum_{j_{\ell-1}\in\Omega_\ell} \exp(\theta_{j_{\ell-1}})\big)/|\Omega_\ell|} \qquad (66) \\
= & \frac{\sum_{j_{\ell-1}\in\Omega_\ell} \exp(\theta_{j_{\ell-1}})}{\exp(\theta_i) + \exp(\theta_{i'}) + \sum_{j_{\ell-1}\in\Omega_\ell} \exp(\theta_{j_{\ell-1}}) - \big(\sum_{j_{\ell-1}\in\Omega_\ell} \exp(\theta_{j_{\ell-1}})\big)/|\Omega_\ell|} \\
= & \left(\frac{\exp(\theta_i) + \exp(\theta_{i'})}{\sum_{j_{\ell-1}\in\Omega_\ell} \exp(\theta_{j_{\ell-1}})} + 1 - \frac{1}{\kappa-\ell}\right)^{-1} \\
\ge & \left(\frac{\widetilde{\alpha}_1}{\kappa-\ell} + 1 - \frac{1}{\kappa-\ell}\right)^{-1} \qquad (67) \\
= & \frac{\kappa-\ell}{\widetilde{\alpha}_1 + \kappa - \ell - 1} \\
= & \sum_{j_{\ell-1}\in\Omega_\ell} \frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W} - \sum_{k=j_1}^{j_{\ell-2}} \exp(\widetilde{\theta}_k) - \exp(\widetilde{\theta}_{j_{\ell-1}})}\,, \qquad (68)
\end{aligned}
$$

where (66) follows from the Jensen's inequality and the fact that for any $c > 0$, $0 < x < c$, $\frac{x}{c-x}$ is convex in x. Equation (67) follows from the definition of $\widetilde{\alpha}_{i,i',\ell,\theta}$, (50), and the fact that $|\Omega_\ell| = \kappa - \ell$. Equation (68) uses the definition of $\{\widetilde{\theta}_j\}_{j\in S}$.

Consider $\{\Omega_{\widetilde{\ell}}\}_{2\le\widetilde{\ell}\le\ell-1}$, $|\Omega_{\widetilde{\ell}}| = \kappa - \widetilde{\ell}$, corresponding to the subsequent summation terms in (65). Observe that $\frac{\exp(\theta_i)+\exp(\theta_{i'})}{\sum_{j\in\Omega_{\widetilde{\ell}}}\exp(\theta_j)} \le \widetilde{\alpha}_{i,i',\ell,\theta}/|\Omega_{\widetilde{\ell}}|$. Therefore, each summation term in equation (65) can

be lower bounded by the corresponding term where $\{\theta_j\}_{j\in S}$ is replaced by $\{\widetilde{\theta}_j\}_{j\in S}$. Hence, we have

$$\begin{aligned}
&\mathbb{P}_\theta\Big[\sigma^{-1}(i)=\ell,\sigma^{-1}(i')>\ell\Big]\\
\geq\ & \frac{\exp(\theta_i)}{W}\sum_{\substack{j_1\in S\\ j_1\neq i,i'}}\left(\frac{\exp(\widetilde{\theta}_{j_1})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})}\sum_{\substack{j_2\in S\\ j_2\neq i,i',j_1}}\left(\frac{\exp(\widetilde{\theta}_{j_2})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})-\exp(\widetilde{\theta}_{j_2})}\cdots\right.\right.\\
&\left.\left.\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,i',\\ j_1,\cdots,j_{\ell-2}}}\left(\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-1}}\exp(\widetilde{\theta}_k)}\right)\right)\right)\\
\geq\ & \frac{e^{-4b}\exp(\widetilde{\theta}_i)}{\widetilde{W}}\sum_{\substack{j_1\in S\\ j_1\neq i,i'}}\left(\frac{\exp(\widetilde{\theta}_{j_1})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})}\sum_{\substack{j_2\in S\\ j_2\neq i,i',j_1}}\left(\frac{\exp(\widetilde{\theta}_{j_2})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})-\exp(\widetilde{\theta}_{j_2})}\cdots\right.\right.\\
&\left.\left.\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,i',\\ j_1,\cdots,j_{\ell-2}}}\left(\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-1}}\exp(\widetilde{\theta}_k)}\right)\right)\right)\\
=\ & \big(e^{-4b}\big)\mathbb{P}_{\widetilde{\theta}}\Big[\sigma^{-1}(i)=\ell,\sigma^{-1}(i')>\ell\Big]\,. \qquad (69)
\end{aligned}$$

The second inequality uses $\frac{\exp(\theta_i)}{W}\geq e^{-2b}/\kappa$ and $\frac{\exp(\widetilde{\theta}_i)}{\widetilde{W}}\leq e^{2b}/\kappa$. Observe that $\exp(\widetilde{\theta}_j)=1$ for all $j\neq i,i'$ and $\exp(\widetilde{\theta}_i)+\exp(\widetilde{\theta}_{i'})=\widetilde{\alpha}_{i,i',\ell,\theta}\leq\lceil\widetilde{\alpha}_{i,i',\ell,\theta}\rceil=\alpha_{i,i',\ell,\theta}\geq 1$. Therefore, we have

$$\begin{aligned}
&\mathbb{P}_{\widetilde{\theta}}\Big[\sigma^{-1}(i)=\ell,\sigma^{-1}(i')>\ell\Big]\\
=\ & \binom{\kappa-2}{\ell-1}\frac{(\widetilde{\alpha}_{i,i',\ell,\theta}/2)(\ell-1)!}{(\kappa-2+\widetilde{\alpha}_{i,i',\ell,\theta})(\kappa-2+\widetilde{\alpha}_{i,i',\ell,\theta}-1)\cdots(\kappa-2+\widetilde{\alpha}_{i,i',\ell,\theta}-(\ell-1))}\\
\geq\ & \frac{(\kappa-2)!}{(\kappa-\ell-1)!}\frac{e^{-2b}}{(\kappa+\alpha_{i,i',\ell,\theta}-2)(\kappa+\alpha_{i,i',\ell,\theta}-3)\cdots(\kappa+\alpha_{i,i',\ell,\theta}-(\ell+1))} \qquad (70)\\
=\ & \frac{e^{-2b}(\kappa-\ell+\alpha_{i,i',\ell,\theta}-2)(\kappa-\ell+\alpha_{i,i',\ell,\theta}-3)\cdots(\kappa-\ell)}{(\kappa+\alpha_{i,i',\ell,\theta}-2)(\kappa+\alpha_{i,i',\ell,\theta}-3)\cdots(\kappa-1)}\\
=\ & \frac{e^{-2b}}{(\kappa-1)}\frac{(\kappa-\ell+\alpha_{i,i',\ell,\theta}-2)(\kappa-\ell+\alpha_{i,i',\ell,\theta}-3)\cdots(\kappa-\ell)}{(\kappa+\alpha_{i,i',\ell,\theta}-2)(\kappa+\alpha_{i,i',\ell,\theta}-3)\cdots(\kappa)}\\
\geq\ & \frac{e^{-2b}}{(\kappa-1)}\left(1-\frac{\ell}{\kappa}\right)^{\alpha_{i,i',\ell,\theta}-1}\\
=\ & \frac{e^{-2b}(\kappa-\ell)}{\kappa(\kappa-1)}\left(1-\frac{\ell}{\kappa}\right)^{\alpha_{i,i',\ell,\theta}-2}, \qquad (71)
\end{aligned}$$

where (70) follows from the fact that $\widetilde{\alpha}_{i,i',\ell,\theta}\geq 2e^{-2b}$. Claim (49) follows by combining Equations (69) and (71).

8.2.6 Proof of Lemma 15

Analogous to the proof of Lemma 14, we construct a new set of parameters $\{\widetilde{\theta}_j\}_{j\in[d]}$ from the original θ. We denote the sum of the weights by $W \equiv \sum_{j\in S}\exp(\theta_j)$. We define a new set of parameters $\{\widetilde{\theta}_j\}_{j\in S}$:

$$\widetilde{\theta}_j \;=\; \begin{cases} \log(\widetilde{\alpha}_{i,\ell,\theta}) & \text{for } j = i\ , \\ 0 & \text{otherwise}\ . \end{cases} \tag{72}$$

Similarly define $\widetilde{W} \equiv \sum_{j\in S}\exp(\widetilde{\theta}_j) = \kappa - 1 + \widetilde{\alpha}_{i,\ell,\theta}$. We have,

$$\begin{aligned}
&\mathbb{P}_\theta\Big[\sigma^{-1}(i) = \ell\Big] \\
&= \sum_{\substack{j_1\in S\\ j_1\neq i}}\left(\frac{\exp(\theta_{j_1})}{W}\sum_{\substack{j_2\in S\\ j_2\neq i, j_1}}\left(\frac{\exp(\theta_{j_2})}{W-\exp(\theta_{j_1})}\cdots\left(\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,\\ j_1,\cdots,j_{\ell-2}}}\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-2}}\exp(\theta_k)}\frac{\exp(\theta_i)}{W-\sum_{k=j_1}^{j_{\ell-1}}\exp(\theta_k)}\right)\right)\right) \\
&\le \sum_{\substack{j_1\in S\\ j_1\neq i}}\left(\frac{\exp(\theta_{j_1})}{W}\sum_{\substack{j_2\in S\\ j_2\neq i, j_1}}\left(\frac{\exp(\theta_{j_2})}{W-\exp(\theta_{j_1})}\cdots\left(\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,\\ j_1,\cdots,j_{\ell-2}}}\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-2}}\exp(\theta_k)}\right)\right)\right)\frac{e^{2b}}{\kappa-\ell+1}
\end{aligned} \tag{73}$$

Consider the last summation term in the equation (73), and let $\Omega_\ell = S\setminus\{i, j_1, \ldots, j_{\ell-2}\}$, such that $|\Omega_\ell| = \kappa - \ell + 1$. Observe that from equation (53), $\frac{\exp(\theta_i)}{\sum_{j\in\Omega_\ell}\exp(\theta_j)} \ge \frac{\widetilde{\alpha}_{i,\ell,\theta}}{\kappa-\ell+1}$. We have,

$$\begin{aligned}
\sum_{j_{\ell-1}\in\Omega_\ell}\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-2}}\exp(\theta_k)} &= \frac{\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})}{\exp(\theta_i)+\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})} \\
&\le \left(\frac{\widetilde{\alpha}_{i,\ell,\theta}}{\kappa-\ell+1}+1\right)^{-1} \\
&= \frac{\kappa-\ell+1}{\widetilde{\alpha}_{i,\ell,\theta}+\kappa-\ell+1} \\
&= \sum_{j_{\ell-1}\in\Omega_\ell}\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-2}}\exp(\widetilde{\theta}_k)},
\end{aligned} \tag{74}$$

where (74) follows from the definition of $\{\widetilde{\theta}\}_{j\in S}$.

Consider $\{\Omega_{\widetilde{\ell}}\}_{2\le\widetilde{\ell}\le\ell-1}$, $|\Omega_{\widetilde{\ell}}| = \kappa - \widetilde{\ell} + 1$, corresponding to the subsequent summation terms in (73). Observe that $\frac{\exp(\theta_i)}{\sum_{j\in\Omega_{\widetilde{\ell}}}\exp(\theta_j)} \ge \widetilde{\alpha}_{i,\ell,\theta}/|\Omega_{\widetilde{\ell}}|$. Therefore, each summation term in equation (65) can be lower bounded by the corresponding term where $\{\theta_j\}_{j\in S}$ is replaced by $\{\widetilde{\theta}_j\}_{j\in S}$. Hence, we

have

$$\begin{aligned}
&\mathbb{P}_\theta\Big[\sigma^{-1}(i)=\ell\Big]\\
\leq\ &\sum_{\substack{j_1\in S\\ j_1\neq i}}\left(\frac{\exp(\widetilde{\theta}_{j_1})}{\widetilde{W}}\sum_{\substack{j_2\in S\\ j_2\neq i,j_1}}\left(\frac{\exp(\widetilde{\theta}_{j_2})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})}\cdots\left(\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,\\ j_1,\cdots,j_{\ell-2}}}\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-2}}\exp(\widetilde{\theta}_k)}\right)\right)\right)\frac{e^{2b}}{\kappa-\ell+1}\\
\leq\ &e^{4b}\sum_{\substack{j_1\in S\\ j_1\neq i}}\left(\frac{\exp(\widetilde{\theta}_{j_1})}{\widetilde{W}}\sum_{\substack{j_2\in S\\ j_2\neq i,j_1}}\left(\frac{\exp(\widetilde{\theta}_{j_2})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})}\cdots\right.\right.\\
&\left.\left.\left(\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,\\ j_1,\cdots,j_{\ell-2}}}\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-2}}\exp(\widetilde{\theta}_k)}\frac{\exp(\widetilde{\theta}_i)}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-1}}\exp(\widetilde{\theta}_k)}\right)\right)\right)\\
\leq\ &e^{4b}\mathbb{P}_{\widetilde{\theta}}\Big[\sigma^{-1}(i)=\ell\Big] \qquad (75)
\end{aligned}$$

The second inequality uses $\widetilde{\alpha}_2/(\kappa-\ell+\widetilde{\alpha}_{i,\ell,\theta})\geq e^{-2b}/(\kappa-\ell+1)$. Observe that $\exp(\widetilde{\theta}_j)=1$ for all $j\neq i$ and $\exp(\widetilde{\theta}_i)=\widetilde{\alpha}_{i,\ell,\theta}\geq\lfloor\widetilde{\alpha}_{i,\ell,\theta}\rfloor=\alpha_{i,\ell,\theta}\geq 0$. Therefore, we have

$$\begin{aligned}
\mathbb{P}_{\widetilde{\theta}}\Big[\sigma^{-1}(i)=\ell\Big] &= \binom{\kappa-1}{\ell-1}\frac{\widetilde{\alpha}_{i,\ell,\theta}(\ell-1)!}{(\kappa-1+\widetilde{\alpha}_{i,\ell,\theta})(\kappa-2+\widetilde{\alpha}_{i,\ell,\theta})\cdots(\kappa-\ell+\widetilde{\alpha}_{i,\ell,\theta})}\\
&\leq \frac{(\kappa-1)!}{(\kappa-\ell)!}\frac{e^{2b}}{(\kappa-1+\alpha_{i,\ell,\theta})(\kappa-2+\alpha_{i,\ell,\theta})\cdots(\kappa-\ell+\alpha_{i,\ell,\theta})}\\
&\leq \frac{e^{2b}}{\kappa}\left(1-\frac{\ell}{\kappa+\alpha_{i,\ell,\theta}}\right)^{\alpha_{i,\ell,\theta}-1}, \qquad (76)
\end{aligned}$$

Note that equation (76) holds for all values of $\alpha_{i,\ell,\theta}\geq 0$. Claim 52 follows by combining Equations (75) and (76).

8.3 Proof of Theorem 4

Let $H(\theta)\in\mathcal{S}^d$ be Hessian matrix such that $H_{ii'}(\theta)=\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta_i\partial\theta_{i'}}$. The Fisher information matrix is defined as $I(\theta)=-\mathbb{E}_\theta[H(\theta)]$. Fix any unbiased estimator $\widehat{\theta}$ of $\theta\in\Omega_b$. Since, $\widehat{\theta}\in\mathcal{U}$, $\widehat{\theta}-\theta$ is orthogonal to $\mathbf{1}$. The Cramér-Rao lower bound then implies that $\mathbb{E}[\|\widehat{\theta}-\theta^*\|^2]\geq\sum_{i=2}^d\frac{1}{\lambda_i(I(\theta))}$. Taking the supremum over both sides gives

$$\sup_\theta\mathbb{E}[\|\widehat{\theta}-\theta\|^2] \ \geq\ \sup_\theta\sum_{i=2}^d\frac{1}{\lambda_i(I(\theta))}\geq\sum_{i=2}^d\frac{1}{\lambda_i(I(\mathbf{0}))}\,.$$

The following lemma provides a lower bound on $\mathbb{E}_\theta[H(\mathbf{0})]$, where $\mathbf{0}$ indicates the all-zeros vector.

Lemma 16. *Under the hypotheses of Theorem 4,*

$$\mathbb{E}_\theta[H(\mathbf{0})] \ \succeq\ -\sum_{j=1}^n\frac{2p\log(\kappa_j)^2}{\kappa_j(\kappa_j-1)}\sum_{i'<i\in S_j}(e_i-e_{i'})(e_i-e_{i'})^\top\,. \qquad (77)$$

Observe that $I(\mathbf{0})$ is positive semi-definite. Moreover, $\lambda_1(I(\mathbf{0}))$ is zero and the corresponding eigenvector is the all-ones vector. It follows that

$$\begin{aligned} I(0) &\preceq \sum_{j=1}^{n} \frac{2p\log(\kappa_j)^2}{\kappa_j(\kappa_j-1)} \sum_{i'<i\in S_j} (e_i-e_{i'})(e_i-e_{i'})^\top \\ &\preceq 2p\log(\kappa_{\max})^2 \underbrace{\sum_{j=1}^{n} \frac{1}{\kappa_j(\kappa_j-1)} \sum_{i'<i\in S_j} (e_i-e_{i'})(e_i-e_{i'})^\top}_{=L}, \end{aligned}$$

where L is the Laplacian defined for the comparison graph $\mathcal{H}$, Definition 1, as $\ell_j = 1$ for all $j \in [n]$ in this setting. By Jensen's inequality, we have

$$\sum_{i=2}^{d} \frac{1}{\lambda_i(L)} \geq \frac{(d-1)^2}{\sum_{i=2}^{d}\lambda_i(L)} = \frac{(d-1)^2}{\mathrm{Tr}(L)} = \frac{(d-1)^2}{n}.$$

8.3.1 Proof of Lemma 16

Define $\mathcal{L}_j(\theta)$ for $j \in [n]$ such that $\mathcal{L}(\theta) = \sum_{j=1}^{n}\mathcal{L}_j(\theta)$. Let $H^{(j)}(\theta) \in \mathcal{S}^d$ be the Hessian matrix such that $H^{(j)}_{ii'}(\theta) = \frac{\partial^2\mathcal{L}_j(\theta)}{\partial\theta_i\partial\theta_{i'}}$ for $i, i' \in S_j$. We prove that for all $j \in [n]$,

$$\mathbb{E}_\theta[H^{(j)}(\mathbf{0})] \succeq -\frac{2p\log(\kappa_j)^2}{\kappa_j(\kappa_j-1)} \sum_{i'<i\in S_j} (e_i-e_{i'})(e_i-e_{i'})^\top. \tag{78}$$

In the following, we omit superscript/subscript j for brevity. With a slight abuse of notation, we use $\mathbb{I}_{\{\Omega^{-1}(i)=a\}} = 1$ if item i is ranked at the a-th position in all the orderings $\sigma \in \Omega$. Let $\mathbb{P}[\theta]$ be the likelihood of observing $\Omega^{-1}(p) = i^{(p)}$ and the set Λ (the set of the items that are ranked before the p-th position). We have,

$$\mathbb{P}(\theta) = \sum_{\sigma\in\Omega} \left(\frac{\exp\left(\sum_{m=1}^{p}\theta_{\sigma(m)}\right)}{\prod_{a=1}^{p}\left(\sum_{m'=a}^{\kappa}\exp\left(\theta_{\sigma(m')}\right)\right)} \right). \tag{79}$$

For $i, i' \in S_j$, we have

$$H_{ii'}(\theta) = \frac{1}{\mathbb{P}(\theta)}\frac{\partial^2\mathbb{P}(\theta)}{\partial\theta_i\partial\theta_{i'}} - \frac{\nabla_i\mathbb{P}(\theta)\nabla_{i'}\mathbb{P}(\theta)}{\left(\mathbb{P}(\theta)\right)^2} \tag{80}$$

We claim that at $\theta = \mathbf{0}$,

$$-H_{ii'}(\mathbf{0}) = \begin{cases} C_1 & \text{if } i=i',\ \left\{\Omega^{-1}(i) \geq p\right\} \\ C_2 + A_3^2 - C_3 & \text{if } i=i',\ \left\{\Omega^{-1}(i) < p\right\} \\ -B_1 & \text{if } i\neq i',\ \left\{\Omega^{-1}(i) \geq p,\ \Omega^{-1}(i') \geq p\right\} \\ -B_2 & \text{if } i\neq i',\ \left\{\Omega^{-1}(i) \geq p,\ \Omega^{-1}(i') < p\right\} \\ -B_2 & \text{if } i\neq i',\ \left\{\Omega^{-1}(i) < p,\ \Omega^{-1}(i') \geq p\right\} \\ -(B_3+B_4-A_3^2) & \text{if } i\neq i',\ \left\{\Omega^{-1}(i) < p,\ \Omega^{-1}(i') < p\right\}. \end{cases} \tag{81}$$

where constants $A_3, B_1, B_2, B_3, B_4, C_1, C_2$ and C_3 are defined in Equations (88), (90), (91), (92), (93), (95), (96) and (97) respectively. From this computation of the Hessian, note that we have

$$H(\mathbf{0}) = \sum_{i'<i\in S} (e_i - e_{i'})(e_i - e_{i'})^\top \Big(H_{ii'}(\mathbf{0})\Big) \,. \tag{82}$$

which follows directly from the fact that the diagonal entries are summations of the off-diagonals, i.e. $C_1 = B_1(\kappa-p)+B_2(p-1)$ and $C_2+A_3^2-C_3 = B_2(\kappa-p+1)+(B_3+B_4-A_3^2)(p-2)$. The second equality follows from the fact that $C_2 = B_2(\kappa-p+1)+B_3(p-2)$ and $A_3^2(p-1) = B_4(p-2)+C_3$. Note that since $\theta = \mathbf{0}$, all items are exchangeable. Hence, $\mathbb{E}[H_{ii'}(\mathbf{0})] = \mathbb{E}[H_{ii}(\mathbf{0})]/(\kappa-1)$, and substituting this into (82) and using Equations (81), we get

$$\begin{aligned}
&\mathbb{E}\Big[H(\mathbf{0})\Big] \\
=\;& -\frac{1}{\kappa-1}\Bigg(\mathbb{P}\big[\Omega^{-1}(i)\geq p\big]C_1 + \mathbb{P}\big[\Omega^{-1}(i)<p\big](C_2+A_3^2-C_3)\Bigg)\sum_{i'<i\in S}(e_i-e_{i'})(e_i-e_{i'})^\top \\
\succeq\;& -\frac{1}{\kappa(\kappa-1)}\sum_{i'<i\in S}(e_i-e_{i'})(e_i-e_{i'})^\top \\
& \left((\kappa-p+1)\log\left(\frac{\kappa}{\kappa-p}\right)+(p-1)\left(\log\left(\frac{\kappa}{\kappa-p+1}\right)+\log\left(\frac{\kappa}{\kappa-p+1}\right)^2\right)\right) \qquad (83)\\
\succeq\;& -\frac{2p\log(\kappa)^2}{\kappa(\kappa-1)}\sum_{i'<i\in S}(e_i-e_{i'})(e_i-e_{i'})^\top \,, \qquad (84)
\end{aligned}$$

where (83) uses $\sum_{a=1}^{p}\frac{1}{\kappa-a+1} \leq \log\left(\frac{\kappa}{\kappa-p}\right)$ and $C_3 \geq 0$. Equation (84) follows from the fact that for any $x>0$, $\log(1+x)\leq x$. To prove (81), we have the first order partial derivative of $\mathbb{P}(\theta)$ given by

$$\nabla_i\mathbb{P}(\theta) \;=\; \mathbb{I}_{\{\Omega^{-1}(i)\leq p\}}\mathbb{P}(\theta) - \sum_{\sigma\in\Omega}\left(\frac{\exp\left(\sum_{m=1}^{p}\theta_{\sigma(m)}\right)}{\prod_{a=1}^{p}\left(\sum_{m'=a}^{\kappa}\exp\left(\theta_{\sigma(m')}\right)\right)}\left(\sum_{a=1}^{p}\frac{\mathbb{I}_{\{\sigma^{-1}(i)\geq a\}}\exp(\theta_i)}{\sum_{m'=a}^{\kappa}\exp\left(\theta_{\sigma(m')}\right)}\right)\right) \tag{85}$$

Define constants A_1, A_2 and A_3 such that

$$\begin{aligned}
A_1 \;&\equiv\; \mathbb{P}(\theta)\big|_{\{\theta=\mathbf{0}\}} = \frac{(p-1)!}{\kappa(\kappa-1)\cdots(\kappa-p+1)}, &(86)\\
A_2 \;&\equiv\; \left(\sum_{a=1}^{p}\frac{\exp(\theta_i)}{\sum_{m'=a}^{\kappa}\exp\left(\theta_{\sigma(m')}\right)}\right)\Bigg|_{\{\theta=\mathbf{0}\}} = \left(\frac{1}{\kappa}+\frac{1}{\kappa-1}+\cdots+\frac{1}{\kappa-p+1}\right), &(87)\\
A_3 \;&\equiv\; \left(\frac{(p-1)(p-2)!}{(p-1)!(\kappa)}+\frac{(p-2)(p-2)!}{(p-1)!(\kappa-1)}+\cdots+\frac{(p-2)!}{(p-1)!(\kappa-p+2)}\right). &(88)
\end{aligned}$$

Observe that, for all $i\in[d]$,

$$\nabla_i\mathbb{P}(\theta)\big|_{\{\theta=\mathbf{0}\}} = A_1\Big(\mathbb{I}_{\{\Omega_j^{-1}(i)=p\}}(1-A_2)+\mathbb{I}_{\{\Omega_j^{-1}(i)<p\}}(1-A_3)-\mathbb{I}_{\{\Omega_j^{-1}(i)>p\}}A_2\Big) \;. \tag{89}$$

Further define constants B_1, B_2, B_3 and B_4 such that

$$B_1 \equiv \left(\frac{1}{\kappa^2} + \frac{1}{(\kappa-1)^2} + \cdots + \frac{1}{(\kappa-p+1)^2}\right), \tag{90}$$

$$B_2 \equiv \left(\frac{p-1}{(p-1)\kappa^2} + \frac{p-2}{(p-1)(\kappa-1)^2} + \cdots + \frac{1}{(p-1)(\kappa-p+2)^2}\right), \tag{91}$$

$$B_3 \equiv \left(\frac{(p-1)(p-2)(p-3)!}{(p-1)!\kappa^2} + \frac{(p-2)(p-3)(p-3)!}{(p-1)!(\kappa-1)^2} + \cdots + \frac{2(p-3)!}{(p-1)!(\kappa-p+3)^2}\right), \tag{92}$$

$$B_4 \equiv \frac{(p-3)!}{(p-1)!}\left(\sum_{a,b\in[p-1],b\neq a}\left(\frac{1}{\kappa} + \frac{1}{\kappa-1} + \cdots + \frac{1}{\kappa-a+1}\right)\left(\frac{1}{\kappa} + \frac{1}{\kappa-1} + \cdots + \frac{1}{\kappa-b+1}\right)\right) \tag{93}$$

Observe that,

$$\begin{aligned}
&\left.\frac{\partial^2 \mathbb{P}(\theta)}{\partial\theta_i\partial\theta_{i'}}\right|_{\theta=\mathbf{0}} \\
=\;& \mathbb{I}_{\left\{\Omega^{-1}(i),\Omega^{-1}(i')>p\right\}} A_1\Big((-A_2)(-A_2)+B_1\Big) \\
&+ \Big(\mathbb{I}_{\left\{\Omega^{-1}(i)>p,\Omega^{-1}(i')=p\right\}} + \mathbb{I}_{\left\{\Omega^{-1}(i)=p,\Omega^{-1}(i')>p\right\}}\Big) A_1\Big((-A_2)(1-A_2)+B_1\Big) \\
&+ \Big(\mathbb{I}_{\left\{\Omega^{-1}(i)=p,\Omega^{-1}(i')<p\right\}} + \mathbb{I}_{\left\{\Omega^{-1}(i)<p,\Omega^{-1}(i')=p\right\}}\Big) A_1\Big((1-A_3)+(-A_2)(1-A_3)+B_2\Big) \\
&+ \Big(\mathbb{I}_{\left\{\Omega^{-1}(i)>p,\Omega^{-1}(i')<p\right\}} + \mathbb{I}_{\left\{\Omega^{-1}(i)<p,\Omega^{-1}(i')>p\right\}}\Big) A_1\Big((-A_2)(1-A_3)+B_2\Big) \\
&+ \mathbb{I}_{\left\{\Omega^{-1}(i)<p,\Omega^{-1}(i')<p\right\}} A_1\Big((1-A_3)+(-A_3)+B_4+B_3\Big).
\end{aligned} \tag{94}$$

The claims in (81) are easy to verify by combining Equations (89) and (94) with (80). Also, define constants C_1, C_2 and C_3 such that,

$$C_1 \equiv \left(\frac{\kappa-1}{(\kappa)^2} + \frac{\kappa-2}{(\kappa-1)^2} + \cdots + \frac{\kappa-p}{(\kappa-p+1)^2}\right), \tag{95}$$

$$C_2 \equiv \left(\frac{(p-1)(p-2)!(\kappa-1)}{(p-1)!(\kappa)^2} + \frac{(p-2)(p-2)!(\kappa-2)}{(p-1)!(\kappa-1)^2} + \cdots + \frac{(p-2)!(\kappa-p+1)}{(p-1)!(\kappa-p+2)^2}\right), \tag{96}$$

$$C_3 \equiv \frac{(p-2)!}{(p-1)!}\left(\sum_{a,b\in[p-1],b=a}\left(\frac{1}{\kappa} + \frac{1}{\kappa-1} + \cdots + \frac{1}{\kappa-a+1}\right)\left(\frac{1}{\kappa} + \frac{1}{\kappa-1} + \cdots + \frac{1}{\kappa-b+1}\right)\right), \tag{97}$$

such that,

$$\begin{aligned}
\left.\frac{\partial^2 \mathbb{P}(\theta)}{\partial\theta_i^2}\right|_{\theta=\mathbf{0}} =\;& \mathbb{I}_{\{\Omega^{-1}(i)>p\}} A_1\Big((-A_2)(-A_2)-C_1\Big) + \mathbb{I}_{\{\Omega^{-1}(i)=p\}} A_1\Big((1-A_2)-A_2(1-A_2)-C_1\Big) \\
&+ \mathbb{I}_{\{\Omega^{-1}(i)<p\}} A_1\Big((1-A_3)-A_3-C_2+C_3\Big).
\end{aligned} \tag{98}$$

The claims (81) is easy to verify by combining Equations (89) and (98) with (80).

8.4 Proof of Theorem 5

The proof is analogous to the proof of Theorem 8. It differs primarily in the lower bound that is achieved for the second smallest eigenvalue of the Hessian matrix $H(\theta)$, (35).

Lemma 17. *Under the hypotheses of Theorem 5, if* $\sum_{j=1}^n \ell_j \geq (2^{12}e^{6b}/\beta\alpha^2)d\log d$ *then with probability at least* $1-d^{-3}$,

$$\lambda_2(-H(\theta)) \;\geq\; \frac{\alpha}{2(1+e^{2b})^2}\frac{1}{d-1}\sum_{j=1}^n \ell_j\,. \tag{99}$$

Using Lemma 10 that is derived for the general value of $\lambda_{j,a}$ and $p_{j,a}$, and by substituting $\lambda_{j,a}=1/(\kappa_j-1)$ and $p_{j,a}=a$ for each $j\in[n]$, we get that with probability at least $1-2e^3d^{-3}$,

$$\|\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)\|_2 \;\leq\; \sqrt{16\log d\sum_{j=1}^n \ell_j}\,. \tag{100}$$

Theorem 5 follows from Equations (100), (99) and (40).

8.4.1 Proof of Lemma 17

Define $M^{(j)}\in\mathcal{S}^d$ as

$$M^{(j)} \;=\; \frac{1}{\kappa_j-1}\sum_{i<i'\in S_j}\sum_{a=1}^{\ell_j}\mathbb{I}_{\{(i,i')\,\in\, G_{j,a}\}}(e_i-e_{i'})(e_i-e_{i'})^\top, \tag{101}$$

and let $M=\sum_{j=1}^n M^{(j)}$. Similar to the analysis carried out in the proof of Lemma 11, we have $\lambda_2(-H(\theta))\geq\frac{e^{2b}}{(1+e^{2b})^2}\lambda_2(M)$, when $\lambda_{j,a}=1/(\kappa_j-1)$ is substituted in the Hessian matrix $H(\theta)$, Equation (35). From Weyl's inequality we have that

$$\lambda_2(M) \;\geq \lambda_2(\mathbb{E}[M]) - \|M-\mathbb{E}[M]\|\,. \tag{102}$$

We will show in (107) that $\lambda_2(\mathbb{E}[M])\geq e^{-2b}(\alpha/(d-1))\sum_{j=1}^n\ell_j$ and in (112) that $\|M-\mathbb{E}[M]\|\leq 32e^b\sqrt{\frac{\log d}{\beta d}\sum_{j=1}^n\ell_j}$.

$$\lambda_2(M) \;\geq\; \frac{\alpha e^{-2b}}{d-1}\sum_{j=1}^n\ell_j \;-\; 32e^b\sqrt{\frac{\log d}{\beta d}\sum_{j=1}^n\ell_j} \;\geq\; \frac{\alpha e^{-2b}}{2(d-1)}\sum_{j=1}^n\ell_j\,, \tag{103}$$

where the last inequality follows from the assumption that $\sum_{j=1}^n\ell_j\geq(2^{12}e^{6b}/\beta\alpha^2)d\log d$. This proves the desired claim.

To prove the lower bound on $\lambda_2(\mathbb{E}[M])$, notice that

$$\mathbb{E}[M] \;=\; \sum_{j=1}^n\frac{1}{\kappa_j-1}\sum_{i<i'\in S_j}\mathbb{E}\left[\sum_{a=1}^{\ell_j}\mathbb{I}_{\{(i,i')\in G_{j,a}\}}\Big|(i,i'\in S_j)\right](e_i-e_{i'})(e_i-e_{i'})^\top\,. \tag{104}$$

Using the fact that $p_{j,a} = a$ for each $j \in [n]$, and the definition of rank-breaking graph $G_{j,a}$, we have that

$$\begin{aligned}\mathbb{E}\left[\sum_{a=1}^{\ell_j} \mathbb{I}_{\{(i,i')\in G_{j,a}\}}\Big|(i,i' \in S_j)\right] &= \mathbb{P}\Big[\mathbb{I}_{\{\sigma_j^{-1}(i)\le \ell_j\}} + \mathbb{I}_{\{\sigma_j^{-1}(i')\le \ell_j\}} \ge 1\Big|(i,i' \in S_j)\Big] \\ &\ge \mathbb{P}\Big[(\sigma^{-1}(i) \le \ell_j\Big|(i,i' \in S_j)\Big]. \end{aligned} \tag{105}$$

The following lemma provides a lower bound on $\mathbb{P}[(\sigma^{-1}(i) \le \ell_j|(i,i' \in S_j)]$.

Lemma 18. *Consider a ranking σ over a set of items S of size κ. For any item $i \in S$,*

$$\mathbb{P}[(\sigma^{-1}(i) \le \ell] \ge e^{-2b}\frac{\ell}{\kappa}\,. \tag{106}$$

Therefore, using the fact that $(e_i - e_{i'})(e_i - e_{i'})^\top$ is positive semi-definite, and Equations (104), (105) and (106) we have

$$\mathbb{E}[M] \;\succeq\; e^{-2b}\sum_{j=1}^{n}\frac{\ell_j}{\kappa_j(\kappa_j-1)}\sum_{i<i'\in S_j}(e_i-e_{i'})(e_i-e_{i'})^\top = e^{-2b}L, \tag{107}$$

where L is the Laplacian defined for the comparison graph $\mathcal{H}$, Definition 1. Using $\lambda_2(L) = (\alpha/(d-1))\sum_{j=1}^n \ell_j$ from (5), we get the desired bound $\lambda_2(\mathbb{E}[M]) \ge e^{-2b}(\alpha/(d-1))\sum_{j=1}^n \ell_j$.

For top-ℓ_j rank breaking, $M^{(j)}$ is also given by

$$M^{(j)} = \frac{1}{\kappa_j - 1}\Big((\kappa_j-\ell_j)\mathrm{diag}(e_{\{I_j\}}) + \ell_j\mathrm{diag}(e_{\{S_j\}}) - e_{\{I_j\}}e_{\{S_j\}}^\top - e_{\{S_j\}}e_{\{I_j\}}^\top + e_{\{I_j\}}e_{\{I_j\}}^\top\Big), \tag{108}$$

where $e_{\{S_j\}}, e_{\{I_j\}} \in \mathbb{R}^d$ are zero-one vectors, $e_{\{S_j\}}$ has support corresponding to the set of items S_j and $e_{\{I_j\}}$ has support corresponding to the random top-ℓ_j items in the ranking σ_j. $I_j = \{\sigma_j(1), \sigma_j(2), \cdots, \sigma_j(\ell_j)\}$ for $j \in [n]$. $(M^{(j)})^2$ is given by

$$\begin{aligned}(M^{(j)})^2 = \frac{1}{(\kappa_j-1)^2}\Big((\kappa_j^2-\ell_j^2)\mathrm{diag}(e_{\{I_j\}}) + \ell_j{}^2\mathrm{diag}(e_{\{S_j\}}) - \\ (\kappa_j+\ell_j)(e_{\{I_j\}}e_{\{S_j\}}^\top + e_{\{S_j\}}e_{\{I_j\}}^\top - e_{\{I_j\}}e_{\{I_j\}}^\top) + \ell_j e_{\{S_j\}}e_{\{S_j\}}^\top\Big).\end{aligned}$$

Note that $\mathbb{P}[i \in I_j | i \in S_j] \le \ell_j e^{2b}/\kappa_j$ for all $i \in S_j$. Its proof is similar to the proof of Lemma 18. Therefore, we have $\mathbb{E}[\mathrm{diag}(e_{\{I_j\}})] \preceq \ell_j e^{2b}/\kappa_j \mathrm{diag}(e_{\{\mathbf{1}\}})$. To bound $\|\sum_{j=1}^n \mathbb{E}[(M^{(j)})^2]\|$, we use the fact that for $J \in \mathbb{R}^{d\times d}$, $\|J\| \le \max_{i\in[d]}\sum_{i'=1}^d |J_{ii'}|$. Maximum of row sums of $\mathbb{E}[e_{\{I_j\}}e_{\{I_j\}}^\top]$ is upper

bounded by $\max_{i\in[d]}\left\{\ell_j\mathbb{P}[i\in I_j|i\in S_j]\right\}\leq {\ell_j}^2e^{2b}/\kappa_j$. Therefore using triangle inequality, we have,

$$\begin{aligned}
&\left\|\sum_{j=1}^{n}\mathbb{E}\big[(M^{(j)})^2\big]\right\| \\
\leq\ & \max_{i\in[d]}\left\{\sum_{j:i\in S_j}\frac{1}{(\kappa_j-1)^2}\left(\frac{(\kappa_j^2-{\ell_j}^2)\ell_j e^{2b}}{\kappa_j}+{\ell_j}^2+e^{2b}(\kappa_j+\ell_j)(2\ell_j+{\ell_j}^2/\kappa_j)+\ell_j\kappa_j\right)\right\} \\
\leq\ & \max_{i\in[d]}\left\{\sum_{j:i\in S_j}\frac{\ell_j e^{2b}}{\kappa_j}\left(\frac{(\kappa_j^2-{\ell_j}^2)}{(\kappa_j-1)^2}+\frac{\ell_j\kappa_j}{(\kappa_j-1)^2}+\frac{2(\kappa_j+\ell_j)\kappa_j}{(\kappa_j-1)^2}+\frac{(\kappa_j+\ell_j)\ell_j}{(\kappa_j-1)^2}+\frac{\kappa_j^2}{(\kappa_j-1)^2}\right)\right\} \\
\leq\ & \max_{i\in[d]}\left\{\sum_{j:i\in S_j}\frac{\ell_j e^{2b}}{\kappa_j}\left(\frac{(\kappa_j^2-1)}{(\kappa_j-1)^2}+\frac{\kappa_j(\kappa_j-1)}{(\kappa_j-1)^2}+\frac{4\kappa_j^2}{(\kappa_j-1)^2}+\frac{2\kappa_j(\kappa_j-1)}{(\kappa_j-1)^2}+\frac{\kappa_j^2}{(\kappa_j-1)^2}\right)\right\} \\
\leq\ & \max_{i\in[d]}\left\{\sum_{j:i\in S_j}\frac{\ell_j e^{2b}}{\kappa_j}\Big(3+2+16+4+4\Big)\right\} \qquad (109)\\
\leq\ & 29e^{2b}\max_{i\in[d]}\left\{\sum_{j:i\in S_j}\frac{\ell_j}{\kappa_j}\right\} \\
=\ & 29e^{2b}D_{\max} \qquad (110)\\
=\ & \frac{29e^{2b}}{\beta d}\sum_{j=1}^{n}\ell_j\,, \qquad (111)
\end{aligned}$$

where (109) uses the fact that $\kappa_j\geq 2$ and $1\leq\ell_j\leq\kappa_j-1$ for all $j\in[n]$. (110) follows from the definition of $D_{\max}$, Definition 1 and (111) follows from the Equation (6). Also, note that $\|M_j\|\leq 2$ for all $j\in[n]$. Applying matrix Bernstien inequality, we have,

$$\mathbb{P}\Big[\|M-\mathbb{E}[M]\|\geq t\Big]\leq d\,\exp\left(\frac{-t^2/2}{\frac{29e^{2b}}{\beta d}\sum_{j=1}^{n}\ell_j+4t/3}\right).$$

Therefore, with probability at least $1-d^{-3}$, we have,

$$\|M-\mathbb{E}[M]\|\leq 22e^b\sqrt{\frac{\log d}{\beta d}\sum_{j=1}^{n}\ell_j}+\frac{64\log d}{3}\leq 32e^b\sqrt{\frac{\log d}{\beta d}\sum_{j=1}^{n}\ell_j}\,, \qquad (112)$$

where the second inequality follows from the assumption that $\sum_{j=1}^{n}\ell_j\geq 2^{12}d\log d$ and $\beta\leq 1$.

8.4.2 Proof of Lemma 18

Define $i_{\min}\equiv\arg\min_{i\in S}\theta_i$. We claim the following. For all $i\in S$ and any $1\leq\ell\leq|S|-1$,

$$\mathbb{P}[\sigma^{-1}(i)>\ell]\ \leq\ \mathbb{P}[\sigma^{-1}(i_{\min})>\ell]\ \text{ and }\ \mathbb{P}[\sigma^{-1}(i_{\min})=\ell]\ \geq\ \mathbb{P}[\sigma^{-1}(i_{\min})=1]\,. \qquad (113)$$

Therefore $\mathbb{P}[\sigma^{-1}(i)\leq\ell]\ \geq\ \mathbb{P}[\sigma^{-1}(i_{\min})\leq\ell]$. Using $\mathbb{P}[\sigma^{-1}(i_{\min})=1]>e^{-2b}/\kappa$, we get the desired bound $\mathbb{P}[\sigma^{-1}(i)\leq\ell]>e^{-2b}\ell/\kappa$.

To prove the claim (113), let $\widehat{\sigma}_1^\ell$ denote a ranking of top-ℓ items of the set S and $\mathbb{P}[\widehat{\sigma}_1^\ell]$ be the probability of observing $\widehat{\sigma}_1^\ell$. Let $i \in (\widehat{\sigma}_1^\ell)^{-1}$ denote that $i = (\widehat{\sigma}_1^\ell)^{-1}(j)$ for some $1 \le j \le \ell$. Let

$$\Omega_1 = \Big\{\widehat{\sigma}_1^\ell : i \notin (\widehat{\sigma}_1^\ell)^{-1}, i_{\min} \in (\widehat{\sigma}_1^\ell)^{-1}\Big\} \text{ and } \Omega_2 = \Big\{\widehat{\sigma}_1^\ell : i \in (\widehat{\sigma}_1^\ell)^{-1}, i_{\min} \notin (\widehat{\sigma}_1^\ell)^{-1}\Big\}.$$

We have $\mathbb{P}[\sigma^{-1}(i) > \ell] - \mathbb{P}[\sigma^{-1}(i_{\min}) > \ell] = \sum_{\widehat{\sigma}_1^\ell \in \Omega_1} \mathbb{P}[\widehat{\sigma}_1^\ell] - \sum_{\widehat{\sigma}_1^\ell \in \Omega_2} \mathbb{P}[\widehat{\sigma}_1^\ell]$. Now, take any ranking $\widehat{\sigma}_1^\ell \in \Omega_1$ and construct another ranking $\widetilde{\sigma}_1^\ell$ from $\widehat{\sigma}_1^\ell$ by replacing $i_{\min}$ with i-th item. Observe that $\mathbb{P}[\widehat{\sigma}_1^\ell] \le \mathbb{P}[\widetilde{\sigma}_1^\ell]$ and $\widetilde{\sigma}_1^\ell \in \Omega_2$. Moreover, such a construction gives a bijective mapping between Ω_1 and Ω_2. Hence, the first claim is proved. For the second claim, let

$$\widehat{\Omega}_1 = \Big\{\widehat{\sigma}_1^\ell : (\widehat{\sigma}_1^\ell)^{-1}(i_{\min}) = 1\Big\} \text{ and } \widehat{\Omega}_2 = \Big\{\widehat{\sigma}_1^\ell : (\widehat{\sigma}_1^\ell)^{-1}(i_{\min}) = \ell\Big\}.$$

We have $\mathbb{P}[\sigma^{-1}(i_{\min}) = 1] - \mathbb{P}[\sigma^{-1}(i_{\min}) = \ell] = \sum_{\widehat{\sigma}_1^\ell \in \widehat{\Omega}_1} \mathbb{P}[\widehat{\sigma}_1^\ell] - \sum_{\widehat{\sigma}_1^\ell \in \widehat{\Omega}_2} \mathbb{P}[\widehat{\sigma}_1^\ell]$. Now, take any ranking $\widehat{\sigma}_1^\ell \in \widehat{\Omega}_1$ and construct another ranking $\widetilde{\sigma}_1^\ell$ from $\widehat{\sigma}_1^\ell$ by swapping items at 1st position and ℓ-th position. Observe that $\mathbb{P}[\widehat{\sigma}_1^\ell] \le \mathbb{P}[\widetilde{\sigma}_1^\ell]$ and $\widetilde{\sigma}_1^\ell \in \widehat{\Omega}_2$. Moreover, such a construction gives a bijective mapping between $\widehat{\Omega}_1$ and $\widehat{\Omega}_2$. Hence, the claim is proved.

8.5 Proof of Theorem 6

The first order partial derivative of $\mathcal{L}(\theta)$, Equation (15), is given by

$$\begin{aligned}
&\nabla_i \mathcal{L}(\theta) \\
&= \sum_{j: i \in S_j} \sum_{m=1}^{\ell_j} \mathbb{I}_{\{\sigma_j^{-1}(i) \ge m\}} \Big[\mathbb{I}_{\{\sigma_j(m) = i\}} - \frac{\exp(\theta_i)}{\exp(\theta_{\sigma_j(m)}) + \exp(\theta_{\sigma_j(m+1)}) + \cdots + \exp(\theta_{\sigma_j(\kappa_j)})}\Big], \ \forall i \in [d]
\end{aligned}$$

and the Hessian matrix $H(\theta) \in \mathcal{S}^d$ with $H_{ii'}(\theta) = \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_i \partial \theta_{i'}}$ is given by

$$H(\theta) = -\sum_{j=1}^{n} \sum_{i<i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \sum_{m=1}^{\ell_j} \frac{\exp(\theta_i + \theta_{i'}) \mathbb{I}_{\{\sigma_j^{-1}(i), \sigma_j^{-1}(i') \ge m\}}}{[\exp(\theta_{\sigma_j(m)}) + \exp(\theta_{\sigma_j(m+1)}) + \cdots + \exp(\theta_{\sigma_j(\kappa_j)})]^2}. \tag{114}$$

It follows from the definition that $-H(\theta)$ is positive semi-definite for any $\theta \in \mathbb{R}^n$.

The Fisher information matrix is defined as $I(\theta) = -\mathbb{E}_\theta[H(\theta)]$ and given by

$$I(\theta) = \sum_{j=1}^{n} \sum_{i<i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \sum_{m=1}^{\ell_j} \mathbb{E}\left[\frac{\mathbb{I}_{\{\sigma_j^{-1}(i), \sigma_j^{-1}(i') \ge m\}}}{[\exp(\theta_{\sigma_j(m)}) + \cdots + \exp(\theta_{\sigma_j(\kappa_j)})]^2}\right] \exp(\theta_i + \theta_{i'}).$$

Since $-H(\theta)$ is positive semi-definite, it follows that $I(\theta)$ is positive semi-definite. Moreover, $\lambda_1(I(\theta))$ is zero and the corresponding eigenvector is the all-ones vector. Fix any unbiased estimator $\widehat{\theta}$ of $\theta \in \Omega_b$. Since, $\widehat{\theta} \in \mathcal{U}$, $\widehat{\theta} - \theta$ is orthogonal to $\mathbf{1}$. The Cramér-Rao lower bound then implies that $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2] \ge \sum_{i=2}^{d} \frac{1}{\lambda_i(I(\theta))}$. Taking the supremum over both sides gives

$$\sup_\theta \mathbb{E}[\|\widehat{\theta} - \theta\|^2] \ge \sup_\theta \sum_{i=2}^{d} \frac{1}{\lambda_i(I(\theta))} \ge \sum_{i=2}^{d} \frac{1}{\lambda_i(I(\mathbf{0}))} \ .$$

If θ equals the all-zero vector, then

$$\mathbb{P}_\theta[\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq m] = \frac{\binom{\kappa_j - m + 1}{2}}{\binom{\kappa_j}{2}} = \frac{(\kappa_j - m + 1)(\kappa_j - m)}{\kappa_j(\kappa_j - 1)}.$$

It follows from the definition that

$$\begin{aligned} I(0) &= \sum_{j=1}^{n} \sum_{i<i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \sum_{m=1}^{\ell_j} \frac{(\kappa_j - m)}{\kappa_j(\kappa_j - 1)(\kappa_j - m + 1)} \\ &\preceq \ell\Big(1 - \frac{1}{\ell_j} \sum_{m=1}^{\ell_j} \frac{1}{\kappa_{\max} - m + 1}\Big) \underbrace{\sum_{j=1}^{n} \frac{1}{\kappa_j(\kappa_j - 1)} \sum_{i<i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top}_{=L}, \end{aligned}$$

where L is the Laplacian defined for the comparison graph $\mathcal{H}$, Definition 1. By Jensen's inequality, we have

$$\sum_{i=2}^{d} \frac{1}{\lambda_i(L)} \geq \frac{(d-1)^2}{\sum_{i=2}^{d} \lambda_i(L)} = \frac{(d-1)^2}{\mathrm{Tr}(L)} = \frac{(d-1)^2}{n}.$$

8.6 Proof of Theorem 7

We prove a slightly more general result that implies the desired theorem. For $\ell \geq 4$, we can choose $\beta_1 = 1/2$. Then, the condition that $\gamma_{\beta_1} \leq 1$ implies $\widetilde{d} \leq (\ell/2 + 1)(d-2)/(\kappa - 2)$, which implies $\widetilde{d} \leq \ell d/(2\kappa)$. With the choice of $\widetilde{d} = \ell d/(2\kappa)$, this implies Theorem 7.

Theorem 19. *Under the bottom-ℓ separators scenario and the PL model, n partial orderings are sampled over d items parametrized by $\theta^* \in \Omega_b$. For any β_1 with $0 \leq \beta_1 \leq \frac{\ell-2}{\ell}$, define*

$$\gamma_{\beta_1} \equiv \frac{\widetilde{d}(\kappa - 2)}{(\lfloor \ell\beta_1 \rfloor + 1)(d-2)}, \tag{115}$$

and for $\gamma_{\beta_1} \leq 1$,

$$\chi_{\beta_1} \equiv \big(1 - \lfloor \ell\beta_1 \rfloor / \ell\big)^2 \left(1 - \exp\left(-\frac{(\lfloor \ell\beta_1 \rfloor + 1)^2 (1 - \gamma_{\beta_1})^2}{2(\kappa - 2)}\right)\right). \tag{116}$$

If

$$n\ell \geq \left(\frac{2^{12} e^{8b}}{\chi_{\beta_1}^2} \frac{d^2}{\widetilde{d}^2} \frac{\kappa}{\ell}\right) d \log d\,, \tag{117}$$

then the rank-breaking *estimator in* (18) *achieves*

$$\frac{1}{\sqrt{\widetilde{d}}} \big\| \widehat{\widetilde{\theta}} - \widetilde{\theta}^* \big\|_2 \leq \frac{32\sqrt{2}(1 + e^{4b})^2}{\chi_{\beta_1}} \frac{d^{3/2}}{\widetilde{d}^{3/2}} \sqrt{\frac{d \log d}{n\ell}}\,, \tag{118}$$

with probability at least $1 - 3e^3 d^{-3}$.

Proof is very similar to the proof of Theorem 8. It mainly differs in the lower bound that is achieved for the second smallest eigenvalue of the Hessian matrix $H(\widetilde{\theta})$ of $\mathcal{L}_{\mathrm{RB}}(\widetilde{\theta})$, Equation (17). Equation (17) can be rewritten as

$$\mathcal{L}_{\mathrm{RB}}(\widetilde{\theta}) = \sum_{j=1}^{n}\sum_{a=1}^{\ell}\sum_{\substack{i<i'\in S_j\\ :i,i'\in[\widetilde{d}]}} \mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}}\lambda_{j,a}\Big(\widetilde{\theta}_i\mathbb{I}_{\left\{\sigma_j^{-1}(i)<\sigma_j^{-1}(i')\right\}} + \widetilde{\theta}_{i'}\mathbb{I}_{\left\{\sigma_j^{-1}(i)>\sigma_j^{-1}(i')\right\}} - \log\left(e^{\widetilde{\theta}_i}+e^{\widetilde{\theta}_{i'}}\right)\Big)\,, \tag{119}$$

where $(i,i')\in G_{j,a}$ implies either (i,i') or (i',i) belong to $E_{j,a}$. The Hessian matrix $H(\widetilde{\theta})\in\mathcal{S}^{\widetilde{d}}$ with $H_{ii'}(\widetilde{\theta}) = \frac{\partial^2\mathcal{L}_{\mathrm{RB}}(\widetilde{\theta})}{\partial\widetilde{\theta}_i\partial\widetilde{\theta}_{i'}}$ is given by

$$H(\widetilde{\theta}) = -\sum_{j=1}^{n}\sum_{a=1}^{\ell}\sum_{\substack{i<i'\in S_j:\\ i,i'\in[\widetilde{d}]}} \mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}}\left((\widetilde{e}_i-\widetilde{e}_{i'})(\widetilde{e}_i-\widetilde{e}_{i'})^\top\frac{\exp(\widetilde{\theta}_i+\widetilde{\theta}_{i'})}{[\exp(\widetilde{\theta}_i)+\exp(\widetilde{\theta}_{i'})]^2}\right). \tag{120}$$

The following lemma gives a lower bound for $\lambda_2(-H(\widetilde{\theta}))$.

Lemma 20. *Under the hypothesis of Theorem 19, with probability at least* $1-d^{-3}$,

$$\lambda_2(-H(\widetilde{\theta})) \geq \frac{\chi_{\beta_1}}{8(1+e^{4b})^2}\frac{n\widetilde{d}\ell^2}{d^2}\,. \tag{121}$$

Observe that although $\widetilde{\theta}^*\in\mathbb{R}^{\widetilde{d}}$, Lemma 10 can be directly applied to upper bound $\|\nabla\mathcal{L}_{\mathrm{RB}}(\widetilde{\theta}^*)\|_2$. It might be possible to tighten the upper bound, given that $\widetilde{d}\leq d$. However, for $\ell\ll\kappa$, for the smallest preference score item, $i_{\min}\equiv\arg\min_{i\in[d]}\widetilde{\theta}_i^*$, the upper bound $\mathbb{P}[\sigma^{-1}(i_{\min})>\kappa-\ell]\leq 1$ is tight upto constant factor (Lemma 15). Substituting $\lambda_{j,a}=1$ and $p_{j,a}=\kappa-\ell+a$ for each $j\in[n]$, $a\in[\ell]$, in Lemma 10, we have that with probability at least $1-2e^3d^{-3}$,

$$\|\nabla\mathcal{L}_{\mathrm{RB}}(\widetilde{\theta}^*)\|_2 \;\leq\; (\ell-1)\sqrt{8n\ell\log d}. \tag{122}$$

Theorem 19 follows from Equations (40), (121) and (122).

8.6.1 Proof of Lemma 20

Define $\widetilde{M}^{(j)}\in\mathcal{S}^{\widetilde{d}}$,

$$\widetilde{M}^{(j)} = \sum_{i<i'\in S_j:i,i'\in[\widetilde{d}]}\sum_{a=1}^{\ell}\mathbb{I}_{\{(i,i')\in G_{j,a}\}}(\widetilde{e}_i-\widetilde{e}_{i'})(\widetilde{e}_i-\widetilde{e}_{i'})^\top, \tag{123}$$

and let $\widetilde{M}=\sum_{j=1}^n\widetilde{M}^{(j)}$. Similar to the analysis in Lemma 11, we have $\lambda_2(-H(\widetilde{\theta}))\geq\frac{e^{4b}}{(1+e^{4b})^2}\lambda_2(\widetilde{M})$. Note that we have e^{4b} instead of e^{2b} as $\widetilde{\theta}\in\widetilde{\Omega}_{2b}$. We will show a lower bound on $\lambda_2(\mathbb{E}[\widetilde{M}])$ in (129) and an upper bound on $\|\widetilde{M}-\mathbb{E}[\widetilde{M}]\|$ in (133). Therefore using $\lambda_2(\widetilde{M})\geq\lambda_2(\mathbb{E}[\widetilde{M}])-\|\widetilde{M}-\mathbb{E}[\widetilde{M}]\|$,

$$\lambda_2(\widetilde{M}) \;\geq\; \frac{e^{-4b}}{4}\underbrace{(1-\beta_1)^2\left(1-\exp\left(-\frac{(\lfloor\ell\beta_1\rfloor+1)^2(1-\gamma_{\beta_1})^2}{2(\kappa-2)}\right)\right)}_{\equiv\chi_{\beta_1}}\frac{n\widetilde{d}\ell^2}{d^2} - 8\ell\sqrt{\frac{n\kappa\log d}{d}}\,. \tag{124}$$

The desired claim follows from the assumption that $n\ell \geq \big(\frac{2^{12}e^{8b}}{\chi^2_{\beta_1}}\frac{d^2}{\widetilde{d}^2}\frac{\kappa}{\ell}\big)d\log d$, where χ_{β_1} is defined in (117). To prove the lower bound on $\lambda_2(\mathbb{E}[\widetilde{M}])$, notice that

$$\mathbb{E}\big[\widetilde{M}\big] \;=\; \sum_{j=1}^{n}\sum_{i<i'\in[\widetilde{d}]}\mathbb{E}\Bigg[\sum_{a=1}^{\ell}\mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}}\Big|(i,i'\in S_j)\Bigg]\mathbb{P}\Big[i,i'\in S_j\Big](\widetilde{e}_i-\widetilde{e}_{i'})(\widetilde{e}_i-\widetilde{e}_{i'})^\top\,. \quad (125)$$

Since the sets S_j are chosen uniformly at random, $\mathbb{P}[i,i'\in S_j]=\kappa(\kappa-1)/d(d-1)$. Using the fact that $p_{j,a}=\kappa-\ell+a$ for each $j\in[n]$, and the definition of rank breaking graph $G_{j,a}$, we have that

$$\mathbb{E}\Bigg[\sum_{a=1}^{\ell}\mathbb{I}_{\left\{(i,i')\in G_{j,a}\right\}}\Big|(i,i'\in S_j)\Bigg]=\mathbb{P}\Big[\big(\sigma_j^{-1}(i),\sigma_j^{-1}(i')>\kappa-\ell\big)\Big|(i,i'\in S_j)\Big]\,. \quad (126)$$

The following lemma provides a lower bound on $\mathbb{P}[(\sigma_j^{-1}(i),\sigma_j^{-1}(i'))>\kappa-\ell|(i,i'\in S_j)]$.

Lemma 21. *Under the hypotheses of Theorem 19, for any two items $i,i'\in[\widetilde{d}]$, the following holds:*

$$\mathbb{P}\Big[\sigma^{-1}(i),\sigma^{-1}(i')>\kappa-\ell\;\Big|\;i,i'\in S\Big]\;\geq\;\frac{e^{-4b}(1-\beta_1)^2(1-\exp(-\eta_{\beta_1}(1-\gamma_{\beta_1})^2))}{2}\frac{\ell^2}{\kappa^2}\,, \quad (127)$$

where $\gamma_{\beta_1}\equiv\widetilde{d}(\kappa-2)/(\lfloor\ell\beta_1\rfloor+1)(d-2)$ and $\eta_{\beta_1}\equiv(\lfloor\ell\beta_1\rfloor+1)^2/2(\kappa-2)$.

Therefore, using Equations (125), (126) and (127) we have,

$$\mathbb{E}\big[\widetilde{M}\big]\;\succeq\;\frac{e^{-4b}(1-\beta_1)^2(1-\exp(-\eta_{\beta_1}(1-\gamma_{\beta_1})^2))}{2}\frac{\ell^2}{\kappa^2}\frac{\kappa(\kappa-1)}{d(d-1)}\sum_{j=1}^{n}\sum_{i<i'\in[\widetilde{d}]}(\widetilde{e}_i-\widetilde{e}_{i'})(\widetilde{e}_i-\widetilde{e}_{i'})^\top \quad (128)$$

Define $\widetilde{L}=\sum_{j=1}^{n}\sum_{i<i'\in[\widetilde{d}]}(\widetilde{e}_i-\widetilde{e}_{i'})(\widetilde{e}_i-\widetilde{e}_{i'})^\top$. We have, $\lambda_1(\widetilde{L})=0$ and $\lambda_2(\widetilde{L})=\lambda_3(\widetilde{L})=\cdots=\lambda_{\widetilde{d}}(\widetilde{L})$. Therefore, using $\lambda_2(\widetilde{L})=\mathrm{Tr}(\widetilde{L})/(\widetilde{d}-1)=n\widetilde{d}$. Using the fact that $\mathbb{E}[\widetilde{M}]$ and $\widetilde{L}$ are symmetric matrices, we have,

$$\lambda_2(\mathbb{E}\big[\widetilde{M}\big])\geq\frac{e^{-4b}(1-\beta_1)^2(1-\exp(-\eta_{\beta_1}(1-\gamma_{\beta_1})^2))}{4}\frac{n\widetilde{d}\ell^2}{d^2}. \quad (129)$$

To get an upper bound on $\|\widetilde{M}-\mathbb{E}[\widetilde{M}]\|$, notice that $\widetilde{M}^{(j)}$ is also given by,

$$\widetilde{M}^{(j)}\;=\;\ell\,\mathrm{diag}(\widetilde{e}_{\{I_j\}})-\widetilde{e}_{\{I_j\}}\widetilde{e}_{\{I_j\}}^\top\,, \quad (130)$$

where $\widetilde{e}_{\{I_j\}}\in\mathbb{R}^{\widetilde{d}}$ is a zero-one vector, with support corresponding to the bottom-ℓ subset of items in the ranking σ_j. $I_j=\{\sigma_j(\kappa-\ell+1),\cdots,\sigma_j(\kappa)\}$ for $j\in[n]$. $(\widetilde{M}^{(j)})^2$ is given by

$$(\widetilde{M}^{(j)})^2\;=\;\ell^2\,\mathrm{diag}(\widetilde{e}_{\{I_j\}})-\ell\,\widetilde{e}_{\{I_j\}}\widetilde{e}_{\{I_j\}}^\top\,. \quad (131)$$

Using the fact that sets $\{S_j\}_{j\in[n]}$ are chosen uniformly at random and $\mathbb{P}[i\in\mathbb{I}_j|i\in S_j]\leq 1$, we have $\mathbb{E}[\mathrm{diag}(\widetilde{e}_{\{I_j\}})]\preceq(\kappa/d)\mathrm{diag}(\widetilde{e}_{\{\mathbf{1}\}})$. Maximum of row sums of $\mathbb{E}\big[\widetilde{e}_{\{I_j\}}\widetilde{e}_{\{I_j\}}^\top\big]$ is upper bounded by

$\ell\kappa/d$. Therefore, from triangle inequality we have $\|\sum_{j=1}^{n}\mathbb{E}[(\widetilde{M}^{(j)})^2]\| \le 2n\ell^2\kappa/d$. Also, note that $\|\widetilde{M}^{(j)}\| \le 2\ell$ for all $j \in [n]$. Applying matrix Bernstien inequality, we have that

$$\mathbb{P}\Big[\|\widetilde{M} - \mathbb{E}[\widetilde{M}]\| \ge t\Big] \le d\,\exp\Big(\frac{-t^2/2}{2n\ell^2\kappa/d + 4\ell t/3}\Big). \tag{132}$$

Therefore, with probability at least $1 - d^{-3}$, we have,

$$\|\widetilde{M} - \mathbb{E}[\widetilde{M}]\| \le 4\ell\sqrt{\frac{2n\kappa\log d}{d}} + \frac{64\ell\log d}{3} \le 8\ell\sqrt{\frac{n\kappa\log d}{d}}\,, \tag{133}$$

where the second inequality follows from the assumption that $n\ell \ge 2^{12}d\log d$.

8.6.2 Proof of Lemma 21

Without loss of generality, assume that $i' < i$, i.e., $\widetilde{\theta}^*_{i'} \le \widetilde{\theta}^*_i$. Define Ω such that $\Omega = \{j : j \in S, j \ne i, i'\}$. For any $\beta_1 \in [0, (\ell-2)/\ell]$, define event E_{β_1} that occurs if in the randomly chosen set S there are at most $\lfloor \ell\beta_1 \rfloor$ items that have preference scores less than $\widetilde{\theta}^*_i$, i.e.,

$$E_{\beta_1} \;\equiv\; \Big\{ \textstyle\sum_{j\in\Omega}\mathbb{I}_{\{\widetilde{\theta}^*_i > \widetilde{\theta}^*_j\}} \le \lfloor \ell\beta_1 \rfloor \Big\}\,. \tag{134}$$

We have,

$$\begin{aligned} & \mathbb{P}\Big[\sigma^{-1}(i), \sigma^{-1}(i') \;>\; \kappa - \ell \;\Big|\; i, i' \in S\Big] \\ > \;& \mathbb{P}\Big[\sigma^{-1}(i), \sigma^{-1}(i') > \kappa - \ell \;\Big|\; i, i' \in S; E_{\beta_1}\Big]\mathbb{P}\Big[E_{\beta_1} \;\Big|\; i, i' \in S\Big] \end{aligned} \tag{135}$$

The following lemma provides a lower bound on $\mathbb{P}[\sigma^{-1}(i), \sigma^{-1}(i') > \kappa - \ell \mid i, i' \in S; E_{\beta_1}]$.

Lemma 22. *Under the hypotheses of Lemma 21,*

$$\mathbb{P}\Big[\sigma^{-1}(i), \sigma^{-1}(i') \;>\; \kappa - \ell \;\Big|\; i, i' \in S; E_{\beta_1}\Big] \;\ge\; \frac{e^{-4b}(1 - \lfloor \ell\beta_1 \rfloor/\ell)^2}{2}\frac{\ell^2}{\kappa^2}\,. \tag{136}$$

Next, we provide a lower bound on $\mathbb{P}[E_{\beta_1} \mid i, i' \in S]$. Fix i, i' such that $i, i' \in S$. Selecting a set uniformly at random is probabilistically equivalent to selecting items one at a time uniformly at random without replacement. Without loss of generality, assume that i, i' are the 1st and 2nd pick. Define Bernoulli random variables $Y_{j'}$ for $3 \le j' \le \kappa$ corresponding to the outcome of the j'-th random pick from the set of $(d - j' - 1)$ items to generate the set Ω such that $Y_{j'} = 1$ if and only if $\widetilde{\theta}^*_i > \widetilde{\theta}^*_{j'}$.

Recall that $\gamma_{\beta_1} \equiv \widetilde{d}(\kappa-2)/(\lfloor \ell\beta_1 \rfloor + 1)(d-2)$ and $\eta_{\beta_1} \equiv (\lfloor \ell\beta_1 \rfloor + 1)^2/2(\kappa-2)$. Construct Doob's martingale $(Z_2, \cdots, Z_\kappa)$ from $\{Y_{k'}\}_{3\le k'\le\kappa}$ such that $Z_{j'} = \mathbb{E}[\sum_{k'=3}^{\kappa} Y_{k'} \mid Y_3, \cdots, Y_{j'}]$, for $2 \le j' \le \kappa$. Observe that, $Z_2 = \mathbb{E}[\sum_{k'=3}^{\kappa} Y_{k'}] \le \frac{(i-2)(\kappa-2)}{d-2} \le \gamma_{\beta_1}(\lfloor \ell\beta_1 \rfloor + 1)$, where the last inequality follows

from the assumption that $i \leq \widetilde{d}$. Also, $|Z_{j'} - Z_{j'-1}| \leq 1$ for each j'. Therefore, we have

$$\begin{aligned}
\mathbb{P}\Big[\textstyle\sum_{j\in\Omega} \mathbb{I}_{\{\widetilde{\theta}^*_i > \widetilde{\theta}^*_j\}} \leq \lfloor \ell\beta_1 \rfloor \Big] &= \mathbb{P}\Big[\textstyle\sum_{j'=3}^{\kappa} Y_{j'} \leq \lfloor \ell\beta_1 \rfloor \Big] \\
&= 1 - \mathbb{P}\Big[\textstyle\sum_{j'=3}^{\kappa} Y_{j'} \geq \lfloor \ell\beta_1 \rfloor + 1 \Big] \\
&\geq 1 - \mathbb{P}\Big[Z_{\kappa-2} - Z_2 \geq (\ell\beta_1 + 1) - \gamma(\lfloor \ell\beta_1 \rfloor + 1) \Big] \\
&\geq 1 - \exp\Big(- \frac{(\lfloor \ell\beta_1 \rfloor + 1)^2 (1-\gamma_1)^2}{2(\kappa - 2)} \Big) \\
&= 1 - \exp\Big(- \eta_{\beta_1} (1 - \gamma_{\beta_1})^2 \Big), \qquad (137)
\end{aligned}$$

where the inequality follows from the Azuma-Hoeffding bound. Since, the above inequality is true for any fixed $i, i' \in S$, for random indices i, i' we have $\mathbb{P}[E_{\beta_1} \mid i, i' \in S] \geq 1 - \exp(-\eta_{\beta_1}(1-\gamma_{\beta_1})^2)$. Claim (127) follows by combining Equations (135), (136) and (137).

8.6.3 Proof of Lemma 22

Without loss of generality, assume that $i' < i$, i.e., $\widetilde{\theta}^*_{i'} \leq \widetilde{\theta}^*_i$. Define $\Omega = \{j : j \in S, j \neq i, i'\}$, and event $E_{\beta_1} = \{i, i' \in S; \sum_{j\in\Omega} \mathbb{I}_{\{\widetilde{\theta}^*_i > \widetilde{\theta}^*_j\}} \leq \lfloor \ell\beta_1 \rfloor\}$. Since set S is chosen randomly, i, i' and $j \in \Omega$ are random. Throughout this section, we condition on the random indices i, i' and the set Ω such that event E_{β_1} holds. To get a lower bound on $\mathbb{P}[\sigma^{-1}(i), \sigma^{-1}(i') > \kappa - \ell]$, define independent exponential random variables $X_j \sim \exp(e^{\widetilde{\theta}^*_j})$ for $j \in S$. Observe that given event E_{β_1} holds, there exists a set $\Omega_1 \subseteq \Omega$ such that

$$\Omega_1 = \Big\{ j \in S : \widetilde{\theta}^*_i \leq \widetilde{\theta}^*_j \Big\}, \qquad (138)$$

and $|\Omega_1| = \kappa - \lfloor \ell\beta_1 \rfloor - 2$. In fact there can be many such sets, and for the purpose of the proof we can choose one such set arbitrarily. Note that $\lfloor \ell\beta_1 \rfloor + 2 \leq \ell$ by assumption on β_1, so $|\Omega_1| \geq \kappa - \ell$. From the Random Utility Model (RUM) interpretation of the PL model, we know that the PL model is equivalent to ordering the items as per *random cost* of each item drawn from exponential random variable with mean $e^{\widetilde{\theta}^*_i}$. That is, we rank items according to X_j's such that the lower cost items are ranked higher. From this interpretation, we have that

$$\begin{aligned}
\mathbb{P}\Big[\sigma^{-1}(i), \sigma^{-1}(i') > \kappa - \ell \Big] &= \mathbb{P}\Big[\sum_{j\in\Omega} \mathbb{I}_{\left\{ \min\{X_i, X_{i'}\} > X_j \right\}} \geq \kappa - \ell \Big] \\
&> \mathbb{P}\Big[\sum_{j'\in\Omega_1} \mathbb{I}_{\left\{ \min\{X_i, X_{i'}\} > X_{j'} \right\}} \geq \kappa - \ell \Big] \qquad (139)
\end{aligned}$$

The above inequality follows from the fact that $\Omega_1 \subseteq \Omega$ and $|\Omega_1| \geq \kappa - \ell$. It excludes some of the rankings over the items of the set S that constitute the event $\{\sigma^{-1}(i), \sigma^{-1}(i') > \kappa - \ell\}$. Define $\Omega_2 = \{\Omega_1, i, i'\}$. Observe that items i, i' have the least preference scores among all the items in the set Ω_2. Therefore, the term in Equation (139) is the probability of the least two preference score items in the set Ω_2, that is of size $(\kappa - \lfloor \ell\beta_1 \rfloor)$, being ranked in bottom $(\ell - \lfloor \ell\beta_1 \rfloor)$ positions.

The following lemma shows that the probability of the least two preference score items in a set being ranked at any two positions is lower bounded by their probability of being ranked at 1st and 2nd position.

Lemma 23. *Consider a set of items S and a ranking σ over it. Define $i_{\min_1} \equiv \arg\min_{i\in S} \theta_i$, $i_{\min_2} \equiv \arg\min_{i\in S\setminus i_{min_1}} \theta_i$. For all $1 \le i_1, i_2 \le |S|$, $i_1 \neq i_2$, following holds:*

$$\mathbb{P}\Big[\sigma^{-1}(i_{\min_1}) = i_1, \sigma^{-1}(i_{\min_2}) = i_2\Big] \ge \mathbb{P}\Big[\sigma^{-1}(i_{\min_1}) = 1, \sigma^{-1}(i_{\min_2}) = 2\Big]. \tag{140}$$

Using the fact that $i' = \arg\min_{j\in\Omega_2} \widetilde{\theta}^*_j$, $i = \arg\min_{j\in\Omega_2\setminus i'} \widetilde{\theta}^*_j$, for all $1 \le i_1, i_2 \le \kappa - \lfloor \ell\beta_1 \rfloor$, $i_1 \neq i_2$, we have that

$$\mathbb{P}\Big[\sigma^{-1}(i') = i_1, \sigma^{-1}(i) = i_2\Big] \ge \mathbb{P}\Big[\sigma^{-1}(i') = 1, \sigma^{-1}(i) = 2\Big] \ge e^{-4b}\frac{1}{\kappa^2}\,, \tag{141}$$

where the second inequality follows from the definition of the PL model and the fact that $\widetilde{\theta}^* \in \widetilde{\Omega}_{2b}$. Together with Equation (141) and the fact that there are a total of $(\ell - \lfloor \ell\beta \rfloor)(\ell - \lfloor \ell\beta \rfloor - 1) \ge (\ell - \lfloor \ell\beta \rfloor)^2/2$ pair of positions that i, i' can occupy in order to being ranked in bottom $(\ell - \lfloor \ell\beta \rfloor)$, we have,

$$\mathbb{P}\Big[\sigma^{-1}(i), \sigma^{-1}(i') > \kappa - \ell\Big] \ge \frac{e^{-4b}(1 - \lfloor \ell\beta_1 \rfloor / \ell)^2}{2}\frac{\ell^2}{\kappa^2}. \tag{142}$$

Since, the above inequality is true for any fixed i, i' and $j \in \Omega$ such that event E holds, it is true for random indices i, i' and $j \in \Omega$ such that event E holds, hence the claim is proved.

8.6.4 Proof of Lemma 23

Let $\widehat{\sigma}$ denote a ranking over the items of the set S and $\mathbb{P}[\widehat{\sigma}]$ be the probability of observing $\widehat{\sigma}$. Let

$$\widehat{\Omega}_1 = \Big\{\widehat{\sigma} : \widehat{\sigma}^{-1}(i_{\min_1}) = i_1, \widehat{\sigma}^{-1}(i_{\min_2}) = i_2\Big\} \text{ and } \widehat{\Omega}_2 = \Big\{\widehat{\sigma} : \sigma^{-1}(i_{\min_1}) = 1, \sigma^{-1}(i_{\min_2}) = 2\Big\}. \tag{143}$$

Now, take any ranking $\widehat{\sigma} \in \widehat{\Omega}_1$ and construct another ranking $\widetilde{\sigma}$ from $\widehat{\sigma}$ as following. If $i_1 = 2, i_2 = 1$, then swap the items at i_1-th and i_2-th position in ranking $\widehat{\sigma}$ to get $\widetilde{\sigma}$. Else, if $i_1 < i_2$, then first: swap items at i_1-th position and 1st position, and second: swap items at i_2-th position and 2nd position, to get $\widetilde{\sigma}$; if $i_2 < i_1$, then first: swap items at i_2-th position and 2nd position, and second: swap items at i_1-th position and 1st position, to get $\widetilde{\sigma}$.

Observe that $\mathbb{P}[\widetilde{\sigma}] \le \mathbb{P}[\widehat{\sigma}]$ and $\widetilde{\sigma}^\ell_1 \in \widehat{\Omega}_2$. Moreover, such a construction gives a bijective mapping between $\widehat{\Omega}_1$ and $\widehat{\Omega}_2$. Hence, the claim is proved.

Acknowledgements

The authors thank the anonymous reviewers for their constructive feedback. This work was partially supported by National Science Foundation Grants MES-1450848, CNS-1527754, and CCF-1553452.

References

N. Ailon. Active learning ranking from pairwise preferences with almost optimal query complexity. In *Advances in Neural Information Processing Systems*, pages 810–818, 2011.

A. Ammar and D. Shah. Ranking: Compare, don't score. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 776–783. IEEE, 2011.

A. Ammar, S. Oh, D. Shah, and L. Voloch. What's your choice? learning the mixed multi-nomial logit model. In *Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, 2014.

H. Azari Soufiani, D. C. Parkes, and L. Xia. Random utility theory for social choice. In *NIPS*, pages 126–134, 2012.

H. Azari Soufiani, W. Chen, D. C Parkes, and L. Xia. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems 26*, pages 2706–2714, 2013.

H. Azari Soufiani, D. Parkes, and L. Xia. Computing parametric ranking models via rank-breaking. In *Proceedings of The 31st International Conference on Machine Learning*, pages 360–368, 2014.

M. E. Ben-Akiva and S. R. Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

J. Blanchet, G. Gallego, and V. Goyal. A Markov chain approximation to choice modeling. In *EC*, pages 103–104, 2013.

M. Braverman and E. Mossel. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.

Y. Chen and C. Suh. Spectral mle: Top-k rank aggregation from pairwise comparisons. *arXiv preprint arXiv:1504.07218*, 2015.

C. Cortes, M. Mohri, and A. Rastogi. Magnitude-preserving ranking algorithms. In *Proceedings of the 24th international conference on Machine learning*, pages 169–176. ACM, 2007.

J. C. de Borda. Mémoire sur les élections au scrutin. 1781.

N. De Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale, 1785.

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

W. Ding, P. Ishwar, and V. Saligrama. Learning mixed membership mallows models from pairwise comparisons. *arXiv preprint arXiv:1504.00757*, 2015.

J. C. Duchi, L. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the ICML Conference*, Haifa, Israel, June 2010.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.

O. Dykstra. Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs. *Biometrics*, 16(2):176–188, 1960.

V. F. Farias, S. Jagabathula, and D. Shah. A data-driven approach to modeling choice. In *NIPS*, pages 504–512, 2009.

V. F. Farias, S. Jagabathula, and D. Shah. A nonparametric approach to modeling choice with limited data. *Management Science*, 59(2):305–322, 2013.

J. B. Feldman and H. Topaloglu. Revenue management under the markov chain choice model. 2014.

L. R. Ford Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

Ryan G. Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. Crowdclustering. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 558–566. Curran Associates, Inc., 2011. URL `http://papers.nips.cc/paper/4187-crowdclustering.pdf`.

P. M. Guadagni and J. D. Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238, 1983.

B. Hajek, S. Oh, and J. Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems 27*, pages 1475–1483, 2014.

T. P. Hayes. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2005.

D. R. Hunter. Mm algorithms for generalized bradley-terry models. *Annals of Statistics*, pages 384–406, 2004.

K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2240–2248, 2011.

T. 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. ACM, 2003.

T. Le Van, M. van Leeuwen, S. Nijssen, and L. De Raedt. Rank matrix factorisation. In *Advances in Knowledge Discovery and Data Mining*, pages 734–746. Springer, 2015.

G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. In *Advances in neural information processing systems*, pages 857–864, 2007.

T. Lu and C. Boutilier. Learning mallows models with pairwise preferences. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 145–152, 2011a.

T. Lu and C. Boutilier. Budgeted social choice: From consensus to personalized decision making. In *IJCAI*, volume 11, pages 280–286, 2011b.

Y. Lu and S. Negahban. Individualized rank aggregation using nuclear norm regularization. *arXiv preprint arXiv:1410.0860*, 2014.

J. Lundell. Second report of the irish commission on electronic voting. *Voting Matters*, 23:13–17, 2007.

L. Maystre and M. Grossglauser. Fast and accurate inference of plackett-luce models. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015a.

L. Maystre and M. Grossglauser. Robust active ranking from sparse noisy comparisons. *arXiv preprint arXiv:1502.05556*, 2015b.

D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1973.

D. McFadden. Econometric models for probabilistic choice among products. *Journal of Business*, 53(3):S13–S29, 1980.

D. McFadden and K. Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

K. Miyahara and M. J. Pazzani. Collaborative filtering with the simple bayesian classifier. In *PRICAI 2000 Topics in Artificial Intelligence*, pages 679–689. Springer, 2000.

S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *NIPS*, pages 2483–2491, 2012.

S. Negahban, S. Oh, and D. Shah. Rank centrality: Ranking from pair-wise comparisons. preprint arXiv:1209.1688, 2014.

S. Oh and D. Shah. Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems*, pages 595–603, 2014.

S. Oh, K. K. Thekumparampil, and J. Xu. Collaboratively learning preferences from ordinal data. In *Advances in Neural Information Processing Systems 28*, pages 1900–1908, 2015.

D. Park, J. Neeman, J. Zhang, S. Sanghavi, and I. S. Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1907–1916, 2015.

H. Polat and W. Du. Svd-based collaborative filtering with privacy. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 791–795. ACM, 2005.

A. Rajkumar and S. Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of The 31st International Conference on Machine Learning*, pages 118–126, 2014.

P. Ray. Independence of irrelevant alternatives. *Econometrica: Journal of the Econometric Society*, pages 987–991, 1973.

N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.

N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *arXiv preprint arXiv:1505.01462*, 2015a.

N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*, 2015b.

P. Sham and D. Curtis. An extended transmission/disequilibrium test (tdt) for multi-allele marker loci. *Annals of human genetics*, 59(3):323–336, 1995.

J. Walker and M. Ben-Akiva. Generalized random utility model. *Mathematical Social Sciences*, 43 (3):303–343, 2002.

R. Wu, J. Xu, R. Srikant, L. Massoulié, M. Lelarge, and B. Hajek. Clustering and inference from pairwise comparisons. *arXiv preprint arXiv:1502.04631*, 2015.

J. Yi, R. Jin, S. Jain, and A. Jain. Inferring users? preferences from crowdsourced pairwise comparisons: A matrix completion approach. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.