Data-driven Rank Breaking for Efficient Rank Aggregation

Ashish Khetan KHETAN2@ILLINOIS.EDU
ISE Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA

Sewoong Oh SWOH@ILLINOIS.EDU
ISE Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA

Abstract

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

1. Introduction

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

Traditionally, learning to rank has been studied under the following data collection scenarios: pairwise comparisons, best-out-of-k comparisons, and k-way comparisons. *Pairwise comparisons* are commonly studied in the classical context of sports matches as well as more recent applications in crowdsourcing, where each worker is presented with a pair of choices and asked to choose the more favorable one. *Best-out-of-k comparisons* datasets are commonly available from purchase history of customers. Typically, a set of k alternatives are offered among which one is chosen or purchased by each customer. This has been widely studied in operations research in the context of modeling customer choices for revenue management and assortment optimization. The *k-way comparisons* are assumed in traditional rank aggregation scenarios, where each person reveals his/her preference as a ranked list over a set of k items. In some real-world elections, voters provide ranked preferences over the whole set of candidates (Lundell, 2007). We refer to these three types of ordinal data collection scenarios as 'traditional' throughout.

For such traditional datasets, there are several computationally efficient inference algorithms for finding the Maximum Likelihood (ML) estimates that provably achieve the minimax optimal performance (Negahban et al., 2012; Shah et al., 2015a; Hajek et al., 2014). However, modern datasets can be unstructured. This calls for a more flexible approaches for rank aggregation that can take such diverse forms of ordinal data into account. For such non-traditional datasets, finding the ML estimate can become significantly more challenging, requiring run-time exponential in the problem parameters.

To avoid such a computational bottleneck, a common heuristic is to resort to *rank-breaking*. The collected ordinal data is first transformed into a bag of pairwise com-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

parisons, ignoring the dependencies that were present in the original data. This is then processed via existing inference algorithms tailored for *independent* pairwise comparisons, hoping that the dependency present in the input data does not introduce bias. This idea is one of the main motivations for numerous approaches specializing in learning to rank from pairwise comparisons, e.g. (Ford Jr., 1957; Negahban et al., 2014; Azari Soufiani et al., 2013). However, such a heuristic of full rank-breaking, where all pairs are weighted and treated equally, has been recently shown to introduce estimation bias (Azari Soufiani et al., 2014).

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

Problem formulation. Users' revealed preferences are expressed in the form of *partial orderings*. The data from each individual can be represented by a *partially ordered set (poset)*. Assuming consistency in a user's revealed preferences, any ordered relations can be seamlessly translated into a poset, represented by a directed acyclic graph (DAG). The DAG below represents ordered relations $a > \{b, d\}$, $b > c$, $\{c, d\} > e$, and $e > f$. For example, this could have been translated from two sources: a five star rating on a and a three star ratings on b, c, d, a two star rating on e, and a one star rating on f; and the item b being purchased after reviewing c as well.

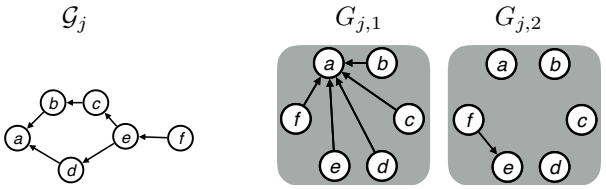

Figure 1. A DAG representation of consistent partial ordering of a user j (top). Two rank-breaking graphs extracted from $\mathcal{G}_j$ for the separator item a and e, respectively (bottom).

There are n users or agents, and each agent j provides his/her ordinal evaluation on a subset S_j of d items or alternatives. We refer to $S_j \subseteq \{1, 2, \ldots, d\}$ as *offerings* provided to j, and use $\kappa_j = |S_j|$ to denote the size of the offerings. We assume that the partial ordering over the offerings is a manifestation of her preferences as per a popular choice model known as Plackett-Luce (PL) model.

The PL model is a special case of *random utility models*, defined as follows (Walker & Ben-Akiva, 2002; Azari Soufiani et al., 2012). Each item i has a real-valued latent utility θ_i. When presented with a set of items, a user's reveled preference is a partial ordering according to noisy manifestation of the utilities, i.e. i.i.d. noise added to the true utility θ_i's. The PL model is a special case where the noise follows the standard Gumbel distribution, and is one of the most popular model in social choice theory (McFadden, 1973; McFadden & Train, 2000). PL has several important properties, making this model realistic in various domains, including marketing (Guadagni & Little, 1983), transportation (McFadden, 1980; Ben-Akiva & Lerman, 1985), biology (Sham & Curtis, 1995), and natural language processing (Mikolov et al., 2013). Precisely, each user j, when presented with a set S_j of items, draws a noisy utility of each item i according to

$$u_i \quad = \quad \theta_i + Z_i \ , \tag{1}$$

where Z_i's follow the independent standard Gumbel distribution. Then we observe the ranking resulting from sorting the items as per noisy observed utilities u_j's.

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

$$\begin{aligned} \mathbb{P}(a > b > c) = \ & \mathbb{P}(a > \{b, c\}) \ \mathbb{P}(b > c) \\ = \ & \frac{e^{\theta_a}}{(e^{\theta_a} + e^{\theta_b} + e^{\theta_c})} \ \frac{e^{\theta_b}}{(e^{\theta_b} + e^{\theta_c})} \ . \end{aligned} \tag{2}$$

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

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

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

We have n users each providing a partial ordering of a set of offerings S_j according to the PL model. Let $\mathcal{G}_j$ denote

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

$$\underset{\theta \in \Omega_b}{\text{maximize}} \quad \sum_{j=1}^{n} \log \mathbb{P}(\mathcal{G}_j) \; . \tag{3}$$

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

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

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

$$\widehat{\theta} \;\in\; \arg\max_{\theta \in \Omega_b} \sum_{(i > i') \in \mathcal{C}} \left(\theta_i - \log\left(e^{\theta_i} + e^{\theta_{i'}} \right) \right) . \tag{4}$$

There are several efficient implementation tailored for this problem (Ford Jr., 1957; Hunter, 2004; Negahban et al., 2012; Maystre & Grossglauser, 2015a), and under the traditional scenarios, these approaches provably achieve the minimax optimal rate (Hajek et al., 2014; Shah et al., 2015a). For general non-traditional datasets, there is a significant gain in computational complexity. In the case of position-p ranking, where each of the n users report his/her p-th ranking item among κ items, the computational complexity reduces from $O(n\,(p-1)!)$ for the MLE in (3) to $O(n\,p\,(\kappa - p))$ for the full rank-breaking estimator in (4). However, this gain comes at the cost of accuracy. It is known that the full-rank breaking estimator is inconsistent (Azari Soufiani et al., 2014); the error is strictly bounded away from zero even with infinite samples.

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

Recall that a user j provides his/her preference as a poset represented by a DAG $\mathcal{G}_j$. Consistent rank-breaking first identifies all *separators* in the DAG. A node in the DAG is a separator if one can partition the rest of the nodes into two parts. A partition A_{top} which is the set of items that are preferred over the separator item, and a partition A_{bottom} which is the set of items that are less preferred than the separator item. One caveat is that we allow A_{top} to be empty, but A_{bottom} must have at least one item. In the example in Figure 1, there are two separators: the item a and the item e. Using these separators, one can extract the following partial ordering from the original poset: $(a > \{b, c, d\} > e > f)$. The items a and e separate the set of offerings into partitions, hence the name separator. We use ℓ_j to denote the number of separators in the poset $\mathcal{G}_j$ from user j. We let $p_{j,a}$ denote the ranked position of the a-th separator in the poset $\mathcal{G}_j$, and we sort the positions such that $p_{j,1} \leq p_{j,2} \leq \ldots \leq p_{j,\ell_j}$. The set of separators is denoted by $\mathcal{P}_j = \{p_{j,1}, p_{j,2}, \cdots, p_{j,\ell_j}\}$. For example, the separator a is ranked at position 1 and e is at 5. Then, $\ell_j = 2$, $p_{j,1} = 1$, and $p_{j,2} = 5$. Note that f is not a separator (whereas a is) since corresponding A_{bottom} is empty.

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

This rank-breaking graphs were introduced in (Azari Soufiani et al., 2013), where it was shown that the pairwise

ordinal relations that is represented by edges in the rank-breaking graphs are sufficient information for using any estimation based on the idea of rank-breaking. Precisely, on the converse side, it was proved in (Azari Soufiani et al., 2014) that any pairwise outcomes that is not present in the rank-breaking graphs $G_{j,a}$'s introduces bias for a general θ^*. On the achievability side, it was proved that all pairwise outcomes that are present in the rank-breaking graphs are unbiased, as long as all the paired comparisons in each $G_{j,a}$ are weighted equally. In the algorithm described in (29), we satisfy this sufficient condition for consistency by restricting to a class of convex optimizations that use the same weight $\lambda_{j,a}$ for all $(\kappa - p_{j,a})$ paired comparisons in the objective function, as opposed to allowing more general weights that defer from a pair to another pair in a rank-breaking graph $G_{j,a}$.

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

$$\mathcal{L}_{\rm RB}(\theta) = \sum_{j=1}^{n}\sum_{a=1}^{\ell_j}\lambda_{j,a}\sum_{(i,i')\in E_{j,a}}\log\Big(\frac{e^{\theta_{i'}}}{e^{\theta_i}+e^{\theta_{i'}}}\Big), \quad (5)$$

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

$$\widehat{\theta} \in \arg\max_{\theta\in\Omega_b} \mathcal{L}_{\rm RB}(\theta)\,. \quad (6)$$

By changing how we weigh each rank-breaking graph (by choosing the $\lambda_{j,a}$'s), the convex program (6) spans the entire set of consistent rank-breaking estimators, as characterized in (Azari Soufiani et al., 2014). However, only asymptotic consistency was known, which holds independent of the choice of the weights $\lambda_{j,a}$'s. A uniform choice of $\lambda_{j,a} = \lambda$ was proposed in (Azari Soufiani et al., 2014).

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

Contributions. In this paper, we provide an upper bound on the error achieved by the rank-breaking estimator of (6) for any choice of the weights (Theorem 5 in the Appendix). This explicitly shows how the error depends on the choice of the weights, and provides a guideline for choosing the optimal weights $\lambda_{j,a}$'s in a data-driven manner. We provide the explicit formula for the optimal choice of the weights and provide the the error bound (Theorem 2). The analysis shows the dependence of the error in the dimension d and the number of users n that matches the experiments.

If we are designing surveys then we want to maximize the accuracy for a given number of questions asked. Our analysis provides how the accuracy depends on the topology of the collected data, and provides a guidance when we do have some control over which questions to ask and which data to collect. One should maximize the spectral gap of corresponding comparison graph. Further, for some canonical scenarios, we quantify the price of rank-breaking by comparing the error bound of the proposed data-driven rank-breaking with the lower bound on the MLE, which can have a significantly larger computational cost (Theorem 4). All the proofs and technical lemmas are provided in the Appendix in the included supplementary material.

Notations. For any positive integer N, let $[N] = \{1,\cdots,N\}$. For a ranking σ over S, i.e., σ is a mapping from $[|S|]$ to S, let σ^{-1} denote the inverse mapping. Let $\mathcal{S}^d$ denote the set of $d\times d$ symmetric matrices.

2. Comparison graph

In the analysis of the convex program (6), we show that, with high probability, the objective function is strictly concave with $\lambda_2(H(\theta)) \le -C_b\,\gamma\,\lambda_2(L) < 0$ (Lemma 8 in the Appendix) for all $\theta\in\Omega_b$ and the gradient is bounded by $\|\nabla\mathcal{L}_{\rm RB}(\theta^*)\|_2 \le C'_b\sqrt{\log d\,\sum_{j\in[n]}\ell_j}$ (Lemma 7 in the Appendix). Shortly, we will define γ and $\lambda_2(L)$, which captures the dependence on the topology of the data, and C'_b and C_b are constants that only depend on b. Putting these together, we show that there exists a $\theta\in\Omega_b$ such that

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

Here $\lambda_2(H(\theta))$ denotes the second largest eigenvalue of a negative semi-definite Hessian matrix $H(\theta)$ of the ob-

jective function. The reason the second largest eigenvalue shows up is because the top eigenvector is always the all-ones vector which by the definition of Ω_b is infeasible. The accuracy depends on the topology of the collected data via the comparison graph of given data.

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

$$A_{ii'} \quad = \quad \sum_{j\in[n]:i,i'\in S_j} \frac{\ell_j}{\kappa_j(\kappa_j-1)}\,, \tag{7}$$

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

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

$$L \quad = \quad \sum_{j=1}^{n} \frac{\ell_j}{\kappa_j(\kappa_j-1)} \sum_{i<i'\in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \tag{8}$$

Let $0 = \lambda_1(L) \le \lambda_2(L) \le \cdots \le \lambda_d(L)$ denote the (sorted) eigenvalues of L. Of special interest is $\lambda_2(L)$, also called the spectral gap, which measured how well-connected the graph is. Intuitively, one can expect better accuracy when the spectral gap is larger, as evidenced in previous learning to rank results in simpler settings (Negahban et al., 2014; Shah et al., 2015a; Hajek et al., 2014). This is made precise in (7), and in the main result of Theorem 2, we appropriately rescale the spectral gap and use $\alpha \in [0, 1]$ defined as

$$\alpha \quad \equiv \quad \frac{\lambda_2(L)(d-1)}{\mathrm{Tr}(L)} \quad = \quad \frac{\lambda_2(L)(d-1)}{\sum_{j=1}^n \ell_j}\,. \tag{9}$$

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

$$\beta \quad \equiv \quad \frac{\mathrm{Tr}(L)}{dD_{\max}} \quad = \quad \frac{\sum_{j=1}^n \ell_j}{dD_{\max}}\,. \tag{10}$$

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

$$\gamma \quad \equiv \quad \min_{j\in[n]} \left\{ \left(1 - \frac{p_{j,\ell_j}}{\kappa_j}\right)^{\lceil 2e^{2b}\rceil - 2} \right\}. \tag{11}$$

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

$$\eta \equiv \max_{j\in[n]}\{\eta_j\}, \text{ where } \eta_j = \frac{\kappa_j}{\max\{\ell_j, \kappa_j - p_{j,\ell_j}\}}\,. \tag{12}$$

Note that $1 < \eta_j \le \kappa_j/\ell_j$. We discuss the role of the topology of data captures by these parameters in Section 4.

3. Main results

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

3.1. Upper bound on the achievable error

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

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

$$\sum_{j=1}^{n} \ell_j \quad \ge \quad \frac{2^{11} e^{18b} \eta \log(\ell_{\max}+2)^2}{\alpha^2\gamma^2\beta} d\log d\,, \tag{13}$$

where $b \equiv \max_i |\theta_i^*|$ is the dynamic range, $\ell_{\max} \equiv \max_{j\in[n]} \ell_j$, α is the (rescaled) spectral gap defined in (9), β is the (rescaled) maximum degree defined in (10), γ and η are defined in Eqs. (11) and (12), then the rank-breaking estimator *in (6) with the choice of*

$$\lambda_{j,a} \quad = \quad \frac{1}{\kappa_j - p_{j,a}}\,, \tag{14}$$

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

$$\frac{1}{\sqrt{d}} \left\|\widehat{\theta} - \theta^*\right\|_2 \quad \le \quad \frac{4\sqrt{2}e^{4b}(1+e^{2b})^2}{\alpha\gamma} \sqrt{\frac{d\,\log d}{\sum_{j=1}^n \ell_j}}\,, \tag{15}$$

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

We refer to the proposed optimal choice of the weights in (14) as *data-driven rank-breaking estimator*. In the ideal

case where the spectral gap is large such that α is a strictly positive constant and the dynamic range b is finite and $\max_{j\in[n]} p_{j,\ell_j}/\kappa_j = C$ for some constant $C < 1$ such that γ is also a constant independent of the problem size d. Then the upper bound in (15) implies that we need the effective sample size to scale as $O(d \log d)$, which is only a logarithmic factor larger than the number of parameters to be estimated.

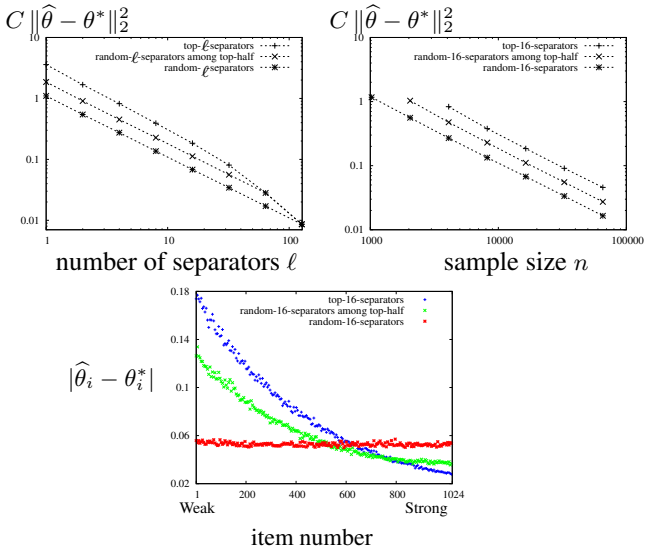

Figure 2. $\|\theta^* - \widehat{\theta}\|_2^2 \propto 1/(\ell\, n)$ as theoretically predicted and smaller error is achieved for separators that are well spread out.

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

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

3.2. The price of rank-breaking

Rank-breaking achieves computational efficiency at the cost of estimation accuracy. In this section, we quantify this tradeoff for a canonical example of position-p ranking, where each sample provides the following information: an unordered set of $p-1$ items that are ranked high, one item that is ranked at the p-th position, and the rest of $\kappa_j - p$ items that are ranked on the bottom. Since each sample has only one separator for $2 < p$, Theorem 2 simplifies to the following Corollary.

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

$$\frac{1}{\sqrt{d}}\left\|\widehat{\theta}-\theta^*\right\|_2 \;\leq\; \frac{c}{\alpha\gamma}\sqrt{\frac{d\,\log d}{n}}\,. \tag{16}$$

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

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

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

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

Note that the second inequality is tight up to a constant factor, when the graph is an expander with a large spectral gap. For expanders, α in the bound (16) is also a strictly positive constant. This suggests that rank-breaking gains in computational efficiency by a super-exponential factor of $(p-1)!$, at the price of increased error by a factor of p, ignoring poly-logarithmic factors.

3.3. Optimality of the choice of the weights

We propose the optimal choice of the weights $\lambda_{j,a}$'s in Theorem 2. In this section, we show numerical results comparing the proposed approach to other naive choices of the weights under various scenarios. The offering sets S_j's are chosen independently and uniformly at random from $[d]$.

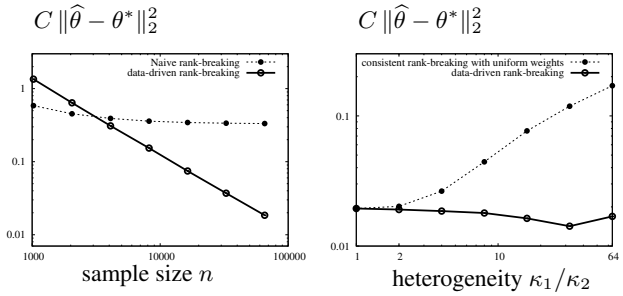

Figure 3. Left: Data-driven rank-breaking is consistent, while a random rank-breaking results in inconsistency. Right: The gain of choosing optimal $\lambda_{j,a}$'s is significant when κ_j's are highly heterogeneous.

In the left panel, Figure 3 illustrates that a naive choice of rank-breakings can result in inconsistency. We create partial orderings dataset by fixing $\kappa = 128$ and select $\ell = 8$ random positions in $\{1, \ldots, 127\}$. Each dataset consists of partial orderings with separators at those 8 random positions, over 128 randomly chosen subset of items. We vary the sample size n and plot the resulting mean squared error for the two approaches. The data-driven rank-breaking, which uses the optimal choice of the weights, achieves error scaling as $1/n$ as predicted by Theorem 2, which implies consistency. For fair comparisons, we feed the same number of pairwise orderings to a naive rank-breaking estimator. This estimator uses randomly chosen pairwise orderings with uniform weights, and is generally inconsistent. However, when sample size is small, inconsistent estimators can achieve smaller variance leading to smaller error. We use the minorization-maximization algorithm from (Hunter, 2004) for computing the estimates.

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

Figure 4 illustrates that the optimal choice of the weights improves over consistent rank-breaking with uniform weights by a constant factor. We fix $\kappa = 128$ and $n = 128000$. As illustrated by a figure on the right, the position of the separators are chosen such that there is one separator at position one, and the rest of $\ell - 1$ separators are at the bottom. Precisely, $(p_{j,1}, p_{j,2}, p_{j,3}, \ldots, p_{j,\ell}) = (1, 128 - \ell + 1, 128 - \ell + 2, \ldots, 127)$. We consider this scenario to emphasize the gain of optimal weights. Observe that the MSE does not decrease at a rate of $1/\ell$ in this case. The parameter γ which appears in the bound of Theorem 2 is very small when the breaking positions $p_{j,a}$ are of the order κ_j as is the case here, when ℓ is small.

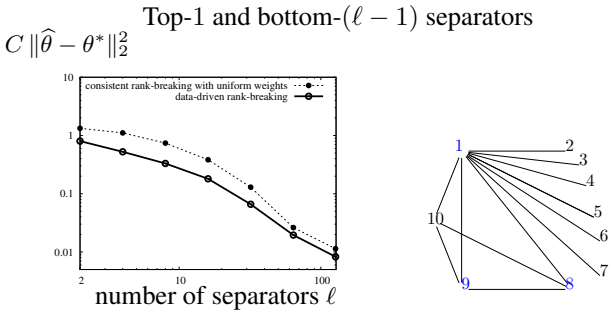

Figure 4. There is a constant factor gain of choosing optimal $\lambda_{j,a}$'s when the size of offerings are fixed, i.e. $\kappa_j = \kappa$ (left). We choose a particular set of separators where one separators is at position one and the rest are at the bottom. On right is an example of $\ell = 3$ and $\kappa = 10$ with separators indicated by blue.

The gain of optimal weights is significant when the size of S_j's are highly heterogeneous. In the right panel, Figure 3 compares performance of the proposed algorithm, for the optimal choice and uniform choice of weights $\lambda_{j,a}$ when the comparison sets S_j's are of different sizes. We consider the case when n_1 agents provide their top-ℓ_1 choices over the sets of size κ_1, and n_2 agents provide their top-1 choice over the sets of size κ_2. We take $n_1 = 1024$, $\ell_1 = 8$, and $n_2 = 10 n_1 \ell_1$. Figure 3, right panel, shows MSE for the two choice of weights, when we fix $\kappa_1 = 128$, and vary κ_2 from 2 to 128. As predicted from our bounds, when optimal choice of $\lambda_{j,a}$ is used MSE is not sensitive to sample set sizes κ_2.

4. The role of the topology of the data

Using the same number of samples, comparison graphs with larger spectral gap achieve better accuracy, compared to those with smaller spectral gaps. We consider a scenario where we fix the size of offerings as $\kappa_j = \kappa = O(1)$ and each agent provides partial ranking with ℓ separators, positions of which are chosen uniformly at random. The resulting spectral gap α of different choices of the set S_j's are provided below. The total number edges in the comparisons graph (counting hyper-edges as multiple edges) is defined as $|E| \equiv \binom{\kappa}{2} n$.

Complete graphs have a spectral gap of one, which is the maximum possible value. Hence, complete graph is optimal for rank aggregation. *Sparse random graphs*, in the regime with $n = \Omega(\log d)$ to ensure connectivity, has a strictly positive spectral gap. Hence, $\alpha = \Theta(1)$ and sparse random graphs are near-optimal for rank-aggregation. *Chain graphs* have $\alpha = \Theta(1/d^2)$. Hence, a chain graph is strictly sub-optimal for rank aggregation. *Star-like graphs* have $\alpha = \Theta(1)$ and star-like graphs are near-optimal for rank-aggregation. *Barbell-like graphs* have $\alpha = \Theta(1/d^2)$. Hence, a chain graph is strictly sub-optimal for rank aggregation.

Figure 5 illustrates how graph topology effects the accuracy. When θ^* is chosen uniformly at random, the accuracy does not change with d (left), and the accuracy is better for those graphs with larger spectral gap. However, for a certain worst-case θ^*, the error increases with d for the chain graph and the barbell-like graph, as predicted by the above analysis of the spectral gap. We use $\ell = 4$, $\kappa = 17$ and vary d from 129 to 2049. κ is kept small to make the resulting graphs more like the above discussed graphs. Figure on left shows accuracy when θ^* is chosen i.i.d. uniformly over $[-b, b]$ with $b = 2$. Error in this case is roughly same for each of the graph topologies with chain graph being the worst. However, when θ^* is chosen carefully error for chain graph and barbell-like graph increases with d as shown in the figure right. We chose θ^* such that all the items of a set have same weight, either $\theta_i = 0$ or $\theta_i = b$ for chain graph and barbell-like graph.

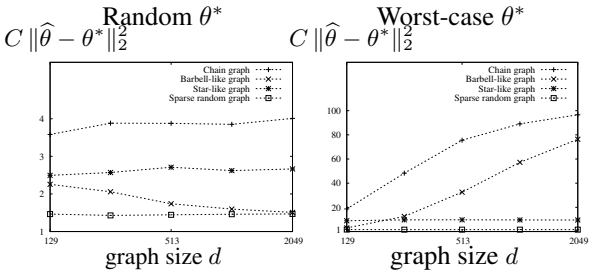

Figure 5. For random θ^* the error does not change with d (left). For a particular worst-case θ^* the error increases with d for the Chain and Barbell-like graphs as predicted by the theorem (right).

5. Real-world datasets

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

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

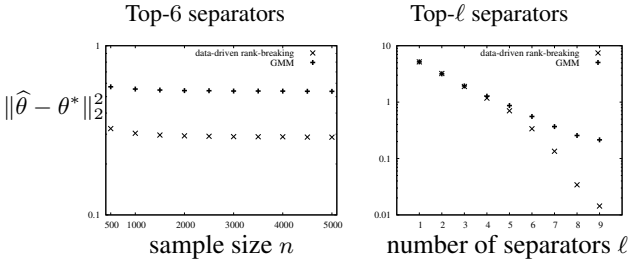

Figure 6. The data-driven rank-breaking achieves smaller error compared to the state-of-the-art GMM approach.

6. Discussion

We study the problem of learning the PL model from ordinal data. Under the traditional data collection scenarios, several efficient algorithms find the maximum likelihood estimates and at the same time provably achieve minimax optimal performance. However, for some non-traditional scenarios, computational complexity of finding the maximum likelihood estimate can scale super-exponentially in the problem size. We provide the first finite-sample analysis of computationally efficient estimators known as rank-breaking estimators. This provides guidelines for choosing the weights in the estimator to achieve optimal performance, and also explicitly shows how the accuracy depends on the topology of the data. An interesting future direction is relating this work to non-parametric learning from paired comparisons, initiated in several recent papers such as (Duchi et al., 2010; Rajkumar & Agarwal, 2014; Shah et al., 2015b; Shah & Wainwright, 2015). Another interesting future direction is adaptively choosing the offerings as in (Braverman & Mossel, 2009; Ailon, 2011; Jamieson & Nowak, 2011; Maystre & Grossglauser, 2015b).

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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