Journal of Machine Learning Research 19 (2018) 1-42 Submitted 8/16; Published 9/18

Generalized Rank-Breaking: Computational and Statistical Tradeoffs

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: Guy Lebanon

Abstract

For massive and heterogeneous modern datasets, it is of fundamental interest to provide guarantees on the accuracy of estimation when computational resources are limited. In the application of rank aggregation, for the Plackett-Luce model, we provide a hierarchy of rank-breaking mechanisms ordered by the complexity in thus generated sketch of the data. This allows the number of data points collected to be gracefully traded off against computational resources available, while guaranteeing the desired level of accuracy. Theoretical guarantees on the proposed generalized rank-breaking implicitly provide such trade-offs, which can be explicitly characterized under certain canonical scenarios on the structure of the data. Further, the proposed generalized rank-breaking algorithm involves set-wise comparisons as opposed to traditional pairwise comparisons. The maximum likelihood estimate of pairwise comparisons is computed efficiently using the celebrated minorization maximization algorithm (Hunter, 2004). To compute the pseudo-maximum likelihood estimate of the set-wise comparisons, we provide a generalization of the minorization maximization algorithm and give guarantees on its convergence.

Keywords: Rank aggregation, Plackett-Luce model, Sample and Computational tradeoff.

1. Introduction

In classical statistical inference, we are typically interested in characterizing how more data points improve the accuracy, with little restrictions or considerations on computational aspects of solving the inference problem. However, with massive growths of the amount of data available and also the complexity and heterogeneity of the collected data, computational resources, such as time and memory, are major bottlenecks in many modern applications. As a solution, recent advances in learning theory introduce hierarchies of algorithmic solutions, ordered by the respective computational complexity, for several fundamental machine learning applications in Bousquet and Bottou (2008); Shalev-Shwartz and Srebro (2008); Chandrasekaran and Jordan (2013); Agarwal et al. (2012); Lucic et al. (2015). Guided by sharp analyses on the sample complexity, these approaches provide theoretically sound guidelines that allow the analyst the flexibility to fall back to simpler algorithms to enjoy the full merit of the improved run-time.

Inspired by these advances, we study the time-data tradeoff in rank aggregation. In many applications such as election, policy making, polling, and recommendation systems, we want to aggregate individual preferences to produce a global ranking that best represents the collective

social preference. We assume that the data comes from a parametric family of choice models, and learns the parameters that determine the global ranking. Traditionally, each revealed preference is assumed to have one of the following three structures. *Pairwise comparison*, where one item is preferred over another, is common in sports and chess matches. *Best-out-of-κ comparison*, where one is chosen among a set of κ alternatives, is common in historical purchase data. *κ-way comparison*, where we observe a linear ordering of a set of κ candidates, is used in some elections and surveys. We will refer to such structures as *traditional* in comparisons to modern datasets with non-traditional structures. For such traditional preferences, efficient schemes for rank aggregation have been proposed, such as Ford Jr. (1957); Hunter (2004); Hajek et al. (2014); Chen and Suh (2015), which we explain in detail in Section 2.1. However, modern datasets are unstructured and heterogeneous. As Khetan and Oh (2016) show, this can lead to significant increase in the computational complexity, requiring exponential run-time in the size of the problem in the worst case.

To alleviate this computational challenge, we propose a hierarchy of estimators which we call *generalized rank-breaking*, ordered in increasing computational complexity and achieving increasing accuracy. The key idea is to break down the heterogeneous revealed preferences into simpler pieces of ordinal relations, and apply an estimator tailored for those simple structures treating each piece as independent. Several aspects of rank-breaking makes this problem interesting and challenging. A priori, it is not clear which choices of the simple ordinal relations are rich enough to be statistically efficient and yet lead to tractable estimators. Even if we identify which ordinal relations to extract, the ignored correlations among those pieces can lead to an inconsistent estimate, unless we choose carefully which pieces to include and which to omit in the estimation. We further want sharp analysis on the sample complexity, which reveals how computational and statistical efficiencies trade off. We would like to address all these challenges in providing generalized rank-breaking methods.

2. Problem formulation.

We study the problem of aggregating ordinal data based on users' preferences that are expressed in the form of *partially ordered sets (poset)*. A poset is a collection of ordinal relations among items. For example, consider a poset $\{(i_6 \prec \{i_5, i_4\}), (i_5 \prec i_3), (\{i_3, i_4\} \prec \{i_1, i_2\})\}$ over items $\{i_1, \ldots, i_6\}$, where $(i_6 \prec \{i_5, i_4\})$ indicates that item i_5 and i_4 are both preferred over item i_6. Such a relation is extracted from, for example, the user giving a 2-star rating to i_5 and i_4 and a 1-star to i_6.

We assume there are n users and d items. We denote the set of n users by $[n] = \{1, \ldots, n\}$ and the set of d items by $[d]$. We assume that each user $j \in [n]$ is presented with a subset of items $S_j \subseteq [d]$, and independently provides her ordinal preference in the form of a poset, where the ordering is drawn from the Plackett-Luce (PL) model. Since, an ordering drawn from the PL model is consistent, a poset can be represented as a directed acyclic graph (DAG). Let $\mathcal{G}_j$ denote the DAG representation of the poset provided by the user j over $S_j \subseteq [d]$ according to the PL model with weights θ^*. The task is to learn $\widehat{\theta}$, an estimate of the true weights θ^*. Below is an example of a DAG $\mathcal{G}_j$. We use index i to denote items and j to denote users.

Plackett-Luce model. The PL model is a popular choice model from operations research and psychology, used to model how people make choices under uncertainty. It is a special case of *random utility models*, where each item i is parametrized by a latent true utility $\theta_i \in \mathbb{R}$. When offered with

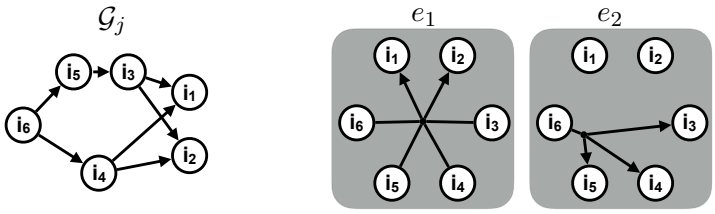

Figure 1: An example of $\mathcal{G}_j$ for user j's consistent poset, and two rank-breaking hyper edges extracted from it: $e_1 = (\{i_6, i_5, i_4, i_3\} \prec \{i_2, i_1\})$ and $e_2 = (\{i_6\} \prec \{i_5, i_4, i_3\})$.

S_j, the user samples the perceived utility U_i for each item independently according to $U_i = \theta_i + Z_i$, where Z_i's are i.i.d. noise. In particular, the PL model assumes Z_i's follow the standard Gumbel distribution. The observed poset is a partial observation of the ordering according to this perceived utilities. We discuss possible extensions to general class of random utility models in Section 2.1.

The particular choice of the Gumbel distribution has several merits, largely stemming from the fact that the Gumbel distribution has a log-concave pdf and is inherently memoryless. In our analyses, we use the log-concavity to show that our proposed algorithm is a concave maximization (Remark 1) and the memoryless property forms the basis of our rank-breaking idea. Precisely, the PL model is statistically equivalent to the following procedure. Consider a ranking as a mapping from a position in the rank to an item, i.e. $\sigma_j : [|S_j|] \to S_j$. It can be shown that the PL model is generated by first independently assigning each item $i \in S_j$ an unobserved value Y_i, exponentially distributed with mean $e^{-\theta_i}$, and the resulting ranking σ_j is inversely ordered in Y_i's so that $Y_{\sigma_j(1)} \le Y_{\sigma_j(2)} \le \cdots \le Y_{\sigma_j(|S_j|)}$.

This inherits the memoryless property of exponential variables, such that $\mathbb{P}(Y_1 < Y_2 < Y_3) = \mathbb{P}(Y_1 < \{Y_2, Y_3\})\mathbb{P}(Y_2 < Y_3)$, leading to a simple interpretation of the PL model as sequential choices:

$$\mathbb{P}_\theta(i_3 \prec i_2 \prec i_1) \;=\; \mathbb{P}_\theta(\{i_3, i_2\} \prec i_1)\mathbb{P}_\theta(i_3 \prec i_2) \;=\; \frac{e^{\theta_{i_1}}}{e^{\theta_{i_1}} + e^{\theta_{i_2}} + e^{\theta_{i_3}}} \times \frac{e^{\theta_{i_2}}}{e^{\theta_{i_2}} + e^{\theta_{i_3}}} \,.$$

In general, for true utility θ^*, we have

$$\mathbb{P}_{\theta^*}[\sigma_j] \;=\; \prod_{i=1}^{|S_j|-1} \frac{e^{\theta^*_{\sigma_j(i)}}}{\sum_{i'=i}^{|S_j|} e^{\theta^*_{\sigma_j(i')}}} \,.$$

We assume that the true utility $\theta^* \in \Omega_b$ where

$$\Omega_b \;=\; \left\{ \theta \in \mathbb{R}^d \,\middle|\, \sum_{i \in [d]} \theta_i = 0, |\theta_i| \le b \text{ for all } i \in [d] \right\} . \tag{1}$$

Notice that centering of θ ensures its uniqueness as PL model is invariant under shifting of θ. The bound b on θ_i is written explicitly to capture the dependence in our main results. We interchangeably refer θ as utilities and weights.

Maximum Likelihood Estimate of DAG. Probability of observing a DAG $\mathcal{G}_j$ is the sum of probabilities of all possible rankings that are consistent with it. Precisely, under the PL model, for

a DAG $\mathcal{G}_j$, we have,

$$\mathbb{P}_\theta[\mathcal{G}_j] \quad = \quad \sum_{\sigma \in \mathcal{G}_j} \mathbb{P}_\theta[\sigma]\,,$$

where we slightly abuse the notation $\mathcal{G}_j$ to denote the set of all rankings σ that are consistent with the observation. For example, if $\mathcal{G}_j$ consists of only one hyper edge $e_1 = (\{i_3\} \prec \{i_2, i_1\})$ then $\mathbb{P}[\mathcal{G}_j] = \mathbb{P}(i_3 \prec i_2 \prec i_1) + \mathbb{P}(i_3 \prec i_1 \prec i_2)$. The maximum likelihood estimate (MLE) maximizes log-likelihood of observing $\mathcal{G}_j$ for each j:

$$\widehat{\theta} \quad \in \quad \arg\max_{\theta \in \Omega_b} \left\{ \sum_{j=1}^{n} \log \mathbb{P}_\theta[\mathcal{G}_j] \right\}. \tag{2}$$

When $\mathcal{G}_j$ has a *traditional* structure as explained earlier in this section, then the optimization is a simple multinomial logit regression, that can be solved efficiently with off-the-shelf convex optimization tools. Hajek et al. (2014) provides full analysis of the statistical complexity of this MLE under traditional structures. For general posets, it can be shown that the above optimization is a concave maximization, using similar techniques as Remark 1. However, the summation over rankings in $\mathcal{G}_j$ can involve number of terms super exponential in the size $|S_j|$, in the worst case. This renders MLE intractable and impractical.

Pairwise rank-breaking. A common remedy to this computational blow-up is to use rank-breaking. Rank-breaking traditionally refers to *pairwise rank-breaking*, where a bag of all the pairwise comparisons is extracted from observations $\{\mathcal{G}_j\}_{j \in [n]}$ and is applied to estimators that are tailored for pairwise comparisons, treating each paired outcome as independent. This is one of the motivations behind the algorithmic advances in the popular topic of aggregation from pairwise comparisons in (Ford Jr., 1957; Hunter, 2004; Negahban et al., 2014; Shah et al., 2015a; Maystre and Grossglauser, 2015).

It is computationally efficient to apply maximum likelihood estimator assuming independent pairwise comparisons, which takes $O(d^2)$ operations to evaluate. However, this computational gain comes at the cost of statistical efficiency. Azari Soufiani et al. (2014) showed that if we include all paired comparisons, then the resulting estimate can be statistically inconsistent due to the ignored correlations among the paired orderings, even with infinite samples. In the example from Figure 1, there are 12 paired relations implied by the DAG: $(i_6 \prec i_5), (i_6 \prec i_4), (i_6 \prec i_3), \ldots, (i_3 \prec i_1), (i_4 \prec i_1)$. In order to get a consistent estimate, Azari Soufiani et al. (2014) provide a rule for choosing which pairs to include, and Khetan and Oh (2016) provide an estimator that optimizes how to weigh each of those chosen pairs to get the best finite sample complexity bound. However, such a consistent pairwise rank-breaking results in throwing away many of the ordered relations, resulting in significant loss in accuracy. For example, including any paired relation from $\mathcal{G}_j$ in the example results in a biased estimator. None of the pairwise orderings can be used from $\mathcal{G}_j$, without making the estimator inconsistent as shown in Azari Soufiani et al. (2013). Whether we include all paired comparisons or only a subset of consistent ones, there is a significant loss in accuracy as illustrated in Figure 4. For the precise condition for consistent rank-breaking we refer to (Azari Soufiani et al., 2013, 2014; Khetan and Oh, 2016).

The state-of-the-art approaches operate on either one of the two extreme points on the computational and statistical trade-off. The MLE in (2) requires $O(\sum_{j \in [n]} |S_j|!)$ summations to just

evaluate the objective function, in the worst case. On the other hand, the pairwise rank-breaking requires only $O(d^2)$ summations, but suffers from significant loss in the sample complexity. Ideally, we would like to give the analyst the flexibility to choose a target computational complexity she is willing to tolerate, and provide an algorithm that achieves the optimal trade-off at the chosen operating point.

Contribution. We introduce a novel *generalized rank-breaking* that bridges the gap between MLE and pairwise rank-breaking. Our approach allows the user the freedom to choose the level of computational resources to be used, and provides an estimator tailored for the desired complexity. We prove that the proposed estimator is tractable and consistent, and provide an upper bound and a lower bound on the error rate in the finite sample regime. The analysis explicitly characterizes the dependence on the topology of the data. This in turn provides a guideline for designing surveys and experiments in practice, in order to maximize the sample efficiency. The proposed generalized rank-breaking mechanism involves set-wise comparisons as opposed to traditional pairwise comparisons. In order to compute the rank-breaking estimate, we generalize the celebrated minorization maximization algorithm for computing maximum likelihood estimate of pairwise comparisons (Hunter, 2004) to more general set-wise comparisons and give guarantees on its convergence.

2.1 Related work

In classical statistics, one is interested in the tradeoff between the sample size and the accuracy, with little considerations to the computational complexity or time. As more computations are typically required with increasing availability of data, the computational resources are often the bottleneck. Recently, a novel idea known as "algorithmic weakening" has been investigated to overcome such a bottleneck, in which a hierarchy of algorithms is proposed to allow for faster algorithms at the expense of decreased accuracy. When guided by sound theoretical analyses, this idea allows the statistician to achieve the same level of accuracy and *save* time when more data is available. This is radically different from classical setting where processing more data typically requires more computational time.

Depending on the application, several algorithmic weakenings have been studied. In the application of supervised learning, Bousquet and Bottou (2008) proposed the idea that weaker approximate optimization algorithms are sufficient for learning when more data is available. Various gradient based algorithms are analyzed that show the time-accuracy-sample tradeoff. In a similar context, Shalev-Shwartz and Srebro (2008) analyze a particular implementation of support vector machine and show that the target accuracy can be achieved faster when more data is available, by running the iterative algorithm for shorter amount of time. In the application of de-noising, Chandrasekaran and Jordan (2013) provide a hierarchy of convex relaxations where constraints are defined by convex geometry with increasing complexity. For unsupervised learning, Lucic et al. (2015) introduce a hierarchy of data representations that provide more representative elements when more data is available at no additional computation. Standard clustering algorithms can be applied to thus generated summary of the data, requiring less computational complexity.

In the application of rank aggregation, we follow the principle of algorithmic weakening and propose a novel rank-breaking to allow the practitioner to navigate gracefully the time-sample trade off as shown in the Figure 2. We propose a hierarchy of estimators indexed by $M \in \mathbb{Z}^+$ indicating how complex the estimator is (defined formally in Section 3). Figure 2 shows the result of a experiment on synthetic datasets on how much time (in seconds) and how many samples are

required to achieve a target accuracy. If we are given more samples, then it is possible to achieve the target accuracy, which in this example is MSE$\leq 0.3d^2 \times 10^{-6}$, with fewer operations by using a simpler estimator with smaller M. The details of the experiment is explained in Figure 4.

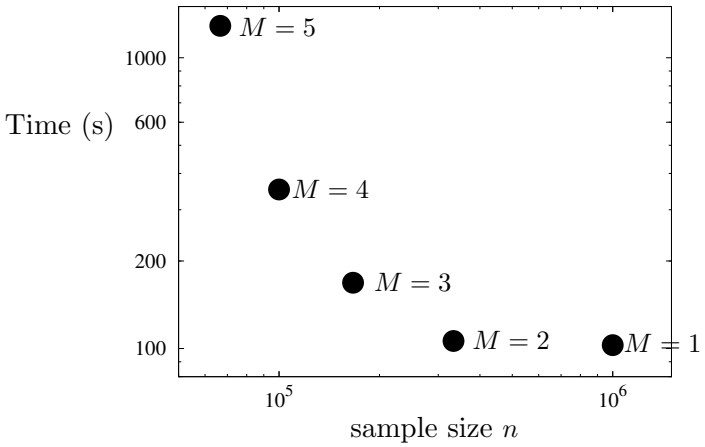

Figure 2: Depending on how much computational resources are available, the various choices of M achieve different operating points on the time-data trade-off to achieve some fixed target accuracy $\varepsilon > 0$. If more samples are available, one can resort to faster methods with smaller M while achieving the same level of accuracy.

Rank aggregation under the PL model has been studied extensively under the *traditional* scenario dating back to Zermelo (1929) who first introduced the PL model for pairwise comparisons. Various approaches for estimating the PL weights from traditional samples have been proposed. The problem can be formulated as a convex optimization that can be solved efficiently using the off-the-shelf solvers. However, tailored algorithms for finding the optimal solution have been proposed in Ford Jr. (1957) and Hunter (2004), which iteratively finds the fixed point of the KKT condition. Negahban et al. (2014) introduce Rank Centrality, a novel spectral ranking algorithm which formulates a random walk from the given data, and show that the stationary distribution provides accurate estimates of the PL weights. Maystre and Grossglauser (2015) provide a connection between those previous approaches, and give a unified random walk approach that finds the fixed point of the KKT conditions.

On the theoretical side, when samples consist of pairwise comparisons, Simons and Yao (1999) first established consistency and asymptotic normality of the maximum likelihood estimate when all teams play against each other. For a broader class of scenarios where we allow for sparse observations, where the number of total comparisons grow linearly in the number of teams, Negahban et al. (2014) show that Rank Centrality achieves optimal sample complexity by comparing it to a lower bound on the minimax rate. For a more general class of traditional observations, including pairwise comparisons, Hajek et al. (2014) provide similar optimal guarantee for the maximum likelihood estimator. Chen and Suh (2015) introduced Spectral MLE that applies Rank Centrality followed by MLE, and showed that the resulting estimate is optimal in L_∞ error as well as the previously analyzed L_2 error. Shah et al. (2015a) study a new measure of the error induced by the

Laplacian of the comparisons graph and prove a sharper upper and lower bounds that match up to a constant factor.

However, in modern applications, the computational complexity of the existing approaches blow-up due to the heterogeneity of modern datasets. Although, statistical and computational tradeoffs have been investigated under other popular choice models such as the Mallows models by Betzler et al. (2014) or stochastically transitive models by Shah et al. (2015b), the algorithmic solutions do not apply to random utility models and the analysis techniques do not extend. We provide a novel rank-breaking algorithms and provide finite sample complexity analysis under the PL model. This approach readily generalizes to some RUMs such as the flipped Gumbel distribution. However, it is also known from Azari Soufiani et al. (2014), that for general RUMs there is no consistent rank-breaking, and the proposed approach does not generalize.

3. Generalized rank-breaking

Given $\mathcal{G}_j$'s representing the users' preferences, *generalized rank-breaking* extracts a set of ordered relations and applies an estimator treating each ordered relation as independent. Concretely, for each $\mathcal{G}_j$, we first extract a maximal ordered partition $\mathcal{P}_j$ of S_j that is consistent with $\mathcal{G}_j$. An ordered partition is a partition with a linear ordering among the subsets, e.g. $\mathcal{P}_j = (\{i_6\} \prec \{i_5, i_4, i_3\} \prec \{i_2, i_1\})$ for $\mathcal{G}_j$ from Figure 1. This is maximal, since we cannot further partition any of the subsets without creating artificial ordered relations that are not present in the original $\mathcal{G}_j$.

To precisely define maximal ordered partition $\mathcal{P}_j$, first, let's define an ordered partition $\widetilde{\mathcal{P}}_j$ of S_j that is consistent with $\mathcal{G}_j$. Consider disjoint subsets $\mathcal{C}_1, \cdots, \mathcal{C}_{\ell_j} \subseteq S_j$ such that their union is S_j that is $\cup_{a=1}^{\ell_j} C_a = S_j$. The subsets $\mathcal{C}_1, \cdots, \mathcal{C}_{\ell_j}$ define an ordered partition

$$\widetilde{\mathcal{P}}_j = \mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_{\ell_j} ,$$

if each ordered relation that can be read from this linear ordering of subsets is present in the DAG $\mathcal{G}_j$. Let $|\mathcal{P}_j|$ denote the size of the partition, that is $|\mathcal{P}_j| = \ell_j$. A maximal ordered partition $\mathcal{P}_j$ is the one which has the largest size.

$$\mathcal{P}_j = \arg\max_{\widetilde{\mathcal{P}}_j} \left\{ |\widetilde{\mathcal{P}}_j| \right\} .$$

In the following we provide an algorithm to find a maximal ordered partition $\mathcal{P}_j$ of S_j that is consistent with a given $\mathcal{G}_j$.

Finding Maximal Ordered Partition. Given a DAG $\mathcal{G}_j$, a corresponding maximal ordered partition $\mathcal{P}_j$ can be extracted by recursively finding common ancestors of the sink-nodes of the vertex induced sub-graph starting with the DAG $\mathcal{G}_j$. Algorithm 1 gives a pseudocode to find $\mathcal{P}_j$'s. Common ancestors of all the sink nodes of a DAG can be found in time $O(d^{2.6})$ using fast algorithms given in (Czumaj et al., 2007; Bender et al., 2005). Therefore, computational complexity of the Algorithm 1 is $O(d^{3.6})$. In line 2, Algorithm 1, $V(\mathcal{G})$ denotes the set of vertices of DAG $\mathcal{G}$. In line 5, $\mathcal{G}(S)$ denote the vertex induced subgraph of graph $\mathcal{G}$ corresponding to vertex set S. Note that Algorithm 1 returns a unique maximal ordered partition $\mathcal{P}_j$ for a given DAG $\mathcal{G}_j$.

In general there is no one-to-one mapping from a DAG $\mathcal{G}_j$ to its maximal ordered partition $\mathcal{P}_j$. There may be many ordered relations present in $\mathcal{G}_j$ that are not represented in the ordered

Algorithm 1 Finding Maximal Ordered Partition

Require: DAG $\mathcal{G}_j$
Ensure: maximal ordered partition $\mathcal{P}_j$
1: $\mathcal{G} \leftarrow \mathcal{G}_j$, $\mathcal{P}_j = \{\}$
2: **while** $|V(\mathcal{G})| > 0$ **do**
3: $S \leftarrow$ Common ancestors of all sink-nodes of DAG $\mathcal{G}$ (Czumaj et al., 2007)
4: $\mathcal{P}_j \leftarrow \mathcal{P}_j \succ \{V(\mathcal{G}) \setminus S\}$
5: $\mathcal{G} \leftarrow \mathcal{G}(S)$
6: **end while**

partition $\mathcal{P}_j$. This gives a many-to-one mapping from $\mathcal{G}_j$ to $\mathcal{P}_j$. In our generalized rank-breaking framework, we can only use those ordered relations that can be represented in an ordered partition. This is required for the estimator to be consistent. This is the cost we pay to reduce computational complexity from $O(|S_j|!)$, complexity of MLE of DAG (2), to $O(M!)$ for a suitably desired $M \in \mathbb{Z}^+$ as explained below. However, if the DAG $\mathcal{G}_j$ represents a full ranking or a traditional structure then its maximal ordered partition $\mathcal{P}_j$ will represent all the ordered relations present in $\mathcal{G}_j$ and our rank-breaking will reduce to MLE of the DAG $\mathcal{G}_j$. In such a case, all the subsets of $\mathcal{P}_j$ will have cardinality one except the least preferred set which can have more than one item in case of best-out-of κ comparison.

Rank-Breaking Graph. The extracted maximal ordered partition $\mathcal{P}_j$ is represented by a directed hypergraph $G_j(S_j, E_j)$, which we call a *rank-breaking graph*. Each edge $e = (B(e), T(e)) \in E_j$ is a directed hyper edge from a subset of nodes $B(e) \subseteq S_j$ to another subset $T(e) \subseteq S_j$. The number of edges in E_j is $|\mathcal{P}_j| - 1$. For each subset in $\mathcal{P}_j$ except for the least preferred subset, there is a corresponding edge whose *top-set* $T(e)$ is the subset, and the *bottom-set* $B(e)$ is the set of all items less preferred than $T(e)$. For the example in Figure 1, we have $E_j = \{e_1, e_2\}$ where $e_1 = (B(e_1), T(e_1)) = (\{i_6, i_5, i_4, i_3\}, \{i_2, i_1\})$ and $e_2 = (B(e_2), T(e_2) = (\{i_6\}, \{i_5, i_4, i_3\})$ extracted from $\mathcal{G}_j$. Algorithm 2 gives the precise method to construct a rank-breaking graph.

Algorithm 2 Constructing Rank-Breaking Graph

Require: maximal ordered partition $\mathcal{P}_j = \mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_{\ell_j}$ of set S_j
Ensure: directed hypergraph $G_j(S_j, E_j)$
1: construct directed hypergraph $G_j(S_j, E_j = \{\})$
2: **for** $a = 2$ to ℓ_j **do**
3: construct hyper edge e between top-set $T(e) = \mathcal{C}_a$ and bottom-set $B(e) = \cup_{a'=1}^{a-1} \mathcal{C}_{a'}$
4: $E_j \leftarrow E_j \cup e$
5: **end for**
6: Return $G_j(S_j, E_j)$

Denote the probability that $T(e)$ is preferred over $B(e)$ when $T(e) \cup B(e)$ is offered as

$$\mathbb{P}_\theta(e) \;=\; \mathbb{P}_\theta\big(B(e) \prec T(e)\big) \;=\; \sum_{\sigma \in \Lambda_{T(e)}} \frac{\exp\left(\sum_{c=1}^{|T(e)|} \theta_{\sigma(c)}\right)}{\prod_{u=1}^{|T(e)|} \left(\sum_{c'=u}^{|T(e)|} \exp\left(\theta_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta_i\right)\right)}, \qquad (3)$$

which follows from the definition of the PL model, where $\Lambda_{T(e)}$ is the set of all rankings over $T(e)$. The computational complexity of evaluating this probability is determined by the size of the *top-set* $|T(e)|$, as it involves $(|T(e)|!)$ summations.

In the subsequent results, we show that by maximizing likelihood of the hyper edges assuming they are independent we get a consistent estimator. Therefore, our approach provides flexibility to choose which hyper edge to include in the likelihood maximization function. We let the analyst choose the order $M \in \mathbb{Z}^+$ depending on how much computational resource is available, and include only those edges with $|T(e)| \leq M$ in the likelihood objective function.

If in a given G_j there are no hyper edges with top sets of size less than M, then the analyst does not get any ordered relations from that rank-breaking graph under her computational constraint reflected in the particular choice of M. Artificially reducing the size of the top-sets so as to get the hyper edges with top sets of size less than M implies we need to add new ordered relations that are not present in the DAG provided by the user. Such an estimator could result in a non-zero bias. A concrete example of such cases has been studied in Azari Soufiani et al. 2013, where the authors showed that for $M = 1$, applying rank-breaking to those comparisons with top-set larger than one results in non-zero bias.

We emphasize that M is chosen by the analyst and our estimator works for any choice of $M \in \mathbb{Z}^+$. Given unlimited computational resources, an analyst would chose $M = d$, and all the hyper edges would be included in the likelihood objective function.

Sampling. We assume that for each $j \in [n]$, the topology of DAG $\mathcal{G}_j$ which represent the partial preference order provided by the j-th user is fixed apriori. Also, the set of hyper edges $e \in E_j$ of each rank breaking graph $G_j(S_j, E_j)$ that are included in the likelihood objective function are fixed apriori. The randomness that we observe is in the position of the S_j items in the DAG $\mathcal{G}_j$. For an hyper edge $e \in E_j$, the randomness is in which items of the set S_j appear in the bottom $|B(e)|$ positions and the bottom $|T(e)| + |B(e)|$ positions in the preference order of the user j. Note that this precisely captures the randomness due to the PL model in the observed DAG $\mathcal{G}_j$. We do not impose any restrictions on the topology of the DAG $\mathcal{G}_j$'s and each of them can be different. Further, our analysis captures effect of their topologies on the statistical efficiency of the estimation.

Pseudo-MLE of Rank-Breaking Graph. We apply the MLE for comparisons over paired subsets, assuming all hyper edges in the rank-breaking graph G_j are independently drawn. Precisely, for any choice of $M \in \mathbb{Z}^+$, we propose *order-M rank-breaking estimate*, which is the solution that maximizes the log-likelihood under the independence assumption:

$$\begin{aligned} \widehat{\theta} &\in \arg\max_{\theta \in \Omega_b} \mathcal{L}_{\mathrm{RB}}(\theta)\,, \text{ where} \\ \mathcal{L}_{\mathrm{RB}}(\theta) &= \sum_{j \in [n]} \sum_{e \in E_j : |T(e)| \leq M} \ln \mathbb{P}_\theta(e)\,. \end{aligned} \tag{4}$$

Due to independence assumption, we refer to it as pseudo-MLE. In a special case when $M = 1$, this can be transformed into the traditional pairwise rank-breaking, where (i) this is a concave maximization; (ii) the estimate is (asymptotically) unbiased and consistent as shown in Azari Soufiani et al. (2013, 2014); and (iii) the finite sample complexity have been analyzed in Khetan and Oh (2016). Although, this order-1 rank-breaking provides a significant gain in computational efficiency, the information contained in higher-order edges are unused, resulting in a significant loss in accuracy.

We provide the analyst the freedom to choose the computational complexity he/she is willing to tolerate. However, for general M, it has not been known if the optimization in (4) is tractable and/or if the solution is consistent. Since $\mathbb{P}_\theta(B(e) \prec T(e))$ as explicitly written in (3) is a sum of log-concave functions, it is not clear if the sum is also log-concave. Due to the ignored dependency in the formulation (4), it does not follow immediately that the resulting estimate is consistent.

We first establish that (4) is a concave maximization, Section 3.1. Though one can use any off-the-shelf convex maximization tool to compute $\widehat{\theta}$, we provide an efficient minorization-maximization (MM) algorithm for estimating $\widehat{\theta}$, Section 3.2. In Section 3.3, we show that the MM algorithm converges to the unique global optimal solution $\widehat{\theta}$ under the standard assumption given by Ford Jr. (1957) for pairwise comparisons. Under the same assumption, we show that the estimate $\widehat{\theta}$ is consistent, Section 3.4. In Section 3.5, we give the complete algorithm to compute $\widehat{\theta}$ using the proposed MM algorithm, given $\mathcal{G}_j$'s representing users' preferences. In Section 4 and Section 6, we provide a sharp analysis of the performance in the finite sample regime, characterizing the trade-off between computation and sample size, and verify the results from the numerical experiments.

3.1 Concavity of likelihood of rank-breaking graph

In the following, we show that likelihood of a hyper edge is log-concave for a family of Random Utility Models including the PL model.

Remark 1 *$\mathcal{L}_{\mathrm{RB}}(\theta)$ is concave in $\theta \in \mathbb{R}^d$.*

Proof Recall that $\mathbb{P}_\theta(B(e) \prec T(e))$ is the probability that an agent ranks the collection of items $T(e)$ above $B(e)$ when offered $S = B(e) \cup T(e)$. We want to show that $\mathbb{P}_\theta(B(e) \prec T(e))$ is log-concave under the PL model. We prove a slightly general result which works for a family of RUMs in the location family. RUM are defined as a probabilistic model where there is a real-valued utility parameter θ_i associated with each items $i \in S$, and an agent independently samples random utilities $\{U_i\}_{i \in S}$ for each item i with conditional distribution $\mu_i(\cdot|\theta_i)$. Then the ranking is obtained by sorting the items in decreasing order as per the observed random utilities U_i's. *Location family* is a subset of RUMs where the shapes of μ_i's are fixed and the only parameters are the means of the distributions. For location family, the noisy utilities can be written as $U_i = \theta_i + Z_i$ for i.i.d. random variable Z_i's. In particular, it is PL model when Z_i's follow the independent standard Gumbel distribution. We will show that for the location family if the probability density function for each Z_i's is log-concave then $\log \mathbb{P}_\theta(B(e) \prec T(e))$ is concave. The desired claim follows as the pdf of standard Gumbel distribution is log-concave. We use the following Theorem from Prékopa (1980). A similar technique was used to prove concavity when $|T(e)| = 1$ in Azari Soufiani et al. (2012).

Lemma 2 (Extension of Theorem 9 in Prékopa (1980)) *Suppose $g_1(\theta, Y), \cdots, g_r(\theta, Y)$ are concave functions in $\mathbb{R}^{2q}$, where $\theta, Y \in \mathbb{R}^q$, and Z is a $q-$component random vector whose probability distribution is logarithmic concave in $\mathbb{R}^q$, then the function*

$$h(\theta) = \mathbb{P}[g_1(\theta, Z) \geq 0, \cdots, g_r(\theta, Z) \geq 0], \qquad \text{for } \theta \in \mathbb{R}^q$$

is logarithmic concave on $\mathbb{R}^q$. Moreover, concavity is strict if the probability density function of Z is strictly logarithmic concave and $\theta \neq \tilde{\theta}$ implies $H(\theta) \neq H(\tilde{\theta})$. Where $H(\theta)$ is

$$H(\theta) \quad \equiv \quad \left\{ Y \;\middle|\; g_i(\theta, Y) \geq 0, \quad i = 1, \cdots, r \right\}.$$

Proof Theorem 9 in Prékopa (1980) proves concavity. The strict concavity follows from the fact that for a strictly logarithmic concave measure the following inequality is strict if $H(\theta) \neq H(\tilde{\theta})$.

$$\mathbb{P}[Z \in \lambda H(\theta) + (1-\lambda)H(\tilde{\theta})] \;\; \geq \;\; \mathbb{P}[Z \in \lambda H(\theta)]^{\lambda} \mathbb{P}[Z \in (1-\lambda)H(\tilde{\theta})]^{1-\lambda},$$

where $\lambda \in (0,1)$. For a detailed proof, we refer the reader to the proof of Theorem 9 in Prékopa (1980). ■

To apply the above lemma to get concavity, let $q = |S|$, $r = 1$, $g_1(\theta, Y) = \min_{i \in T(e)}\{\theta_i + Y_i\} - \max_{i' \in B(e)}\{\theta_{i'} + Y_{i'}\}$. Observe that $g_1(\theta, Y)$ is concave in $\mathbb{R}^{2q}$, and $\mathbb{P}_\theta(B(e) \prec T(e)) = \mathbb{P}(g_1(\theta, Z) \geq 0)$. We use strict concavity part of the lemma in the subsequent section. ■

3.2 Minorization-maximization algorithm for pseudo-MLE of rank-breaking graph

We give a minorization-maximization algorithm for computing $\widehat{\theta}$ defined in (4). It is inspired from the MM algorithm given by Hunter (2004) for the case of pairwise comparisons and full-ranking. For any fixed parameter $\theta^{(t)} \in \mathbb{R}^d$, and a hyper edge e in a rank breaking graph G, define $Q(e, \theta; \theta^{(t)})$ as

$$Q(e,\theta;\theta^{(t)}) \;\; \equiv \;\; \sum_{\sigma \in \Lambda_{T(e)}} \left(\frac{\mathbb{P}_{\theta^{(t)}}(e,\sigma)}{\mathbb{P}_{\theta^{(t)}}(e)} \sum_{u=1}^{|T(e)|} \left(\theta_{\sigma(u)} - \frac{\sum_{c'=u}^{|T(e)|} \exp\left(\theta_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta_i\right)}{\sum_{c'=u}^{|T(e)|} \exp\left(\theta^{(t)}_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta^{(t)}_i\right)} \right) \right),$$

where $\mathbb{P}_\theta(e, \sigma)$ is defined such that $\mathbb{P}_\theta(e) = \sum_{\sigma \in \Lambda_{T(e)}} \mathbb{P}_\theta(e, \sigma)$. Recall from Equation (3) that $\Lambda_{T(e)}$ is the set of all rankings over $T(e)$.

$$\mathbb{P}_\theta(e,\sigma) \;\; \equiv \;\; \frac{\exp\left(\sum_{c=1}^{|T(e)|} \theta_{\sigma(c)}\right)}{\prod_{u=1}^{|T(e)|} \left(\sum_{c'=u}^{|T(e)|} \exp\left(\theta_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta_i\right) \right)}.$$

We show that $Q(e, \theta; \theta^{(t)})$ minorizes $\ln(\mathbb{P}_\theta(e))$ at $\theta^{(t)}$. It is equal to $\ln(\mathbb{P}_\theta(e))$, up to a constant, if and only if $\theta^{(t)} = \theta$.

Lemma 3

$$Q(e,\theta;\theta^{(t)}) + f(e,\theta^{(t)}) \;\; \leq \;\; \ln(\mathbb{P}_\theta(e)) \quad \textit{with equality if } \theta = \theta^{(t)},$$

where $f(e, \theta^{(t)})$ is a function of the hyper edge e and the parameter $\theta^{(t)}$, it does not depend upon θ.

We give a proof of the Lemma in Section 7.1. It follows that for any $Q(e, \theta; \theta^{(t)})$ satisfying minorizing condition in the above lemma,

$$Q(e,\theta;\theta^{(t)}) \;\geq\; Q(e,\theta^{(t)};\theta^{(t)}) \quad \text{implies} \;\; \ln(\mathbb{P}_\theta(e)) \;\geq\; \ln(\mathbb{P}_{\theta^{(t)}}(e)). \tag{5}$$

Property (5) suggests an iterative algorithm in which we let $\theta^{(t)}$ be the parameter vector before the t-th iteration and define $\theta^{(t+1)}$ to be the maximizer of the $Q(e, \theta; \theta^{(t)})$. Since this algorithm consists of alternately creating a minorizing function $Q(e, \theta; \theta^{(t)})$ and then maximizing it, it is called an MM

algorithm (Hunter and Lange, 2000). To compute $\widehat{\theta}$ in (4), starting from an arbitrary initialization $\theta^{(1)}$, we estimate $\theta^{(t+1)}$ by maximizing

$$\theta^{(t+1)} \;=\; \arg\max_{\theta\in\mathbb{R}^d} \left\{ \sum_{j=1}^{n} \sum_{e\in E_j:|T(e)|\le M} Q(e,\theta;\theta^{(t)}) \right\}.$$

Since the parameters $\{\theta_i\}_{i\in[d]}$ are separated in $Q(e,\theta;\theta^{(t)})$, its maximization can be explicitly accomplished as, for $i\in[d]$

$$e^{\theta_i^{(t+1)}} = \frac{N_i}{\sum_{j=1}^{n}\sum_{e\in E_j:|T(e)|\le M}\sum_{\sigma\in\Lambda_{T(e)}}\frac{\mathbb{P}_{\theta^{(t)}}(e,\sigma)}{\mathbb{P}_{\theta^{(t)}}(e)}\sum_{u=1}^{|T(e)|}\delta_{i,e,\sigma,u}\left(\sum_{c'=u}^{|T(e)|}\exp\left(\theta^{(t)}_{\sigma(c')}\right)+\sum_{i\in B(e)}\exp\left(\theta_i^{(t)}\right)\right)^{-1}}, \tag{6}$$

where N_i is the total number of hyper edges in which the i-th item is in the top set.

$$N_i \;=\; \sum_{j=1}^{n}\sum_{e\in E_j:|T(e)|\le M} \mathbb{I}[i\in T(e)]\,.$$

$\delta_{i,e,\sigma,u}$ is the indicator variable defined as

$$\delta_{i,e,\sigma,u} = \begin{cases} 1\,, & \text{if } i\in\{T(e)\cup B(e)\} \text{ and } \sigma^{-1}(i)\ge u\,,\\ 0\,, & \text{otherwise.}\end{cases}$$

3.3 Convergence properties of the MM algorithm

In the following we show that $\lim_{t\to\infty}\theta^{(t)}$, (6), converges to the global optimal solution of the pseudo-likelihood objective given in (4) under standard assumption on the observed comparisons.

For pairwise comparisons, Ford Jr. (1957) noted that if it is possible to partition the set of items into two subsets A and B such that there are never any inter-set comparisons, then there is no basis for rating any item in set A with respect to any item in set B. On the other hand, if in all the inter-set comparisons, items from set A are preferred over the items in set B, then if all the parameters θ_i belonging to set A are doubled and the resulting vector θ renormalized, the likelihood must increase; thus the likelihood has no maximizer. The following assumption (Ford Jr., 1957) eliminates these possibilities.

Assumption 4 *In every possible partition of the items into two nonempty subsets, some item in the second set is preferred over some item in the first set at least once.*

Assumption 4 has a graph-theoretic interpretation. Suppose, items are denoted by nodes of a graph G and a directed edge (i,j) represents that the there is at least one user who prefers item i over item j. Then the Assumption 4 is equivalent to the statement that there is a path from i to j for all nodes i and j of the graph G. This assumption implies that there exists a unique maximizer of the log-likelihood function of pairwise comparisons.

In our setting, Assumption 4 makes sense if we interpret item i being preferred over item j to mean that there exists a hyper edge e such that the item i is in its top set $T(e)$ and the item j is in its bottom set $B(e)$. In the following theorem, we prove the convergence properties of the MM algorithm.

Theorem 5 *Under Assumption 4 the iterative minorization-maximization algorithm given in Equation* (6) *produces a sequence* $\theta^{(1)}, \theta^{(2)}, \cdots$ *guaranteed to converge to the unique estimate of the optimization problem given in Equation* (4)*.*

Proof In general, it is always not possible to prove that the sequence of parameters $\theta^{(t)}$ defined by an MM algorithm converges at all. Nonetheless, it is often possible to obtain convergence results in specific cases. For pairwise comparisons, using property of stationary point, Ford Jr. (1957) showed that the MM algorithm converges to the unique maximum likelihood estimate under Assumption 4. Hunter (2004) established strict concavity of the likelihood function under Assumption 4 and proved the same result using the Liapounov's theorem. We follow the approach used by Hunter (2004). The following Liapounov's theorem guarantees that the MM algorithm converges to the stationary point of the pseudo log-likelihood objective (4). In Lemma 7, we show that the likelihood function (4) has a unique stationary point, namely the global maximizer. Which concludes that the MM algorithm converges to the unique global optimal solution irrespective of its starting point.

Lemma 6 (Liapounov's theorem) *Suppose* $M : \Omega \to \Omega$ *is continuous and* $\mathcal{L}_{\mathrm{RB}} : \Omega \to \mathbb{R}$ *is differentiable and for all* $\theta \in \Omega$ *we have* $\mathcal{L}_{\mathrm{RB}}(M(\theta)) \geq \mathcal{L}_{\mathrm{RB}}(\theta)$*, with equality only if* θ *is a stationary point of* $\mathcal{L}_{\mathrm{RB}}(\cdot)$*. Then, for arbitrary* $\theta^{(1)} \in \Omega$*, any limit point of the sequence* $\{\theta^{(t+1)} = M(\theta^{(t)})\}_{t \geq 1}$ *is a stationary point of* $\mathcal{L}_{\mathrm{RB}}(\theta)$*.*

Let $\Omega = \{\theta \in \mathbb{R}^d | \sum_{i \in [d]} \theta_i = 0\}$ be the parameter space Ω_b defined in (1) with $b = \infty$. Taking M to be the map implicitly defined by one iteration of the MM algorithm, we have $\mathcal{L}_{\mathrm{RB}}(M(\theta)) \geq \mathcal{L}_{\mathrm{RB}}(\theta)$ from (5). $\mathcal{L}_{\mathrm{RB}}(M(\theta)) = \mathcal{L}_{\mathrm{RB}}(\theta)$ implies that θ is a stationary point which follows from the fact that the differentiable minorizing function Q is a tangent to the log-likelihood $\mathcal{L}_{\mathrm{RB}}(\theta)$ at the current iterate $\theta^{(t)}$. Therefore, $\lim_{t \to \infty} \theta^{(t)}$, defined by the MM algorithm in (6) converges to the stationary point of the pseudo-likelihood objective (4). It remains to prove that $\mathcal{L}_{\mathrm{RB}}(\theta)$ has a unique stationary point, the global maximizer.

Lemma 7 *Under Assumption 4* $\mathcal{L}_{\mathrm{RB}}(\theta)$ *has a unique stationary point.*

Proof First, in Lemma 8, we show that $\mathcal{L}_{\mathrm{RB}}(\theta)$ is an upper compact function under the Assumption 4. $\mathcal{L}_{\mathrm{RB}}(\theta)$ is defined to be upper compact if, for any constant c, the set $\{\theta \in \Omega : \mathcal{L}_{\mathrm{RB}}(\theta) \geq c\}$ is a compact set of the parameter space Ω. Second, in Lemma 9, we show that $\mathcal{L}_{\mathrm{RB}}(\theta)$ is strictly concave. Since upper compactness implies the existence of at least one limit point and strict concavity implies the existence of at most one stationary point, we conclude that $\mathcal{L}_{\mathrm{RB}}(\theta)$ has a unique stationary point. ■

■

Lemma 8 $\mathcal{L}_{\mathrm{RB}}(\theta)$ *defined in Equation* (4) *is an upper compact function of* θ *under the Assumption 4.*

Proof We prove upper compactness following the idea of Hunter (2004). Consider what happens to $\mathcal{L}_{\mathrm{RB}}(\theta)$ when θ approaches the boundary of Ω. If $\tilde{\theta}$ lies on the boundary of Ω, then $\tilde{\theta}_i \to -\infty$ and $\tilde{\theta}_j \to \infty$ for some items i and j. If items are nodes of a directed graph in which edge (i, i') represent that there is at least one user who prefers i over i', then Assumption 4 implies that a directed path exists from i to j. Therefore, there must be some item a with $\tilde{\theta}_a \to -\infty$ which is preferred over item b with $\tilde{\theta}_b > C$, for some constant C. That is there exists an hyper edge e with $a \in T(e)$ and $b \in B(e)$. Which means that for $\theta \in \Omega$, taking limits in

$$\begin{aligned}\mathcal{L}_{\mathrm{RB}}(\theta) &\leq \ln \mathbb{P}_\theta(e) \\ &= \ln\left(\sum_{\sigma\in\Lambda_{T(e)}} \frac{\exp\left(\sum_{c=1}^{|T(e)|}\theta_{\sigma(c)}\right)}{\prod_{u=1}^{|T(e)|}\left(\sum_{c'=u}^{|T(e)|}\exp\left(\theta_{\sigma(c')}\right)+\sum_{i\in B(e)}\exp\left(\theta_i\right)\right)}\right)\end{aligned}$$

gives $\lim_{\theta\to\tilde{\theta}} \mathcal{L}_{\mathrm{RB}}(\theta) = -\infty$. Thus, for any given constant c, the set $\{\theta \in \Omega : \mathcal{L}_{\mathrm{RB}}(\theta) \geq c\}$ is a closed and bounded set, and hence a compact set. ■

Lemma 9 *$\mathcal{L}_{\mathrm{RB}}(\theta)$ defined in Equation (4) is strictly concave in θ.*

Proof To prove strict concavity, we use Lemma 2. Define $\widetilde{\Omega} = \{\theta \in \mathbb{R}^d | \theta_1 = 0\}$, a reparameterization of the set $\Omega = \{\theta \in \mathbb{R}^d | \sum_{i\in[d]} \theta_i = 0\}$. To apply Lemma 2 to prove strict concavity of log-likelihood of an hyper edge e, take $g_{ij}(\theta, Y) = (\theta_i + Y_i) - (\theta_j + Y_j)$, for all $i \in T(e)$ and $j \in B(e)$. Consider $\theta, \tilde{\theta} \in \widetilde{\Omega}$. $H(\theta) = H(\tilde{\theta})$ implies that $\theta_i - \theta_j = \tilde{\theta}_i - \tilde{\theta}_j$, for all $i \in T(e)$ and $j \in B(e)$. This follows from the fact that for a fixed parameter θ, the hyper planes $\{g_{ij}(\theta, Y) \geq 0\}_{ij}$ are linearly independent. Thus, Assumption 4 combined with the fact that $\theta_1 = \tilde{\theta}_1$ means that $\theta = \tilde{\theta}$. Since the Gumbel distribution has strictly logarithmic concave density function, we conclude that $\mathcal{L}_{\mathrm{RB}}(\theta)$ is strictly concave. ■

3.4 Consistency of pseudo-MLE of rank-breaking graph

In order to discuss consistency of the proposed approach, we need to specify how we sample the set of items to be offered S_j and also which partial ordering over S_j is to be observed. Here, we consider a simple but canonical scenario for sampling ordered relations, and show the proposed method is consistent for all non-degenerate cases. Later, in Section 4, we study a more general sampling scenario, when we analyze the order-M estimator in the finite sample regime.

We define a canonical sampling scenario in the following. There is a set of ℓ integers $(m_1, \ldots, m_\ell)$ whose sum is strictly less than d. A new arriving user is presented with all d items and is asked to provide her top m_1 items as an unordered set, and then the next m_2 items, and so on. This is sampling from the PL model and observing an ordered partition with $(\ell + 1)$ subsets of sizes m_a's, and the last subset includes all remaining items. We apply the generalized rank-breaking to get rank-breaking graphs $\{G_j\}$ with ℓ edges each, and order-M estimate is computed. We show that this is consistent, i.e. asymptotically unbiased in the limit of the number of users n.

Remark 10 *Under the* PL *model and the above sampling scenario, the order-M rank-breaking estimate $\widehat{\theta}$ in (4) is consistent for all choices of $M \geq \min_{a\in\ell} m_a$.*

Proof It is sufficient to show that (a) the estimate $\widehat{\theta}$, (4) is unique under the above sampling scenario, and (b) expectation of the gradient of $\mathcal{L}_{\mathrm{RB}}(\theta^*)$ is zero, i.e, $\mathbb{E}_{\theta^*}[\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)] = 0$, (Azari Soufiani et al., 2013). For the above sampling scenario in the limit of the number of users n, Assumption 4 is satisfied. Therefore, from Lemma 7, the estimate $\widehat{\theta}$, (4) is unique. In Lemma 13, we show that $\mathbb{E}_{\theta^*}[\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)] = 0$. We would like to mention that the Lemma 13 crucially relies on the memoryless property of the PL model.

■

3.5 Algorithm to estimate $\widehat{\theta}$ given DAG $\mathcal{G}_j$'s

Summarizing the rank-breaking approach explained in the previous sections, we give Algorithm 3, an algorithm to compute $\widehat{\theta}$, (4), an estimate of θ^*. Algorithm 3 takes as input DAG $\mathcal{G}_j$'s generated under PL model with parameter θ^*, rank-breaking order $M \in \mathbb{Z}^+$, a desired error threshold ϵ, and returns $\widehat{\theta}$.

Algorithm 3 Estimate θ^* given DAG $\mathcal{G}_j$'s.

Require: DAG $\{\mathcal{G}_j\}_{1\leq j\leq n}$ generated under PL model with parameter θ^*, rank-breaking order M, error threshold ϵ
Ensure: $\widehat{\theta}$ - an estimate of θ^*
find maximal ordered partitions $\{\mathcal{P}_j\}_{1\leq j\leq n}$ consistent with $\{\mathcal{G}_j\}_{1\leq j\leq n}$ [Algorithm 1]
construct rank breaking graph $\{G_j(S_j, E_j)\}_{1\leq j\leq n}$ from $\{\mathcal{P}_j\}_{1\leq j\leq n}$ [Algorithm 2]
$\widehat{\theta} \leftarrow \mathbf{0}_{d\times 1}$
repeat
  $\widetilde{\theta} \leftarrow \widehat{\theta}$
  for $i = 1$ to d **do**
    $\widehat{\theta}_i \leftarrow$ from minorizing maximizing Equation (6) using $\widetilde{\theta}$, $\{G_j(S_j, E_j)\}_{1\leq j\leq n}$, M
  end for
until $\|\widehat{\theta} - \widetilde{\theta}\|_\infty \leq \epsilon$
return $\widehat{\theta}$

4. Analysis of the Algorithm

We first summarize the notations defined so far and introduce some new notations that are used in our theoretical results. We define a *comparison graph* that captures the topology of the offer sets S_j. Our upper and lower bounds both depend on the spectral properties of the comparison graph. Then, we present main theoretical analyses and numerical simulations confirming the theoretical results.

Notations. Following is a summary of all the notations defined above. Also, we introduce some new notations that are used in our theoretical results. We use n to denote the number of users providing partial rankings, indexed by $j \in [n]$ where $[n] = \{1, 2, \ldots, n\}$. We use d to denote the number of items, indexed by $i \in [d]$. Given rank-breaking graphs $\{G_j(S_j, E_j)\}_{j\in[n]}$ extracted from the DAGs $\{\mathcal{G}_j\}$, we first define the order M rank-breaking graphs $\{G_j^{(M)}(S_j, E_j^{(M)})\}$, where $E_j^{(M)}$

is a subset of E_j that includes only those edges $e_j \in E_j$ with $|T(e_j)| \leq M$. This represents those edges that are included in the estimation for a choice of M. For finite sample analysis, the following quantities capture how the error depends on the topology of the data collected. Let $\kappa_j \equiv |S_j|$ and $\ell_j \equiv |E_j^{(M)}|$. We index each edge e_j in $E_j^{(M)}$ by $a \in [\ell_j]$ and define $m_{j,a} \equiv |T(e_{j,a})|$, size of top-set, for the a-th hyper edge of the j-th rank-breaking graph, and $r_{j,a} \equiv |T(e_{j,a})| + |B(e_{j,a})|$, sum of size of the top-set and the bottom-set. We let $p_j \equiv \sum_{a\in[\ell_j]} m_{j,a}$ denote the effective sample size for the observation $G_j^{(M)}$.

$$m_{j,a} \equiv |T(e_{j,a})|\,, \quad \text{size of top-set for the } e_{j,a} \text{ hyper edge of rank-breaking graph } G_j^{(M)}. \tag{7}$$

$$r_{j,a} \equiv |T(e_{j,a})| + |B(e_{j,a})|\,, \quad \text{sum of size of the top-set and the bottom-set for the } E_{j,a}. \tag{8}$$

$$p_j \equiv \sum_{a\in[\ell_j]} m_{j,a}\,, \quad \text{sum of size of all top-sets of } G_j^{(M)} \text{ (which are smaller than } M). \tag{9}$$

Notice that although we do not explicitly write the dependence on M, all of the above quantities implicitly depend on the choice of M. For ease of notations, we remove the superscript M from $G_j^{(M)}$ in the following.

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 eigenvalues of X sorted in increasing order. Let $\mathrm{Tr}(X) = \sum_{i=1}^d \lambda_i(X)$ denote trace of X and $\|X\| = \max\{|\lambda_1(X)|, |\lambda_d(X)|\}$ denote spectral norm of X. 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.

4.1 Comparison graph

We define a comparison graph $\mathcal{H}([d], E)$ as a weighted undirected graph with weights

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

That is we put an edge (i, i') if there exists a user j whose offerings is a set S_j such that $i, i' \in S_j$. Define a diagonal matrix $D = \mathrm{diag}(A\mathbf{1})$, and the corresponding graph Laplacian $L = D - A$ such that

$$L \equiv \sum_{j=1}^{n} \frac{p_j}{\kappa_j(\kappa_j - 1)} \sum_{i<i'\in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top\,. \tag{10}$$

It is immediate that $\lambda_1(L) = 0$ with $\mathbf{1}$ as the eigenvector. There are remaining $d-1$ eigenvalues that sum to $\mathrm{Tr}(L) = \sum_j p_j$. The rescaled $\lambda_2(L)$ and $\lambda_d(L)$ capture the dependency on the topology:

$$\alpha \equiv \frac{\lambda_2(L)(d-1)}{\mathrm{Tr}(L)}\,, \quad \beta \equiv \frac{\mathrm{Tr}(L)}{\lambda_d(L)(d-1)}\,. \tag{11}$$

In an ideal case where the graph is well connected, then the spectral gap of the Laplacian is large. The chosen rescaling ensures that if all the non-zero eigenvalues are of the same order then there

exists constants $0 \leq c_1, c_2 \leq 1$ such that $c_1 \leq \alpha \leq 1$ and $c_2 < \beta \leq 1$. If the graph is connected then c_1 is strictly greater than zero. If $\lambda_2(L) = \cdots = \lambda_d(L)$ then $\alpha = \beta = 1$. We will show that the performance of our estimator depends upon topology of the comparison graph through these two parameters. The larger the rescaled spectral gap α the smaller the error we get with the same effective sample size. The rescaled largest eigenvalue β along with α determine how many samples are required for the analysis to hold. In general, α and β depend upon both the topology of the offer sets S_j and the topology of the rank-breaking graphs G_j, through the edge weights $A_{ii'}$. However, if topology of all the rank-breaking graphs G_j's is same then the comparison graph $\mathcal{H}$ and α, β depend only upon the topology of the offer sets S_j. For such a comparison graph, Khetan and Oh (2016) provides a detailed discussion on the spectral gap for various canonical graphs following the setup given in Shah et al. (2015a).

The concavity of $\mathcal{L}_{\mathrm{RB}}(\theta)$ also depends on the following quantities.

$$\gamma_1 \equiv \min_{j,a} \left\{ \left(\frac{r_{j,a} - m_{j,a}}{\kappa_j} \right)^{2e^{2b}-2} \right\}, \quad \gamma_2 \equiv \min_{j,a} \left\{ \left(\frac{r_{j,a} - m_{j,a}}{r_{j,a}} \right)^{2} \right\}. \tag{12}$$

The parameter γ_1 incorporates asymmetry in probabilities of items being ranked at different positions depending upon their weight θ^*_i. Recall that b is the upper bound on $\|\theta^*\|_\infty$, Equation 1. The parameter γ_1 is 1 for $b = 0$ that is when all the items have the same weight θ^*_i, and it decreases exponentially with increase in b. The exponential decrease is tight and reflects the fact that under PL model probability of the highest weight item being ranked last is exponentially smaller than its probability of being ranked first.

Suppose the number of items d grows to infinity and the rank-breaking graphs G_j's are determined such that size of the offered subsets κ_j's are increasing with d that is $\kappa_j = \Theta(d)$. If all the top-set sizes are much smaller than the size of the rank-breaking edge such that, $m_{j,a} = o(r_{j,a})$ and $r_{j,a} = \Theta(\kappa_j)$, then for $b = O(1)$, γ_1 can be made arbitrarily close to one, for large enough problem size d. On the other hand, when either $r_{j,a}$ is much smaller than κ_j or if $r_{j,a} = \Theta(\kappa_j)$ but $r_{j,a} - m_{j,a} = O(1)$ then γ_1 can be arbitrarily close to zero and accuracy of the parameter estimation can degrade significantly as stronger alternatives will have small chance of showing up in the bottom set. The value of γ_1 is quite sensitive to b. The parameter γ_2 controls the range of the size of the top-set with respect to the size of the bottom-set for which the error decays with the rate of 1/(size of the top-set). It does not depend upon the size of of the rank-breaking edge, $r_{j,a}$, in comparison to the offer set size κ_j. It only depends upon the size of the top sets $m_{j,a}$ in comparison to the size of the rank breaking edge $r_{j,a}$. If size of top sets $m_{j,a} = o(r_j, a)$ then γ_2 would be close to one. The dependence of accuracy on γ_1, γ_2 is demonstrated in simulations in Figure 5.

We define the following additional quantities that control our upper bound. The dependence in γ_3 and ν are due to weakness in the analysis, and ensures that the Hessian matrix is strictly negative definite.

$$\gamma_3 \equiv 1 - \max_{j,a} \left\{ \frac{4e^{16b}}{\gamma_1} \frac{m^2_{j,a} r^2_{j,a} \kappa^2_j}{(r_{j,a} - m_{j,a})^5} \right\}, \quad \nu \equiv \max_{j,a} \left\{ \frac{m_{j,a} \kappa^2_j}{(r_{j,a} - m_{j,a})^2} \right\}. \tag{13}$$

For our analysis to hold we need $\gamma_3 > 0$ which in addition to the conditions needed for γ_1 being close to one require that $m_{j,a} \leq c\sqrt{r_{j,a}}$, for a sufficiently small positive constant c. We believe this is a limitation on our analysis and the results should hold for any values of $m_{j,a} = o(r_{j,a})$. For

the special case when $m_{j,a} \leq 3$ for all j, a, we provide a tighter result that does not depend upon γ_3. However, in general getting rid of γ_3 is challenging. ν shows up in the number of samples required for our analysis to hold. Note that the quantities defined in this section implicitly depend on the choice of M, which controls the necessary computational power, via the definition of the rank-breaking graphs $\{G_j^{(M)}\}_{j\in[n]}$.

4.2 Upper bound on the achievable error

We provide an upper bound on the error for the order-M rank-breaking Algorithm 3, showing the explicit dependence on the topology of the offered sets $\{S_j\}_{j\in[n]}$. Recall from the sampling assumptions in Section 3 that we assume the topology of the observed DAG $\mathcal{G}_j$'s, and the rank-breaking order M is fixed apriori. The randomness that we observe is in the position of S_j items in the DAG $\mathcal{G}_j$. For an hyper edge $e \in E_j$, the randomness is in which items of the set S_j appear in the bottom $|B(e)|$ positions and the bottom $|T(e)| + |B(e)|$ positions in the preference order of the user j. This precisely captures the randomness due to the PL model in the observed DAG $\mathcal{G}_j$. The following theorem provides an upper bound on the achieved error, and a proof is provided in Section 7.

Theorem 11 *Suppose there are n users, d items parametrized by $\theta^* \in \Omega_b$, and each user $j \in [n]$ is presented with a set of offerings $S_j \subseteq [d]$ and the user provides a partial ordering under the PL model consistent with the topology of the apriori fixed DAG $\mathcal{G}_j$. For a choice of $M \in \mathbb{Z}^+$, if $\gamma_3 > 0$ and the effective sample size $\sum_{j=1}^n p_j$ is large enough such that*

$$\sum_{j=1}^{n} p_j \;\geq\; \frac{2^{14} e^{20b} \nu^2}{(\alpha\gamma_1\gamma_2\gamma_3)^2 \beta} \frac{p_{\max}}{\kappa_{\min}} d \log d \,, \tag{14}$$

where $b \equiv \max_i |\theta_i^*|$ is the dynamic range, $p_{\max} = \max_{j\in[n]} p_j$, $\kappa_{\min} = \min_{j\in[n]} \kappa_j$, α is the (rescaled) spectral gap, β is the (rescaled) spectral radius in (11)*, and γ_1, γ_2, γ_3, and ν are defined in* (12) *and* (13)*, then the generalized rank-breaking estimator in* (4) *achieves*

$$\frac{1}{\sqrt{d}} \|\widehat{\theta} - \theta^*\|_2 \;\;\leq\;\; \frac{40 e^{7b}}{\alpha\gamma_1\gamma_2^{3/2}\gamma_3} \sqrt{\frac{d \log d}{\sum_{j=1}^n p_j}} \,, \tag{15}$$

with probability at least $1 - 3e^3 d^{-3}$. Moreover, for $M \leq 3$ the above bound holds with γ_3 replaced by one, giving a tighter result.

Note that the dependence on the choice of M is not explicit in the bound, but rather is implicit in the construction of the comparison graph and the number of effective samples.

Suppose the number of items d is large enough and the size of the offered subsets κ_j's and the size of the rank breaking edges $r_{j,a}$'s are increasing with d, that is there exists positive constants c_1, c_2 such that $\kappa_j \geq c_1 d$, and $r_{j,a} \geq c_2 \kappa_j$. Then for $b = O(1)$ there exists a universal constant c_3 such that if top-set sizes $m_{j,a} \leq c_3 (r_{j,a})^{1/3}$ then there exists constants $0 < c_4, c_5, c_6 \leq 1$ such that $c_4 \leq \gamma_1 < 1$, $c_5 \leq \gamma_2 < 1$ and $c_6 \leq \gamma_3 < 1$. Further, if the comparison graph $\mathcal{H}$ is well connected then there exists a constant $0 < c_7, c_8 \leq 1$ such that the rescaled spectral gap $c_7 \leq \alpha \leq 1$ and rescaled largest eigenvalue $c_8 \leq \beta \leq 1$. In this ideal case, the condition on the effective sample size is met with $\sum_j p_j = O(d \log d)$, (14). Therefore the effective sample size $\sum_{j=1}^n p_j = \Omega(d \log d)$ is sufficient to

ensure $\|\widehat{\theta} - \theta^*\|_2 = o(\sqrt{d})$ which is only a logarithmic factor larger than the number of parameters. We need $m_{j,a} \leq c_3 (r_{j,a})^{1/3}$ to satisfy $(\nu^2 p_{\max})/\kappa_{\min} = O(1)$, otherwise $m_{j,a} \leq c_3 (r_{j,a})^{1/2}$ is sufficient to ensure $\gamma_3 > 0$. We believe that dependence in γ_3 is weakness of our analysis and there is no dependence as long as $m_{j,a} < r_{j,a}$. For, rank-breaking order $M \leq 3$, we are able to give tighter results where there is no dependence on γ_3.

As explained above, in the ideal case, for large enough problem size d, there exists a positive constant C such that $\|\widehat{\theta} - \theta^*\|_2^2 \leq C d^2 \log d / (\sum_{j=1}^n p_j)$. Recall from the construction of the likelihood objective function, $\mathcal{L}_{\mathrm{RB}}(\theta) = \sum_{j \in [n]} \sum_{e \in E_j : |T(e)| \leq M} \ln \mathbb{P}_\theta(e)$. If we fix all the problem parameters including topology of the DAG $\mathcal{G}_j$'s and increase M then $p_j = \sum_{a \in [\ell_j]} m_{j,a}$ increases. Therefore, by increasing M we can get the same number of effective samples $\sum_{j=1}^n p_j$ with smaller number of rankings n. However, increasing M increases computational complexity as $M!$. Therefore, to achieve a fixed target accuracy $\|\widehat{\theta} - \theta^*\|_2$, an analyst can trade-off the required number of rankings with the budgeted computational complexity.

If the DAG $\mathcal{G}_j$'s are complete graph that is each user provides a full ranking over the offered subset S_j, we get $m_{j,a} = 1$, $\ell_j = \kappa_j - 1$, and the total effective sample size $\sum_j p_j = \sum_{j \in [n]} (\kappa_j - 1)$. Therefore, from the above theorem, $\sum_{j \in [n]} (\kappa_j - 1) = \Omega(d \log d)$ is sufficient to ensure $\|\widehat{\theta} - \theta^*\|_2 = o(\sqrt{d})$. It matches with the results for full rankings given in Hajek et al. (2014); Khetan and Oh (2016).

Unordered vs. ordered top-m ranking. In the ideal case, a perhaps surprising observation is that, for a ranking j, sizes of the top-sets $\{m_{j,a}\}_{a \in [\ell_j]}$ impacts estimation accuracy only via $p_j = \sum_{a \in [\ell_j]} m_{j,a}$, when $m_{j,a}$'s are sufficiently small in comparison to $r_{j,a}$'s, sum of the top-set size and the bottom-set size. In particular, for estimation accuracy it does not matter whether users reveal their top-m choices in the ordered way $\{i_1\} \succ \{i_2\} \succ \cdots \succ \{i_m\} \succ \{i_{m+1}, \cdots, i_k\}$ or the unorderd way $\{i_1, i_2, \cdots, i_m\} \succ \{i_{m+1}, \cdots i_k\}$, when m is sufficiently small in comparison to k. Numerical results in Figure 5 confirm this.

Proof idea. The analysis of the optimization in (4) shows that, with high probability, $\mathcal{L}_{\mathrm{RB}}(\theta)$ is strictly concave with $\lambda_2(H(\theta)) \leq -C_b \gamma_1 \gamma_2 \gamma_3 \lambda_2(L) < 0$ for all $\theta \in \Omega_b$ (Lemma 15), and the gradient is also bounded with $\|\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)\| \leq C_b' \gamma_2^{-1/2} (\sum_j p_j \log d)^{1/2}$ (Lemma 14). This leads to Theorem 11:

$$\|\widehat{\theta} - \theta^*\|_2 \;\leq\; \frac{2\|\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)\|}{-\lambda_2(H(\theta))} \;\leq\; C_b'' \frac{\sqrt{\sum_j p_j \log d}}{\gamma_1 \gamma_2^{3/2} \gamma_3 \lambda_2(L)} \,,$$

where C_b, C_b', and C_b'' are constants that only depend on b, and $\lambda_2(H(\theta))$ is the second largest eigenvalue of a negative semidefinite Hessian matrix $H(\theta)$ of $\mathcal{L}_{\mathrm{RB}}(\theta)$. Recall that $\theta^\top \mathbf{1} = 0$ since we restrict our search in Ω_b. Hence, the error depends on $\lambda_2(H(\theta))$ instead of $\lambda_1(H(\theta))$ whose corresponding eigenvector is the all-ones vector.

4.3 Lower bound on computationally unbounded estimators

Suppose $M = d$. We prove a fundamental lower bound on the achievable error rate that holds for any *unbiased* estimator with no restrictions on the computational complexity. For each (j, a),

define $\eta_{j,a}$ as

$$\eta_{j,a} \equiv \sum_{u=0}^{m_{j,a}-1} \Big(\frac{1}{r_{j,a}-u} + \frac{u(m_{j,a}-u)}{m_{j,a}(r_{j,a}-u)^2} \Big) + \sum_{u<u' \in [m_{j,a}-1]} \frac{2u}{m_{j,a}(r_{j,a}-u)} \frac{m_{j,a}-u'}{r_{j,a}-u'} \quad (16)$$

$$< \sum_{u=0}^{m_{j,a}-1} \Big(\frac{1}{m_{j,a}-u} + \frac{u}{m_{j,a}(m_{j,a}-u)} \Big) + \sum_{u<u' \in [m_{j,a}-1]} \frac{2u}{m_{j,a}(m_{j,a}-u)} \quad (17)$$

$$= \sum_{u=0}^{m_{j,a}-1} \Big(\frac{1}{m_{j,a}-u} + \frac{u}{m_{j,a}(m_{j,a}-u)} + \frac{2u(m_{j,a}-1-u)}{m_{j,a}(m_{j,a}-u)} \Big) = m_{j,a}\,,$$

where (17) follows from the fact that (16) is monotonically strictly decreasing in $r_{j,a}$ for $r_{j,a} \geq m_{j,a}$. Since by definition $r_{j,a} > m_{j,a}$, we substitute $r_{j,a} = m_{j,a}$ to get a strict upper bound.

Theorem 12 *Let $\mathcal{U}$ denote the set of all unbiased estimators of θ^* that are centered such that $\widehat{\theta}\mathbf{1} = 0$, and let $\mu = \max_{j\in[n],a\in[\ell_j]}\{m_{j,a} - \eta_{j,a}\}$. For all $b > 0$,*

$$\inf_{\widehat{\theta}\in\mathcal{U}} \sup_{\theta^*\in\Omega_b} \mathbb{E}[\|\widehat{\theta} - \theta^*\|^2] \geq \max\left\{ \frac{(d-1)^2}{\sum_{j=1}^n \sum_{a=1}^{\ell_j} (m_{j,a} - \eta_{j,a})}\,,\ \frac{1}{\mu}\sum_{i=2}^{d} \frac{1}{\lambda_i(L)} \right\}. \quad (18)$$

The proof relies on the Cramer-Rao bound and is provided in Section 7.6. Since $0 < \eta_{j,a} < m_{j,a}$, the mean squared error is lower bounded by $(d-1)^2/(\sum_{j=1}^n \sum_{a=1}^{\ell_j} m_{j,a}) = (d-1)^2/(\sum_{j=1}^n p_j)$, where $\sum_{j=1}^n p_j$ is the effective sample size. Comparing it to the upper bound in (15), this is tight up to a logarithmic factor when (a) the topology of the data is well-behaved such that all the quantities $\gamma_1, \gamma_2, \gamma_3, \alpha, \beta$ are greater than a positive constant $c \leq 1$; and (b) there is no limit on the computational power and M can be made as large as we need. For full-rankings, this bound reduces to the one given in Hajek et al. (2014); Khetan and Oh (2016). For full rankings, $\sum_{a=1}^{\ell_j}(m_{j,a} - \eta_{j,a}) = (\kappa_j - 1)^2/\kappa_j$.

The bound in Eq. (18) further gives a tighter lower bound, capturing the dependency in $\eta_{j,a}$'s and $\lambda_i(L)$'s. The second term in (18) implies we get a tighter bound when $\lambda_2(L)$ is smaller. If the comparison graph $\mathcal{H}$ is disconnected that is $\lambda_2(L) = 0$, the bound shows that θ^* can not be estimated.

To understand the impact of $\eta_{j,a}$ on MSE, we plot $(m_{j,a} - \eta_{j,a})/r_{j,a}$ as a function of $m_{j,a}/r_{j,a}$ for different values of $r_{j,a}$ in Figure 3. Recall that $m_{j,a}$ is the size of the top-set, (7) and $r_{j,a}$ is the sum of size of the top-set and the bottom-set, (8). We vary $m_{j,a}$ from 1 to $r_{j,a} - 1$, for $r_{j,a}$ in $\{2, 4, 8, 16, 32, 256, 1024\}$. From the Theorem 12, contribution of an hyper edge $e_{j,a}$ to the effective samples is $(m_{j,a} - \eta_{j,a})$. Since $\eta_{j,a}$ increases with $m_{j,a}$, a natural question is what is the optimal value of $m_{j,a}$ that gives the smallest MSE, for a fixed $r_{j,a}$. Figure 3 shows that $(m_{j,a} - \eta_{j,a})/r_{j,a}$ achieves its maximum value at $m/r \approx 0.8$ when r is sufficiently large. It also shows that $(m_{j,a} - \eta_{j,a}) \geq c\, m_{j,a}$, for $m_{j,a}/r_{j,a} \leq c_1 (\approx 0.8)$, for positive constants $c, c_1 < 1$, when $r_{j,a}$ is large. That is the contribution of an hyper edge $e_{j,a}$ to the effective sample size is at least $c\, m_{j,a}$ for $m_{j,a}/r_{j,a} \leq c_1$. Comparing this with the lower bound for top-$m_{j,a}$ ranking given in Khetan and Oh (2016), it can be concluded that the (unobserved) relative ordering among the items in the top-set of the hyper edge $e_{j,a}$ has limited impact on the MSE. Khetan and Oh (2016) show in the their lower bound that the contribution of top-m ranking on the effective sample size is m.

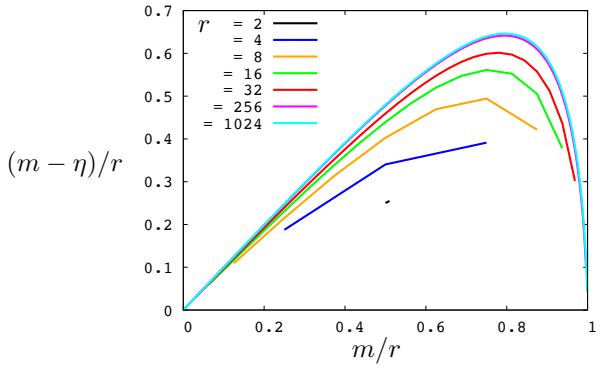

Figure 3: It shows how η varies as a function of m, size of the top-set, for a fixed value of r, sum of top-set and the bottom-set sizes, Equation (16).

Note that the lower bound is derived for the easiest case, $b = 0$, when all the items $i \in [d]$ have the same weight $\theta_i^* = \theta_0^*$. Therefore, the above conclusion that the relative ordering among the items in the top-set of the hyper edge $e_{j,a}$ has limited impact on the accuracy can be made only for this case when all the items have the same PL weight. However, the upper bound shows that this conclusion holds true in general. The 'unordered vs. ordered top-m ranking' paragraph in the previous section explains that for the ideal case when $m_{j,a}$ is sufficiently small in comparison to $r_{j,a}$, the relative ordering has limited impact.

Recall that in the upper bound, γ_1 and γ_2 capture the impact of $m_{j,a}/r_{j,a}$ on the effective number of samples. However, for $b = 0$, $\gamma_1 = 1$, and for $b > 0$ it captures asymmetry in the probability of the highest weight item appearing in bottom set. $\gamma_2 = \min_{j,a} \left\{ \left(\frac{r_{j,a} - m_{j,a}}{r_{j,a}} \right)^2 \right\}$ captures the role played by $\eta_{j,a}$ in the lower bound.

5. Numerical results

We provide extensive numerical results on simulated and real-world datasets confirming our theoretical results and performance gains of the generalized rank-breaking algorithm over the pairwise rank-breaking algorithm.

5.1 Simulated datasets

In the following, we give numerical results confirming our theoretical results. Our numerical experiments show that the dependence of MSE on $n, d, \kappa_j, r_{j,a}, m_{j,a}, \ell_j$ as given in Theorem 11, Equation (15) holds true, even when the conditions for the theorem to hold are not met. For the theorem to hold, it is required that the number of items d, the set sizes κ_j and the hyper edge sizes $r_{j,a}$ are sufficiently large such that $\gamma_3 > 0$ and the number of effective samples satisfies (14). However, in all our experiments the number of items $d <= 512$ and $b = 2$, therefore from (13) $\gamma_3 < 0$, and also the condition in (14) is not met.

In particular, in our numerical setting $\gamma_1 =$

Impact of the number of independent rankings n and the number of rank-breaking hyper edges ℓ_j on accuracy. Figure 4 (first panel) shows the accuracy-sample tradeoff for

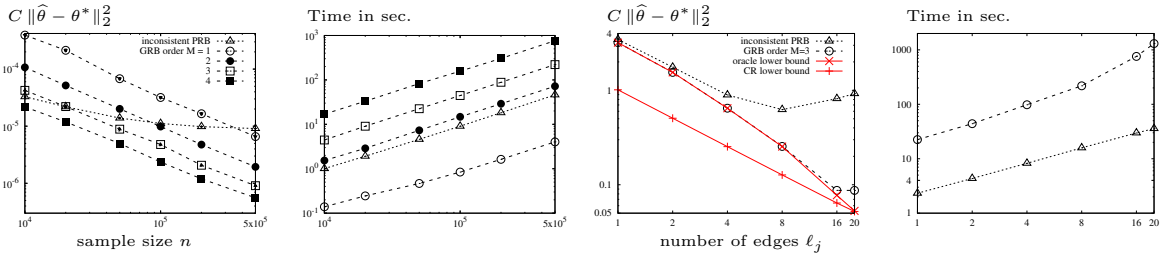

Figure 4: On the first panel, we fix $d = 256$, $\kappa = 32$, $\ell = 4$, $m_a = a$ for $a \in \{1, 2, 3, 4\}$, and sample posets from the canonical scenario explained in Section 3.4. On the third panel, we let $n = 10^5$, $d = 512$, $\kappa = 64$, $m_a = 3$ for all $a \in [\ell]$ and vary $\ell \in \{1, 2, 4, 8, 16\}$. The second and the fourth panel show the computation time for the first and the third panel respectively. The PL weights are chosen uniformly spaced over $[-2, 2]$. Smaller error is achieved when using more computational resources with larger M and using all paired comparisons results in an inconsistent Pairwise Rank-Breaking (PRB) whose error does not vanish with sample size (first panel). Generalized Rank-Breaking (GRB) utilizes all the observations achieving the oracle lower bound (third panel).

increasing computation M on the same data. As predicted by the anlaysis, generalized rank-breaking (GRB) is consistent (Remark 10) and the mean square error (MSE) decays at the rate $(1/n)$, and decreases with increase in M, order of rank-breaking (Theorem 11). For comparison, we also plot the MSE achieved by pairwise rank-breaking (PRB) approach where we include all paired relations derived from data, which we call *inconsistent PRB*. As predicted by Azari Soufiani et al. (2014), this results in an inconsistent estimate, whose MSE does not vanish as we increase the sample size. Notice that including all paired comparisons increases bias, but also decreases variance of the estimate. Hence, when sample size is limited and variance is dominating the bias, it is actually beneficial to include those biased paired relations to gain in variance at the cost of increased bias. Theoretical analysis of such a bias-variance tradeoff is outside the scope of this paper, but proposes an interesting research direction.

In the third panel, the GRB with $M = 3$ achieves decreasing MSE, whereas for PRB the increased bias dominates the MSE. For comparisons, we provide the error achieved by an oracle estimator who knows the exact ordering among those items belonging to the top-sets and runs MLE. For example, if $\ell = 2$, the GRB observes an ordering $(\{i_1, i_2, i_4, i_5, \ldots\} \prec \{i_{17}, i_3, i_6\} \prec \{i_9, i_2, i_{11}\})$ whereas the oracle estimator has extra information on the ordering among those top sets, i.e. $(\{i_1, i_2, i_4, i_5, \ldots\} \prec i_{17} \prec i_3 \prec i_6 \prec i_9 \prec i_2 \prec i_{11}\})$. Perhaps surprisingly, GRB is able to achieve a similar performance without this significant extra information, unless ℓ is large. The performance degradation in large ℓ regime stems from the fact that the ratio of m_a and r_a approaches 1 for a close to ℓ when ℓ is large. Therefore the parameters γ_1 and γ_2 become small, and the upper bound MSE increases consequentially. The normalization constant C is $1/d^2$ for the first panel and nm/d^2 for the third panel. All the numerical results in this paper are averaged over 10 instances. Standard error is very small in all the results, therefore we do not give error bars, except in the first panel in Figure 4.

Impact of the top-set size m and the set-size κ on accuracy. In Figure 5, the first and the third panel, we compare performance of our algorithm with pairwise breaking, Cramer Rao

lower bound and oracle MLE lower bound. Oracle MLE knows relative ordering of items in the top-sets $T(e)$ and hence is strictly better than the GRB. For the settings chosen, Oracle MLE gets the ordered ranking of top-m items whereas GRB gets unordered top-m items. As predicted by our analysis, GRB matches with the oracle MLE which means relative ordering of top-m items among themselves is statistically insignificant when m is sufficiently small in comparison to $r = \kappa$. For $r = \kappa = 32$ in the first panel, MSE decays as m increases from 1 to 5. However, when $r = \kappa = 16$ in the third panel, for the same increase of m from 1 to 5 MSE starts increasing when m grows beyond 4. The reason is that the quantities γ_1 and γ_2 get smaller as m increases, and the upper bound increases consequently. The normalization constant C is n/d^2 for these two panels.

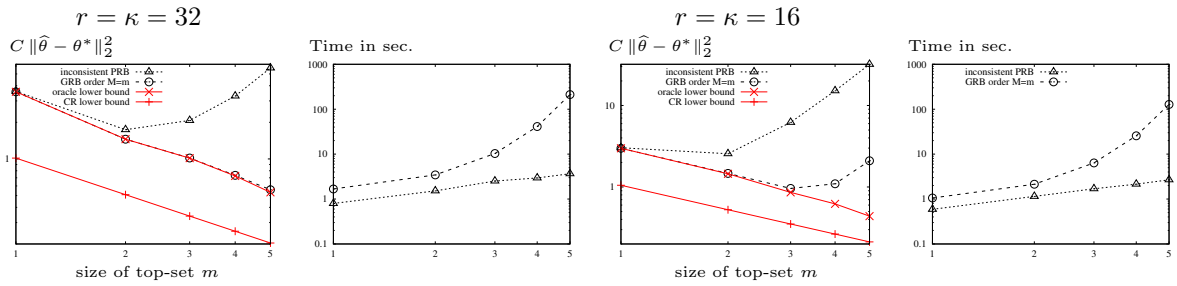

Figure 5: PRB: pairwise rank-breaking, GRB: generalized rank-breaking. θ^* is chosen uniformly spaced over $[-2, 2]$ and $d = 512$, $n = 10^5$ and number of hyperedges $\ell = 1$. The second and the fourth panel show the computation time for the first and the third panel respectively. MSE decreases as m increases when r, sum of the size of the top-set and the bottom-set is sufficiently large (first panel). When r is small, with increase in m MSE initially decreases but as m grows large MSE starts increasing (third panel).

Impact of the dynamic range b on accuracy. In Figure 6, we show the impact of b and $r = \kappa$ on the accuracy for fixed $m = 4$. When κ is small, γ_2 is small, and hence error is large; when b is large γ_1 is exponentially small, and hence error is significantly large. This is different from learning Mallows models in Ali and Meilă (2012) where peaked distributions are easier to learn, and is related to the fact that we are not only interested in recovering the (ordinal) ranking but also the (cardinal) weight. The normalization constant C is nm/d^2.

5.2 Real-world datasets

On sushi preferences (Kamishima, 2003) and jester dataset (Goldberg et al., 2001), we improve over pairwise breaking and achieve same performance as the oracle MLE.

Sushi dataset. There are $d = 100$ types of sushi. Full rankings over subsets S_j of size $\kappa = 10$ are provided by $n = 5000$ individuals. The offering subsets S_j are chosen uniformly at random from the entire set d. We set the ground truth θ^* to be the MLE of the PL weights over the entire data. In the left panel of Figure 7, for each $m \in \{3, 4, 5, 6\}$, we remove the known ordering among the top-m and bottom-$(10 - m)$ sushi in each set, and run our estimator with one rank-breaking hyper edge between top-m and bottom-$(10 - m)$ items. We compare our algorithm with inconsistent pairwise breaking (using optimal choice of parameters from Khetan and Oh (2016)) and the oracle MLE. For $m \leq 6$, the proposed rank-breaking performs as good as the oracle who knows the relative

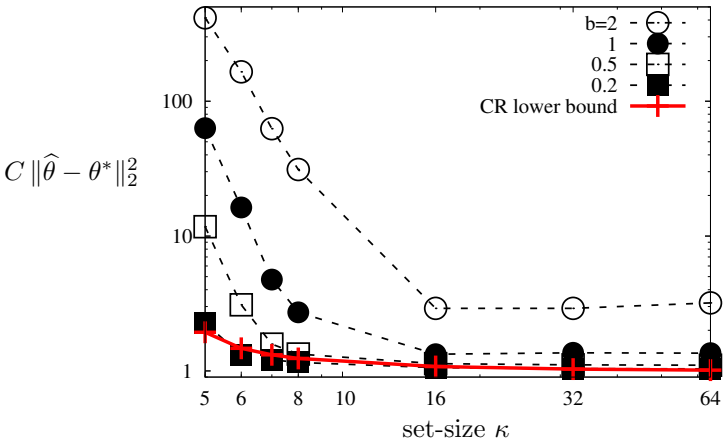

Figure 6: $d = 512$, $n = 10^5$ and θ^* is chosen uniformly spaced over $[-2, 2]$. Number of hyper edges $\ell = 1$ with $r = \kappa$ and $m = 4$. MSE increases as the dynamic range b gets large.

ordering among the top m items. In other words, an individual providing a set of ordered top-6 sushi or a set of unordered top-6 sushi statistically reveals the same information, for the purpose of estimating the ground truth parameters. As predicted by our theory, error decreases with increase in top-set size m.

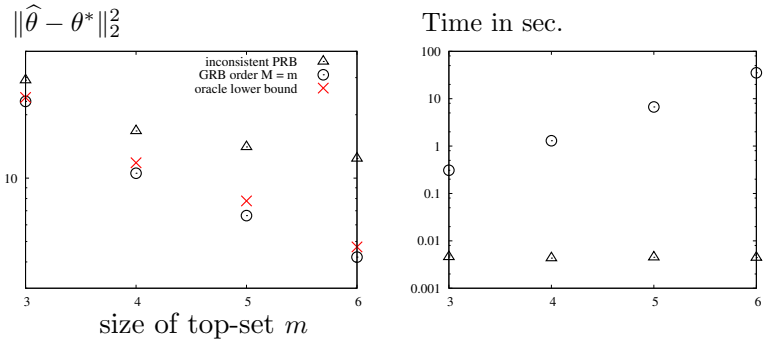

Figure 7: The sushi dataset has $d = 100$, $n = 5000$, and $\kappa = 10$. The right panel shows computation time. Generalized rank-breaking improves over pairwise RB and performs as good as oracle MLE on sushi dataset.

Jester dataset. It consists of continuous ratings between -10 to $+10$ of 100 jokes on sets of size κ, $36 \le \kappa \le 100$, by $24,983$ users. We convert cardinal ratings into ordinal full rankings. The ground truth θ^* is set to be the MLE of the PL weights over the entire data. For $m \in \{2, 3, 4, 5\}$, we convert each full ranking into a poset that has $\ell = \lfloor \kappa/m \rfloor$ partitions of size m, by removing known relative ordering from each partition. The leads to total number of effective samples $\sum_j p_j = \sum_j \sum_{a\in[\ell_j]} m_{j,a} = \sum_j (\kappa_j - m)$, which is approximately equal for each $m \in \{2, 3, 4, 5\}$. However, with increasing m, the quantities $\gamma_1, \gamma_2, \gamma_3$ become smaller and hence the error increases (third panel in Figure 8). Figure 8 compares the three algorithms for two different settings. In the first

panel, we fix $m = 4$ and vary the number of samples n. Mean square error decreases with increase in the number of samples. In the third panel, we use $n = 5000$ samples, and vary $m \in \{2, 3, 4, 5\}$.

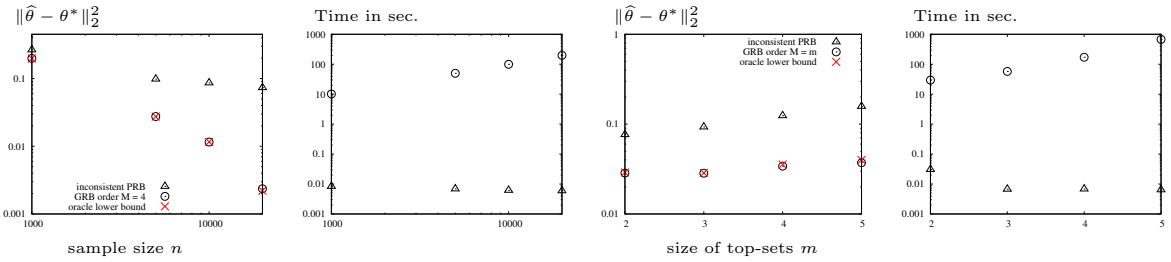

Figure 8: The jester dataset which has $d = 100$, $n = 24,983$, and $36 \leq \kappa_j \leq 100$. The second and the fourth panel show the computation time for the first and the third panel respectively. Generalized rank-breaking improves over pairwise RB and performs as good as oracle MLE on jester dataset.

6. Computational and statistical tradeoff

For estimators with limited computational power, however, the lower bound Theorem 12 fails to capture the dependency on the allowed computational power. Understanding such fundamental trade-offs is a challenging problem, which has been studied only in a few special cases, e.g. planted clique problem (Deshpande and Montanari, 2015; Meka et al., 2015). This is outside the scope of this paper, and we instead investigate the trade-off achieved by the proposed rank-breaking approach. When we are limited on computational power, Theorem 11 implicitly captures this dependence when order-M rank-breaking is used. The dependence is captured indirectly via the resulting rank-breaking $\{G_{j,a}\}_{j\in[n],a\in[\ell_j]}$ and the topology of it. We make this trade-off explicit by considering a simple but canonical example. Suppose $\theta^* \in \Omega_b$ with $b = O(1)$. Each user gives an i.i.d. partial ranking, where all items are offered and the partial ranking is based on an ordered partition with $\ell_j = \lfloor \sqrt{2c} d^{1/3} \rfloor$ subsets for a constant c. The top subset has size $m_{j,1} = 1$, and the a-th subset has size $m_{j,a} = a$, up to $a < \ell_j$. The choice of ℓ_j with a sufficiently small constant c ensures that all the conditions of the ideal case explained in the previous section for holding the Theorem 11 are satisfied.

Computation. For a choice of M such that $M \leq \ell_j - 1$, we consider the computational complexity in computing $\theta^{(t)}$, (6) in one iteration of the minorization-maximization algorithm, which scales as $T(M, n) = O(M! \times dn)$. A detailed analysis of the convergence rate of the MM algorithm is outside the scope of this paper.

Accuracy. Under the canonical setting, for $M \leq \ell_j - 1$, Laplacian matrix L of the comparison graph $\mathcal{H}$ is $L = nM(M+1)/(2d(d-1))\big(d\mathbb{I} - \mathbf{1}\mathbf{1}^\top\big)$. All the non-zero eigenvalues of this complete graph are equal, $\lambda_2(L) = \cdots = \lambda_d(L) = \mathrm{Tr}(L)/(d-1)$. Therefore, the resclaed spectral gap $\alpha = 1$, and the rescaled largest eigenvalue $\beta = 1$. Since the effective sample size is $\sum_{j,a} m_{j,a}\mathbb{I}\{m_{j,a} \leq M\} = nM(M+1)/2$, it follows from Theorem 11 that the (rescaled) root mean squared error is $O(\sqrt{(d\log d)/(nM^2)})$. In order to achieve a smaller target error rate of ε for a fixed problem size d, an analyst can increase the rank-breaking order M and/or increase n that is collect more i.i.d. rankings. Fixing the rank-breaking order M, we need to collect $n = \Omega((d\log d)/(\varepsilon^2 M^2))$

i.i.d. rankings. The resulting trade-off between run-time and root mean squared error ε is $T(\varepsilon) \propto (M!(d^2 \log d)/(\varepsilon^2 M^2)$. The computational complexity is quadratic in the target error ϵ, when we can collect more rankings. On the other hand, fixing the number of rankings n, we need to choose $M = \Omega((1/\varepsilon)\sqrt{(d \log d)/n})$. The resulting trade-off between run-time and root mean squared error ε is $T(\epsilon) \propto (\lceil (1/\varepsilon)\sqrt{(d \log d)/n} \rceil)! dn$. The computational complexity is super exponential in the target error ϵ, for a fixed problem size d and the number of rankings n. Super exponential complexity is unavoidable as computing likelihood is super exponential in M. However, our approach provides flexibility to the analyst to choose between collecting more rankings n or increasing the rank-breaking order M to achieve the desired target error. We show numerical experiment under this canonical setting in Figure 4 (left) with $d = 256$ and $M \in \{1, 2, 3, 4, 5\}$, illustrating the trade-off in practice.

7. Proofs

We provide the proofs of the main results.

7.1 Proof of Lemma 3

In the following, we show that $Q(e, \theta; \theta^{(t)})$ minorizes $\ln(\mathbb{P}_\theta(e))$ at $\theta^{(t)}$. Using Jensen's inequality $\ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln(X)]$, for any given parameter $\theta^{(t)} \in \mathbb{R}^d$, we have,

$$
\begin{aligned}
& \ln(\mathbb{P}_\theta(e)) \\
= \; & \ln\left(\mathbb{P}_\theta\big(B(e) \prec T(e)\big)\right) \\
= \; & \ln\left(\sum_{\sigma \in \Lambda_{T(e)}} \frac{\exp\left(\sum_{c=1}^{|T(e)|} \theta_{\sigma(c)}\right)}{\prod_{u=1}^{|T(e)|}\left(\sum_{c'=u}^{|T(e)|} \exp\left(\theta_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta_i\right)\right)}\right) \\
\geq \; & \sum_{\sigma \in \Lambda_{T(e)}} \frac{\mathbb{P}_{\theta^{(t)}}(e, \sigma)}{\mathbb{P}_{\theta^{(t)}}(e)} \ln\left(\frac{\exp\left(\sum_{c=1}^{|T(e)|} \theta_{\sigma(c)}\right)}{\prod_{u=1}^{|T(e)|}\left(\sum_{c'=u}^{|T(e)|} \exp\left(\theta_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta_i\right)\right)} \frac{\mathbb{P}_{\theta^{(t)}}(e)}{\mathbb{P}_{\theta^{(t)}}(e, \sigma)}\right) \qquad (19) \\
= \; & \sum_{\sigma \in \Lambda_{T(e)}} \frac{\mathbb{P}_{\theta^{(t)}}(e, \sigma)}{\mathbb{P}_{\theta^{(t)}}(e)} \sum_{u=1}^{|T(e)|} \left(\theta_{\sigma(u)} - \ln\left(\sum_{c'=u}^{|T(e)|} \exp\left(\theta_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta_i\right)\right)\right) \\
& + \sum_{\sigma \in \Lambda_{T(e)}} \frac{\mathbb{P}_{\theta^{(t)}}(e, \sigma)}{\mathbb{P}_{\theta^{(t)}}(e)} \ln\left(\frac{\mathbb{P}_{\theta^{(t)}}(e)}{\mathbb{P}_{\theta^{(t)}}(e, \sigma)}\right) \\
\geq \; & \sum_{\sigma \in \Lambda_{T(e)}} \frac{\mathbb{P}_{\theta^{(t)}}(e, \sigma)}{\mathbb{P}_{\theta^{(t)}}(e)} \sum_{u=1}^{|T(e)|} \left(\theta_{\sigma(u)} - \frac{\sum_{c'=u}^{|T(e)|} \exp\left(\theta_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta_i\right)}{\sum_{c'=u}^{|T(e)|} \exp\left(\theta^{(t)}_{\sigma(c')}\right) + \sum_{i \in B(e)} \exp\left(\theta^{(t)}_i\right)}\right) + f(e, \theta^{(t)}) \\
\equiv \; & Q(e, \theta; \theta^{(t)}) \,.
\end{aligned}
$$

Note that inequality in (19) is tight if $\theta_t = \theta$. The last inequality follows from the fact that for any positive x and y, we have

$$
-\ln x \geq 1 - \ln y - (x/y) \quad \text{with equality if and only if } x = y.
$$

Therefore, $Q(e, \theta; \theta^{(t)})$ minorizes $\ln(\mathbb{P}_\theta(e))$ and is equal to $\ln(\mathbb{P}_\theta(e))$ if and only if $\theta^{(t)} = \theta$.

7.2 Proof of Theorem 11

We define few additional notations. $p \equiv (1/n)\sum_{j=1}^{n} p_j$. $V(e_{j,a}) \equiv T(e_{j,a}) \cup B(e_{j,a})$ for all $j \in [n]$ and $a \in [\ell_j]$. Note that by definition of rank-breaking edge $e_{j,a}$, $V(e_{j,a})$ is a random set of items that are ranked in bottom $r_{j,a}$ positions in a set of S_j items by the user j.

The proof sketch is inspired from Khetan and Oh (2016). The main difference and technical challenge is in showing the strict concavity of $\mathcal{L}_{\mathrm{RB}}(\theta)$ when restricted to Ω_b. We want to prove an upper bound on $\Delta = \widehat{\theta} - \theta^*$, where $\widehat{\theta}$ is the sample dependent solution of the optimization (4) and θ^* is the true utility parameter from which the samples are drawn. Since $\widehat{\theta}, \theta^* \in \Omega_b$, it follows that $\Delta \mathbf{1} = 0$. Since $\widehat{\theta}$ is the maximizer of $\mathcal{L}_{\mathrm{RB}}(\theta)$, we have the following inequality,

$$\mathcal{L}_{\mathrm{RB}}(\widehat{\theta}) - \mathcal{L}_{\mathrm{RB}}(\theta^*) - \langle \nabla \mathcal{L}_{\mathrm{RB}}(\theta^*), \Delta \rangle \;\geq\; -\langle \nabla \mathcal{L}_{\mathrm{RB}}(\theta^*), \Delta \rangle \;\geq\; -\|\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)\|_2 \|\Delta\|_2, \tag{20}$$

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

$$\mathcal{L}_{\mathrm{RB}}(\widehat{\theta}) - \mathcal{L}_{\mathrm{RB}}(\theta^*) - \langle \nabla \mathcal{L}_{\mathrm{RB}}(\theta^*), \Delta \rangle \;=\; \frac{1}{2}\Delta^\top H(\theta)\Delta \leq -\frac{1}{2}\lambda_2(-H(\theta))\|\Delta\|_2^2, \tag{21}$$

where $\lambda_2(-H(\theta))$ is the second smallest eigen value of $-H(\theta)$. We will show in Lemma 15 that $-H(\theta)$ is positive semi definite with one eigenvalue at zero with a corresponding eigen vector $\mathbf{1} = [1, \ldots, 1]^\top$. The last inequality follows since $\Delta^\top \mathbf{1} = 0$. Combining Equations (20) and (21),

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

where we used the fact that $\lambda_2(-H(\theta)) > 0$ from Lemma 15. The following technical lemmas prove that the norm of the gradient is upper bounded by $\gamma_2^{-1/2} e^b \sqrt{6np\log d}$ with high probability and the second smallest eigen value of negative of the Hessian is lower bounded by $(1/8)\, e^{-6b}\alpha\gamma_1\gamma_2\gamma_3(np/(d-1))$. This finishes the proof of Theorem 11.

The (random) gradient of the log likelihood in (4) can be written as the following, where the randomness is in which items ended up in the top set $T(e_{j,a})$ and the bottom set $B(e_{j,a})$:

$$\nabla_i \mathcal{L}_{\mathrm{RB}}(\theta) \;=\; \sum_{j=1}^{n}\sum_{a=1}^{\ell_j} \sum_{\substack{\mathcal{C}\subseteq S_j,\\ |\mathcal{C}|=r_{j,a}-1}} \mathbb{I}\{\, V(e_{j,a}) = \{\mathcal{C}, i\}\,\} \frac{\partial \log \mathbb{P}_\theta(e_{j,a})}{\partial \theta_i}\,.$$

Note that we are intentionally decomposing each summand as a summation over all $\mathcal{C}$ of size $r_{j,a}-1$, such that we can separate the analysis of the expectation in the following lemma. The random variable $\mathbb{I}\{\{\mathcal{C}, i\} = V(e_{j,a})\}$ indicates that we only include one term for any given instance of the sample. Note that the event $\mathbb{I}\{\{\mathcal{C}, i\} = V(e_{j,a})\}$ is equivalent to the event that the $\{\mathcal{C}, i\}$ items are ranked in bottom $r_{j,a}$ positions in the set S_j, that is $V(e_{j,a})$ items are ranked in bottom $r_{j,a}$ positions in the set S_j.

Lemma 13 *If the j-th poset is drawn from the PL model with weights θ^* then for any given $\mathcal{C}' \subseteq S_j$ with $|\mathcal{C}'| = r_{j,a}$,*

$$\mathbb{E}\left[\mathbb{I}\{\mathcal{C}' = V(e_{j,a})\}\frac{\partial \log \mathbb{P}_{\theta^*}(e_{j,a})}{\partial \theta_i^*}\,\middle|\, \{e_{j,a'}\}_{a'<a}\right] \;=\; 0\,.$$

First, this lemma implies that $\mathbb{E}\big[\mathbb{I}\{\mathcal{C}' = V(e_{j,a})\}\frac{\partial \log \mathbb{P}_{\theta^*}(e_{j,a})}{\partial \theta_i^*}\big] = 0$. Secondly, the above lemma allows us to construct a vector-valued martingale and apply a generalization of Azuma-Hoeffding's tail bound on the norm to prove the following concentration of measure. This proves the desired bound on the gradient.

Lemma 14 *If n posets are independently drawn over d items from the PL model with weights θ^* then with probability at least $1 - 2e^3 d^{-3}$,*

$$\|\nabla \mathcal{L}_{\mathrm{RB}}(\theta^*)\| \le \gamma_2^{-1/2} e^b \sqrt{6np \log d}\,,$$

where γ_2 depend on the choice of the rank-breaking and are defined in Section 4.1.

We will prove in (26) that the Hessian matrix $H(\theta) \in \mathcal{S}^d$ with $H_{ii'}(\theta) = \frac{\partial^2 \mathcal{L}_{\mathrm{RB}}(\theta)}{\partial \theta_i \partial \theta_{i'}}$ can be expressed as

$$-H(\theta) = \sum_{j=1}^{n} \sum_{a=1}^{\ell_j} \sum_{i<i' \in S_j} \mathbb{I}\{(i,i') \subseteq V(e_{j,a})\} \left(\frac{\partial^2 \log \mathbb{P}_\theta(e_{j,a})}{\partial \theta_i \partial \theta_{i'}} (e_i - e_{i'})(e_i - e_{i'})^\top \right). \tag{22}$$

It is easy to see that $H(\theta)\mathbf{1} = 0$. The following lemma proves a lower bound on the second smallest eigenvalue $\lambda_2(-H(\theta))$ in terms of re-scaled spectral gap α of the comparison graph $\mathcal{H}$ defined in Section 4.1.

Lemma 15 *Under the hypothesis of Theorem 11, if the assumptions in Equation* (14) *are satisfied then with probability at least $1 - d^{-3}$, the following holds for any $\theta \in \Omega_b$:*

$$\lambda_2(-H(\theta)) \;\ge\; \frac{e^{-6b}\alpha\gamma_1\gamma_2\gamma_3}{8} \frac{np}{(d-1)}\,,$$

and $\lambda_1(-H(\theta)) = 0$ with corresponding eigenvector $\mathbf{1}$.

This finishes the proof of the desired claim.

7.3 Proof of Lemma 13

Recall that $e_{j,a}$ is a random event where randomness is in which items ended up in the top-set $T(e_{j,a})$ and the bottom-set $B(e_{j,a})$, and $\mathbb{P}_{\theta^*}(e_{j,a}) = \mathbb{P}_{\theta^*}[B(e_{j,a}) \prec T(e_{j,a})]$ that is the probability of observing $B(e_{j,a}) \prec T(e_{j,a})$ when the offer set is $B(e_{j,a}) \cup T(e_{j,a})$ as defined in (3). Define, $\mathbb{P}_{\theta^*,S_j}[e_{j,a}|V(e_{j,a}) = \mathcal{C}']$ to be the conditional probability of observing $B(e_{j,a}) \prec T(e_{j,a})$, when the offer set is S_j, conditioned on the event that $V(e_{j,a}) = \mathcal{C}'$. Note that we have put subscript S_j in $\mathbb{P}_{\theta^*}$ to specify that the offer set is S_j. Observe that for any set $\mathcal{C}' \subseteq S_j$, the event $\{\mathcal{C}' = V(e_{j,a})\}$ is equivalent to $\mathcal{C}'$ items being ranked in bottom $r_{j,a}$ positions when the offer set is S_j. In other words, it is conditioned on the event that the subset $V(e_{j,a})$ items are ranked in bottom $r_{j,a}$ positions when the offer set is S_j. In Equation (23), we show that under PL model

$$\mathbb{P}_{\theta^*,S_j}[e_{j,a}|V(e_{j,a}) = \mathcal{C}'] = \mathbb{P}_{\theta^*}[e_{j,a}]\,.$$

Also, by conditioning on any outcome of $\{e_{j,a'}\}_{a'<a}$ it can be checked that

$$\mathbb{P}_{\theta^*,S_j}[e_{j,a}|V(e_{j,a}) = \mathcal{C}', \{e_{j,a'}\}_{a'<a}] = \mathbb{P}_{\theta^*,S_j}[e_{j,a}|V(e_{j,a}) = \mathcal{C}'].$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial \log \mathbb{P}_{\theta^*}\left[e_{j,a}\right]}{\partial \theta_i^*}\middle| V(e_{j,a})=\mathcal{C}', \{e_{j,a'}\}_{a'<a}\right] \\
&= \mathbb{E}\left[\frac{\partial \log \mathbb{P}_{\theta^*,S_j}\left[e_{j,a}|V(e_{j,a})=\mathcal{C}', \{e_{j,a'}\}_{a'<a}\right]}{\partial \theta_i^*}\middle| V(e_{j,a})=\mathcal{C}', \{e_{j,a'}\}_{a'<a}\right] \\
&= \sum_{\substack{e_{j,a}:V(e_{j,a})=\mathcal{C}' \\ \{e_{j,a'}\}_{a'<a}}} \mathbb{P}_{\theta^*,S_j}\left[e_{j,a}\middle|V(e_{j,a})=\mathcal{C}', \{e_{j,a'}\}_{a'<a}\right] \frac{\partial}{\partial \theta_i^*} \log \mathbb{P}_{\theta^*,S_j}\left[e_{j,a}\middle|V(e_{j,a})=\mathcal{C}', \{e_{j,a'}\}_{a'<a}\right] \\
&= \frac{\partial}{\partial \theta_i^*} \sum_{e_{j,a}:V(e_{j,a})=\mathcal{C}'} \mathbb{P}_{\theta^*,S_j}\left[e_{j,a}\middle|V(e_{j,a})=\mathcal{C}'\right] = \frac{\partial}{\partial \theta_i^*} 1 = 0\,,
\end{aligned}$$

where we used $\{e_{j,a} : V(e_{j,a}) = \mathcal{C}'\} = \{e_{j,a} : V(e_{j,a}) = \mathcal{C}', \{e_{j,a'}\}_{a'<a}\}$ which follows from the definition of rank-breaking edges $e_{j,a}$. This proves the desired claim. It remains to show that

$$\mathbb{P}_{\theta^*,S_j}[e_{j,a}|V(e_{j,a})=\mathcal{C}'] = \mathbb{P}_{\theta^*}[e_{j,a}]\,.$$

This follows from the fact that under PL model for any disjoint set of items $\{\mathcal{C}_i\}_{i\in[\ell]}$ such that $\cup_{i=1}^{\ell}\mathcal{C}_i = S$,

$$\mathbb{P}\big(\mathcal{C}_\ell \prec \mathcal{C}_{\ell-1} \prec \cdots \prec \mathcal{C}_1\big) = \mathbb{P}\big(\mathcal{C}_\ell \prec \mathcal{C}_{\ell-1}\big)\mathbb{P}\big(\{\mathcal{C}_\ell, \mathcal{C}_{\ell-1}\} \prec \mathcal{C}_{\ell-2}\big)\cdots\mathbb{P}\big(\{\mathcal{C}_\ell, \mathcal{C}_{\ell-1}, \cdots, \mathcal{C}_2\} \prec \mathcal{C}_1\big)\,, \quad (23)$$

where $\mathbb{P}(\mathcal{C}_{i_1} \prec \mathcal{C}_{i_2})$ is the probability that $\mathcal{C}_{i_2}$ items are ranked higher than $\mathcal{C}_{i_1}$ items when the offer set is $\{\mathcal{C}_{i_1} \cup \mathcal{C}_{i_2}\}$.

7.4 Proof of Lemma 14

We 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 $\nabla\mathcal{L}_{\mathrm{RB}}^{(e_{j,a})} \in \mathbb{R}^d$ as the gradient vector arising out of each rank-breaking edge $\{e_{j,a}\}_{j\in[n],a\in[\ell_j]}$ as

$$\nabla_i\mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}(\theta^*) \;\equiv\; \sum_{\mathcal{C}\subseteq S_j} \mathbb{I}\big\{V(e_{j,a}) = \{\mathcal{C}, i\}\big\}\nabla_i \log \mathbb{P}_{\theta^*}(e_{j,a})\,,$$

such that $\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*) = \sum_{j\in[n]}\sum_{a\in[\ell_j]}\nabla\mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}$. We take $\nabla\mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}$ as the incremental random vector in a martingale of $\sum_{j=1}^{n}\ell_j$ time steps. Let $H_{j,a}$ denote (the sigma algebra of) the history up to $e_{j,a}$ and define a sequence of random vectors in $\mathbb{R}^d$:

$$Z_{j,a} \;\equiv\; \mathbb{E}[\nabla\mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}(\theta^*)|H_{j,a}]\,,$$

with the convention that $Z_{1,1} = \mathbb{E}[\nabla\mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}(\theta^*)] = 0$ as proved in Lemma 13. It also follows from Lemma 13 that $\mathbb{E}[Z_{j,a+1}|Z_{j,a}] = Z_{j,a}$ for $a < \ell_j$. Also, from the independence of samples, it follows that $\mathbb{E}[Z_{j+1,1}|Z_{j,\ell_j}] = Z_{j,\ell_j}$. Applying a generalized version of the vector Azuma-Hoeffding inequality which readily follows from [Theorem 1.8, Hayes (2005)], we have

$$\mathbb{P}\big[\,\|\nabla\mathcal{L}_{\mathrm{RB}}(\theta^*)\| \geq \delta\,\big] \;\leq\; 2e^3\exp\left(-\frac{\delta^2}{\sum_{j=1}^{n}\sum_{a=1}^{\ell_j} m_{j,a}2\gamma_2^{-1}e^{2b}}\right),$$

where we used $\|\nabla \mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}\|^2 \le m_{j,a} 2\gamma_2^{-1} e^{2b}$. Choosing $\delta = \gamma_2^{-1} e^b \sqrt{6np \log d}$ gives the desired bound.

Now we are left to show that $\|\nabla \mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}\|^2 \le 2m_{j,a}\gamma_2^{-1} e^{2b}$ for any $\theta \in \Omega_b$. Recall that $\sigma \in \Lambda_{T(e_{j,a})}$ is the set of all full rankings over $T(e_{j,a})$ items. In rest of the proof, with a slight abuse of notations, we extend each of these ranking σ over $T(e_{j,a}) \cup B(e_{j,a})$ items in the following way. Consider any full ranking $\tilde{\sigma}$ over $B(e_{j,a})$ items. Then for each $\sigma \in \Lambda_{T(e_{j,a})}$, the extension is such that $\sigma(|T(e_{j,a})| + c) = \tilde{\sigma}(c)$ for $1 \le c \le |B(e_{j,a})|$. The choice of ranking $\tilde{\sigma}$ will have no impact on any of the following mathematical expressions. From the definition of $\mathbb{P}_\theta(e_{j,a})$ (3), we have, for any $i \in V(e_{j,a})$,

$$\frac{\partial \mathbb{P}_\theta(e_{j,a})}{\partial \theta_i} = \mathbb{I}\{i \in T(e_{j,a})\}\mathbb{P}_\theta(e_{j,a}) - \underbrace{\sum_{\sigma \in \Lambda_{T(e_{j,a})}} \underbrace{\frac{\exp\left(\sum_{c=1}^{m_{j,a}} \theta_{\sigma(c)}\right)}{\prod_{u=1}^{m_{j,a}} \left(\sum_{c'=u}^{r_{j,a}} \exp\left(\theta_{\sigma(c')}\right)\right)}}_{\equiv A_\sigma} \underbrace{\left(\sum_{u'=1}^{m_{j,a}} \frac{\mathbb{I}\{\sigma^{-1}(i) \ge u'\} \exp(\theta_i)}{\sum_{c'=u'}^{r_{j,a}} \exp\left(\theta_{\sigma(c')}\right)}\right)}_{\equiv B_{\sigma,i}}}_{\equiv E_i}. \tag{24}$$

Note that $A_\sigma, B_{\sigma,i}$ and E_i depend on $e_{j,a}$. Observe that for any $1 \le u' \le m_{j,a}$ and any $\sigma \in \Lambda_{T(e_{j,a})}$,

$$\sum_{i \in V(e_{j,a})} \mathbb{I}\{\sigma^{-1}(i) \ge u'\} \exp(\theta_i) = \sum_{c'=u'}^{r_{j,a}} \exp\left(\theta_{\sigma(c')}\right).$$

Therefore, $\sum_{i \in V(e_{j,a})} B_{\sigma,i} = m_{j,a}$. It follows that

$$\sum_{i \in V(e_{j,a})} E_i = \sum_{\sigma \in \Lambda_{T(e_{j,a})}} A_\sigma \left(\sum_{i \in V(e_{j,a})} B_{\sigma,i}\right) = m_{j,a} \sum_{\sigma \in \Lambda_{T(e_{j,a})}} A_\sigma = m_{j,a}\mathbb{P}_\theta(e_{j,a}), \tag{25}$$

where the last equality follows from the definition of $\mathbb{P}_\theta(e_{j,a})$ (4). Also, since for any i, i', $e^{(\theta_i - \theta_{i'})} \le e^{2b}$; for any i, $B_{\sigma,i} \le e^{2b} \sum_{k=r_{j,a}-m_{j,a}+1}^{r_{j,a}} (1/k) \le e^{2b}(1 + \log(r_{j,a}/(r_{j,a} - m_{j,a} + 1))) \le \gamma_2^{-1} e^{2b}$, where the last inequality follows from the definition of γ_2 (12) and the fact that $x \le \sqrt{1 + \log x}$ for all $x \ge 1$. Therefore, $E_i \le \gamma_2^{-1} e^{2b} \sum_{\sigma \in \Lambda_{T(e_{j,a})}} A_\sigma = \gamma_2^{-1} e^{2b} \mathbb{P}_\theta(e_{j,a})$. We have $\partial \log \mathbb{P}_\theta(e_{j,a})/\partial\theta_i = (1/\mathbb{P}_\theta(e_{j,a}))\partial \mathbb{P}_\theta(e_{j,a})/\partial\theta_i = \mathbb{I}\{i \in T(e_{j,a})\} - E_i/\mathbb{P}_\theta(e_{j,a})$. Since $|T(e_{j,a})| = m_{j,a}$, $\|\nabla \mathcal{L}_{\mathrm{RB}}^{(e_{j,a})}\|^2 \le m_{j,a} + \sum_{i \in V(e_{j,a})} (E_i/\mathbb{P}_\theta(e_{j,a}))^2 \le 2m_{j,a}\gamma_2^{-1} e^{2b}$, where we used (25) and the fact that $\gamma_2^{-1} \ge 1$.

7.4.1 Proof of Lemma 15

First, we prove (22). For brevity, remove $\{j, a\}$ from $\mathbb{P}_\theta(e_{j,a})$. From Equations (24) and (25), and $|T(e_{j,a})| = m_{j,a}$, we have $\sum_{i \in V(e_{j,a})} \frac{\partial}{\partial \theta_i} \mathbb{P}_\theta(e) = m_{j,a}\mathbb{P}_\theta(e) - m_{j,a}\mathbb{P}_\theta(e) = 0$. It follows that

$$\sum_{i \in V(e_{j,a})} \left(\frac{\partial^2 \log \mathbb{P}_\theta(e)}{\partial \theta_{i'} \partial \theta_i}\right) = \frac{1}{\mathbb{P}_\theta(e)} \frac{\partial}{\partial \theta_{i'}} \left(\sum_{i \in V(e_{j,a})} \left(\frac{\partial \mathbb{P}_\theta(e)}{\partial \theta_i}\right)\right) - \frac{1}{(\mathbb{P}_\theta(e))^2} \frac{\partial \mathbb{P}_\theta(e)}{\partial \theta_{i'}} \left(\sum_{i \in V(e_{j,a})} \left(\frac{\partial \mathbb{P}_\theta(e)}{\partial \theta_i}\right)\right) = 0. \tag{26}$$

Since by definition $\mathcal{L}_{\mathrm{RB}}(\theta) = \sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\log \mathbb{P}_\theta(e_{j,a})$, and $H_{ii'}(\theta) = \frac{\partial^2 \mathcal{L}_{\mathrm{RB}}(\theta)}{\partial\theta_i \partial\theta_{i'}}$ which is a symmetric matrix, Equation (26) implies that it can be expressed as given in Equation (22). It follows that all-ones is an eigenvector of $H(-\theta)$ with the corresponding eigenvalue being zero.

To get a lower bound on $\lambda_2(-H(\theta))$, we apply Weyl's inequality

$$\lambda_2(-H(\theta)) \geq \lambda_2(\mathbb{E}[-H(\theta)]) - \|H(\theta) - \mathbb{E}[H(\theta)]\|\,.$$

We will show in (27) that $\lambda_2(\mathbb{E}[-H(\theta)]) \geq e^{-6b}\alpha\gamma_1\gamma_2\gamma_3(np/(4(d-1)))$ and in (39) that $\|H(\theta) - \mathbb{E}[H(\theta)]\| \leq 16e^{4b}\nu\sqrt{\frac{p_{\max}}{\kappa_{\min}}\frac{np}{\beta(d-1)}\log d}$. Putting these together,

$$\begin{aligned}\lambda_2(-H(\theta)) &\geq e^{-6b}\alpha\gamma_1\gamma_2\gamma_3\frac{np}{4(d-1)} - 16e^{4b}\nu\sqrt{\frac{p_{\max}}{\kappa_{\min}}\frac{np}{\beta(d-1)}\log d}\\ &\geq \frac{e^{-6b}\alpha\gamma_1\gamma_2\gamma_3}{8}\frac{np}{(d-1)}\,,\end{aligned}$$

where the last inequality follows from the assumption on $n\kappa_{\min}$ given in (14).

To prove a lower bound on $\lambda_2(\mathbb{E}[-H(\theta)])$, we claim that for $\theta \in \Omega_b$,

$$\begin{aligned}\mathbb{E}\big[-H(\theta)\big] &\succeq e^{-6b}\gamma_1\gamma_2\gamma_3\sum_{j=1}^{n}\frac{p_j}{4\kappa_j(\kappa_j-1)}\sum_{i<i'\in S_j}(e_i-e_{i'})(e_i-e_{i'})^\top \qquad (27)\\ &= \frac{e^{-6b}\gamma_1\gamma_2\gamma_3}{4}L\,,\end{aligned}$$

where $L \in \mathcal{S}^d$ is defined in (10). Using $\lambda_2(L) = np\alpha/(d-1)$ from (11), we have $\lambda_2(-H(\theta)) \geq e^{-6b}\alpha\gamma_1\gamma_2\gamma_3(np/(4(d-1)))$. To prove (27), notice that

$$\mathbb{E}[-H(\theta)_{ii'}] = \mathbb{E}\Big[\sum_{j\in[n]}\sum_{a\in[\ell_j]}\mathbb{I}\big\{(i,i')\subseteq V(e_{j,a})\big\}\frac{\partial^2\log\mathbb{P}_\theta(e_{j,a})}{\partial\theta_i\partial\theta_{i'}}\Big]\,, \qquad (28)$$

when $i \neq i'$. We will show that for any $i \neq i' \in V(e_{j,a})$,

$$\frac{\partial^2\log\mathbb{P}_\theta(e_{j,a})}{\partial\theta_i\partial\theta_{i'}} \geq \begin{cases}\frac{e^{-2b}m_{j,a}}{r_{j,a}^2} & \text{if } i,i' \in B(e_{j,a})\\ -\frac{e^{4b}m_{j,a}^2}{(r_{j,a}-m_{j,a}+1)^2} & \text{otherwise}\,.\end{cases} \qquad (29)$$

We need to bound the probability of two items appearing in the bottom-set $B(e_{j,a})$ and in the top-set $T(e_{j,a})$.

Lemma 16 *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 $1 \leq \ell, \ell_1, \ell_2 \leq \kappa - 1$,*

$$\mathbb{P}_\theta\big[\sigma^{-1}(i), \sigma^{-1}(i') > \ell\big] \geq \frac{e^{-4b}(\kappa-\ell)(\kappa-\ell-1)}{\kappa(\kappa-1)}\left(1-\frac{\ell}{\kappa}\right)^{2e^{2b}-2}\,, \qquad (30)$$

$$\mathbb{P}_\theta\big[\sigma^{-1}(i) = \ell\big] \leq \frac{e^{6b}}{\kappa-\ell}\,, \qquad (31)$$

$$\mathbb{P}_\theta\big[\sigma^{-1}(i) = \ell_1, \sigma^{-1}(i') = \ell_2\big] \leq \frac{e^{10b}}{(\kappa-\ell_1-1)(\kappa-\ell_2)}\,. \qquad (32)$$

where the probability $\mathbb{P}_\theta$ is with respect to the sampled ranking resulting from PL weights $\theta \in \Omega_b$.

Substituting $\ell = \kappa_j - r_{j,a} + m_{j,a}$ in (30), and $\ell, \ell_1, \ell_2 \leq \kappa_j - r_{j,a} + m_{j,a}$ in (31) and (32), we have,

$$\begin{aligned}
\mathbb{P}_\theta\big[(i,i') \subseteq B(e_{j,a})\big] &\geq \frac{e^{-4b}(r_{j,a}-m_{j,a})^2}{4\kappa_j(\kappa_j-1)} \Big(\frac{r_{j,a}-m_{j,a}}{\kappa_j}\Big)^{2e^{2b}-2}, &(33)\\
\mathbb{P}_\theta\big[i \in T(e_{j,a}), i' \in B(e_{j,a})\big] &\leq m_{j,a} \max_{\ell \in [\kappa_j - r_{j,a} + m_{j,a}]} \mathbb{P}(\sigma^{-1}(i) = \ell)\\
&\leq \frac{e^{6b} m_{j,a}}{r_{j,a}-m_{j,a}}, &(34)\\
\mathbb{P}_\theta\big[(i,i') \subseteq T(e_{j,a})\big] &\leq m_{j,a}^2 \max_{\ell_1,\ell_2 \in [\kappa_j - r_{j,a} + m_{j,a}]} \mathbb{P}(\sigma^{-1}(i) = \ell_1, \sigma^{-1}(i') = \ell_2)\\
&\leq \frac{e^{10b} m_{j,a}^2}{2\,(r_{j,a}-m_{j,a}-1)(r_{j,a}-m_{j,a})}, &(35)
\end{aligned}$$

where (33) uses $r_{j,a} - m_{j,a} - 1 \geq (r_{j,a} - m_{j,a})/4$, (34) uses $\mathbb{P}_\theta[i \in T(e_{j,a}), i' \in B(e_{j,a})] \leq \mathbb{P}_\theta[i \in T(e_{j,a})]$, and (34)-(35) uses counting on the possible choices. The bound in (35) is smaller than the one in (34) as per our assumption that $\gamma_3 > 0$.

Using Equations (28)-(29) and (33)-(35), and the definitions of $\gamma_1, \gamma_2, \gamma_3$ from Section 4.1, we get

$$\begin{aligned}
&\mathbb{E}[-H(\theta)_{ii'}] \geq\\
&\sum_{j\in[n]} \sum_{a \in [\ell_j]} \Big\{ \underbrace{\Big(\frac{r_{j,a}-m_{j,a}}{\kappa_j}\Big)^{2e^{2b}-2}}_{\geq \gamma_1} \underbrace{\Big(\frac{r_{j,a}-m_{j,a}}{r_{j,a}}\Big)^{2}}_{\geq \gamma_2} \frac{e^{-6b} m_{j,a}}{4\kappa_j(\kappa_j-1)} \;-\; \frac{e^{6b} m_{j,a}}{r_{j,a}-m_{j,a}} \frac{e^{4b} m_{j,a}^2}{(r_{j,a}-m_{j,a}+1)^2} \Big\}\\
&\geq \sum_{j,a} \frac{\gamma_1\gamma_2 e^{-6b} m_{j,a}}{4\kappa_j(\kappa_j-1)} \underbrace{\Big(1 - \frac{4e^{16b}}{\gamma_1} \frac{m_{j,a}^2 r_{j,a}^2 \kappa_j^2}{(r_{j,a}-m_{j,a})^5}\Big)}_{\geq \gamma_3}.
\end{aligned}$$

This combined with (22) proves the desired claim (27). Further, in Appendix 7.7, we show that if $m_{j,a} \leq 3$ for all $\{j,a\}$ then $\frac{\partial^2 \log \mathbb{P}_\theta(e_{j,a})}{\partial\theta_i \partial\theta_{i'}}$ is non-negative even for $i \neq i' \in T(e_{j,a})$, and $i \in T(e_{j,a}), i' \in B(e_{j,a})$ as opposed to a negative lower-bound given in (29). Therefore, bound on $\mathbb{E}[-H(\theta)]$ in (27) can be tightened by a factor of γ_3.

To prove claim (29), define the following for $\sigma \in \Lambda_{T(e_{j,a})}$,

$$\begin{aligned}
A_\sigma &\equiv \frac{\exp\big(\sum_{c=1}^{m_{j,a}} \theta_{\sigma(c)}\big)}{\prod_{u=1}^{m_{j,a}} \big(\sum_{c'=u}^{r_{j,a}} \exp\big(\theta_{\sigma(c')}\big)\big)}, B_\sigma \equiv \sum_{u'=1}^{m_{j,a}} \frac{1}{\sum_{c'=u'}^{r_{j,a}} \exp\big(\theta_{\sigma(c')}\big)},\\
B_{\sigma,i} &\equiv \sum_{u'=1}^{m_{j,a}} \frac{\mathbb{I}\{\sigma^{-1}(i) \geq u'\}}{\sum_{c'=u'}^{r_{j,a}} \exp\big(\theta_{\sigma(c')}\big)}, C_\sigma \equiv \sum_{u'=1}^{m_{j,a}} \frac{1}{\big(\sum_{c'=u'}^{r_{j,a}} \exp\big(\theta_{\sigma(c')}\big)\big)^2},\\
C_{\sigma,i} &\equiv \sum_{u'=1}^{m_{j,a}} \frac{\mathbb{I}\{\sigma^{-1}(i) \geq u'\}}{\big(\sum_{c'=u'}^{r_{j,a}} \exp\big(\theta_{\sigma(c')}\big)\big)^2}, C_{\sigma,i,i'} \equiv \sum_{u'=1}^{m_{j,a}} \frac{\mathbb{I}\{\sigma^{-1}(i), \sigma^{-1}(i') \geq u'\}}{\big(\sum_{c'=u'}^{r_{j,a}} \exp\big(\theta_{\sigma(c')}\big)\big)^2}.
\end{aligned} \tag{36}$$

First, a few observations about the expression of A_σ. For any $\sigma \in \Lambda_{T(e_{j,a})}$ and any $i \in V(e_{j,a})$, θ_i is in the numerator if and only if $i \in T(e_{j,a})$, since in all the rankings that are consistent with

the observation $e_{j,a}$, $T(e_{j,a})$ items are ranked in top $m_{j,a}$ positions. For any $\sigma \in \Lambda_{T(e_{j,a})}$ and any $i \in B(e_{j,a})$, θ_i is in all the product terms $\prod_{u=1}^{m_{j,a}}(\cdot)$ of the denominator, since in all the consistent rankings these items are ranked below $m_{j,a}$ position. For any $i \in T(e_{j,a})$, θ_i appears in product term corresponding to index u if and only if item i is ranked at position u or lower than u in the ranking $\sigma \in \Lambda_{T(e_{j,a})}$. Now, observe that B_σ is defined such that the partial derivative of A_σ with respect to any $i \in B(e_{j,a})$ is $-A_\sigma B_\sigma e^{\theta_i}$, and $B_{\sigma,i}$ is defined such that the partial derivative of A_σ with respect to any $i \in T(e_{j,a})$ is $A_\sigma - A_\sigma B_\sigma e^{\theta_i}$. Further, observe that $-C_\sigma e^{\theta_i}$ is the partial derivative of B_σ with respect to $i \in B(e_{j,a})$, $-C_{\sigma,i}e^{\theta_i}$ is the partial derivative of $B_{\sigma,i}$ with respect to $i \in T(e_{j,a})$, and $-C_{\sigma,i}e^{\theta_{i'}}$ is the partial derivative of $B_{\sigma,i}$ with respect to $i' \in B(e_{j,a})$. $-C_{\sigma,i,i'}e^{\theta_{i'}}$ is the partial derivative of $B_{\sigma,i}$ with respect to $i' \neq i \in T(e_{j,a})$.

For ease of notation, we omit subscript (j,a) whenever it is clear from the context. Also, we use $\sum_\sigma$ to denote $\sum_{\sigma \in \Lambda_{T(e_{j,a})}}$. With the above defined notations, from (4), we have, $\mathbb{P}_\theta(e) = \sum_\sigma A_\sigma$. With the above given observations for the notations in (36), first partial derivative of $\mathbb{P}_\theta(e)$ can be expressed as following:

$$\frac{\partial \mathbb{P}_\theta(e)}{\partial \theta_i} = \begin{cases} \sum_\sigma \left(A_\sigma - A_\sigma B_{\sigma,i} e^{\theta_i}\right) & \text{if } \ i \in T(e_{j,a}) \\ \sum_\sigma \left(-A_\sigma B_\sigma e^{\theta_i}\right) & \text{if } \ i \in B(e_{j,a})\,. \end{cases} \tag{37}$$

It follows that for $i \neq i' \in V(e_{j,a})$,

$$\begin{aligned} &\frac{\partial^2 \mathbb{P}_\theta(e)}{\partial \theta_i \partial \theta_{i'}} \\ &= \begin{cases} \sum_\sigma \left((A_\sigma (B_\sigma)^2 + A_\sigma C_\sigma) e^{(\theta_i + \theta_{i'})}\right) & \text{if } i, i' \in B(e_{j,a}) \\ \sum_\sigma \left(A_\sigma - A_\sigma B_{\sigma,i'} e^{\theta_{i'}} + (A_\sigma B_{\sigma,i} B_{\sigma,i'} + A_\sigma C_{\sigma,i,i'}) e^{(\theta_i + \theta_{i'})} - A_\sigma B_{\sigma,i} e^{\theta_i}\right) & \text{if } i, i' \in T(e_{j,a}) \\ \sum_\sigma \left((A_\sigma B_\sigma B_{\sigma,i} + A_\sigma C_{\sigma,i}) e^{(\theta_i + \theta_{i'})} - A_\sigma B_\sigma e^{\theta_{i'}}\right) & \text{otherwise}\,. \end{cases} \end{aligned}$$

Using $\frac{\partial^2 \log \mathbb{P}_\theta(e)}{\partial \theta_i \partial \theta_{i'}} = \frac{1}{\mathbb{P}_\theta(e)} \frac{\partial^2 \mathbb{P}_\theta(e)}{\partial \theta_i \partial \theta_{i'}} - \frac{1}{(\mathbb{P}_\theta(e))^2} \frac{\partial \mathbb{P}_\theta(e)}{\partial \theta_i} \frac{\partial \mathbb{P}_\theta(e)}{\partial \theta_{i'}}$, with above derived first and second derivatives, and after following some algebra, we have

$$\begin{aligned} &\frac{(\mathbb{P}_\theta(e))^2}{e^{(\theta_i + \theta_{i'})}} \frac{\partial^2 \log \mathbb{P}_\theta(e)}{\partial \theta_i \partial \theta_{i'}} \\ &= \begin{cases} (\sum_\sigma A_\sigma)(\sum_\sigma A_\sigma (B_\sigma)^2) - (\sum_\sigma A_\sigma B_\sigma)^2 + (\sum_\sigma A_\sigma)(\sum_\sigma A_\sigma C_\sigma) & \text{if } \ i, i' \in B(e_{j,a}) \\ (\sum_\sigma A_\sigma)(\sum_\sigma A_\sigma B_{\sigma,i} B_{\sigma,i'} + A_\sigma C_{\sigma,i,i'}) - (\sum_\sigma A_\sigma B_{\sigma,i})(\sum_\sigma A_\sigma B_{\sigma,i'}) & \text{if } \ i, i' \in T(e_{j,a}) \\ (\sum_\sigma A_\sigma)(\sum_\sigma A_\sigma B_\sigma B_{\sigma,i} + A_\sigma C_{\sigma,i}) - (\sum_\sigma A_\sigma B_\sigma)(\sum_\sigma A_\sigma B_{\sigma,i}) & \text{otherwise}\,. \end{cases} \end{aligned} \tag{38}$$

Observe that from Cauchy-Schwartz inequality $(\sum_\sigma A_\sigma)(\sum_\sigma A_\sigma (B_\sigma)^2) - (\sum_\sigma A_\sigma B_\sigma)^2 \geq 0$. Also, we have $e^{(\theta_i + \theta_{i'})} C_\sigma \geq e^{-2b}(m/r^2)$ and $e^{\theta_i} B_{\sigma,i} \leq e^{\theta_i} B_\sigma \leq e^{2b}(m/(r-m+1))$ for any $i \in V(e_{j,a})$. This proves the desired claim (29).

Next we need to upper bound deviation of $-H(\theta)$ from its expectation. From (38), we have, $\left|\frac{\partial^2 \log \mathbb{P}_\theta(e_{j,a})}{\partial \theta_i \partial \theta_{i'}}\right| \leq 3e^{4b} m_{j,a}^2/(r_{j,a} - m_{j,a} + 1)^2 \leq 3e^{4b} \nu m_{j,a}/(\kappa_j(\kappa_j - 1))$, where the last inequality

follows from the definition of ν (13). Therefore,

$$\begin{aligned} -H(\theta) &\preceq 3e^{4b}\nu \sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\sum_{i<i'\in S_j} \mathbb{I}\{(i,i')\subseteq V(e_{j,a})\}\frac{m_{j,a}}{\kappa_j(\kappa_j-1)}(e_i-e_{i'})(e_i-e_{i'})^\top \\ &\preceq 3e^{4b}\nu \sum_{j=1}^{n}\sum_{i<i'\in S_j} \frac{\sum_{a=1}^{\ell_j} m_{j,a}}{\kappa_j(\kappa_j-1)}(e_i-e_{i'})(e_i-e_{i'})^\top \equiv \sum_{j=1}^{n} y_j L_j\,, \end{aligned}$$

where $y_j = (3e^{4b}\nu p_j)/(\kappa_j(\kappa_j-1))$ and $L_j = \sum_{i<i'\in S_j}(e_i-e_{i'})(e_i-e_{i'})^\top = \kappa_j \mathrm{diag}(e_{S_j}) - e_{S_j}e_{S_j}^\top$ for $e_{S_j} = \sum_{i\in S_j} e_i$. Observe that $\|y_jL_j\| \le (3e^{4b}\nu p_{\max})/\kappa_{\min}$. Moreover, $L_j^2 \preceq \kappa_j L_j$, and it follows that

$$\sum_{j=1}^{n} y_j^2 L_j^2 \preceq 9e^{8b}\nu^2 \sum_{j=1}^{n}\frac{p_j^2}{\kappa_j^2(\kappa_j-1)^2}\kappa_j L_j \preceq \frac{9e^{8b}\nu^2 p_{\max}}{\kappa_{\min}} L\,,$$

where we used the fact that $L = (p_j/(\kappa_j(\kappa_j-1)))\sum_{j=1}^n L_j$, for L defined in (10). Using $\lambda_d(L) = np/(\beta(d-1))$ from (11), it follows that $\|\sum_{j=1}^n \mathbb{E}_\theta[y_j^2 Y_j^2]\| \le \frac{9e^{8b}\nu^2 p_{\max}}{\kappa_{\min}}\frac{np}{\beta(d-1)}$. By the matrix Bernstien inequality, with probability at least $1-d^{-3}$,

$$\begin{aligned} \|H(\theta)-\mathbb{E}[H(\theta)]\| &\le 12e^{4b}\nu\sqrt{\frac{p_{\max}}{\kappa_{\min}}\frac{np}{\beta(d-1)}\log d} + \frac{8e^{4b}\nu p_{\max}\log d}{\kappa_{\min}} \\ &\le 16e^{4b}\nu\sqrt{\frac{p_{\max}}{\kappa_{\min}}\frac{np}{\beta(d-1)}\log d}\,, \end{aligned} \tag{39}$$

where the last inequality follows from the assumption on $n\kappa_{\min}$ given in (14).

7.5 Proof of Lemma 16

Claim (30)**:** Since providing a lower bound on $\mathbb{P}_\theta\big[\sigma^{-1}(i), \sigma^{-1}(i') > \ell\big]$ 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. Define $\widetilde{\alpha}_{i,i',\ell,\theta}$ as

$$\widetilde{\alpha}_{i,i',\ell,\theta} \equiv \max_{\ell'\in[\ell]}\max_{\substack{\Omega\subseteq S\setminus\{i,i'\}\\ :|\Omega|=\kappa-\ell'}} \left\{\frac{\exp(\theta_i)+\exp(\theta_{i'})}{\big(\sum_{j\in\Omega}\exp(\theta_j)\big)/|\Omega|}\right\}, \tag{40}$$

and $\alpha_{i,i',\ell,\theta} = \big\lceil \widetilde{\alpha}_{i,i',\ell,\theta}\big\rceil$. For ease of notation we remove the subscript from α and $\widetilde{\alpha}$. We denote the sum of the weights by $W \equiv \sum_{j\in S}\exp(\theta_j)$. We define a new set of parameters $\{\widetilde{\theta}_j\}_{j\in S}$:

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

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

$$\begin{aligned}
&\mathbb{P}_\theta\Big[\sigma^{-1}(i), \sigma^{-1}(i') > \ell\Big] \\
= &\sum_{\substack{j_1\in S\\ j_1\neq i,i'}} \left(\frac{\exp(\theta_{j_1})}{W}\sum_{\substack{j_2\in S\\ j_2\neq i,i',j_1}}\left(\frac{\exp(\theta_{j_2})}{W-\exp(\theta_{j_1})}\cdots\left(\sum_{\substack{j_\ell\in S\\ j_\ell\neq i,i',\\ j_1,\cdots,j_{\ell-1}}}\frac{\exp(\theta_{j_\ell})}{W-\sum_{k=j_1}^{j_{\ell-1}}\exp(\theta_k)}\right)\cdots\right)\right) \\
= &\sum_{\substack{j_1\in S\\ j_1\neq i,i'}}\left(\frac{\exp(\theta_{j_1})}{W-\exp(\theta_{j_1})}\cdots\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,i',\\ j_1,\cdots,j_{\ell-2}}}\left(\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-1}}\exp(\theta_k)}\sum_{\substack{j_\ell\in S\\ j_\ell\neq i,i',\\ j_1,\cdots,j_{\ell-1}}}\left(\frac{\exp(\theta_{j_\ell})}{W}\right)\cdots\right)\right)
\end{aligned} \tag{41}$$

Consider the second-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 (40), $\frac{\exp(\theta_i)+\exp(\theta_{i'})}{\sum_{j\in\Omega_\ell}\exp(\theta_j)} \le \frac{\widetilde{\alpha}}{\kappa-\ell}$. We have,

$$\begin{aligned}
&\sum_{j_{\ell-1}\in\Omega_\ell}\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-1}}\exp(\theta_k)} \\
= &\sum_{j_{\ell-1}\in\Omega_\ell}\frac{\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-2}}\exp(\theta_k)-\exp(\theta_{j_{\ell-1}})} \\
\ge &\frac{\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})}{W-\sum_{k=j_1}^{j_{\ell-2}}\exp(\theta_k)-\big(\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})\big)/|\Omega_\ell|} \qquad (42)\\
= &\frac{\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})}{\exp(\theta_i)+\exp(\theta_{i'})+\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})-\big(\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})\big)/|\Omega_\ell|} \\
= &\left(\frac{\exp(\theta_i)+\exp(\theta_{i'})}{\sum_{j_{\ell-1}\in\Omega_\ell}\exp(\theta_{j_{\ell-1}})}+1-\frac{1}{\kappa-\ell}\right)^{-1} \\
\ge &\left(\frac{\widetilde{\alpha}}{\kappa-\ell}+1-\frac{1}{\kappa-\ell}\right)^{-1} \qquad (43)\\
= &\frac{\kappa-\ell}{\widetilde{\alpha}+\kappa-\ell-1} = \sum_{j_{\ell-1}\in\Omega_\ell}\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-1}}\exp(\widetilde{\theta}_k)}\,, \qquad (44)
\end{aligned}$$

where (42) 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 (43) follows from the definition of $\widetilde{\alpha}_{i,i',\ell,\theta}$, (40), and the fact that $|\Omega_\ell| = \kappa-\ell$. Equation (44) 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 (41). Observe that $\frac{\exp(\theta_i)+\exp(\theta_{i'})}{\sum_{j\in\Omega_{\widetilde{\ell}}}\exp(\theta_j)} \le \alpha/|\Omega_{\widetilde{\ell}}|$. Therefore, each summation term in equation (41) 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

$$\begin{aligned}
&\mathbb{P}_\theta\Big\{\sigma^{-1}(i),\sigma^{-1}(i')>\ell\Big\}\\
\geq\ & \sum_{\substack{j_1\in S\\ j_1\neq i,i'}}\left(\frac{\exp(\widetilde{\theta}_{j_1})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})}\cdots\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,i',\\ j_1,\cdots,j_{\ell-2}}}\left(\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-1}}\exp(\widetilde{\theta}_k)}\sum_{\substack{j_\ell\in S\\ j_\ell\neq i,i',\\ j_1,\cdots,j_{\ell-1}}}\left(\frac{\exp(\theta_{j_\ell})}{W}\right)\cdots\right)\right)\\
\geq\ & e^{-4b}\sum_{\substack{j_1\in S\\ j_1\neq i,i'}}\left(\frac{\exp(\widetilde{\theta}_{j_1})}{\widetilde{W}-\exp(\widetilde{\theta}_{j_1})}\cdots\sum_{\substack{j_{\ell-1}\in S\\ j_{\ell-1}\neq i,i',\\ j_1,\cdots,j_{\ell-2}}}\left(\frac{\exp(\widetilde{\theta}_{j_{\ell-1}})}{\widetilde{W}-\sum_{k=j_1}^{j_{\ell-1}}\exp(\widetilde{\theta}_k)}\sum_{\substack{j_\ell\in S\\ j_\ell\neq i,i',\\ j_1,\cdots,j_{\ell-1}}}\left(\frac{\exp(\widetilde{\theta}_{j_\ell})}{\widetilde{W}}\right)\cdots\right)\right)\\
=\ & \big(e^{-4b}\big)\mathbb{P}_{\widetilde{\theta}}\Big\{\sigma^{-1}(i),\sigma^{-1}(i')>\ell\Big\}\,. \qquad (45)
\end{aligned}$$

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 $\exp(\widetilde{\theta}_i)+\exp(\widetilde{\theta}_{i'})=\widetilde{\alpha}\leq\lceil\widetilde{\alpha}\rceil=\alpha\geq 1$. Therefore, we have

$$\begin{aligned}
\mathbb{P}_{\widetilde{\theta}}\Big\{\sigma^{-1}(i),\sigma^{-1}(i')>\ell\Big\} &= \binom{\kappa-2}{\ell}\frac{\ell\,!}{(\kappa-2+\widetilde{\alpha})(\kappa-2+\widetilde{\alpha}-1)\cdots(\kappa-2+\widetilde{\alpha}-(\ell-1))}\\
&\geq \frac{(\kappa-2)!}{(\kappa-\ell-2)!}\frac{1}{(\kappa+\alpha-2)(\kappa+\alpha-3)\cdots(\kappa+\alpha-(\ell+1))}\\
&\geq \frac{(\kappa-\ell+\alpha-2)(\kappa-\ell+\alpha-3)\cdots(\kappa-\ell-1)}{(\kappa+\alpha-2)(\kappa+\alpha-3)\cdots(\kappa-1)}\\
&\geq \frac{(\kappa-\ell)(\kappa-\ell-1)}{\kappa(\kappa-1)}\left(1-\frac{\ell}{\kappa+1}\right)^{\alpha-2}. \qquad (46)
\end{aligned}$$

Claim (30) follows by combining Equations (45) and (46) and using the fact that $\alpha\leq 2e^{2b}$.

Claim (31)**:** Define,

$$\widetilde{\alpha}_{\ell,\theta}\ \equiv\ \min_{i\in S}\min_{\ell'\in[\ell]}\min_{\substack{\Omega\in S\setminus\{i\}\\ :|\Omega|=\kappa-\ell'+1}}\left\{\frac{\exp(\theta_i)}{\big(\sum_{j\in\Omega}\exp(\theta_j)\big)/|\Omega|}\right\}. \qquad (47)$$

Also, define $\alpha_{\ell,\theta}\equiv\lfloor\widetilde{\alpha}_{\ell,\theta}\rfloor$. Note that $\alpha_{\ell,\theta}\geq 0$ and $\widetilde{\alpha}_{\ell,\theta}\leq e^{2b}$. We denote the sum of the weights by $W\equiv\sum_{j\in S}\exp(\theta_j)$. Analogous to the proof of claim (30), we define the new set of parameters $\{\widetilde{\theta}_j\}_{j\in S}$:

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

Similarly define $\widetilde{W}\equiv\sum_{j\in S}\exp(\widetilde{\theta}_j)=\kappa-1+\widetilde{\alpha}_{\ell,\theta}$. Using the techniques similar to the ones used in proof of claim (30), we have,

$$\mathbb{P}_\theta\Big\{\sigma^{-1}(i)=\ell\Big\}\ \leq\ e^{4b}\mathbb{P}_{\widetilde{\theta}}\Big\{\sigma^{-1}(i)=\ell\Big\}\,. \qquad (48)$$

Observe that $\exp(\widetilde{\theta}_j) = 1$ for all $j \neq i$ and $\exp(\widetilde{\theta}_i) = \widetilde{\alpha}_{\ell,\theta} \geq \lfloor \widetilde{\alpha}_{\ell,\theta} \rfloor = \alpha_{\ell,\theta} \geq 0$. Therefore, we have

$$\begin{aligned}
\mathbb{P}_{\widetilde{\theta}}\Big[\sigma^{-1}(i) = \ell\Big] &= \binom{\kappa-1}{\ell-1} \frac{\widetilde{\alpha}_{\ell,\theta}(\ell-1)!}{(\kappa-1+\widetilde{\alpha}_{\ell,\theta})(\kappa-2+\widetilde{\alpha}_{\ell,\theta})\cdots(\kappa-\ell+\widetilde{\alpha}_{\ell,\theta})} \\
&\leq \frac{(\kappa-1)!}{(\kappa-\ell)!} \frac{e^{2b}}{(\kappa-1+\alpha_{\ell,\theta})(\kappa-2+\alpha_{\ell,\theta})\cdots(\kappa-\ell+\alpha_{\ell,\theta})} \\
&\leq \frac{e^{2b}}{\kappa}\left(1 - \frac{\ell}{\kappa+\alpha_{\ell,\theta}}\right)^{\alpha_{\ell,\theta}-1} \leq \frac{e^{2b}}{\kappa-\ell}\,. \qquad (49)
\end{aligned}$$

Claim 31 follows by combining Equations (48) and (49).

Claim (32)**:** Again, we construct a new set of parameters $\{\widetilde{\theta}_j\}_{j\in[d]}$ from the original θ using $\widetilde{\alpha}_{\ell,\theta}$ defined in (47):

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

Similarly define $\widetilde{W} \equiv \sum_{j\in S} \exp(\widetilde{\theta}_j) = \kappa - 2 + 2\widetilde{\alpha}_{\ell,\theta}$. Using the techniques similar to the ones used in proof of claim (30), we have,

$$\mathbb{P}_{\theta}\Big[\sigma^{-1}(i) = \ell_1, \sigma^{-1}(i') = \ell_2\Big] \leq e^{8b}\mathbb{P}_{\widetilde{\theta}}\Big[\sigma^{-1}(i) = \ell_1, \sigma^{-1}(i') = \ell_2\Big] \qquad (50)$$

Observe that $\exp(\widetilde{\theta}_j) = 1$ for all $j \neq i, i'$ and $\exp(\widetilde{\theta}_i) = \exp(\widetilde{\theta}_{i'}) = \widetilde{\alpha}_{\ell,\theta} \geq \lfloor \widetilde{\alpha} \rfloor_{\ell,\theta} = \alpha_{\ell,\theta} \geq 0$. Therefore, we have

$$\begin{aligned}
&= \mathbb{P}_{\widetilde{\theta}}\Big[\sigma^{-1}(i) = \ell_1, \sigma^{-1}(i') = \ell_2\Big] \\
&= \Bigg(\frac{\binom{\kappa-2}{\ell_2-2}\widetilde{\alpha}^2_{\ell,\theta}(\ell_2-2)!}{(\kappa-2+2\widetilde{\alpha}_{\ell,\theta})(\kappa-1+2\widetilde{\alpha}_{\ell,\theta})\cdots(\kappa-2+2\widetilde{\alpha}_{\ell,\theta}-(\ell_1-1))} \\
&\frac{1}{(\kappa-2+\widetilde{\alpha}_{\ell,\theta}-(\ell_1-1))\cdots(\kappa-2+\widetilde{\alpha}_{\ell,\theta}-(\ell_2-2))} \Bigg) \\
&\leq \frac{(\kappa-2)!}{(\kappa-\ell_2)!} \frac{e^{4b}}{(\kappa-2)(\kappa-1)\cdots(\kappa-\ell_1-1)(\kappa-\ell_1-1)\cdots(\kappa-\ell_2)} \\
&\leq \frac{e^{4b}}{(\kappa-\ell_1-1)(\kappa-\ell_2)}\,. \qquad (51)
\end{aligned}$$

Claim 32 follows by combining Equations (50) and (51).

7.6 Proof of Theorem 12

Let $H(\theta) \in \mathcal{S}^d$ be Hessian matrix such that $H_{ii'}(\theta) = \frac{\partial^2 \mathcal{L}_{\mathrm{RB}}(\theta)}{\partial\theta_i \partial\theta_{i'}}$. The Fisher information matrix is defined as $I(\theta) = -\mathbb{E}_\theta[H(\theta)]$. From lemma 1, $\mathcal{L}_{\mathrm{RB}}(\theta)$ is concave. This implies that $I(\theta)$ is positive-semidefinite and from (22) its smallest eigenvalue is zero with all-ones being the corresponding eigenvector. 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

Cramer-Rao lower bound then implies that $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2] \geq \sum_{i=2}^{d} \frac{1}{\lambda_i(I(\theta))}$. Taking supremum over both sides gives

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

In the following, we will show that

$$\begin{aligned} I(\mathbf{0}) \;=\; -\mathbb{E}_\theta[H(\mathbf{0})] \;\; &\preceq \;\; \sum_{j=1}^{n} \sum_{a=1}^{\ell_j} \frac{m_{j,a} - \eta_{j,a}}{\kappa_j(\kappa_j - 1)} \sum_{i<i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \\ &\preceq \;\; \max_{j,a} \big\{ m_{j,a} - \eta_{j,a} \big\}\, L\,. \end{aligned} \tag{52}$$

Using Jensen's inequality, we have $\sum_{i=2}^{d} \frac{1}{\lambda_i(I(\mathbf{0}))} \geq \frac{(d-1)^2}{\sum_{i=2}^{d} \lambda_i(I(\mathbf{0}))} = \frac{(d-1)^2}{\mathrm{Tr}(I(\mathbf{0}))}$. From (52), we have $\mathrm{Tr}(I(\mathbf{0})) \leq \sum_{j,a}(m_{j,a} - \eta_{j,a})$. From (52), we have $\sum_{i=2}^{d} 1/\lambda_i(I(\mathbf{0})) \geq (1/\max\{m_{j,a} - \eta_{j,a}\}) \sum_{i=1}^{d} 1/\lambda_i(L)$. This proves the desired claim.

Now we are left to show claim (52). Consider a rank-breaking edge $e_{j,a}$. Using notations defined in lemma 15, in particular Equation (36), and omitting subscript $\{j, a\}$ whenever it is clear from the context, we have, for any $i \in V(e_{j,a})$,

$$\frac{\partial^2 \mathbb{P}_\theta(e_{j,a})}{\partial^2 \theta_i} = \begin{cases} \sum_\sigma \big(- A_\sigma B_\sigma e^{\theta_i} + A_\sigma (B_\sigma)^2 e^{2\theta_i} + A_\sigma C_\sigma e^{\theta_i} \big) & \text{if } \; i \in B(e_{j,a}) \\ \sum_\sigma \big(A_\sigma - 3 A_\sigma B_{\sigma,i} e^{\theta_i} + A_\sigma C_{\sigma,i}) e^{2\theta_i} + A_\sigma (B_{\sigma,i})^2 e^{2\theta_i} \big) & \text{if } \; i \in T(e_{j,a})\,, \end{cases}$$

and using (37), we have

$$\frac{\partial^2 \log \mathbb{P}_\theta(e_{j,a})}{\partial^2 \theta_i}\Big|_{\theta=\mathbf{0}} = \begin{cases} \big((C_\sigma - B_\sigma) \big)_{\theta=\mathbf{0}} & \text{if } \; i \in B(e_{j,a}) \\ \big(\frac{1}{m_{j,a}!} \sum_\sigma \big(C_{\sigma,i} - B_{\sigma,i} + (B_{\sigma,i})^2 \big) - \big(\sum_\sigma \frac{B_{\sigma,i}}{m_{j,a}!} \big)^2 \big)_{\theta=\mathbf{0}} & \text{if } \; i \in T(e_{j,a})\,, \end{cases}$$

where $\sigma \in \Lambda_{T(e_{j,a})}$ and the subscript $\theta = 0$ indicates the the respective quantities are evaluated at $\theta = 0$. From the definitions given in (36), for $\theta = \mathbf{0}$, we have $B_\sigma - C_\sigma = \sum_{u=0}^{m-1} \frac{(r-u-1)}{(r-u)^2}$ and, $\sum_\sigma (B_{\sigma,i} - C_{\sigma,i})/(m!) = \frac{1}{m} \sum_{u=0}^{m-1} \frac{(m-u)(r-u-1)}{(r-u)^2}$. Also, $\sum_\sigma B_{\sigma,i}/(m!) = \frac{1}{m} \sum_{u=0}^{m-1} \frac{m-u}{r-u}$ and $\sum_\sigma (B_{\sigma,i})^2/(m!) = \frac{1}{m} \sum_{u=0}^{m-1} \big(\sum_{u'=0}^{u} \frac{1}{r-u'} \big)^2$. Combining all these and, using $\mathbb{P}_{\theta=\mathbf{0}}[i \in T(e_{j,a})] = m/\kappa$ and $\mathbb{P}_{\theta=\mathbf{0}}[i \in B(e_{j,a})] = (r-m)/\kappa$, and after following some algebra, we have for any $i \in S_j$,

$$\begin{aligned} & -\mathbb{E}\bigg[\frac{\partial^2 \log \mathbb{P}_\theta(e_{j,a})}{\partial^2 \theta_i}\Big|_{\theta=\mathbf{0}} \bigg] \\ = \;\; & \frac{1}{\kappa} \bigg(m - \sum_{u=0}^{m-1} \frac{1}{r-u} - \frac{1}{m} \sum_{u=0}^{m-1} \frac{u(m-u)}{(r-u)^2} - \frac{1}{m} \sum_{u=0}^{m-2} \frac{2u}{r-u} \bigg(\sum_{u'>u}^{m-1} \frac{m-u'}{r-u'} \bigg) \bigg) \\ = \;\; & \frac{m_{j,a} - \eta_{j,a}}{\kappa_j}\,, \end{aligned} \tag{53}$$

where $\eta_{j,a}$ is defined in (16). Since row-sums of $H(\theta)$ are zeroes, (22), and for $\theta = \mathbf{0}$, all the items are exchangeable, we have for any $i \neq i' \in S_j$,

$$\mathbb{E}\bigg[\frac{\partial^2 \log \mathbb{P}_\theta(e_{j,a})}{\partial \theta_i \partial \theta_{i'}}\Big|_{\theta=\mathbf{0}} \bigg] \;\; = \;\; \frac{m_{j,a} - \eta_{j,a}}{\kappa_j(\kappa_j - 1)}\,,$$

The claim (52) follows from the expression of $H(\theta)$, Equation (22).

To verify (53), observe that $(r-m)(B_\sigma - C_\sigma) + m(\sum_\sigma B_{\sigma,i}/(m!)) = m - \sum_{u=0}^{m-1} \frac{1}{r-u}$. And,

$$\begin{aligned}
&\frac{1}{m}\Big(\sum_{u=0}^{m-1}\frac{m-u}{r-u}\Big)^2 - \sum_{u=0}^{m-1}\Big(\sum_{u'=0}^{u}\frac{1}{r-u'}\Big)^2 \\
= &\sum_{u=0}^{m-1}\left(\frac{(m-u)^2}{m(r-u)^2} - \frac{m-u}{(r-u)^2}\right) + \sum_{0\le u<u'\le m-1}\left(\frac{2(m-u)(m-u')}{m(r-u)(r-u')} - \frac{2(m-u')}{(r-u)(r-u')}\right) \\
= &\sum_{u=0}^{m-1}\frac{-u(m-u)}{m(r-u)^2} + \sum_{0\le u<u'\le m-1}\frac{-2u(m-u')}{m(r-u)(r-u')}\,.
\end{aligned}$$

7.7 Tightening of Lemma 15

Recall that $\mathbb{P}_\theta(e_{j,a})$ is same as probability of $\mathbb{P}_\theta[T(e_{j,a}) \succ B(e_{j,a})]$ that is the probability that an agent ranks $T(e_{j,a})$ items above $B(e_{j,a})$ items when provided with a set comprising $V(e_{j,a})$ items. As earlier, for brevity of notations, we omit subscript $\{j,a\}$ whenever it is clear from the context. For $m = 1$ or 2, it is easy to check that all off-diagonal elements in hessian matrix of $\log \mathbb{P}_\theta(e)$ are non-negative. However, since number of terms in summation in $\mathbb{P}_\theta(e)$ grows as $m!$, for $m \ge 3$ the straight-forward approach becomes too complex. Below, we derive expressions for cross-derivatives in hessian, for general m, using alternate definition (sorting of independent exponential r.v.'s in increasing order) of PL model, where the number of terms grow only as 2^m. However, we are unable to analytically prove that the cross-derivatives are non-negative for $m > 2$. Feeding these expressions in MATLAB and using symbolic computation, for $m = 3$, we can simplify these expressions and it turns out that they are sum of only positive numbers. For $m = 4$, with limited computational power it becomes intractable. We believe that it should hold for any value of $m < r$. Using (29), we need to check only for cross-derivatives for the case when $i \neq i' \in T(e_{j,a})$ or $i \in T(e_{j,a}), i' \in B(e_{j,a})$. Since, minimum of exponential random variables is an exponential random variable, we can assume that $|B(e_{j,a})| = 1$ that is $r = m+1$. Define $\lambda_i \equiv e^{\theta_i}$. Without loss of generality, assume $T(e_{j,a}) = \{2, \cdots, m+1\}$ and $B(e_{j,a}) = \{1\}$. Define $C_x = \prod_{i=3}^{m+1}(1 - e^{-\lambda_i x})$. Then, using the alternate definition of the PL model, we have, $\mathbb{P}_\theta(e) = \int_0^\infty C_x(1-e^{-\lambda_2 x})\lambda_1 e^{-\lambda_1 x}dx$. Following some algebra, $\frac{\partial^2 \log \mathbb{P}_\theta(e)}{\partial\theta_1\partial\theta_2} \ge 0$ is equivalent to $A_1 \ge 0$, where $A_1 \equiv$

$$\left(\int C_x\big(xe^{-\lambda_1 x} - xe^{-\lambda x}\big)dx\right)\left(\int C_x x e^{-\lambda x}dx\right) - \left(\int C_x(e^{\lambda_1 x} - e^{-\lambda x})dx\right)\left(\int C_x x^2 e^{-\lambda x}dx\right),$$

where all integrals are from 0 to ∞ and, $\lambda \equiv \lambda_1 + \lambda_2$. Consider A_1 as a function of λ_1. Since $A_1(\lambda_1) = 0$ for $\lambda_1 = \lambda$, showing $\partial A_1/\partial\lambda_1 \le 0$ for $0 \le \lambda_1 \le \lambda$ would suffice. Following some algebra, and using $\lambda_1 \le \lambda$, $\partial A_1/\partial\lambda_1 \le 0$ is equivalent to $A_2(\lambda_1) \equiv \big(\int_0^\infty C_x x e^{-\lambda_1 x}\big)/\big(\int_0^\infty C_x x^2 e^{-\lambda_1 x}\big)$ being monotonically non-decreasing in λ_1. To further simplify the condition, define $f^{(0)}(y) = 1/y^2$, $g^{(0)}(y) = 1/y^3$ and, $f^{(1)}(y) = f^{(0)}(y) - f^{(0)}(y+\lambda_3)$, and recursively $f^{(m-1)}(y) = f^{(m-2)}(y) - f^{(m-2)}(y+\lambda_{m+1})$. Similarly define $g^{(0)}, \cdots, g^{(m-1)}$. Using these recursively defined functions,

$$\begin{aligned}
2A_2(\lambda_1) &= \frac{f^{(m-1)}(\lambda_1)}{g^{(m-1)}(\lambda_1)}\,, \\
\text{for } m=3,\ 2A_2(\lambda_1) &= \frac{\lambda_1^{-2} - (\lambda_1+\lambda_3)^{-2} - (\lambda_1+\lambda_4)^{-2} + (\lambda_1+\lambda_3+\lambda_4)^{-2}}{\lambda_1^{-3} - (\lambda_1+\lambda_3)^{-3} - (\lambda_1+\lambda_4)^{-3} + (\lambda_1+\lambda_3+\lambda_4)^{-3}}\,.
\end{aligned}$$

Therefore, we need to show that $A_2(\lambda_1)$ is monotonically non-decreasing in $\lambda_1 \geq 0$ for any non-negative $\lambda_3, \cdots, \lambda_m$, and that would suffice to prove that the cross-derivatives arising from $i \in T(e_{j,a}), i' \in B(e_{j,a})$ are non-negative.

For cross-derivatives arising from $i \neq i' \in T(e_{j,a})$, define $B_x = \prod_{i=4}^{m+1}(1-e^{\lambda_i x})e^{-\lambda_1 x}$. $\frac{\partial^2 \log \mathbb{P}_\theta(e)}{\partial \theta_2 \partial \theta_3} \geq 0$ is equivalent to $A_3 \geq 0$, where $A_3 \equiv$

$$\left(\int B_x(1-e^{-\lambda_2 x})(1-e^{-\lambda_3 x})dx\right)\left(\int B_x x^2 e^{-(\lambda_2+\lambda_3)x}dx\right)$$
$$-\left(\int B_x(1-e^{-\lambda_2 x})xe^{-\lambda_3 x}dx\right)\left(\int B_x(1-e^{-\lambda_3 x})xe^{-\lambda_2 x}dx\right),$$

where all integrals are from 0 to ∞. For $m = 3$, using MATLAB one can check that the above stated conditions hold true. Therefore both types of cross-derivatives are non-negative.

Acknowledgements

This work is supported by NSF SaTC award CNS-1527754, and NSF CISE award CCF-1553452.

References

A. Agarwal, P. L. Bartlett, and J. C. Duchi. Oracle inequalities for computationally adaptive model selection. *arXiv preprint arXiv:1208.0129*, 2012.

A. Ali and M. Meilă. Experiments with kemeny ranking: What works when? *Mathematical Social Sciences*, 64(1):28–40, 2012.

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.

Michael A Bender, Martin Farach-Colton, Giridhar Pemmasani, Steven Skiena, and Pavel Sumazin. Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms*, 57(2):75–94, 2005.

N. Betzler, R. Bredereck, and R. Niedermeier. Theoretical and empirical evaluation of data reduction for exact kemeny rank aggregation. *Autonomous Agents and Multi-Agent Systems*, 28(5): 721–748, 2014.

O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

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

Artur Czumaj, MirosIaw Kowaluk, and Andrzej Lingas. Faster algorithms for finding lowest common ancestors in directed acyclic graphs. *Theoretical Computer Science*, 380(1-2):37–46, 2007.

Y. Deshpande and A. Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. *arXiv preprint arXiv:1502.06590*, 2015.

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.

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. *Ann. of Stat.*, pages 384–406, 2004.

David R Hunter and Kenneth Lange. Rejoinder. *Journal of Computational and Graphical Statistics*, 9(1):52–59, 2000.

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.

A. Khetan and S. Oh. Data-driven rank breaking for efficient rank aggregation. In *International Conference on Machine Learning*, 2016.

M. Lucic, M. I. Ohannessian, A. Karbasi, and A. Krause. Tradeoffs for space, time, data and risk in unsupervised learning. In *AISTATS*, 2015.

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

R. Meka, A. Potechin, and A. Wigderson. Sum-of-squares lower bounds for planted clique. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 87–96. ACM, 2015.

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

A. Prékopa. Logarithmic concave measures and related topics. In *Stochastic programming*, 1980.

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: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.

S. Shalev-Shwartz and N. Srebro. Svm optimization: inverse dependence on training set size. In *Proceedings of the 25th international conference on Machine learning*, pages 928–935. ACM, 2008.

G. Simons and Y. Yao. Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.

E. Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeit-srechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929.