Learning from Comparisons and Choices

Sahand Negahban   SAHAND.NEGAHBAN@YALE.EDU
Statistics Department
Yale University

Sewoong Oh   SWOH@ILLINOIS.EDU
Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Kiran K. Thekumparampil   THEKUMP2@ILLINOIS.EDU
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Jiaming Xu   XU972@PURDUE.EDU
Krannert School of Management
Purdue University

Abstract

When tracking user-specific online activities, each user's preference is revealed in the form of choices and comparisons. For example, a user's purchase history is a record of her choices, i.e. which item was chosen among a subset of offerings. A user's preferences can be observed either explicitly as in movie ratings or implicitly as in viewing times of news articles. Given such individualized ordinal data in the form of comparisons and choices, we address the problem of collaboratively learning representations of the users and the items. The learned features can be used to predict a user's preference of an unseen item to be used in recommendation systems. This also allows one to compute similarities among users and items to be used for categorization and search. Motivated by the empirical successes of the MultiNomial Logit (MNL) model in marketing and transportation, and also more recent successes in word embedding and crowdsourced image embedding, we pose this problem as learning the MNL model parameters that best explain the data. We propose a convex relaxation for learning the MNL model, and show that it is minimax optimal up to a logarithmic factor by comparing its performance to a fundamental lower bound. This characterizes the minimax sample complexity of the problem, and proves that the proposed estimator cannot be improved upon other than by a logarithmic factor. Further, the analysis identifies how the accuracy depends on the topology of sampling via the spectrum of the sampling graph. This provides a guideline for designing surveys when one can choose which items are to be compared. This is accompanied by numerical simulations on synthetic and real data sets, confirming our theoretical predictions.

Keywords: Collaborative Ranking, Nuclear Norm Minimization, Multi-Nomial Logit Model

1. Introduction

Given data on how users compared subsets of items, we address the fundamental problem of learning a representation of users and items. Such data can be observed in the form

of choices (e.g. which item was bought) or in the form of comparisons (e.g. which items are rated higher). From such ordinal data on the items, we want to find low dimensional representations, which we call (latent) features, that explain crucial aspects of the users' choices. Once learned, these features can be used to predict each user's preference over items that the user has not seen yet, which can be used in recommendation systems and revenue management. These learned features also provide an embedding of the users and items on the same Euclidean space that allows us to directly quantify similarities via distances, that can be used to categorize and cluster. These embeddings can reveal the underlying structure of data such as images. Such an embedding of a discrete set of objects based on ordinal data has recently gained tremendous attraction mainly due to word embeddings based on co-occurrence data and their successes in numerous downstream natural language processing tasks Mikolov et al. (2013b).

The fundamental question in such a representation learning is: what makes one representation better than the others? Our guiding principle is that a good representation is the one that defines a generative model that best explains the given data in the maximum likelihood sense. To this end, we focus on a parametric generative model known as Multi-Nomial Logit (MNL) model, widely used and studied in revenue management. The MNL model has a natural interpretation of human choices as an outcome of maximizing a utility by agents with noisy perception of the utility, also known as *random utility model* in Walker and Ben-Akiva (2002); Azari Soufiani et al. (2012), defined as follows. Each user and item has a latent low-dimensional feature $u_i \in \mathbb{R}^r$ and $v_j \in \mathbb{R}^r$ respectively. The true utility of an item is the inner product of these two features $\Theta_{ij} \triangleq \langle\langle u_i, v_j \rangle\rangle = \sum_k u_{ik} v_{jk}$. The inherent low-rank structure of $\Theta = [\Theta_{ij}]$ captures the collaborative nature of the problem, where users with similar preferences in the past are likely to prefer similar items in the future.

When presented with a set of items, a user reveals a noisy ordering of the items sorted according to her perceived utilities of the items, each of which is perturbed by an i.i.d. noise added to the true utility Θ_{ij}. The MNL model is a special case where the noise follows the standard Gumbel distribution, and is one of the most popular models in choice theory for its simplicity and empirical success McFadden (1973); McFadden and Train (2000). The MNL model has several important properties, making this model realistic in various domains, including marketing Guadagni and Little (1983), transportation McFadden (1980); Ben-Akiva and Lerman (1985), biology Sham and Curtis (1995), sports games Tsokos et al. (2018) and natural language processing Mikolov et al. (2013a). The MNL model (i) satisfies the 'independence of irrelevant alternatives' in social choice theory Ray (1973); (ii) has a maximum likelihood estimator (MLE) which is a convex program in Θ; and (iii) has a simple characterization of sequential (random) choices as follows. Let $\mathbb{P}\{a > \{b, c, d\}\}$ denote the probability a was chosen as the best alternative among the set $\{a, b, c, d\}$. Then, the probability that user i reveals a linear order ($a > b > c > d$) is $\mathbb{P}\{a > \{b, c, d\}\}\mathbb{P}\{b > \{c, d\}\}\mathbb{P}\{c > d\}$, where $\mathbb{P}\{a > \{b, c, d\}\} = e^{\Theta_{ia}}/(e^{\Theta_{ia}} + e^{\Theta_{ib}} + e^{\Theta_{ic}} + e^{\Theta_{id}})$. Essentially the user is modeled as making a sequence of choices, choosing the best alternative first and then making choices on the remaining ones. We give the precise definition of the MNL model in Section 2 for pairwise comparisons and in Section 4 for higher order comparisons and choices. Beyond its success in classical applications such as transportation and marketing, the MNL model and its variants are being rediscovered and successfully applied in more modern applications

such as embedding images using crowdsourcing Tamuz et al. (2011) and word embedding Mikolov et al. (2013b), whose connections we make precise in Section 6.

Motivated by recent advances in learning low-rank models, e.g. Negahban et al. (2009); Davenport et al. (2014), we ask the fundamental question of learning the MNL model from data on comparisons and choices. We provide a general framework using convex relaxations for learning the model. As data is collected in various forms on modern social computing systems, we consider the following four canonical scenarios:

- *Pairwise comparisons.* The most simple and canonical piece of ordinal data one can collect from a user at a time is a pairwise comparison; given two options, we ask the user which one is better. Such data is prevalent in the real world and is the most popular scenario studied in ranking literature, e.g. Shah et al. (2014). However, one significant aspect of the real data that has not been addressed in the literature is irregularities in the sampling. Consider an online seller with various products, say cars and watches. It does not make sense to ask a user to compare a car and a watch; one cannot sample an outcome of a comparison between a watch and a car. However, knowing a user's preference on cars can help in learning her preference on watches. We want to propose a model and design an inference algorithm that can take into account such restrictions in sampling. We further want to quantify the gain in using all such data together in inference, as opposed to running inference in each category separately. To this end, we propose a new model for sampling that we call *graph sampling*. This model explains such irregularities in the real world data. We propose a novel inference algorithm tailored for the given sampling pattern. Our analysis captures precisely how the accuracy depends on the different topologies of the sampling.

- *Higher order comparisons.* Consider an online market that collects each of its user's preference as a ranking over a subset of items that is 'seen' by the user. Such data can be obtained by directly asking to compare some items, or by indirectly tracking online activities on which items are viewed, how much time is spent on the page or how the user rated the items. However, collecting such comparisons over multiple items might come at a cost. We, therefore, want to quantify the gain in the accuracy of the inference when higher order comparison outcomes are collected. We characterize the optimal trade-off between accuracy and the number of items compared, and show that our proposed algorithm seamlessly generalizes to this setting and also achieves the optimal trade-off.

- *Customer choices.* One of the most widely applicable data collection scenarios is customer purchase history. Online and offline service providers can track each customer on which subset of items is offered and which item is chosen. Given historical data on such choices on best-out-of-a-subset, we extract features on the users and items that best explains the collected data.

- *Bundled choices.* Another data collection scenario that is gaining interest recently is bundled choices Chu et al. (2011); Benson et al. (2018). Typical choice models assume that the willingness to buy an item is independent of what else the user bought. In many cases, however, we make 'bundled' purchases: we buy particular ingredients

together for one recipe or we buy two connecting flights. One choice (the first flight) has a significant impact on the other (the connecting flight). In order to optimize the assortment (which flight schedules to offer) for maximum expected revenue, it is crucial to accurately predict the willingness of the consumers to purchase bundled items, based on past history. We propose a model that can capture such interacting preferences for bundled items (e.g. jeans and shirts), and use this model to extract the features of the items in each category from historical bundled purchase data. Both our inference algorithm and the analyses extend to this setting, achieving the optimal trade-off between sample size and accuracy.

Contribution. We first study the canonical scenario of pairwise comparisons from the MNL model in Section 3. Our contribution in the modeling is a new sampling scenario we call *graph sampling* that captures how different pairs of items have varying likelihood of being compared together. Our algorithmic contribution is a convex relaxation with a new regularizer using a variation of the standard nuclear norm tailored for the graph sampling topology. Our theoretical contribution is in the analysis of the proposed estimator and a matching fundamental lower bound (up to a poly-logarithmic factor). This (a) characterizes the minimax sample complexity of the problem; (b) proves that the proposed estimator cannot be improved upon; and (c) identifies how the accuracy depends on the topology of sampling. This in turn provides a guideline for designing surveys when one has a choice on which pairs are to be compared. This is accompanied by experiments on synthetic and real data sets confirming our theoretical predictions.

This framework is extended to higher order comparisons in Section 4. We establish minimax optimality (up to a poly-logarithmic factor) of our estimator and identify the fundamental trade-off between accuracy and sample size. When each user provides a total linear ordering among k items, we show that the required sample size effectively is reduced by a factor of k. When the user provides her best choice (as in purchase history) instead of the total linear ordering, we extend our framework and establish minimax optimality in Section 5.2. We also consider a bundled purchase scenario in Section 5, where customers buy pairs of items from each of the two categories. We extend our framework and establish minimax optimality under the bundled purchase setting. We present experimental results on both synthetic and real-world data sets confirming our theoretical predictions and showing the improvement of the proposed approach in predicting users' choices[1].

Technically, we borrow analysis tools from 1-bit matrix completion Davenport et al. (2014), matrix completion Negahban and Wainwright (2012), and restricted strong convexity Negahban et al. (2009), and crucially utilize the Random Utility Model (RUM) Thurstone (1927); Marschak (1960); Luce (1959) interpretation (outlined in Section 2.1) of the MNL model to prove both the upper bound and the fundamental limit. This could be of interest to analyzing more general class of RUMs.

Notations. We use $\vert\!\vert\!\vert A \vert\!\vert\!\vert_{\mathrm{F}}$ and $\vert\!\vert\!\vert A \vert\!\vert\!\vert_{\infty}$ to denote the Frobenius norm and the ℓ_∞ norm, $\vert\!\vert\!\vert A \vert\!\vert\!\vert_{\mathrm{nuc}} = \sum_i \sigma_i(A)$ to denote the nuclear norm where $\sigma_i(A)$ denotes the i-th singular value, and $\vert\!\vert\!\vert A \vert\!\vert\!\vert_2 = \sigma_1(A)$ for the spectral norm. We use $\langle\!\langle u, v \rangle\!\rangle = \sum_i u_i v_i$ and $\|u\|$ to

---

1. Code for our experiments are available at https://github.com/POLane16/Nucnorm-Ranking.

denote the inner product and the Euclidean norm. All ones vector is denoted by $\mathbb{1}$, $\mathbf{I}$ denotes the identity matrix and $\mathbb{I}(A)$ is the indicator function of the event A. The set of the first N integers are denoted by $[N] = \{1, \ldots, N\}$.

1.1 Related Work

Bradley-Terry and Plackett-Luce models. The simplest form of the MNL model is when all users are sharing the same feature vector such that each item is parametrized by a scalar value. This is known as Bradley-Terry (BT) model when pairwise comparisons are concerned and Plackett-Luce (PL) model when higher order comparisons are concerned. This has been proposed and rediscovered several times in the last century Zermelo (1929); Thurstone (1927); Bradley and Terry (1955); Luce (1959); Plackett (1975); McFadden (1973, 1980) in the context of ranking teams in sports games, ranking items based on surveys, and ranking routes in transportation systems. Unlike the general MNL model, maximum likelihood estimator for the BT and PL models are naturally convex programs. However, learning the BT model has first been addressed in Jr. (1957) where the convergence of the iterative algorithm is analyzed, without explicitly relying on the convexity of the problem. A new algorithm based on Majorize-Minimize framework was proposed in Hunter (2004). First sample complexity of learning BT model was provided in Negahban et al. (2012) where a novel estimator, called Rank Centrality, of the BT parameters was proposed. The authors construct a random walk over a graph where the nodes are the items and the transition probability is constructed from the comparisons outcomes. This spectral approach is proven to achieve a minimax optimal sample complexity. This has been a building block for several ranking algorithms, which further process the Rank Centrality to get better accuracy on top of it Chen and Suh (2015); Jang et al. (2016, 2017); Chen et al. (2017). For higher order comparisons, the sample complexity of learning PL model was provided in Hajek et al. (2014); Shah et al. (2014), where the Maximum Likelihood (ML) estimator is shown to achieve the minimax optimality. Later, Maystre and Grossglauser (2015) made the connection between the spectral approach of Rank Centrality and the ML estimator precise by providing a unifying random walk view to the problem. This led to a novel Accelerated Spectral Ranking algorithm introduced in Agarwal et al. (2018), which not only finds the parameters of the PL model more efficiently in computation, but also achieves optimal sample complexity under general sampling graphs. Recently, Borkar et al. (2016) treat the learning problem as solving a noisy linear system, and propose an algorithm that is amenable to on-line, distributed and asynchronous variants. Vojnovic and Yun (2016) analyzes a more general class of random utility models known as Thurstone models, and provide the minimax sample complexity by analyzing the ML estimator. Note that the ML estimators for Thurstone models in general are computationally intractable.

Generalized BT and PL models. As studied in Rajkumar and Agarwal (2014), the BT model covers a subset of probabilistic models over comparisons. There is a hierarchy of models with increasing complexity and descriptive power. One popular extension is the mixture of BT or PL models. It is known that any choice model can be approximated arbitrarily close with a mixed PL model with sufficient number of mixture components McFadden and Train (2000). The sample complexity of learning a mixed PL model was analyzed in Oh and Shah (2014) where a tensor decomposition for learning a mixture model

was proposed and analyzed under some separation conditions between the weights of the mixtures. For a mixture of two PL model, Chierichetti et al. (2018) shows identifiability and uniqueness of the mixture weights, when all marginal probability over all possible rankings among two items and three items are known. In a crowdsourced setting, Chen et al. (2013) models pairwise comparisons using a mixture of PL models consisting of hammer distribution, which reports the true output of a comparison, and spammer distribution, which reports the exact opposite of a comparison. A different approach that tackles the problem by learning to cluster the users based on the pairwise comparisons is proposed in Rui et al. (2015). The MNL model we study in this paper can be thought of as a generalization of the mixed PL models, where each user has her own preference. To make learning feasible, we inherently impose similarities among users via a low-rank condition. Note that a mixed PL model with r mixture is a special case of the MNL model with rank r, where each user's membership is encoded as a r-dimensional feature in standard basis. In the context of collaborative ranking, algorithms for learning the MNL model from pairwise comparisons have been proposed in Park et al. (2015). Instead of nuclear norm regularization as we propose in this paper, Park et al. (2015) proposes solving a convex relaxation of maximizing the likelihood over matrices with bounded nuclear norm. Under the standard assumption of uniformly chosen pairs, it is shown that this approach achieves statistically optimal generalization error rate, instead of Frobenius norm error that we analyze.

Beyond BT and PL models. Modeling choice is an important problem where the ultimate goal is to find the right parametric model to capture human choices. Ragain and Ugander (2016); Blanchet et al. (2013) use Markov chains to model choices with the parameters in the transition matrix defining the probability model. Ideal point model Massimino and Davenport (2018); Kazemi et al. (2018) assumes that the pairwise comparisons of two items by a user depends on their distance from an ideal item (ideal point) for the user in some metric embedding space of the items. Novel nonparametric models have also been proposed to model human choices, for example Shah et al. (2016b); Pananjady et al. (2017); Falahatgar et al. (2018) uses strong stochastic transitivity to model pairwise choices and Farias et al. (2009) uses distribution over all permutations with sparse support to model higher order choices. We also note that in the context of (non-collaborative) ranking, Gleich and Lim (2011) has proposed nuclear norm minimization based algorithm when comparisons between all pairs items are modeled as a low-rank skew-symmetric matrix. Other nonparametric approaches to solving ranking include empirical risk minimization. Clémençon et al. (2005) analyses risk minimization of U-statistics and a more feasible surrogate convex loss minimization to estimate ranking. Katz-Samuels and Scott (2017) assumes that rating of an item by a user is a Lipschitz function of the user-item pair and analyses a nonparametric collaborative ranking algorithm from partial observation of such ratings.

While we are interested in the (parameters of) full ranking over all items, there have been several recent works which aim to only approximately rank the items, such as retrieving only the top-m items Chen and Suh (2015); Jang et al. (2016, 2017) or partitioning the items into ordered buckets of fixed size Katariya et al. (2018); Heckel et al. (2018).

2. Model and Approach for Pairwise Comparisons

The MultiNomial Logit (MNL) model is one of the most popular models that explains how people make choices when given multiple options and is widely used in behavioral psychology and revenue management. For brevity, we focus our discussion on data collected in the form of pairwise comparisons in Sections 2 and 3, and defer the discussion of the MNL model in its full generality to Sections 4 and 5 . We give a precise definition of the model for paired comparisons and provide a novel algorithmic solution to learn this model from samples.

2.1 MultiNomial Logit (MNL) Model for Pairwise Comparisons

Let Θ^* be a $d_1 \times d_2$ dimensional matrix capturing preferences of d_1 users on d_2 items. The probability with which a user, $i \subseteq [d_1]$, when presented with two items $j_1, j_2 \subseteq [d_2]$, prefers item j_1 over item j_2 is,

$$\mathbb{P}\{j_1 > j_2\} = \frac{e^{\Theta^*_{ij_1}}}{e^{\Theta^*_{ij_1}} + e^{\Theta^*_{ij_2}}} \, . \tag{1}$$

This implies that, more preferred items (as per the ordering of Θ^*_{ij}) are more likely to be ranked higher, with the randomness in choices captured by the probabilistic model.

If we do not impose any further constraints on Θ^*, one entry of Θ^* is not related in any way to any other entries. This implies that one user's preference is completely independent of others' and no efficient learning is possible. Each user's preference has to be learned separately. On the other hand, in real applications, it is reasonable to say that preferences of users depend only on a handful of factors for example, quality, price, and aesthetics. We do not know which features affect users' choices, but we assume that there are r-dimensional latent features for each of the users and items that govern such choices, and that $r \ll d_1, d_2$. This assumption mathematically captures the conventional belief that when two people have similar preferences over a subset of items, they tend to have similar tastes on other items as well. Formally, MNL model assumes that Θ^* is a rank r matrix with $r \ll d_1, d_2$. In this paper, we do not impose a hard constraint on the rank and provide general results for matrices of any rank. In this case, we identify how the accuracy depends on the rate of decay of the singular values.

This MNL model has many roots. In revenue management, this has been proposed as a special case of Random Utility Model (RUM). RUM explains choices that a person makes as the result of maximizing perceived random utilities associated with the set of alternatives presented. In the case of MNL, each decision maker and each alternative are associated with an r-dimensional vector, u_i and v_j, resulting in a low-rank Θ^* if $\Theta^*_{ij} = \langle\!\langle u_i, v_j \rangle\!\rangle$. The perceived utility of the item j for decision maker i is,

$$U_{ij} = \langle\!\langle u_i, v_j \rangle\!\rangle + \xi_{ij} \, , \tag{2}$$

where ξ_{ij}'s are i.i.d. random variables following the standard Gumbel distribution. Different choices of distributions give different variants of RUMs. In our analyses, we utilize this RUM interpretation of the MNL model to prove a particular concentration in Section C.4, for example. The model in Equation. (1) has also been re-discoverd several times in the literature Zermelo (1929); Thurstone (1927); Luce (1959); Bradley and Terry (1955) in several domains.

2.2 Low-rank Regularization using Nuclear Norm Minimization

Given the low-rank structure of the model, a natural but inefficient approach is to minimize the negative of the log likelihood, $\mathcal{L}(\cdot)$, regularized by the rank:

$$\widehat{\Theta} \ \in \ \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} -\mathcal{L}(\Theta) + \lambda \operatorname{rank}(\Theta), \tag{3}$$

for some parameter $\lambda > 0$. As this rank minimization is a notoriously challenging problem, we instead solve a convex relaxation of it. Note that the nuclear norm ball is the convex hull of rank-1 matrices Recht et al. (2010). Analogous to l_1-norm in the case of sparse vectors, nuclear norm is a tight convex surrogate for low-rank solutions. We propose the following nuclear norm regularized optimization problem,

$$\widehat{\Theta} \ \in \ \arg \min_{\Theta \in \Omega} -\mathcal{L}(\Theta) + \lambda \, \|\|\Theta\|\|_{\mathrm{nuc}}, \tag{4}$$

where Ω is a convex constraint which takes care of identifiability and Lipschitz smoothness conditions. Nuclear norm regularization has been widely used Recht et al. (2010) for rank minimization; however, provable guarantees exist only for quadratic loss functions $\mathcal{L}(\Theta)$ Candès and Recht (2009); Negahban and Wainwright (2012). Our analyses extend such results to a convex loss, by first proving that $-\mathcal{L}(\cdot)$ satisfies restricted strong convexity property with high probability. Similar to how (non-collaborative) rank aggregation has been generalized to any strongly log-concave distribution in Shah et al. (2014), our analysis can naturally be extended to a general class of strongly log-concave distributions. We give the expression for the log likelihood in Equation. (8) for pairwise comparisons.

3. Learning MNL Model from Pairwise Comparisons under Graph Sampling

Probabilistic model for sampling. In order to provide performance guarantees on the proposed approach, we need to specify how we sample the pairs that are to be compared. We provide a novel sampling model, which we call *graph sampling* with respect to a weighted graph $\mathcal{G}$. This naturally generalizes Bernoulli sampling typically studied under matrix completion literature Candès and Recht (2009); Keshavan et al. (2010a); Negahban and Wainwright (2012); Jain et al. (2013), and the resulting analysis captures how the performance depends on the topology of the samples. Note that the proposed graph sampling is different from deterministic sampling graphs studied in Hajek et al. (2014); Shah et al. (2016a). This is analytically tractable only in the simpler case of estimating the weight vector of the PL model where there is only one user and the ML estimator is a convex program. However, such deterministic sampling is notoriously hard to handle for matrix estimation, even in the simpler case of matrix completion Bhojanapalli and Jain (2014). Hence, we introduce a probabilistic model that allows enough flexibility to capture the interesting aspects of sampling biases, i.e. grouping.

Precisely, we have a weighted undirected graph $\mathcal{G} = ([d_2], E, \{P_{j_1, j_2}\}_{(j_1, j_2) \in E})$ with d_2 nodes, which represent items, a set of edges E and the edge weight P_{j_1, j_2} between nodes j_1 and j_2. The weights can be written in a symmetric matrix $P \in \mathbb{R}^{d_2 \times d_2}$, and $P_{j_1, j_2} + P_{j_2, j_1} = 2P_{j_1, j_2}$ represent the probability with which the pair (j_1, j_2) is chosen for comparison. Note

that $P_{j,j} = 0$, $\forall j \in [d_2]$, $P_{j_1,j_2} = P_{j_2,j_1}$ and $\sum_{j_1,j_2 \in [d_2]} P_{j_1,j_2} = 1$. We assume we get i.i.d. samples from first choosing a random user among $[d_1]$ users, and then choosing a pair (j_1, j_2) of items at random from P, and finally getting a random comparison from the MNL model, i.e. the probability with which user i prefers item j_1 over item j_2 is $\exp \Theta^*_{ij_1} / \left(\exp \Theta^*_{ij_1} + \exp \Theta^*_{ij_2} \right)$.

One of the most important aspects of real-world data that is captured by this graph sampling model is grouping. Consider two groups of items, say, cars and phones. It does not make sense to ask an individual to compare a phone with a brand of a car (i.e. direct comparison is not feasible), but knowing an individual's preference on cars can help in learning her preference on phones. In graph sampling terms, we are sampling from a graph $\mathcal{G}$ consisting of two disjoint cliques: one for cars and another for phones. By analyzing such a sampling scenario, we want to characterize the gain in using the data from both groups of items together, although there are no inter-group comparisons.

In the preference matrix Θ^*, the values in the set of columns corresponding to each connected component in the sampling graph can be arbitrarily shifted together, without changing the pairwise comparisons outcome distributions. This is because adding the same constant to those items that are compared does not change the probability (for those items within the same group), i.e.

$$\mathbb{P}\{j_1 > j_2\} = \frac{e^{\Theta^*_{ij_1}}}{e^{\Theta^*_{ij_1}} + e^{\Theta^*_{ij_2}}} = \frac{e^{\Theta^*_{ij_1}+c}}{e^{\Theta^*_{ij_1}+c} + e^{\Theta^*_{ij_2}+c}} ,$$

and adding different constants to those items that are not in the same group does not change the probability of the outcome as those items are never compared. Hence, to handle this unidentifiability, we let a centered version of Θ^* represent all those shifted versions defining the same probability distribution. Formally, let a zero-one vector $g_k \in \{0,1\}^{d_2}$ denote the group membership such that $g_{i,k} = 1$ if item j is in group k, else $g_{i,k} = 0$. Note that, by definition, no item can be present in more than one group, that is, $\sum_{k=1}^G g_k = \mathbb{1}$, where G is the number of groups. We define an equivalence class of Θ^* which represent the same probabilistic model as

$$[\Theta^*] = \left\{ \Theta^* + \sum_{k=1}^G u_k g_k^T \ \text{ for all } u_k \in \mathbb{R}^{d_1} \right\}. \tag{5}$$

To overcome the identifiability issue, we represent each equivalence class with the centered matrix satisfying

$$\Theta^* g_k = 0, \quad \forall\, k \in \{1, 2, \ldots, G\} \tag{6}$$

As matrices with large "spikiness" are known to be hard to estimate Negahban and Wainwright (2012), we capture the dependence of the sample complexity on the spikiness as measured by $\alpha := \|\|\Theta^*\|\|_\infty$. This captures the dynamic range of the underlying preference matrix. For a related problem of matrix completion, where the loss $\mathcal{L}(\theta)$ is quadratic, either a similar condition on ℓ_∞ norm is required or another condition on incoherence is required.

Graph Laplacian. The performance of our approach depends on the sampling graph P via its graph Laplacian defined as

$$L = \text{diag}(P\mathbb{1}) - P \tag{7}$$

where $\text{diag}(P\mathbb{1})$ is a diagonal matrix with $\sum_v P_{u,v}$ in the diagonals. Notice that, L is singular and the nullspace is spanned by vectors $\{g_k\}_{k=1}^G$. Let $\sigma_{\max}(L) = \|L\|_2$ and $\sigma_{\min}(L)$ be the smallest eigenvalue of L discounting the G zero-valued eigenvalues. Since the graph has G disconnected maximal components and L is real symmetric, by spectral theorem, $L = U\Sigma U^T$, where U is a matrix of size $d_2 \times (d_2 - G)$ and its $d_2 - G$ columns form an orthonormal set, and Σ is a diagonal matrix such that its diagonal elements are the singular values of L. Let $L^\dagger := U\Sigma^{-1}U^T$ and $L^x := U\Sigma^x U^T$ for all $x \in \mathbb{R}$. We also define the Laplacian induced norms of matrices as,

$$\|\Theta\|_{\mathrm{L}} := \left\|\Theta L^{1/2}\right\|_{\mathrm{F}}, \text{ and, } \|\Theta\|_{\mathrm{L\text{-}nuc}} := \left\|\Theta L^{1/2}\right\|_{\mathrm{nuc}}.$$

These Laplacian induced norms are more appropriate to analyze and quantify the distance between the estimated matrix $\widehat{\Theta}$ and Θ^*.

When items $k(i)$, $l(i)$ are chosen for comparison by user $j(i)$ as the i-th pair of items, we capture this choice with the matrix $X^{(i)} = e_{j(i)}(e_{k(i)} - e_{l(i)})^T$. The outcome of the comparison is represented by y_i, with $y_i = 1$ when item $k(i)$ wins over item $l(i)$ and $y_i = 0$ if otherwise. The log-likelihood of the comparison outcomes with respect to a parameter matrix Θ is,

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^n \left[y_i \langle\!\langle \Theta, X^{(i)} \rangle\!\rangle - \log\left(1 + \exp\left(\langle\!\langle \Theta, X^{(i)} \rangle\!\rangle\right)\right) \right]. \tag{8}$$

We propose and analyze the following convex optimization problem,

$$\widehat{\Theta} \in \operatorname*{argmin}_{\Theta \in \Omega_\alpha} - \mathcal{L}(\Theta) + \lambda \|\Theta\|_{\mathrm{L\text{-}nuc}}, \tag{9}$$

where,

$$\Omega_\alpha = \left\{ \Theta \in \mathbb{R}^{d_1 \times d_2} \mid \|\Theta\|_\infty \leq \alpha, \Theta g_k = 0, \ \forall\, k \in [G] \right\}, \tag{10}$$

with an appropriately chosen $\lambda = 8\sqrt{2}\max\left\{ \sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}(L)^{-1/2}\log(2d)}{n} \right\}$ with $\sigma = \max\{(d_2 - G)/d_1, 1\}$, where $d = (d_1 + d_2)/2$. In practice, the sampling probability distribution P and the corresponding Laplacian L might not be known. In those cases, we propose using the empirical sampling probability distribution $\hat{P}$ and corresponding empirical Laplacian $\hat{L}$ instead. We describe this version of the algorithm formally in Section 3.4.4, where we empirically demonstrate the robustness of this approach. Further, in experiments with real data sets, we use the empirical Laplacian Section 3.4.5.

3.1 Performance Guarantee

We consider the graph sampling scenario where each sample is i.i.d., the ℓ-th sample consists of user i_ℓ chosen uniformly at random, pair of items $(j_{1,\ell}, j_{2,\ell})$ chosen according to the sampling graph $\mathcal{G} = ([d_2], E, P)$, and the resulting outcome y_ℓ distributed as the MNL model with parameter Θ^*.

10

Theorem 1 *Under the graph sampling with respect to $\mathcal{G} = ([d_2], E, P)$ with a graph Laplacian L, and under the MNL preference model with preference matrix Θ^*, solving the optimization problem in (9) with n i.i.d. samples achieves, with probability greater than $1 - 1/4d^3$,*

$$\frac{1}{d_1} \left\| \left(\Theta^* - \widehat{\Theta} \right) L^{1/2} \right\|_{\mathrm{F}}^2 \leq 36\lambda \left(\alpha + \frac{1}{\psi(2\alpha)} \right) \left(\sqrt{2r} \left\| \left(\Theta^* - \widehat{\Theta} \right) L^{1/2} \right\|_{\mathrm{F}} + \sum_{j=r+1}^{\min\{d_1, d_2 - G\}} \sigma_j(\Theta^* L^{1/2}) \right), \qquad (11)$$

for any $r \in \{1, 2, \ldots, \min\{d_1, d_2 - G\}\}$, any $\lambda \geq 8\sqrt{2} \max \left\{ \sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}(L)^{-1/2} \log(2d)}{n} \right\}$ where $\sigma = \max\{(d_2 - G)/d_1, 1\}$ and $d = (d_1 + d_2)/2$, $\psi(x) \triangleq e^x/(1 + e^x)^2$, and for $n \leq \min\{2^6 d_1^2 \sigma^2, 2^2 \left(d_1 \sigma_{\min}(L)^{-1} \right)^{2/3}\} \log(2d)$.

We provide a proof in Appendix A. The above bound holds for any r, where r allows us to trade off the two types of errors: the estimation error and the approximation error. Concretely, the above bound shows a natural splitting of the error into two terms; the first term corresponding to the *estimation error* for the top rank-r component of Θ^* and the second term corresponding to the *approximation error* for how well one can approximate Θ^* with a rank-r matrix. If we know the singular values of Θ^*, we can optimize over r to get the tightest bound. If Θ^* is exactly low-rank then applying a matching rank in the bound gives the following guarantee.

Corollary 2 (Exact rank-r matrix) *Under the same hypothesis as in Theorem 1 with a choice of $\lambda = c_0 \max \left\{ \sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}(L)^{-1/2} \log(2d)}{n} \right\}$ for some $c_0 > 0$, if Θ^* is exactly rank r, there exists a positive constant c_1 such that the proposed estimator achieves,*

$$\frac{1}{\sqrt{d_1}} \left\| \left(\Theta^* - \widehat{\Theta} \right) L^{1/2} \right\|_{\mathrm{F}} \leq c_1 \left(\alpha + \frac{1}{\psi(2\alpha)} \right) \sqrt{r} \max \left\{ \sqrt{\frac{\sigma \, d_1 \log(2d)}{n}}, \frac{\sqrt{(\sigma_{\min}(L)^{-1} d_1)} \, \log(2d)}{n} \right\}, \quad (12)$$

with probability at least $1 - 2/(d_1 + d_2)^3$ and $\sigma = \max\{(d_2 - G)/d_1, 1\}$.

The second term in the maximization is an artifact of the weakness of current analysis technique and does not reflect the actual error. This is confirmed in our simulation results on graphs with very small spectral gap in Figures 1.(b), (d), and (f), where the error in Laplacian-induced norm error does not decrease with spectral gap of L as the line graph has a much smaller spectral gap compared to a complete graph, for example. In fact, for a special Θ^* in Figure 1.(d) it is the other way, for which we do not have a theoretical explanation.

The number of entries in Θ^* is $d_1 d_2$ and we want to rescale the Frobenius norm error appropriately by $1/\sqrt{d_1 d_2}$. As a typical scaling of $L^{1/2}$ is $1/\sqrt{d_2}$ in spectral norm, we

only need to rescale the Laplacian-induced norm error by $1/\sqrt{d_1}$ in the left-hand side of the above bound. For a rank-r Θ^*, the number of degrees of freedom in describing it is $r(d_1 + d_2) - r^2 = O(r(d_1 + d_2))$. The above theorem shows that the total number of samples n needs to scale as $O(r(d_1 + d_2) \log d)$ in order to achieve an arbitrarily small error. This is only a poly-logarithmic factor larger than the degrees of freedom. In Section 3.2 we make this comparison precise by providing a lower bound that matches the upper bound up to a logarithmic factor.

The upper-bound constraint in Theorem 1 on the number of samples n can be met for large enough d_1 and d_2. For simplicity, assume that $d_1 = d_2 = d$ and r is a constant. Since $\sigma_{\min}(L) = O(1/d_2)$, the upper-bound on n becomes $O(\max\{d^2, d^{4/3}\})$. For large enough d, the upper bound on the RHS of Eq. (12) can be made arbitrarily small with n only scaling as $O(r\,d)$. This is significantly smaller than the upper-bound of $O(d^2)$ on n. Further, in the experiments in Section 3.4, we show that n has no practical upper-bound constraint since the error decreases at the same rate as predicted, for arbitrarily large values of n. This constraint may not be necessary and might be a by-product of the proof techniques.

The dependence on the dynamic range α, however, is sub-optimal. It is expected that the error increases with α, since the Θ^* scales as α, but the exponential dependence in the bound seems to be a weakness of the analysis (for example as seen from numerical experiments in the right panel of Figure 6). Although the error increase with α, numerical experiments suggest that it only increases at most linearly. However, tightening the scaling with respect to α is a challenging problem, and such sub-optimal dependence is also present in existing literature for learning even simpler models, such as the Bradley-Terry model Negahban et al. (2012) or the Plackett-Luce model Hajek et al. (2014), which are special cases of the MNL model studied in this paper.

Another issue is that the underlying matrix might not be exactly low rank. It is more realistic to assume that it is approximately low rank. Following Negahban and Wainwright (2012) we formalize this notion with "ℓ_q-ball" of matrices defined as

$$\mathbb{B}_q(\rho_q) \equiv \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j \in [\min\{d_1, d_2\}]} |\sigma_j(\Theta^*)|^q \le \rho_q\} . \tag{13}$$

When $q = 0$, this is a set of rank-ρ_0 matrices. For $q \in (0, 1]$, this is set of matrices whose singular values decay relatively fast. By optimizing the choice of r in Theorem 1, we get the following result.

Corollary 3 (Approximately low-rank matrices) *Suppose $\Theta^* \in \mathbb{B}_q(\rho_q)$ for some $q \in (0, 1]$ and $\rho_q > 0$. Under the hypotheses of Theorem 1, with a choice of $\lambda = c_0 \max\Big\{$*

$\sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}(L)^{-1/2} \log(2d)}{n}\Big\}$ for some constant $c_0 > 0$ there exists a constant $c_1 > 0$ such that solving the optimization (9) achieves with probability at least $1 - 2/(d_1 + d_2)^3$,

$$\frac{1}{\sqrt{d_1}} \left\|\left(\Theta^* - \widehat{\Theta}\right) L^{1/2}\right\|_{\mathrm{F}} \le \frac{c_1 \sqrt{\rho_q}}{\sqrt{d_1}} \left(\left(\alpha + \frac{1}{\psi(2\alpha)}\right) \sqrt{\frac{d_1^2 \sigma \log(2d)}{n}}\right)^{\frac{2-q}{2}} , \tag{14}$$

provided $n \ge \sigma \log(2d)/\sigma_{\min}(L)$.

This is a strict generalization of Corollary 2. For $q = 0$ and $\rho_0 = r$, this recovers the exact low-rank estimation bound up to a factor of two. For approximate low-rank matrices in an ℓ_q-ball, we lose in the error exponent, which reduces from one to $(2 - q)/2$.

3.2 Information-theoretic Lower Bound

For a polynomial-time algorithm of convex relaxation, we gave in the previous section a bound on the achievable error. We next compare this to the fundamental limit of this problem, by giving a lower bound on the achievable error by any algorithm (efficient or not). A simple parameter counting argument indicates that it requires the number of samples to scale as the number of degrees of freedom i.e., $n \propto r(d_1 + d_2)$, to estimate a $d_1 \times d_2$ dimensional matrix of rank r. We construct an appropriate packing over the set of low-rank matrices with bounded entries in Ω_α defined as (10), and show that no algorithm can accurately estimate the true matrix with high probability using the generalized Fano's inequality. This provides a constructive argument to lower bound the minimax error rate, which in turn establishes that the bounds in Theorem 1 is sharp up to a logarithmic factor, and proves no other algorithm can significantly improve over the nuclear norm minimization.

Theorem 4 *Suppose Θ^* has a rank r. Under the previously described graph based sampling model, there exists a constant $c > 0$ such that*

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\Big[\frac{1}{\sqrt{d_1}} \Big\|\!\Big\|\Big(\Theta^* - \widehat{\Theta}\Big) L^{1/2}\Big\|\!\Big\|_{\mathrm{F}}\Big] \geq c \min\Big\{ e^{-\alpha}\sqrt{\frac{r\,d_1}{n}},$$

$$\alpha \max\Big\{\sqrt{\frac{r}{\mathrm{tr}\left(L_r^\dagger\right)}}, \frac{d_2}{\sqrt{d_1 \log d}}\Big\}\Big\}\,, \quad (15)$$

where the infimum is taken over all measurable functions over the observed comparison results and $L_r^\dagger$ is the pseudo inverse of the rank r approximation of the graph Laplacian.

A proof of this theorem is provided in Appendix B. The term of primary interest in this bound is the first one, which shows the scaling of the (rescaled) minimax rate as $\sqrt{rd_1/n}$ and matches the upper bound in (12) up to a logarithmic factor. It is the dominant term in the bound whenever the number of samples is larger than $n \geq d_1 \max\{\mathrm{tr}\left(L_r^\dagger\right), d_1 \log d/d_2^2\}$. As suggested in numerical simulations on graphs with very small spectral gap in Figures 1.(b), (d), and (f), the dependence in $\mathrm{tr}\left(L_r^\dagger\right)$ is an artifact of the weakness of the current analysis technique. Here we note that, while the lower bound in Theorem 4 is in expectation, the upper bound in Theorem 1 is a high-probability result. The upper bound can immediately be translated into a bound in expected error with an additional term scaling as $\alpha\sigma_{\max}(L)^{1/2}\sqrt{d_2}d^{-3}$, which is smaller than other terms in the bound.

3.3 Performance Guarantee and Lower Bound for Complete Graph

It follows from a simple relation $\|\!\|(\Theta^* - \widehat{\Theta})L^{1/2}\|\!\|_F \geq \sigma_{min}^{1/2}\|\!\|\Theta^* - \widehat{\Theta}\|\!\|_F$, which is true since Θ^*, $\widehat{\Theta}$ are in the range space of L, that the above upper bounds automatically give the error bound in the Frobenius norm. When the sampling graph is uniform, i.e. a complete graph

with equal weights $P_{j_1,j_2} = 1/d_2(d_2 - 1)$, $\forall j_1 \neq j_2$, Frobenius norm is the right metric and we show matching upper and lower bounds.

Corollary 5 (Complete graph upper-bound) *Under the same hypothesis as in Corollary 2, if $\mathcal{G}$ is a complete graph, with a choice of $\lambda = c_0 \max \left\{ \sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sqrt{(d_2-1)} \log(2d)}{n} \right\}$ for some $c_0 > 0$, if Θ^* is exactly rank r, there exists a positive constant c_1 such that the proposed estimator achieves,*

$$\frac{\left\| \Theta^* - \widehat{\Theta} \right\|_F}{\sqrt{d_1(d_2 - 1)}} \leq c_1 \left(\alpha + \frac{1}{\psi(2\alpha)} \right) \sqrt{r} \max \left\{ \sqrt{\frac{\sigma \, d_1 \log(2d)}{n}}, \frac{\sqrt{(d_2 - 1)d_1} \, \log(2d)}{n} \right\}, \quad (16)$$

with probability at least $1 - 2/(d_1 + d_2)^3$ and $\sigma = \max\{(d_2 - 1)/d_1, 1\}$.

Corollary 6 (Complete graph lower-bound) *Suppose Θ^* has rank r. Under the previously described graph based sampling model with graph being a complete graph, there is a universal numerical constant $c > 0$ such that*

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E} \left[\frac{1}{\sqrt{d_1 (d_2 - 1)}} \left\| \widehat{\Theta} - \Theta^* \right\|_F \right] \geq c \min \left\{ e^{-\alpha} \sqrt{\frac{r \, d_1}{n}}, \right.$$
$$\left. \alpha \max \left\{ \frac{1}{\sqrt{(d_2 - 1)}}, \frac{d_2}{\sqrt{d_1 \log d}} \right\} \right\}, \quad (17)$$

where the infimum is taken over all measurable functions over the observed comparison results.

3.4 Experiments

We provide a first-order method to solve the proposed convex optimization, and provide numerical experiments using this algorithm. We present two simulation results followed by an experiment on real data.

For the synthetic experiments, we generate random rank-r matrices of dimension $d \times d$, of the form $\Theta^* = UV^T$ with $U \in \mathbb{R}^{d \times r}$ and $V \in \mathbb{R}^{d \times r}$ entries generated i.i.d from uniform distribution over $[0, 1]$. Then the connected-component-mean is subtracted form each connected component, and then the whole matrix is scaled such that the largest entry is $\alpha = 5$. Note that this operation does not increase the rank of the matrix Θ. This is because this de-meaning can be written as $\Theta - \sum_k \Theta g g^T / (g^T \mathbb{1})$ and both terms in the operation are of the same column space as Θ which is of rank r.

3.4.1 ALGORITHM

Let $\Theta' \triangleq \Theta L^{1/2}$. As the nuclear norm regularizer in (9) is non differentiable, we use the proximal gradient descent Agarwal et al. (2010); Cai et al. (2010). At each iteration, we apply the following two operations on the current estimate, Θ'_t, of $\Theta^* L^{1/2}$,

$$\widetilde{\Theta}'_{t+1} = \Theta'_t - \eta_t \nabla_\Theta \mathcal{L}(\Theta'_t L^{-1/2}) L^{-1/2} \quad \text{(gradient descent)} \quad (18)$$
$$\Theta'_{t+1} = M_t (\Gamma_t - \eta_t \lambda \mathbf{I})^+ N_t^T \quad \text{(singular value shrinkage and thresholding)} \quad (19)$$

where $M_t \Gamma_t N_t^T := \widetilde{\Theta'}_t$ is the singular value decomposition of $\widetilde{\Theta'}_t$, such that Γ_t is a diagonal matrix with positive entries, $(\cdot)^+$ is the entry-wise thresholding operation $\max(0, x)$, and η_t is an appropriate step-size. Constraint of zero row sum, is taken care of by initializing the descent algorithm with $\Theta'_0 = 0$, since rows of gradients sum to zero. In practice we do not know the value of α, and hence in experiments we do not enforce the $\|\Theta\|_\infty \le \alpha$ constraint.

Another issue in the implementation is that the convergence rate can be significantly slower for some graph topologies. We accelerate the proximal gradient descent with the following (modified) Barzilai-Borwein (BB) rule Barzilai and Borwein (1988) for choosing the step-size η_t,

$$\eta_t = \begin{cases} \frac{\|\Theta'_t - \Theta'_{t-1}\|_2^2}{\langle\!\langle \Theta'_t - \Theta'_{t-1}, \nabla_{\Theta'}\mathcal{L}'(\Theta'_t) - \nabla_{\Theta'}\mathcal{L}'(\Theta'_{t-1}) \rangle\!\rangle}, & \text{when } t \text{ is odd} \\ \frac{\langle\!\langle \Theta'_t - \Theta'_{t-1}, \nabla_{\Theta'}\mathcal{L}'(\Theta'_t) - \nabla_{\Theta'}\mathcal{L}'(\Theta'_{t-1}) \rangle\!\rangle}{\|\nabla_{\Theta'}\mathcal{L}'(\Theta'_t) - \nabla_{\Theta'}\mathcal{L}'(\Theta'_{t-1})\|_2^2}, & \text{when } t \text{ is even} \end{cases}, \quad (20)$$

where $\nabla_{\Theta'}\mathcal{L}'(\Theta') := \nabla_\Theta \mathcal{L}(\Theta' L^{-1/2}) L^{-1/2}$. We stop the descent algorithm whenever an upper bound of the KKT error is smaller than 10^{-5}.

3.4.2 The Role of the Topology of the Sampling Pattern

In figure 1, we plot the error of our nuclear norm minimization based algorithm versus number of samples (in log-scale), n for $d_1 = d_2 = 300$, $r = 4$, $\alpha = 5.0$, $G = 1$. We consider two errors here; root mean squared error (RMSE) $= \left\|\!\left\|\Theta - \widehat{\Theta}\right\|\!\right\|_F / \sqrt{d_1 d_2}$ and Laplacian induced RMSE (L-RMSE)$= \left\|\!\left\|(\Theta - \widehat{\Theta})L^{1/2}\right\|\!\right\|_F / \sqrt{d_1}$. We plot these errors for four topologies of varying spectral gaps. As discussed Section 3.1, we do not expect the L-RMSE error to change much as we change the topology of sampling. However, as seen from the simple relation $\|\!\|(\Theta^* - \widehat{\Theta})L^{1/2}\|\!\|_F \ge \sigma_{min}^{1/2} \|\!\|\Theta^* - \widehat{\Theta}\|\!\|_F$ Frobenius norm error is more sensitive to the topology of the sampling pattern, captured via the spectral gap, i.e. $\sigma_{\min}(L)$. Specifically we use the following graph topologies.

- **Complete graph.** We first consider a uniform sampling over a complete graph where $P_{j_1,j_2} = 1/d_2(d_2-1)$ for all $j_1, j_2 \in [d_2]$. The resulting spectral gap is $1/(d_2-1)$, which is the maximum possible value. Hence, complete graphs are optimal for learning MNL models, compared in the error metric of the Frobenius norm for fairness.

- **Star graph.** Here we choose one item to be the center, and every other items can only be compared to this center item uniformly at random. Let item 1 be the center one, then $P_{j_1,1} = P_{1,j_2} = 1/2(d_2 - 1)$. Standard spectral analysis shows that the spectral gap is $\Theta(1/d_2)$, and thus the graph is near-optimal for learning MNL models.

- **Line graph.** Next, we consider a line graph with d_2-1 edges where $P_{j,j+1} = P_{j+1,j} = 1/2(d_2 - 1)$. It has a spectral gap of $\Theta(1/d_2^2)$, and is strictly sub-optimal for learning MNL models.

- **Barbell graph.** Consider two equal sized groups of items. Within each group the sub-graph is complete, and between the groups there is a single edge connecting one of the node from group one and one of the node from group two. Each edge is chosen
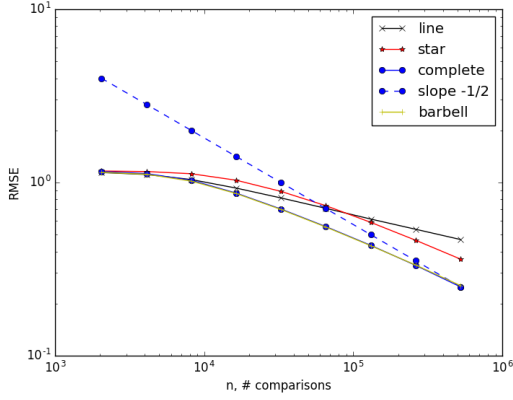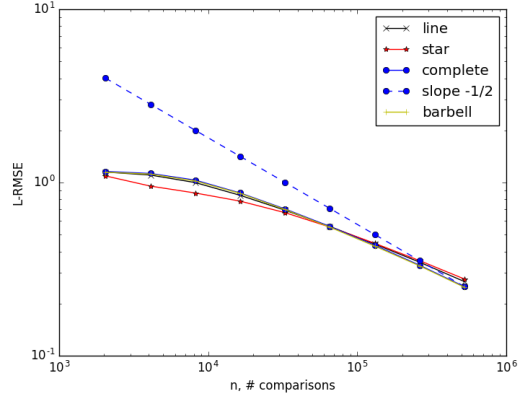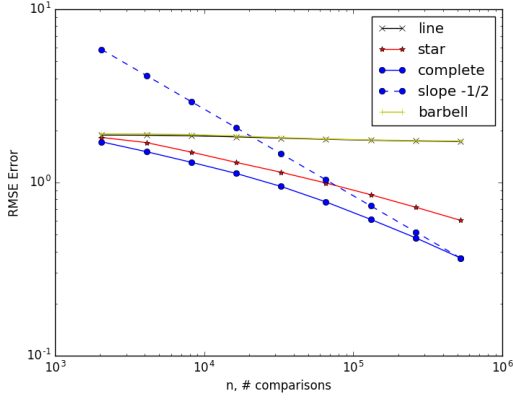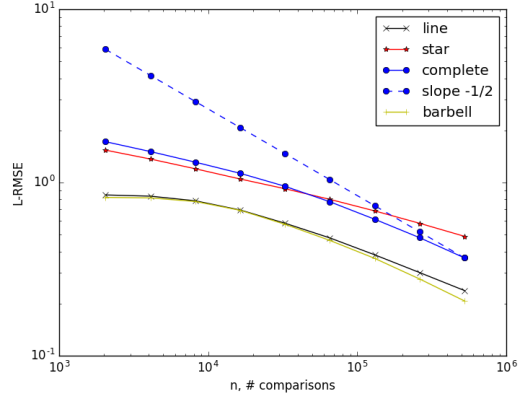
(a) RMSE for i.i.d. Θ_{ij}^*

(b) L-RMSE for i.i.d. Θ_{ij}^*

(c) RMSE for barbell bias Θ_{ij}^*

(d) L-RMSE for barbell bias Θ_{ij}^*
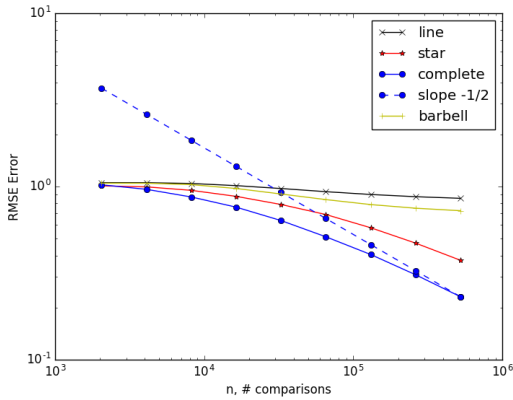
(e) RMSE for line bias Θ_{ij}^*
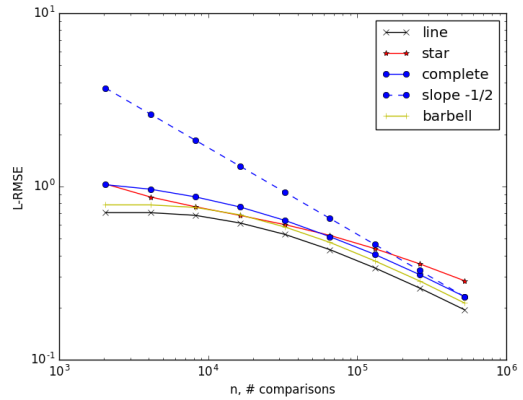
(f) L-RMSE for line bias Θ_{ij}^*

Figure 1: Graphs with small spectral gap achieve significantly larger Frobenius norm error (RMSE) $\left\|\!\left\|\Theta - \widehat{\Theta}\right\|\!\right\|_{\mathrm{F}} / \sqrt{d_1 d_2}$, whereas the Laplacian-induced norm error (L-RMSE) $\left\|\!\left\|(\Theta - \widehat{\Theta}) L^{1/2}\right\|\!\right\|_{\mathrm{F}} / \sqrt{d_1}$ is not sensitive to the spectral gap.

uniformly at random for comparisons. The resulting spectral gap is $\Theta(1/d_2^2)$, and this graph too is strictly sub-optimal for learning MNL models.

First in sub-figures 1a, 1b, we plot RMSE and L-RMSE errors for different graphs using randomly generated Θ_{ij}^*. We see that L-RMSE curves for different graphs are the same (and slopes in log-scale are as expected approaches $-1/2$ with more samples). Further, we do not see any significant difference w.r.t the graph topology even when error is measured in Frobenius norm. The reason is that since Θ_{ij}^*'s are generated i.i.d., the empirical distributions of any large sub-group of items would be similar. Thus, the means of the two cliques of the barbell graph or the means of the items on the two far ends of the line graph are similar. Thus although barbell and line graphs have small spectral gap (high mixing-time), its effect is minimized because these sub-groups can individually be solved without having them to mix since the empirical distributions of the Θ_{ij}^* in the two sub-groups are similar.

To illustrate the role of the topology of the graph, we choose specific Θ^* which depends on the topology of the graph as guided by our analysis on the lower bound (Theorem 4) in sub-figures 1c, 1d. The items are divided into two sets (corresponding to each side of barbell graph), such that corresponding Θ_{ij}^* are i.i.d. inside a set but have similar but shifted means across the sets. We call this type of preference data as *barbell biased*. As expected from theoretical analyses, *L*-RMSE behave similar to the i.i.d. case. However, we see the Frobenius norm error significantly worse in the case of line and barbell shaped graphs, as expected from the Frobenius error bound. In sub-figures 1e, 1f, we simulate *line biased* preference data Θ^*. Items are ordered (in the order of the line graph), such that Θ_{ij}^*'s have similar distributions but their means get shifted in an arithmetic progression as you go down the ordering. Again, Frobenius norm error is significantly larger for line and barbell graphs as spectral gaps are small.

3.4.3 THE GAIN IN INFERENCE OVER MULTIPLE GROUPS OF ITEMS

Consider G groups of items such that, within each group, every pair of items is uniformly likely to get compared, but items from different group are never compared with each each other. As a baseline, one can run inference on each group separately. On the other hand, we propose running inference on all the G groups jointly. Let $\widehat{\Theta}$ be the estimate of Θ^* when solving the groups together, and let $\bar{\Theta}$ be the estimate when groups are estimated separately. Let L, $L^{(k)}$ be the graph Laplacians of the whole graph and k-th connected component (group) respectively. suppose, for simplicity, that $d_1 = d_2$ and the groups are equally sized complete sub-graph components,

$$L = \frac{1}{(d_2 - G)} \left(\mathbf{I}_{d_2 \times d_2} - \frac{G}{d_2} \sum_{k=1}^{G} g_k g_k^T \right) \text{, and,} \tag{21}$$

$$L^{(k)} = \frac{1}{(d_2/G - 1)} \left(\mathbf{I}_{\frac{d_2}{G} \times \frac{d_2}{G}} - \frac{G}{d_2} \mathbb{1}\mathbb{1}^T \right) . \tag{22}$$

According to Theorems 1 and 4, the L-RMSE error of $\widehat{\Theta}$ satisfies,

$$\frac{1}{\sqrt{d_1}} \left\| \left(\Theta^* - \widehat{\Theta} \right) L^{1/2} \right\|_{\mathrm{F}} = \frac{1}{\sqrt{d_1(d_2 - G)}} \left\| \Theta^* - \widehat{\Theta} \right\|_{\mathrm{F}} = \tilde{O} \left(\sqrt{\frac{rd_1}{n}} \right) . \tag{23}$$
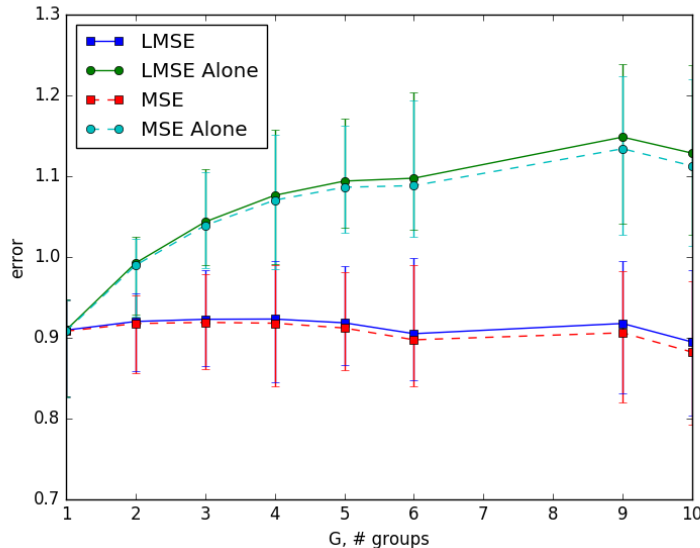
17

Figure 2: As the number of groups increase, the gain in joint inference increases.

Similarly L-RMSE error (with respect to the full Laplacian L) of $\bar{\Theta}$ satisfies,

$$\frac{1}{d_1}\left\|\left|\left(\Theta^* - \bar{\Theta}\right)L^{1/2}\right\|\right\|_{\mathrm{F}}^2 = \frac{1}{d_1(d_2 - G)}\left\|\left|\Theta^* - \bar{\Theta}\right\|\right\|_{\mathrm{F}}^2 \tag{24}$$

$$\overset{(a)}{=} \frac{(d_2/G - 1)}{(d_2 - G)}\sum_{k=1}^{G}\frac{\left\|\left|\left(\Theta_k^* - \bar{\Theta}_k\right)(L^{(k)})^{1/2}\right\|\right\|_{\mathrm{F}}^2}{d_2} \tag{25}$$

$$= \frac{1}{G}\sum_{k=1}^{G}\tilde{O}\left(\frac{rd_1}{n/G}\right) \tag{26}$$

$$= \tilde{O}\left(\frac{Grd_1}{n}\right), \tag{27}$$

where (a) follows from Eq. (22) and assuming $\Theta_k^*, \bar{\Theta}_k$ are sub-matrices restricted to the columns in group k. Thus the estimation errors when running joint inference and separate inference for each group are of the order of $O_G(1)$ and $O_G(\sqrt{G})$ respectively. That is, a user's preference in one group of items will be useful in inferring the same user's preference in another group of items. We illustrate this gain of joint inference in Figure 2. Concretely, the sampling graph $\mathcal{G}$ has G groups where each component is a complete graph and $d_1 = d_2 = 360$, $r = 4$, $\alpha = 5.0$, $n = 2^{14}$. Figure 2 plots the L-RMSE (RMSE) $= \left\|\left|(\Theta - \widehat{\Theta})L^{1/2}\right\|\right\|_{\mathrm{F}}/\sqrt{d_1}$ $\left(\left\|\left|(\Theta - \widehat{\Theta})\right\|\right\|_{\mathrm{F}}/\sqrt{d_1\,d_2}\right)$ errors vs. G, when all the groups are solved together (labelled as LMSE and MSE) or when the groups are solved separately (labelled as LMSE Alone and MSE Alone) using our algorithm. We see that solving the components together keeps the error relatively similar as the number of groups increase, but if we solve
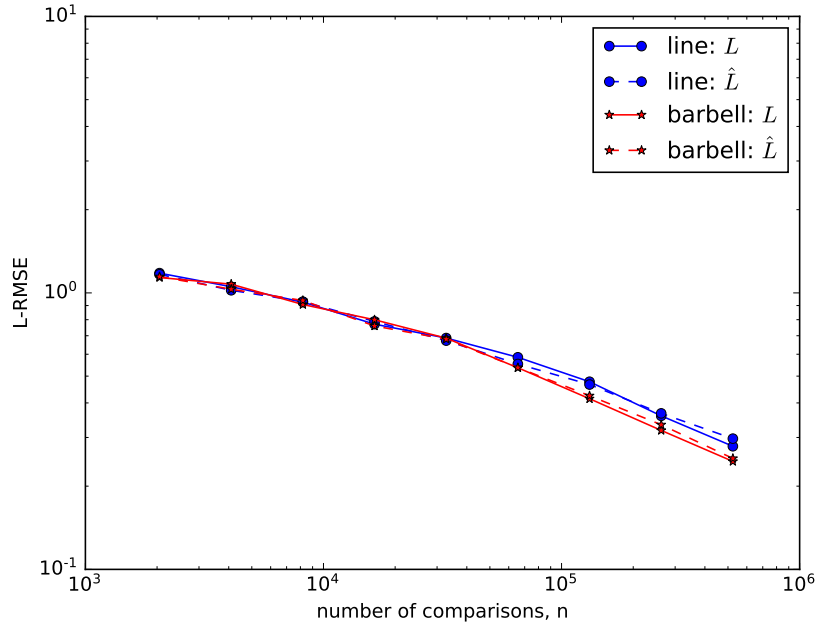
18

Figure 3: L-RMSE for various sampling graphs when true Laplacian L is known (solid) and when empirical Laplacian $\hat{L}$ is used (dashed)

.

the groups separately the error increases with number of groups, although it is at a lower rate than predicted by the upper bound.

3.4.4 ROBUSTNESS TO THE MISMATCHED L-NUC NORM REGULARIZER

In practical scenarios, one might not have access to the sampling graph Laplacian L. We propose using empirical Laplacian $\hat{L}$ defined as

$$\hat{L} \triangleq \operatorname{diag}(\hat{P}\mathbb{1}) - \hat{P}, \tag{28}$$

where $\hat{P} \in \mathbb{R}^{d_2 \times d_2}$ is the empirical distribution of sampled pairs in the given data. Under the experimental setting from Figure 1b, we run additional experiments with this empirical Laplacian $\hat{L}$ in the optimization: minimize $-\mathcal{L}(\Theta) + \lambda \|\|\Theta\|\|_{\hat{L}\text{-nuc}}$. Figure 3 illustrates that the effect on the performance of not knowing the true L is marginal. Both approaches achieve the same error.

3.4.5 REAL DATA: FOOD100

To showcase the practicality of our nuclear norm based algorithm (9) we apply our algorithm to the Food100 Data set[2] Wilber et al. (2014). In the data set, $n = 250320$ triplets, denoted by $\{(a_i, b_i, c_i)\}_i$, of 3 distinct food dishes from a selection of $d = 100$ were sampled. Then in

---

2. Data set is from `https://vision.cornell.edu/se3/projects/cost-effective-hits`.

a crowdsourcing setting, users were asked if, a_i is more similar to b_i than to c_i. The goal is to learn an low-dimensional embedding of the 100 food items where the above similarities are captured. We model the problem as learning an MNL model, parameterized by Θ^*, which gives the following probability distribution for i-th user's answer,

$$\mathbb{P}\left\{a_i \text{ is more similar to } b_i \text{ than to } c_i\right\} = \frac{e^{\Theta^*_{a_i,b_i}}}{e^{\Theta^*_{a_i,b_i}} + e^{\Theta^*_{a_i,c_i}}}.$$

This is the same model as the pairwise comparisons from Section 2, except for the fact that instead of a user (row) comparing two items (columns), here we compare a food item (row) to two other food items (columns). We implement three different algorithms: our nuclear norm based algorithm ('nucnorm'), unregualrized ($\lambda = 0$) likelihood maximization ('fullrank') and maximum likelihood based algorithm to learn rank-1 Plackett-Luce model Luce (1959); Plackett (1975) ('plackett').

In Figure 4, we plot the mean log-likelihood of the learned model versus fraction x of the data used for training for the various algorithms for testing (a) and training (b) data. If x fraction of the data is used for training, we use the rest $(1-x)$ of the data for testing. For the nuclear norm minimization, we estimate the Laplacian L using the empirical distribution of the triplets and λ is chosen to be $0.1\sqrt{\log(d)/2\,d\,xn}$.

In the Fig. 4(b) (to the left) on the testing data set, we see that our MNL model based nuclear norm regularized algorithm clearly outperforms both unregularized algorithm and the Placket-Luce model estimator, especially when there is less training data. In fact, the mean likelihood ($\log(P_{\text{model}}(\text{test data}))$) on the testing data remains relatively the same when we decrease the size of the training data, which supports our claim that real data has low-rank structure. In the Fig. 4(b) (right) on the training data sets, the non-regularized approach of 'fullrank' achieves higher likelihood on the training data, indicating that it overfits to training data.
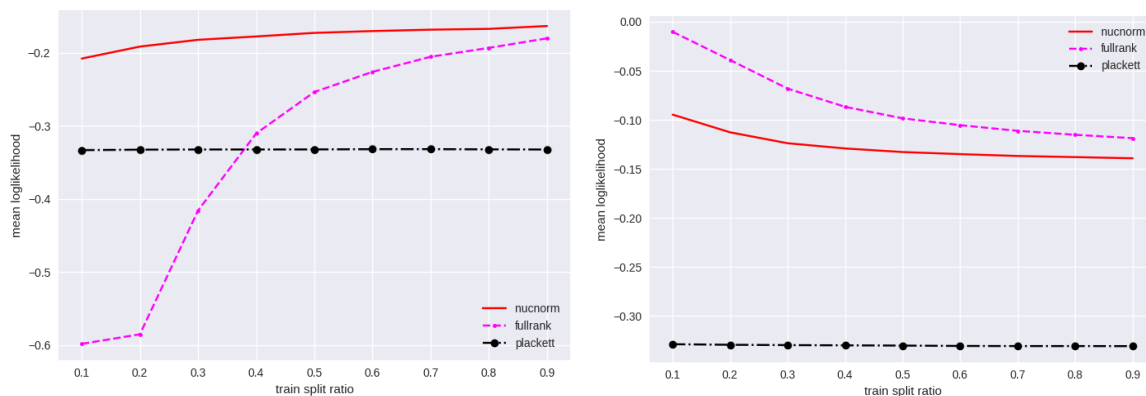
(a) Mean log-likelihood on testing data set   (b) Mean likelihood on training data set

Figure 4: Mean log-likelihood vs fraction of the total data used for training. Our nuclear norm regularized algorithm ('nucnorm') fits the test data better than both unregularized algorithm ('fullrank') and Plackett-Luce model based estimation, especially when training data is small in size.

Figure 5: Food100: t-SNE embedding of the columns of the learned MNL parameter $\widehat{\Theta}$. Desserts (bottom left) are seperated from other dishes. Meat dishes are also separated from vegetable dishes.

In Fig. 5 we plot the t-SNE embedding Maaten and Hinton (2008) of the columns of the estimated MNL parameter matrix $\widehat{\Theta}$ when all the data is used for training. The desserts (left bottom) are separated from other dishes, and meat dishes and salad dishes form two clusters (top right).

4. Learning the MNL Model under Higher Order Comparisons

Higher order comparisons, where a subset of k items are offered to a user who then provides a complete ranking (total linear ordering) of those item, is a natural generalization of pairwise comparisons, that captures some aspect of heterogeneous and complex modern data sets. We refer to such scenarios as k-wise comparisons or k-wise rankings. The MNL model generalizes to such comparisons. Let Θ^* be the $d_1 \times d_2$ dimensional matrix capturing the preference of d_1 users on d_2 items, where the rows and columns correspond to users and items, respectively. In this k-wise ranking setting, when a user i is presented with a set,

$S_i \subseteq [d_2]$, of k alternatives she reveals her preferences as a ranked list over those items. To simplify the notations, we assume that all the users compare the same number k of items, but the analysis naturally generalizes to the case when the size might differ from a user to a user and when each user provides more than one k-wise ranking. Let $v_{i,\ell} \in S_i$ denote the (random) ℓ-th best choice of user i. Each user gives a ranking, independent of other users' rankings, from

$$\mathbb{P}\{v_{i,1}, \ldots, v_{i,k} | S_i \text{ is presented to user } i\} = \prod_{\ell=1}^{k} \frac{e^{\Theta^*_{i,v_{i,\ell}}}}{\sum_{j \in S_{i,\ell}} e^{\Theta^*_{i,j}}} , \qquad (29)$$

where with $S_{i,\ell} \equiv S_i \setminus \{v_{i,1}, \ldots, v_{i,\ell-1}\}$ and $S_{i,1} \equiv S_i$. For a user i, the i-th row of Θ^* represents the underlying preference vector of the user, and the more preferred items are more likely to be ranked higher.

Similar to the pairwise comparisons, the distribution (29) is independent of shifting each row of Θ^* by a constant. Since we can only estimate Θ^* up to this equivalent class, we search for the one whose rows sum to zero, i.e. $\sum_{j \in [d_2]} \Theta^*_{i,j} = 0$ for all $i \in [d_1]$. For capturing the "spikiness" Negahban and Wainwright (2012) of Θ^*, we define $\alpha \equiv \max_{i,j_1,j_2} |\Theta^*_{ij_1} - \Theta^*_{ij_2}|$ to denote the dynamic range of the underlying Θ^*, such that when k items are compared, we always have

$$\frac{1}{k}e^{-\alpha} \leq \frac{1}{1+(k-1)e^{\alpha}} \leq \mathbb{P}\{v_{i,1} = j\} \leq \frac{1}{1+(k-1)e^{-\alpha}} \leq \frac{1}{k}e^{\alpha} , \qquad (30)$$

for all $j \in S_i$, all $S_i \subseteq [d_2]$ satisfying $|S_i| = k$ and all $i \in [d_1]$. We do not make any assumptions on α other than that $\alpha = O(1)$ with respect to d_1 and d_2. Given this definition, we solve the following optimization

$$\widehat{\Theta} \in \arg\min_{\Theta \in \Omega_\alpha} -\mathcal{L}(\Theta) + \lambda \|\|\Theta\|\|_{\mathrm{nuc}}, \qquad (31)$$

where,

$$\mathcal{L}(\Theta) = \frac{1}{k\,d_1} \sum_{i=1}^{d_1} \sum_{\ell=1}^{k} \left(\langle\langle \Theta, e_i e_{v_{i,\ell}}^T \rangle\rangle - \log\left(\sum_{j \in S_{i,\ell}} \exp\left(\langle\langle \Theta, e_i e_j^T \rangle\rangle \right) \right) \right) , \qquad (32)$$

over

$$\Omega_\alpha = \left\{ A \in \mathbb{R}^{d_1 \times d_2} \,\big|\, \|\|A\|\|_\infty \leq \alpha, \text{ and } \forall i \in [d_1] \text{ we have } \sum_{j \in [d_2]} A_{ij} = 0 \right\}. \qquad (33)$$

Note that unlike graph sampling for pairwise comparisons, we assume that each user is presented a subset of k items and provides a complete ranking over those k items. This choice of sampling scenario, together with independent choices of the items in subset S_i's, is crucial for getting a bound that is tight in its scaling with respect to not only d_1, d_2, and r, but also k, as a certain independence is required to apply the symmetrization technique (in Lemma 29) which gives us the desired tight bound on the error. It trivially follows from our analysis that one can relax the assumptions in the sampling scenario significantly (e.g. sampling without replacement, heterogeneous sampling probabilities for each item-user pair, etc.), and the only change in the upper bound of Eq. (34) will be a weaker dependence k.

4.1 Performance Guarantee

We provide an upper bound on the resulting error of our convex relaxation, when a *multi-set* of items S_i presented to user i is drawn uniformly at random with replacement. Precisely, for a given k, $S_i = \{j_{i,1}, \ldots, j_{i,k}\}$ where $j_{i,\ell}$'s are independently drawn uniformly at random over the d_2 items. Further, if an item is sampled more than once, i.e. if there exists $j_{i,\ell_1} = j_{i,\ell_2}$ for some i and $\ell_1 \neq \ell_2$, then we assume that the user treats these two items as if they are two distinct items with the same MNL weights $\Theta^*_{i,j_{i,\ell_1}} = \Theta^*_{i,j_{i,\ell_2}}$. The resulting preference is therefore always over k items (with possibly multiple copies of the same item), and distributed according to (29). For example, if $k = 3$, it is possible to have $S_i = \{j_{i,1} = 1, j_{i,2} = 1, j_{i,3} = 2\}$, in which case the resulting ranking can be $(v_{i,1} = j_{i,1}, v_{i,2} = j_{i,3}, v_{i,3} = j_{i,2})$ with probability $(e^{\Theta^*_{i,1}})/(2\,e^{\Theta^*_{i,1}} + e^{\Theta^*_{i,2}}) \times (e^{\Theta^*_{i,2}})/(e^{\Theta^*_{i,1}} + e^{\Theta^*_{i,2}})$. Such a sampling with replacement is necessary for the analysis, where we require independence in the choice of the items in S_i in order to apply the symmetrization technique (e.g. Boucheron et al. (2013)) to bound the expectation of the deviation (cf. Appendix C.4). Similar sampling assumptions have been made in existing analyses on learning low-rank models from noisy observations, e.g. Negahban and Wainwright (2012). Let $d \equiv (d_1 + d_2)/2$, and let $\sigma_j(\Theta^*)$ denote the j-th singular value of the matrix Θ^*. Define

$$\lambda_0 \;\equiv\; e^{2\alpha} \sqrt{\frac{d_1 \log d + d_2 \, (\log d)^2 (\log 2d)^4}{k \, d_1^2 \, d_2}} \;. \tag{34}$$

Theorem 7 *Under the described sampling model, assume* $24 \leq k \leq \min\{d_1^2 \log d, (d_1^2 + d_2^2)/(2d_1)\} \log d, \; (1/e)\, d_2(4 \log d_2 + 2 \log d_1)\}$, *and* $\lambda \in [480\lambda_0, c_0\lambda_0]$ *with any constant* $c_0 = O(1)$ *larger than 480. Then, solving the optimization (31) achieves*

$$\frac{1}{d_1 d_2} \left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\mathrm{F}}^2 \;\leq\; 288\sqrt{2}\, e^{4\alpha} c_0 \lambda_0 \sqrt{r}\, \left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\mathrm{F}} + 288 e^{4\alpha} c_0 \lambda_0 \sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^*)\,, \tag{35}$$

for any $r \in \{1, \ldots, \min\{d_1, d_2\}\}$ *with probability at least* $1 - 2d^{-3} - d_2^{-3}$ *where* $d = (d_1 + d_2)/2$.

A proof is provided in Appendix C. This bound holds for all values of r and one could potentially optimize over r. We show such results in the following corollaries.

Corollary 8 (Exact low-rank matrices) *Suppose* Θ^* *has rank at most* r. *Under the hypotheses of Theorem 7, solving the optimization (31) with the choice of the regularization parameter* $\lambda \in [480\lambda_0, c_0\lambda_0]$ *achieves with probability at least* $1 - 2d^{-3} - d_2^{-3}$,

$$\frac{1}{\sqrt{d_1 d_2}} \left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\mathrm{F}} \;\leq\; 288\sqrt{2} e^{6\alpha} c_0 \sqrt{\frac{r(d_1 \log d + d_2 \, (\log d)^2 (\log 2d)^4)}{k \, d_1}} \;. \tag{36}$$

The number of entries is $d_1 d_2$ and we rescale the Frobenius norm error appropriately by $1/\sqrt{d_1 d_2}$. For a rank-r matrix Θ^* with $r(d_1 + d_2) - r^2 = O(r(d_1 + d_2))$ degrees of freedom, the above theorem shows that the total number of samples, which is $(k\, d_1)$, needs to scale

as $O(rd_1(\log d) + rd_2 (\log d)^2(\log 2d)^4)$ in order to achieve an arbitrarily small error. This is only poly-logarithmic factor larger than the degrees of freedom. In Section 4.2, we provide a lower bound on the error directly, that matches the upper bound up to a logarithmic factor. The dependence on the dynamic range α is sub-optimal. The exponential dependence in the bound seems to be a weakness of the analysis, as seen from numerical experiments in the right panel of Figure 6. Although the error increase with α, numerical experiments suggests that it only increases at most linearly. A practical issue in achieving the above rate is the choice of λ, since the dynamic range α is not known in advance. Figure 6 illustrates that the error is not sensitive to the choice of λ for a wide range.

For approximately low-rank matrices in ℓ_q-ball defined in (13), optimizing the choice of r in Theorem 7, we get the following result. This is a strict generalization of Corollary 8 and a proof of this Corollary is provided in Appendix D.

Corollary 9 (Approximately low-rank matrices) *Suppose $\Theta^* \in \mathbb{B}_q(\rho_q)$ for some $q \in (0, 1]$ and $\rho_q > 0$. Under the hypotheses of Theorem 7, solving the optimization (31) with the choice of the regularization parameter $\lambda \in [480\lambda_0, c_0\lambda_0]$ achieves with probability at least $1 - 2d^{-3}$,*

$$\frac{1}{\sqrt{d_1 d_2}}\left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}} \;\leq\; \frac{2\sqrt{\rho_q}}{\sqrt{d_1 d_2}}\left(288\sqrt{2}c_0 e^{6\alpha}\sqrt{\frac{d_1 d_2(d_1 \log d + d_2 (\log d)^2(\log 2d)^4)}{k\, d_1}}\right)^{\frac{2-q}{2}}.\tag{37}$$

4.2 Information-theoretic Lower Bound for Low-rank Matrices

A simple parameter counting argument indicates that it requires the number of samples to scale as the degrees of freedom i.e., $kd_1 \propto r(d_1 + d_2)$, to estimate a $d_1 \times d_2$ dimensional matrix of rank r. By applying Fano's inequality with appropriately chosen hypotheses, the following lower bound establishes that the bound in Theorem 7 is sharp up to a logarithmic factor.

Theorem 10 *Suppose Θ^* has rank r. Under the described sampling model, for large enough d_1 and $d_2 \geq d_1$, there is a universal numerical constant $c > 0$ such that*

$$\inf_{\widehat{\Theta}}\sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\left[\frac{1}{\sqrt{d_1 d_2}}\left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}}\right] \;\geq\; c\min\left\{\alpha e^{-\alpha}\sqrt{\frac{r\, d_2}{k\, d_1}}, \frac{\alpha d_2}{\sqrt{d_1 d_2 \log d}}\right\},\tag{38}$$

where the infimum is taken over all measurable functions over the observed ranked lists $\{(v_{i,1}, \ldots, v_{i,k})\}_{i\in[d_1]}$.

A proof of this theorem is provided in Appendix E. The term of primary interest in this bound is the first one, which shows the scaling of the (rescaled) minimax rate as $\sqrt{r(d_1 + d_2)/(kd_1)}$ (when $d_2 \geq d_1$), and matches the upper bound in (35). It is the dominant term in the bound whenever the number of samples is larger than the degrees of freedom by a logarithmic factor, i.e., $kd_1 > r(d_1 + d_2)\log d$, ignoring the dependence on α. This is a typical regime of interest, where the sample size is comparable to the latent dimension of the problem. In this regime, Theorem 10 establishes that the upper bound in Theorem 7 is minimax-optimal up to a logarithmic factor in the dimension d.

4.3 Rank Breaking for Higher Order Comparisons

A common approach in practice to handle higher order comparisons is *rank breaking*, which refers to the practice of breaking the higher order comparisons into a set of pairwise comparisons and applying an estimator tailored for pairwise comparisons treating each pair as independent Azari Soufiani et al. (2013, 2014). When the higher order comparison is given as partial rankings (as opposed to total linear ordering as we assume) then rank breaking can be inconsistent, and special algorithms are needed for weighted rank breaking Khetan and Oh (2016a,b). However, when k-wise rankings (also called total linear orderings) are observed as we assume, simple and standard rank breaking achieves a similar performance as the higher order estimator in (31). Assume that $u_{i,m}$, $i \in [d_1]$, $m \in [k]$, denotes the m-th element observed by the i-th user. Concretely, in rank breaking, we convert the k-wise ranking data into pairwise ranking data and then we solve the following optimization problem:

$$\mathcal{L}(\Theta) = \frac{1}{d_1 \binom{k}{2}} \sum_{i \in [d_1]} \sum_{(m_1,m_2) \in \mathcal{P}_0} \left(\Theta_{i,\,h_i(m_1,m_2)} - \log \left(\exp \left(\Theta_{i,\,u_{i,m_1}} \right) + \exp \left(\Theta_{i,\,u_{i,m_2}} \right) \right) \right) ,$$

(39)

where $\mathcal{P}_0 = \{(i,j) : 1 \leq i < j \leq k\}$, and $h_i(m_1,m_2)$ and $l_i(m_1,m_2)$ is defined as the higher and lower ranked index among u_{i,m_1} and u_{i,m_2} respectively. Then modified optimization problem becomes,

$$\widehat{\Theta} \in \arg \min_{\Theta \in \Omega_\alpha} -\mathcal{L}(\Theta) + \lambda \|\|\|\Theta\|\|\|_{\mathrm{nuc}}$$

(40)

Let $d \equiv (d_1 + d_2)/2$, and let $\sigma_j(\Theta^*)$ denote the j-th singular value of the matrix Θ^*. Define

$$\lambda_0 \equiv \sqrt{\frac{d \log d}{k\, d_1^2\, d_2}} .$$

(41)

With this choice of regularization coefficient, we get the following upper bounds on the rank breaking estimator (40) that are comparable to the upper bounds of k-wise ranking estimator in Theorem 7 and Corollary 8.

Theorem 11 *Under the described sampling model, assume $2(c+4) \log d \leq k \leq \max\{d_1, d_2^2/d_1\} \log d$, $d_1 \geq 4$, and $\lambda \in [2\sqrt{32(c+4)}\lambda_0, c_p\lambda_0]$ with any constant $c = O(1)$ larger than $2\sqrt{32(c+4)}$. Then, solving the optimization (40) achieves*

$$\frac{1}{d_1 d_2} \left\|\left\|\left\|\widehat{\Theta} - \Theta^*\right\|\right\|\right\|_{\mathrm{F}}^2 \leq 144\sqrt{2}\, e^{2\alpha} c \lambda \sqrt{r} \left\|\left\|\left\|\widehat{\Theta} - \Theta^*\right\|\right\|\right\|_{\mathrm{F}} + 144 e^{2\alpha} c \lambda \sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^*) , \quad (42)$$

for any $r \in \{1, \ldots, \min\{d_1, d_2\}\}$ with probability at least $1 - 2d^{-c} - 2d^{-2^{13}}$ where $d = (d_1 + d_2)/2$.

A proof of this theorem is provided in Appendix F, and the following corollary follows for rank-r matrices.

Corollary 12 (Exact low-rank matrices) *Suppose Θ^* has rank at most r. Under the hypotheses of Theorem 11, there exists a constant $c_1 > 0$ such that solving the optimization (40) with the choice of the regularization parameter $\lambda \in [2\sqrt{32(c+4)}\lambda_0, c\lambda_0]$ achieves with probability at least $1 - 2d^{-c} - 2d^{-2^{13}}$,*

$$\frac{1}{\sqrt{d_1 d_2}}\left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}} \leq 144\sqrt{2}e^{2\alpha}c_1\sqrt{\frac{rd\log d}{k\,d_1}}\ . \tag{43}$$

4.4 Experiments

We provide results from numerical experiments on both synthetic and real data sets.

4.4.1 ALGORITHM

Similar to the case of pairwise comparisons in Section 3.4.1, we use proximal gradient descent Agarwal et al. (2010); Cai et al. (2010) along with modified Barzilai-Borwein (BB) step-size selection rule Barzilai and Borwein (1988) with the initial point $\Theta_0 = 0$. Each iteration of the algorithm applies the following two operations on the current estimate, Θ_t, of Θ^*,

$$\widetilde{\Theta}_{t+1} = \Theta_t - \eta_t \nabla_\Theta \mathcal{L}(\Theta_t) \qquad \text{(gradient descent)} \tag{44}$$

$$\Theta_{t+1} = M_t(\Gamma_t - \eta_t \lambda \mathbf{I})^+ N_t^T \qquad \text{(singular value shrinkage and thresholding)} \tag{45}$$

where $M_t \Gamma_t N_t^T := \widetilde{\Theta}_t$ is the singular value decomposition of $\widetilde{\Theta}_t$, such that Γ_t is a diagonal matrix with positive entries, $(\cdot)^+$ is the entry-wise thresholding operation $\max(0, x)$, and η_t is an BB step-size calculated as,

$$\eta_t = \begin{cases} \|\!\|\Theta_t - \Theta_{t-1}\|\!\|_2^2 / \langle\!\langle \Theta_t - \Theta_{t-1}, \nabla_\Theta \mathcal{L}(\Theta_t) - \nabla_\Theta \mathcal{L}(\Theta_{t-1}) \rangle\!\rangle, & \text{when } t \text{ is odd} \\ \langle\!\langle \Theta_t - \Theta_{t-1}, \nabla_\Theta \mathcal{L}(\Theta_t) - \nabla_\Theta \mathcal{L}(\Theta_{t-1}) \rangle\!\rangle / \|\!\|\nabla_\Theta \mathcal{L}(\Theta_t) - \nabla_\Theta \mathcal{L}(\Theta_{t-1})\|\!\|_2^2, & \text{when } t \text{ is even} \end{cases} . \tag{46}$$

4.4.2 SIMULATION: HIGHER ORDER COMPARISONS

The left panel of Figure 6 confirms the scaling of the error rate as predicted by Corollary 8. The lines merge to a single line when the sample size is rescaled appropriately (inset). We make a choice of $\lambda = \sqrt{(\log d)/(kd^2)}$. This choice is independent of α and is smaller than proposed in Theorem 7. We generate the random rank-r true MNL parameters matrices of dimension $d \times d$ using the process mentioned in Section 3.4.1. The root mean squared error (RMSE) is plotted where RMSE $= (1/\sqrt{d_1 d_2})\|\!\|\Theta^* - \widehat{\Theta}\|\!\|_{\mathrm{F}}$. We implement and solve the convex optimization (31) using proximal gradient descent method as analyzed in Agarwal et al. (2010). The right panel in Figure 6 illustrates that the actual error is insensitive to the choice of λ for a broad range of $\lambda \in [\sqrt{(\log d)/(kd^2)}, 2^8\sqrt{(\log d)/(kd^2)}]$, after which it increases with λ.

4.4.3 SIMULATION: RANK BREAKING

In this section we compare the higher order k-wise comparison algorithm (31) ('kwise') with the pairwise rank breaking algorithm (40) ('kbreak'). We use the same setting as
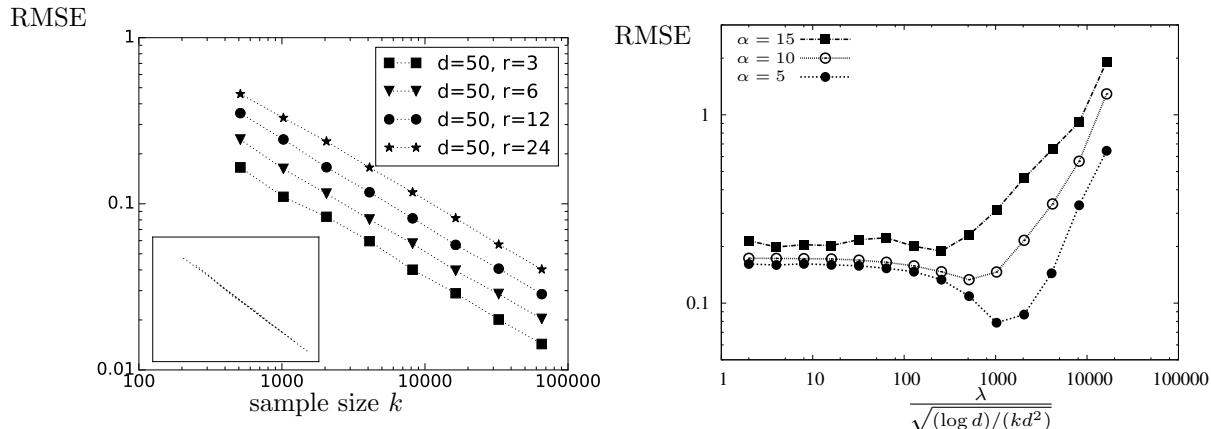
Figure 6: The (rescaled) RMSE scales as $\sqrt{r(\log d)/k}$ as expected from Corollary 8 for fixed $d = 50$ (left). In the inset, the same data is plotted versus rescaled sample size $k/(r \log d)$. The (rescaled) RMSE is stable for a broad range of λ and α for fixed $d = 50$ and $r = 3$ (right).

in Section 4.4.2, where we observe samples from k-wise ranking from an underlying true MNL model and the aim is to recover the true parameter Θ^* of the model. We use $\lambda = 0.45 \sqrt{(\log d)/(kd^2)}$ and $\lambda = 0.1 \sqrt{(\log d)/(kd^2)}$ for k-wise and pairwise rank breaking algorithms respectively. In Fig. 7 we plot the RMSE for both the algorithm for $d = 50$ and $r = 3, 12$. We note that the even though the RMSE decreases in the rate as predicted by the theorem, we see that pairwise rank breaking is worser than the higher order k-wise algorithm which directly uses the k-wise rankings. This is consistent with the experimental observation made previously in Hajek et al. (2014). Further we note that rank breaking is much slower than the other algorithm, since gradient computation of the former takes $O(k^2)$ time whereas for the latter it can be computed in $O(k)$ time.

4.4.4 REAL DATA: JESTER

Jester data set[3] Goldberg et al. (2001) has $24,982$ users, each rating a subset of 100 jokes on continuous scale of $[-10, 10]$. As the scale is continuous, we derive ordinal data from the scores (ties broken uniformly at random). We use only the 7200 users who rated all the jokes for our experiments. For each user, $k = 100x$ jokes were randomly selected uniformly at random for training, rest of the $100 - k = 100(1 - x)$ jokes where used for testing, where x is the fraction of jokes selected for training. We implment four algorithms: nuclear norm minimization ('nucnorm') (31), unregularized ($\lambda = 0$) log-likelihood maximization ('fullrank'), rank-1 Plackett-Luce model estimation ('plackett'), and rank breaking algorithm ('rankbreak') (40). We use $\lambda = 0.7 \sqrt{(0.5 \log(d_1 d_2))/(kd_1 \sqrt{d_1 d_2})}$ and $\lambda = 0.16 \sqrt{(0.5 \log(d_1 d_2))/(kd_1 \sqrt{d_1 d_2})}$ for k-wise and pairwise rank breaking algorithms respectively. In Fig. 8 (a) we plot the multiplicative bias in the mean log-likelihood on the testing data versus the fraction x of training data used. For each model in {'nucnorm',

---

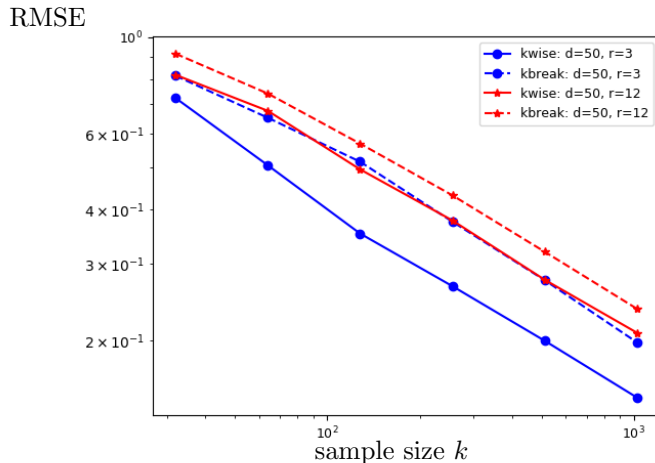3. Data set is from `http://eigentaste.berkeley.edu/dataset/`.

Figure 7: Rank Breaking: RMSE error versus number of samples per user k. k-wise ('kwise') algorithm performs better than the rank breaking ('kbreak') approach.

'fullrank', 'plackett', 'rankbreak'}, we plot in the y-axis

$$\frac{\log(P_{\text{model}}(\text{test data})) - \log(P_{\text{fullrank}}(\text{test data}))}{|\log(P_{\text{fullrank}}(\text{test data}))|} \, ,$$

using fullrank model as a baseline as it has the least test likelihood. Plackett-Luce model achieves the best performance when sample size is small, as this simplest model avoids overfitting. However, for most regimes of sample size, both the nuclear norm minimization and rank breaking achieve similar performance improving upon the others.

The same trend holds when we measure the perfomrance in the normalized Spearman's footule distance Diaconis and Graham (1977) $F(\pi_1, \pi_2) \in [0, 1]$ between two rank-lists π_1, π_2 of length k:

$$F(\pi_1, \pi_2) = \frac{2}{k^2} \sum_{i=1}^{k} |\pi_1(i) - \pi_2(i)|$$

In Fig. 8 (b) we plot the average normalized Spearman's footrule distance between the ground truths and the most likely ranking on the testing data under the estimated model parameters. We see that k-wise nuclear norm minimization and rank breaking algorithms perform the best in recovering the true ranking, except when the fraction of training data used is very small so that the rank-1 Plackett-Luce recovers better ranking.

4.4.5 REAL DATA: IRISH ELECTION

The Irish Election data set[4] is an opinion poll conducted among 1083 participants during the 1997 Irish presidential election campaign Gormley and Murphy (2009). Each participant responded with a ranking the of their top 1, 2, 3, 4, or 5 choices from the 5 candidates:

---

4. Data set is from `https://projecteuclid.org/euclid.aoas/1231424218#supplemental`.

(a) Multiplicative bias in the mean log-likelihood
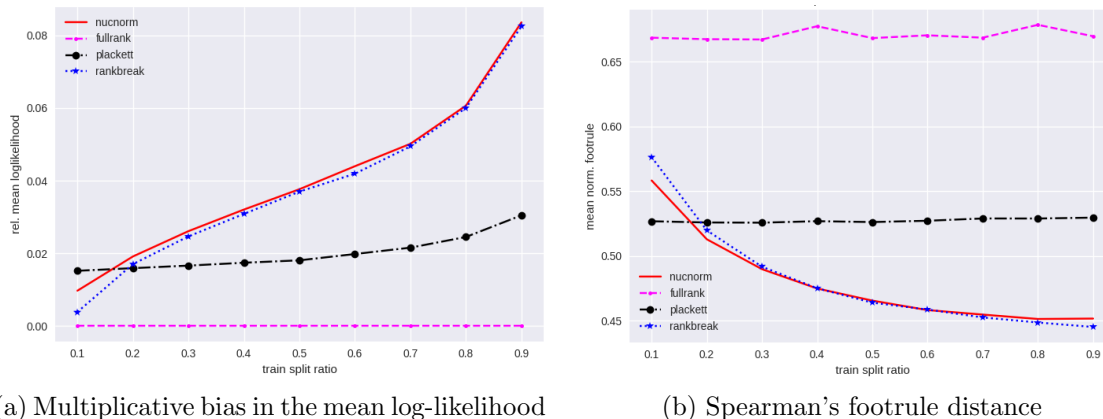

(b) Spearman's footrule distance

Figure 8: Jester data set: Performance on test data vs. fraction of the total data used for training. The proposed nuclear norm regularized algorithm ('nucnorm') and rank breaking ('rankbreak') improves upon both the unregularized algorithm ('fullrank') and the Plackett-Luce model estimation ('plackett') for most regimes of the sample size.

Banotti, McAleese, Nally, Roche, and Scallon. For our experiments we use only the 807 participants who gave their top-5 choices, i.e. full-rankings of all the candidates. Next we divide these participants into 60 (2x3x5x2) group according to a Cartesian product of four categorizations: sex (male/female), marital status (single/married/widowed+divorced), social class (F/AB/C1/C2/DE)[5], location (rural/city+town). We assume that within each group the responses of its member follow the same distriubtion and these distributions of all all the groups are captured by an MNL model with parameter $\Theta^* \in \mathbb{R}^{60 \times 5}$. We implement three algorithms: nuclear norm minimization ('nucnorm') (31), unregularized ($\lambda = 0$) log-likelihood maximization ('fullrank'), and rank-1 Plackett-Luce model estimation ('plackett'). We use $\lambda = 0.8 \sqrt{(0.5 \log(d_1 d_2))/(k d_1 \sqrt{d_1 d_2})}$. If x randomly sampled fraction of the data is used for training, then rest of the data is used for testing. In Fig. 9 we plot the mean log-likelihood ($\log(P_{\mathrm{model}}(\text{test data}))$) on the testing data versus the fraction of training data used. We see that nuclear norm minimization and Plackett-Luce model estimation tie for the first place and both improves significantly upon the un-regularized full-rank MNL model estimation. In Fig. 11 we plot the t-SNE Maaten and Hinton (2008) embedding of the rows of the estimated parameter matrix $\widehat{\Theta}$ when all the data is used for training. In Fig. 11a the markers represent the marital status of the group: single/married/divorced+widowed. In Fig. 11b the markers represent the social class of the groups. We see that married (left) and divorced+widowed (right) groups are clearly separated in the embedding, indicating that marital status influences the preference of candidates. However, we see that the social classes are less influential.

In Fig. 10 we represent the voting characteristics of the top 4 right singular vectors $\{\hat{v}_j\}_{j=1}^4$ of $\widehat{\Theta}$, which has a rank of 4 when all the data is used for training. The each stacked bar corresponds to the singular value σ_j marked on the x-axis. Partition of a bar represents the choice model distribution of the corresponding singular vector: $\exp(v_j)/(\mathbb{1}^T \exp(v_j))$,

---

5. Social classes are F: farmer, AB: middle class. C1: lower middle class, C2: skilled working class, and DE: other working class.
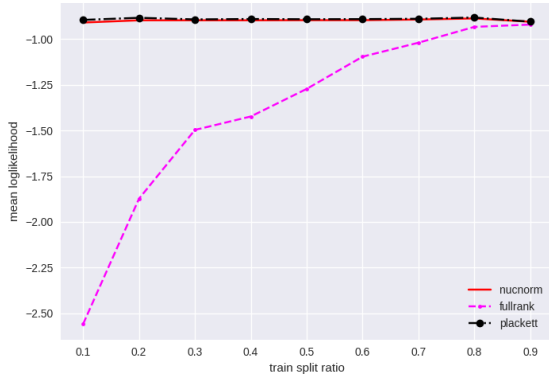
Figure 9: Irish Election: Mean log-likelihood on the test data versus fraction of the data used for training. Nuclear norm minimization and Plackett-Luce model estimator tie for the best performance.
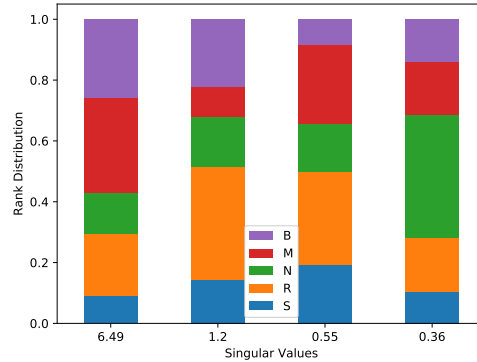
Figure 10: Irish Election: Each bar corresponds to rank distribution of one of the singular 4 values (x-axis) of the $\widehat{\Theta}$. Heights of the partitions represent the probability with which the distribution ranks the corresponding candidate as first (Section 4.4.5).

where $\exp(v_j)$ is the element-wise exponentiation operator. We see that there are 2 majors voting "basis" distributions; one favoring McAleese and another favoring Roche. Similar voting blocs have been observed earlier Gormley and Murphy (2009). Even though Placket-Luce model estimator achieves the similar likelihood as our nuclear norm regularized algorithm, the latter helps us in identifying voter "basis", solely from rank data without using the side-information on the voters as in Gormley and Murphy (2009).
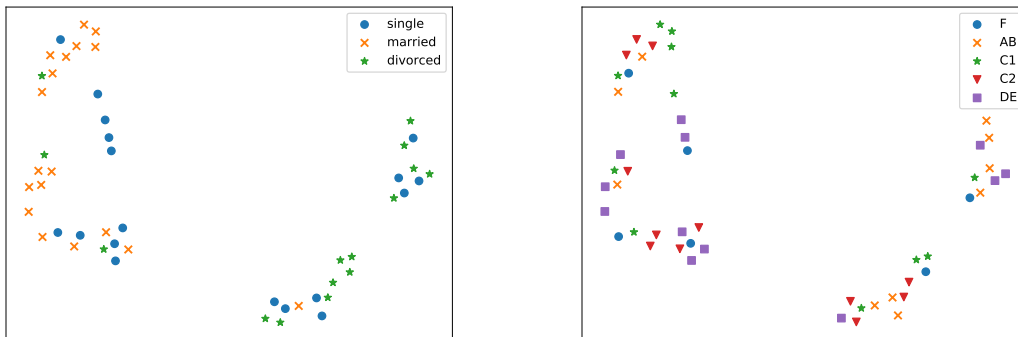
(a) Martial status

(b) Social class

Figure 11: Irish Election: t-SNE embedding of rows of the estimated parameter matrix $\widehat{\Theta}$. The markers correspond to the marital or social status (F: farmer, AB: middle class. C1: lower middle class, C2: skilled working class, DE: other working class) of the rows.

5. Learning the MNL Model from Choices

Choice modeling has had widespread success in numerous application domains such as transportation and marketing Train (1986); Guadagni and Little (1983). Choice models stem from revenue management to tackle the fundamental problem of maximizing expected revenue where the expectation is taken over a probabilistic choice model that is learned from historical purchase data. Revenue management has focused on designing efficient solvers for the optimization problem with exact or approximation guarantees, and has less to do with *learning* the parameters of probabilistic choice model of interest.

In this section, we tackle this unexplored domain of learning choice models from samples with provable guarantees on the sample complexity. In particular, we study learning the MNL model from choices. We study two types of choices under the MNL model that together include all practical scenarios of interest: *bundled choice* and *consumer choice*.

Bundled choice. We consider a novel scenario of significant practical interest: choice modeling from bundled purchase history. In this setting, we assume that we have bundled purchase history data from n users. Precisely, there are two categories of interest with d_1 and d_2 alternatives in each category respectively. For example, there are d_1 tooth pastes to choose from and d_2 tooth brushes to choose from. For the i-th user, a subset $S_i \subseteq [d_1]$ of alternatives from the first category is presented along with a subset $T_i \subseteq [d_2]$ of alternatives from the second category. We use k_1 and k_2 to denote the number of alternatives presented to a single user, i.e. $k_1 = |S_i|$ and $k_2 = |T_i|$, and we assume that the number of alternatives presented to each user is fixed, to simplify notations. However, the analysis naturally generalizes if the number differs from a user to another user. Given these sets of alternatives, each user makes a 'bundled' purchase, of an item from S_i and another item from T_i together, and we use (u_i, v_i) to denote these bundled pair of alternatives (e.g. a tooth brush and a tooth paste) purchased by the i-th user. Each user makes a choice of the best alternative, independent of other users's choices, according to the MNL model as

$$\mathbb{P}\left\{(u_i, v_i) = (j_1, j_2)\right\} \;\; = \;\; \frac{e^{\Theta^*_{j_1, j_2}}}{\sum_{j_1' \in S_i, j_2' \in T_i} e^{\Theta^*_{j_1', j_2'}}} \; , \tag{47}$$

for all $j_1 \in S_i$ and $j_2 \in T_i$. We emphasize here that the preference matrix is indexed by items of type one (in the rows) and items of type two (in the columns). We are taking the existing standard MNL model over user-item pairs to propose a novel choice model for bundled purchases over two types of items. One could go beyond paired bundled choices and include the user identity as another dimension, or add other types of items and consider higher order bundled purchases. This would require MNL model over higher order tensors, which is outside the scope of this paper, but are interesting generalizations. The main challenge in learning such tensor MNL models is that nuclear norm of a higher order tensor is not a computable quantity and hence minimizing the nuclear norm is not algorithmically feasible Yuan and Zhang (2014). Efficient methods exist based on alternating minimizations, but existing analysis tools can handle only quadratic losses Jain and Oh (2014).

The distribution (47) is independent of shifting all the values of Θ^* by a constant. Hence, there is an equivalent class of Θ^* that gives the same distribution for the choices: $[\Theta^*] \equiv \left\{A \in \mathbb{R}^{d_1 \times d_2} \mid A = \Theta^* + c\mathbb{1}\mathbb{1}^T \text{ for some } c \in \mathbb{R}\right\}$. Since we can only estimate Θ^* up to

this equivalent class, we search for the one that sums to zero, i.e. $\sum_{j_1 \in [d_1], j_2 \in [d_2]} \Theta^*_{j_1, j_2} = 0$. Let $\alpha = \max_{j_1, j_1' \in [d_1], j_2, j_2' \in [d_2]} |\Theta^*_{j_1, j_2} - \Theta^*_{j_1', j_2'}|$, denote the dynamic range of the underlying Θ^*, such that when $k_1 \times k_2$ alternatives are presented, we always have

$$\frac{1}{k_1 k_2} e^{-\alpha} \leq \mathbb{P}\{(u_i, v_i) = (j_1, j_2)\} \leq \frac{1}{k_1 k_2} e^{\alpha} , \tag{48}$$

for all $(j_1, j_2) \in S_i \times T_i$ and for all $S_i \subseteq [d_1]$ and $T_i \subseteq [d_2]$ such that $|S_i| = k_1$ and $|T_i| = k_2$. We do not make any assumptions on α other than that $\alpha = O(1)$ with respect to d_1 and d_2. Assuming Θ^* is well approximated by a low-rank matrix, we solve the following convex relaxation, given the observed bundled purchase history $\{(u_i, v_i, S_i, T_i)\}_{i \in [n]}$:

$$\widehat{\Theta} \in \arg \min_{\Theta \in \Omega_\alpha} \mathcal{L}(\Theta) + \lambda \|\|\Theta\|\|_{\text{nuc}} , \tag{49}$$

where the negative log likelihood function according to (47) is

$$\mathcal{L}(\Theta) = -\frac{1}{n} \sum_{i=1}^{n} \left(\langle\!\langle \Theta, e_{u_i} e_{v_i}^T \rangle\!\rangle - \log \left(\sum_{j_1 \in S_i, j_2 \in T_i} \exp\left(\langle\!\langle \Theta, e_{j_1} e_{j_2}^T \rangle\!\rangle\right) \right) \right) , \text{ and} \tag{50}$$

$$\Omega_\alpha \equiv \left\{ A \in \mathbb{R}^{d_1 \times d_2} \,\Big|\, \|\|A\|\|_\infty \leq \alpha, \text{ and} \sum_{j_1 \in [d_1], j_2 \in [d_2]} A_{j_1, j_2} = 0 \right\} . \tag{51}$$

Compared to collaborative ranking, (a) rows and columns of Θ^* correspond to an alternative from the first and second category, respectively; (b) each sample corresponds to the purchase choice of a user which follow the MNL model with Θ^*; (c) each person is presented subsets S_i and T_i of items from each category; (d) each sampled data represents the most preferred bundled pair of alternatives.

Customer choice. The standard customer choice can be thought of as either a special case of *bundled choice* or as a special case of *higher order comparisons*. We consider the standard customer choice data from purchase history. In this setting, we assume that we have purchase history data from d_1 users over d_2 alternatives. The i-th sample is i.i.d. with user u_i chosen uniformly at random and a subset $S_i \subseteq [d_2]$ of alternatives of size k. We fix k in order to be efficient in the notations and any variable size offerings can be handled seamlessly. We assume S_i is chosen uniformly at random with replacement, in a similar way as bundled choice and higher order comparisons.

Given these sets of alternatives, the user u_i makes a 'choice' and we use v_i to denote the purchased alternative by the i-th (sampled) user. Each user makes a choice of the best alternative, independent of other users's choices, according to the MNL model as

$$\mathbb{P}\{v_i = j_2 | u_i = j_1\} = \frac{e^{\Theta^*_{j_1, j_2}}}{\sum_{j_2' \in S_i} e^{\Theta^*_{j_1, j_2'}}} , \tag{52}$$

for all $j_2 \in S_i$. Up to the fact that we index rows by users and not items of one category, this is a special case of the *bundled choice* model where we fix $k_1 = 1$. Mathematically, all of our results under consumer choices are derived as corollaries from our results under bundled choices, but given the prevalent interest in customer choice models, we emphasize the implications of our framework under customer choice models in a separate section (see Section 5.2).

5.1 Learning the MNL Model from Bundled Choices

We provide an upper bound on the error achieved by our convex relaxation, when the *multiset* of alternatives S_i from the first category and T_i from the second category are drawn uniformly at random with replacement from $[d_1]$ and $[d_2]$ respectively. Precisely, for given k_1 and k_2, we let $S_i = \{j_{1,1}^{(i)}, \ldots, j_{1,k_1}^{(i)}\}$ and $T_i = \{j_{2,1}^{(i)}, \ldots, j_{2,k_2}^{(i)}\}$, where $j_{1,\ell}^{(i)}$'s and $j_{2,\ell}^{(i)}$'s are independently drawn uniformly at random over the d_1 and d_2 alternatives, respectively. Similar to the previous section, this sampling with replacement is necessary for the analysis. Define

$$\lambda_0 = \sqrt{\frac{e^{2\alpha} \max\{d_1, d_2\} \log d}{n \, d_1 \, d_2}} \ . \tag{53}$$

Theorem 13 *Under the described sampling model, assume $16e^{2\alpha} \min\{d_1, d_2\} \log d \leq n$ and $n \leq \min\{d^5, k_1 k_2 \max\{d_1^2, d_2^2\}\} \log d$, and $\lambda \in [8\lambda_0, c_1\lambda_0]$ with any constant $c_1 = O(1)$ larger than*
$\max\{8, 128/\sqrt{\min\{k_1, k_2\}}\}$. *Then, solving the optimization (49) achieves*

$$\frac{1}{d_1 d_2} \left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}}^2 \leq 48\sqrt{2}\, e^{2\alpha} c_1 \lambda \sqrt{r} \left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}} + 48e^{2\alpha} c_1 \lambda \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) \ , \tag{54}$$

for any $r \in \{1, \ldots, \min\{d_1, d_2\}\}$ with probability at least $1 - 2d^{-3}$ where $d = (d_1 + d_2)/2$.

A proof is provided in Appendix G. Optimizing over r gives the following corollaries.

Corollary 14 (Exact low-rank matrices) *Suppose Θ^* has rank at most r. Under the hypotheses of Theorem 13, solving the optimization (49) with the choice of the regularization parameter $\lambda \in [8\lambda_0, c_1\lambda_0]$ achieves with probability at least $1 - 2d^{-3}$,*

$$\frac{1}{\sqrt{d_1 d_2}} \left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}} \leq 48\sqrt{2}e^{3\alpha} c_1 \sqrt{\frac{r(d_1 + d_2) \log d}{n}} \ . \tag{55}$$

This corollary shows that the number of samples n needs to scale as $O(r(d_1 + d_2) \log d)$ in order to achieve an arbitrarily small error. This is only a logarithmic factor larger than the number of degrees of freedom. For approximately low-rank matrices in an ℓ_1-ball as defined in (13), we show an upper bound on the error, whose error exponent reduces from one to $(2 - q)/2$.

Corollary 15 (Approximately low-rank matrices) *Suppose $\Theta^* \in \mathbb{B}_q(\rho_q)$ for some $q \in (0, 1]$ and $\rho_q > 0$. Under the hypotheses of Theorem 13, solving the optimization (49) with the choice of the regularization parameter $\lambda \in [8\lambda_0, c_1\lambda_0]$ achieves with probability at least $1 - 2d^{-3}$,*

$$\frac{1}{\sqrt{d_1 d_2}} \left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}} \leq \frac{2\sqrt{\rho_q}}{\sqrt{d_1 d_2}} \left(48\sqrt{2}c_1 e^{3\alpha} \sqrt{\frac{d_1 d_2 (d_1 + d_2) \log d}{n}}\right)^{\frac{2-q}{2}} \ . \tag{56}$$

This follows from the same line of proof as in the proof of Corollary 9 in Appendix D. We next, provide a fundamental lower bound on the error, that matches the upper bound up to a logarithmic factor.

Theorem 16 *Suppose Θ^* has rank r. Under the described sampling model, there is a universal constant $c > 0$ such that that the minimax rate where the infimum is taken over all measurable functions over the observed purchase history $\{(u_i, v_i, S_i, T_i)\}_{i \in [n]}$ is lower bounded by*

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\left[\frac{1}{\sqrt{d_1 d_2}} \left\|\left\|\widehat{\Theta} - \Theta^*\right\|\right\|_F\right] \geq c \min\left\{\sqrt{\frac{e^{-5\alpha}\, r\, (d_1 + d_2)}{n}}, \frac{\alpha(d_1 + d_2)}{\sqrt{d_1 d_2 \log d}}\right\} . \quad (57)$$

We provide a proof in Appendix H. The first term is dominant, and when the sample size is comparable to the latent dimension of the problem, Theorem 13 is minimax optimal up to a logarithmic factor. We emphasize here that the bound in (55) and the matching lower bound in (57) do not depend on the size of the offerings k_1 and k_2. It is independent of how large k_1 and k_2 are because, we only observe one choice, and intuitively the information we get scales at best by a factor of $\log(k_1 k_2)$. The theorems prove that there is no essential gain in learning from large offerings. One might be tempted to stop at proving an upper bound that scales as $O(\sqrt{k_1 k_2 r(d_1 + d_2) \log d/n})$, which is larger than (55) by a factor of $\sqrt{k_1 k_2}$. Such a loose bound follows if one ignores the tight concentration analysis that we do using the symmetrization technique (e.g. in Lemma 29). Getting the tight dependency in k_1 and k_2 is one of the crucial technical challenges we overcome in this paper.

5.2 Learning the MNL Model from Customer Choices

The results for the *customer choice* model follow immediately from the results in *bundled choice* model by simply setting $k_1 = 1$, and we explicitly write those corollaries in this section for completeness. The proposed estimator is minimax optimal up to a logarithmic factor under the standard customer choice model of sampling.

Corollary 17 *Under the described sampling model, assume $16e^{2\alpha} \min\{d_1, d_2\} \log d \leq n \leq \min\{d^5, k \max\{d_1^2, d_2^2\}\} \log d$, and $\lambda \in [8\lambda_0, c_1\lambda_0]$ with any constant $c_1 = O(1)$ larger than 128. Then, solving the optimization (49) achieves*

$$\frac{1}{d_1 d_2} \left\|\left\|\widehat{\Theta} - \Theta^*\right\|\right\|_F^2 \leq 48\sqrt{2}\, e^{2\alpha} c_1 \lambda \sqrt{r} \left\|\left\|\widehat{\Theta} - \Theta^*\right\|\right\|_F + 48e^{2\alpha} c_1 \lambda \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) , \quad (58)$$

for any $r \in \{1, \ldots, \min\{d_1, d_2\}\}$ with probability at least $1 - 2d^{-3}$ where $d = (d_1 + d_2)/2$.

Corollary 18 (Exact low-rank matrices) *Suppose Θ^* has rank at most r. Under the hypotheses of Theorem 17, solving the optimization (49) with the choice of the regularization parameter $\lambda \in [8\lambda_0, c_1\lambda_0]$ achieves with probability at least $1 - 2d^{-3}$,*

$$\frac{1}{\sqrt{d_1 d_2}} \left\|\left\|\widehat{\Theta} - \Theta^*\right\|\right\|_F \leq 48\sqrt{2}e^{3\alpha} c_1 \sqrt{\frac{r(d_1 + d_2) \log d}{n}} . \quad (59)$$

Corollary 19 (Approximately low-rank matrices) *Suppose* $\Theta^* \in \mathbb{B}_q(\rho_q)$ *for some* $q \in (0, 1]$ *and* $\rho_q > 0$. *Under the hypotheses of Theorem 17, solving the optimization* (49) *with the choice of the regularization parameter* $\lambda \in [8\lambda, c_1\lambda]$ *achieves with probability at least* $1 - 2d^{-3}$,

$$\frac{1}{\sqrt{d_1 d_2}} \left\| \left\| \widehat{\Theta} - \Theta^* \right\| \right\|_F \leq \frac{2\sqrt{\rho_q}}{\sqrt{d_1 d_2}} \left(48\sqrt{2} c_1 e^{3\alpha} \sqrt{\frac{d_1 d_2 (d_1 + d_2) \log d}{n}} \right)^{\frac{2-q}{2}}. \tag{60}$$

We emphasize again that the bound in (59) does not depend on the size of the offerings k. It is significantly easier to stop at proving an upper bounds that scale as $O(\sqrt{kr(d_1 + d_2) \log d/n})$, which are larger than (59) by a factor of $\sqrt{k}$. Such a loose bound follows if one ignores the tight concentration analysis that we do using the symmetrization technique (e.g. in Lemma 29). Getting the tight dependency in k is one of the crucial technical challenges we overcome in this paper.

5.3 Experiments

We applied our algorithm to a real world choice data set. The implementation is similar to that of higher order comparisons (see Section 4.4.1).

5.3.1 REAL DATA: EXTENDED BAKERY

Extended Bakery data set [6] Benson et al. (2018) consists of details of 75,000 purchases at a bakery from a selection of 50 items, specifically each purchase is recorded by the set of items bought together. We use only the 13,579 purchases where a pair of items where bought. We divide the items into five categories: cakes (1-10), tarts (11-21), cookies (22-30), pastries (31-40) and drinks (41-50). We study four cases of bundled pairs, cakes and drinks, tarts and drinks, cookies and drinks, and pastries and drinks. These cases have 1503, 910, 500 and 1791 purchases in them respectively. We model the data with an MNL model parameterized by a matrix Θ^*, such that rows and columns corresponds to first and second categories respectively. For every purchase we assume that the subset of alternatives presented is the universal choice set, that is $k_1 = d_1$ and $k_2 = d_2$, so that the purchase of item j_1 from category 1 and j_1 from category 2 has a probability of $\exp(\Theta^*_{j_1,j_2}) / \sum_{j'_1, j'_2 = 1}^{d_1, d_2} \exp(\Theta^*_{j'_1, j'_2})$, where d_1, d_2 are number of items in category 1 and 2 respectively. We also fit the separable model proposed in Benson et al. (2018), which is a simpler model with $\Theta^*_{j, j_2} = a^*_{j_1} + b^*_{j_2}$, where $a \in \mathbb{R}^{d_1}, b \in \mathbb{R}^{d_2}$.

In Fig. 12 we plot the mean log-likelihood on the test data versus the fraction of the data used for our nuclear norm minimization based algorithm ('nucnorm'), un-regularized log-likelihood maximization algorithm ('fullrank'), and maximum likelihood estimator for the separable model ('separable') over 10 trials. We see that nuclear norm minimization outperforms the 'fullrank' and 'separable' algorithms. This is consistent with the marketing practice, of providing different prices for bundled combinations of products, which uses the rationale that, worth of a bundle of products might be different from the sum of the worths

---

6. Data set is from `https://github.com/arbenson/discrete-subset-choice/tree/master/data`.

(a) Cakes and Drinks

(b) Tarts and Drinks

(c) Cookies and Drinks

(d) Pastries and Drinks
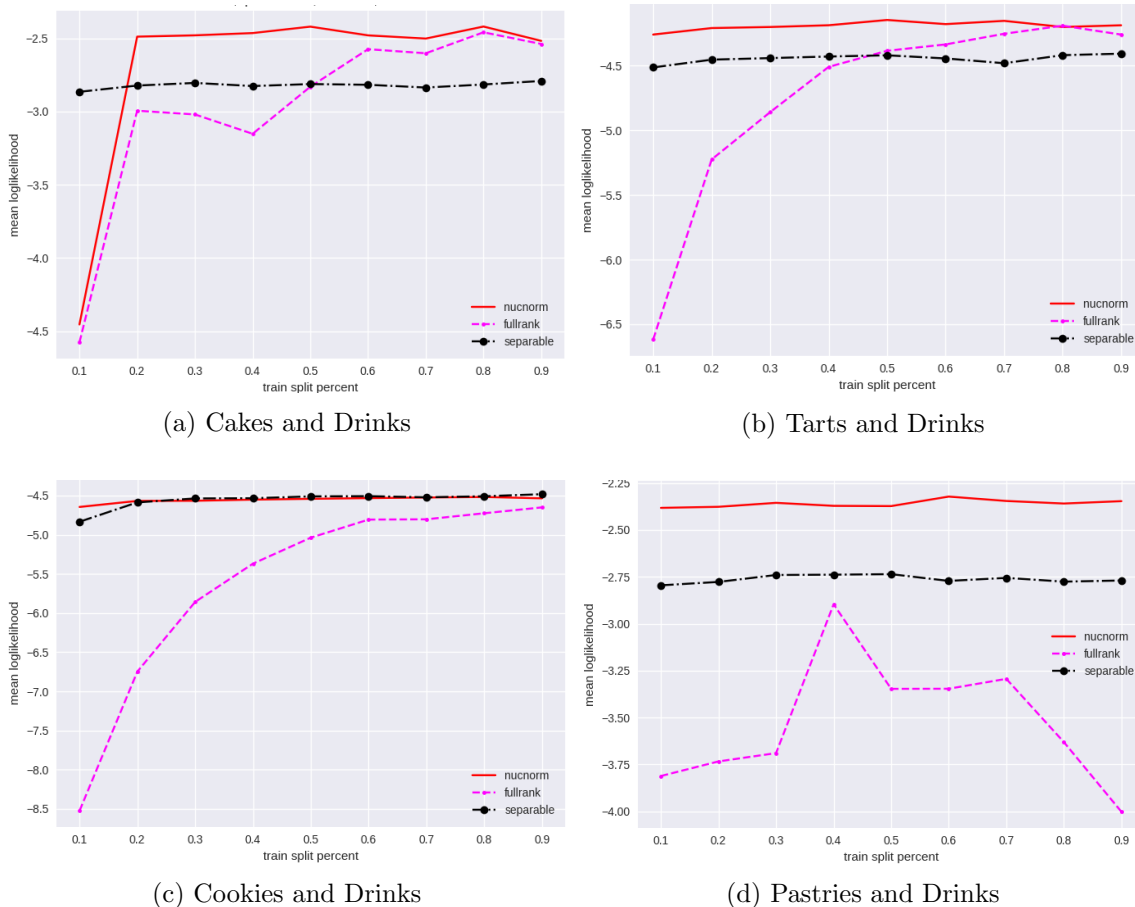
Figure 12: Bakery: Mean log-likelihood on test data versus fraction of data used for training. Nuclear norm minimization improves upon the un-regularized likelihood maximization and separable model, for most of the regimes we consider.

of the individual products constituting the bundle. We also note that the un-regularized algorithm ('fullrank') has the worst performance and high variance and separable model fits the samples almost as good as the nuclear norm minimization in the case of cookies and drinks in the large training data regime.

6. Conclusion

The sample complexity of learning one of the most popular choice models known as Multi-Nomial Logit model has not been addressed in the literature. The main challenge is in the inherent low-rank structure of the parameter to be learned, which leads to a non-convex likelihood maximized problem. Thanks to recent advances in learning low-rank matrices, in particular in 1-bit matrix completion Davenport et al. (2014), matrix completion Negahban and Wainwright (2012), and restricted strong convexity Negahban et al. (2009), we have a polynomial time algorithm and the technical tools to characterize the fundamental sample complexity of learning MNL from samples. This provides a novel algorithm to learn a

low-dimensional representation of users and items from users' historical comparisons and choices. We study three types of data, pairwise comparison, higher order comparison, and choices, and take the first principle approach of identifying the fundamental limits and also developing efficient algorithms matching those fundamental trade offs. We provide a unifying framework to learn the latent preferences by solving a convex program. For each of the data types, accompanied by natural sampling scenarios, we show that our framework achieves a minimax optimal performance, and hence cannot be improved upon other than a small logarithmic factor. This opens a new door to learn representations from comparisons and choices, and we propose new research directions and challenges below. Beyond the low-rank model studied in this paper, recent advances in modeling data in a matrix form such as low algebraic dimension by Ongie et al. (2018) and non-parametric approximation by Borgs et al. (2017) can provide new research directions for modeling choice.

Efficient implementations via non-convex optimization. Nuclear norm minimization, while polynomial-time, is still slow. We want first-order methods that are efficient with provable guarantees. Two main challenges are providing a good initialization to start such non-convex approaches and analyzing gradient descent on the likelihood maximization which is non-convex.

Recent advances in non-convex optimization with rank-constraints have developed via a sequence of innovations that can be summarized as follows, in a number of example problems including matrix completion, robust PCA, matrix sensing, phase retrieval. First, a convex relaxation of nuclear norm minimization is analyzed, e.g. Candès and Recht (2009). Then, a more efficient two-step non-convex optimization approach is proposed with provable guarantees where a global initialization step is followed by a first-order method e.g. Keshavan et al. (2010a,b). Next, first-order methods starting at any initialization point is analyzed via understanding the geometry and checking the stationary points of the objective function e.g. Ge et al. (2016). This recipe, spurred by the advances in the matrix completion problem, has been repeated for several interesting problems involving low rank matrices, over the last decade and over numerous publications by collective effort of the machine learning community.

For the problem of learning MNL, we are at the first stage of this progression where we propose a convex relaxation and provide minimax optimal guarantees. We currently do not have the analysis tools to follow up in analyzing an efficient non-convex optimization problem, although writing the algorithm and implementing is straight forward, and also has been proposed in Park et al. (2015). It is a promising research direction to overcome the challenges in analyzing non-convex optimization methods for the MNL likelihood objective function.

Assumption on sampling with replacement. As mentioned earlier, we assume sampling with replacement, where we can ask a user to compare the same pair more than once, and also we can ask a user to compare two copies of the identical item. Although such sampling with replacement does not happen in practice, the number of such collisions is also very low with high probability under the proposed model. Further, such assumption is critical for getting an upper bound that is tight not only in r, d_1 and d_2, but also in k for higher order comparisons and choices. If, instead, one is interested in sampling without

replacement, then one either can resort to proving a loose bound that is weaker in its dependence in k (and follows trivially as a corollary of the proof of our results) or needs to invent new innovative concentration bounds that do not rely on the powerful symmetrization. The first option is trivial, so we do not provide such corollaries in this paper, and the second option provides an interesting but technically challenging question of resolving between sampling with replacement and sampling without replacement. This we believe is outside the scope of this paper.

Modern data analysis applications. As learning representation from ordinal data is of fundamental interest, there are numerous exciting applications that both the algorithmic framework and also the analysis techniques we develop could be naturally extended to. We present two such examples. First is a recent application of embedding objects with crowdsourced similarity measures, first proposed in Tamuz et al. (2011). Consider a crowdsourcing setting where you have d images and want to learn similarities among those images such that one can embed those images in a lower dimensional Euclidean space. One can show to a person a triplet of images (i_1, i_2, i_3) and ask whether the image in the middle is more similar to the one in the left or the right. A natural model proposed in Tamuz et al. (2011) is to assume that there exists a similarity parameter matrix $\Theta \in \mathbb{R}^{d \times d}$ such that

$$\mathbb{P}\left\{i_1 \text{ is more similar to } i_2 \text{ than } i_3 \text{ is to } i_2\right\} = \frac{e^{\Theta_{i_2,i_1}}}{e^{\Theta_{i_2,i_1}} + e^{\Theta_{i_2,i_3}}} \ .$$

A heuristic algorithm is proposed to learn a low-rank Θ without guarantees. Given the similarity of this model to MNL in (1), both our algorithm and also the analysis will go through to provide a tight characterization of the sample complexity of this problem.

The second application is in word embedding Mikolov et al. (2013b), where the goal is to find embeddings for English words in a lower dimensional Euclidean space. The most successful word embedding has been based on fitting a low-rank matrix $\Theta \in \mathbb{R}^{d \times d}$ where d is the size of the vocabulary, over an MNL-type model:

$$\mathbb{P}\left\{\text{word } i \text{ and word } j \text{ appear within distance ten}|\text{word } j \text{ appear in a sentence}\right\} = \frac{e^{\Theta_{ij}}}{\sum_{i'} e^{\Theta_{i'j}}} \ .$$

As the denominator involves summation over millions of words in the vocabulary, efficient heuristics are proposed to learn such a model from skip-grams; a skip-gram is the count matrix counting how many times words co-appear in the same sentence within a predefined distance. There are several challenges in applying our framework directly to such a setting mainly due to the size of the problem, but nevertheless our analysis can be applied directly to identify the fundamental minimax sample complexity of learning a word embedding from skip-grams.

Acknowledgments

Appendix A. Proof of the Upper Bound for Graph Sampling Theorem 1

The proof of the theorem relies on the following two lemmas. First lemma shows that the negative of the log-likelihood satisfies Restricted Strong Convexity with high probability.

Lemma 20 (Restricted Strong Convexity) *Let* $R = \max\left\{\sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}(L)^{-1/2} \log(2d)}{n}\right\}$ *and the set* $\mathcal{A}(\alpha) = \left\{\Theta \in \mathbf{R}^{d_1 \times d_2}, \|\|\Theta\|\|_\infty \leq \alpha, \|\|\Theta\|\|_{\text{L-nuc}} \leq \frac{\|\|\Theta L^{1/2}\|\|_F^2}{16\alpha d_1 R}\right\}$. *Then we have,*

$$\frac{1}{n}\sum_{i=1}^{n}(\langle\!\langle \Theta, X_i \rangle\!\rangle)^2 \geq \frac{1}{3d_1}\|\|\Theta\|\|_L^2, \quad \forall \, \Theta \in \mathcal{A}(\alpha), \tag{61}$$

with probability at least $1 - 2(2d)^{-4}$, *provided that* $n \leq \log(2d) \min\{2^2(d_1 \sigma_{\min}(L)^{-1})^{2/3}, 2^6 d_1^2 \sigma^2\}, \,$.

Here the upper bound on n may not be necessary, but it is present due to a technical difficulty in using the peeling argument. The intuition behind the above lemma is that the empirical average uniformly concentrates around its expectation. Proof is in Section A.1. The next lemma says that the gradient of the log-likelihood at the actual parameter matrix, Θ^* is controllably small.

Lemma 21 (Bounded Gradient) *Let* $R = \max\left\{\sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}(L)^{-1/2} \log(2d)}{n}\right\}$. *The spectral norm of gradient of the log-likelihood at the actual parameter matrix,* $\nabla \mathcal{L}(\Theta^*)$, *can be upper-bounded with high probability as follows,*

$$\mathbb{P}\left\{\left\|\nabla \mathcal{L}(\Theta^*) L^{-1/2}\right\|_2 \geq \sqrt{32} R\right\} \leq \frac{1}{(d_1 + d_2)^3} \tag{62}$$

Proof the above lemma is in Section A.4. Let $\Delta = \hat{\Theta} - \Theta^*$.
Case 1: $\Delta \notin \mathcal{A}(2\alpha)$ Then,

$$\|\|\Delta\|\|_L^2 \leq 32\alpha d_1 R \|\|\Delta\|\|_{\text{L-nuc}}$$

Case 2: $\Delta \in \mathcal{A}(2\alpha)$ We first write down the second order Taylor series expansion of $\mathcal{L}(\hat{\Theta})$ at around $\Theta = \Theta^*$.

$$-\mathcal{L}(\hat{\Theta}) = -\mathcal{L}(\Theta^*) + \langle\!\langle -\nabla \mathcal{L}(\Theta^*), \Delta \rangle\!\rangle + \frac{1}{2n}\sum_{i=1}^{n} \psi\left(\langle\!\langle \Theta^*, X^{(i)} \rangle\!\rangle + s\langle\!\langle \Delta, X^{(i)} \rangle\!\rangle\right) \langle\!\langle \Delta, X^{(i)} \rangle\!\rangle^2, \tag{63}$$

where $\psi(x) = e^x/(1+e^x)^2$, $x \in [-2\alpha, 2\alpha]$ and $s \in [0,1]$. Next using Lemma 20 and the fact that $\psi(x)$ attains minimum at $x = 2\alpha$ we get,

$$-\mathcal{L}(\hat{\Theta}) + \mathcal{L}(\Theta^*) + \langle\!\langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle\!\rangle \geq \frac{1}{2n}\sum_{i=1}^{n} \psi(2\alpha)\langle\!\langle \Delta, X^{(i)} \rangle\!\rangle^2 \geq \frac{\psi(2\alpha)}{6d_1}\|\|\Delta\|\|_L^2, \tag{64}$$

with probability at least $1 - 1/(d_1 + d_2)^3$. Since $\hat{\Theta}$ is the minimizer for the objective function 9, we have,

$$-\mathcal{L}(\hat{\Theta}) + \lambda \left|\!\left|\!\left| \hat{\Theta} \right|\!\right|\!\right|_{\text{L-nuc}} \leq -\mathcal{L}(\Theta^*) + \lambda \|\!|\Theta^*\|\!|_{\text{L-nuc}},$$

which in turn gives us,

$$\frac{\psi(2\alpha)}{6d_1} \|\!|\Delta\|\!|_{\text{L}}^2 \leq -\mathcal{L}(\hat{\Theta}) + \mathcal{L}(\Theta^*) + \langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle$$

$$(65)$$

$$\leq \lambda \left(\|\!|\Theta^*\|\!|_{\text{L-nuc}} - \left|\!\left|\!\left| \hat{\Theta} \right|\!\right|\!\right|_{\text{L-nuc}} \right) + \langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle$$

$$\leq \lambda \left(\|\!|\Delta\|\!|_{\text{L-nuc}} \right) + \langle\!\langle \nabla\mathcal{L}(\Theta^*)L^{-1/2}, \Delta L^{1/2} \rangle\!\rangle$$

$$\leq \lambda \left(\|\!|\Delta\|\!|_{\text{L-nuc}} \right) + \left\|\nabla\mathcal{L}(\Theta^*)L^{-1/2}\right\|_2 \|\!|\Delta\|\!|_{\text{L-nuc}}, \qquad (66)$$

where last two inequalities follow from the triangle inequality for nuclear norm and generalized Hölder's inequality. Now we put $\lambda = 2\sqrt{32}R$ and use Lemma 21 to get,

$$\|\!|\Delta\|\!|_{\text{L}}^2 \leq \frac{6d_1}{\psi(2\alpha)} \left(\lambda + \frac{\lambda}{2} \right) \|\!|\Delta\|\!|_{\text{L-nuc}} \leq \frac{9d_1\lambda}{\psi(2\alpha)} \|\!|\Delta\|\!|_{\text{L-nuc}}, \qquad (67)$$

with probability at least $1 - 1/(d_1 + d_2)^3$. Combining Case 1 and 2 we get,

$$\|\!|\Delta\|\!|_{\text{L}}^2 \leq 9 \left(\alpha + \frac{1}{\psi(2\alpha)} \right) d_1 \lambda \|\!|\Delta\|\!|_{\text{L-nuc}}$$

Lemma 22 *If* $\lambda \geq 2\|\!|\nabla\mathcal{L}(\Theta^*)\|\!|_2$, *then we have*

$$\|\!|\Delta\|\!|_{\text{L-nuc}} \leq 4\sqrt{2r}\|\!|\Delta\|\!|_{\text{L}} + 4 \sum_{j=r+1}^{\min\{d_1, d_2 - G\}} \sigma_j(\Theta^* L^{1/2}), \qquad (68)$$

for all $r \in [\min\{d_1, d_2 - G\}]$. *(Proof in Section A.5)*

Finally, utilizing the above lemma, we get,

$$\frac{1}{d_1} \|\!|\Delta\|\!|_{\text{L}}^2 \leq 36\lambda \left(\alpha + \frac{1}{\psi(2\alpha)} \right) \left(\sqrt{2r}\|\!|\Delta\|\!|_{\text{L}} + \sum_{j=r+1}^{\min\{d_1, d_2 - G\}} \sigma_j(\Theta^* L^{1/2}) \right)$$

A.1 Proof of Lemma 20

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)} \rangle\!\rangle \right)^2 \geq \frac{1}{3d_1} \|\!|\Theta\|\!|_{\text{L}}^2, \ \forall\, \Theta \in \mathcal{A} \right\}$$

$$= 1 - \mathbb{P}\left\{ \exists\, \Theta \in \mathcal{A}, \text{ such that } \frac{1}{n} \sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)} \rangle\!\rangle \right)^2 < \frac{1}{3d_1} \|\!|\Theta\|\!|_{\text{L}}^2 \right\} \qquad (69)$$

When $\Theta \in \mathcal{A}$,

$$\|\Theta\|_{\mathrm{L}}^2 \geq 16\alpha d_1 R \|\Theta\|_{\mathrm{L\text{-}nuc}} \geq 16\alpha d_1 R \|\Theta\|_{\mathrm{L}} \implies \|\Theta\|_{\mathrm{L}} \geq 16\alpha d_1 R := \mu \,, \tag{70}$$

where the second inequality follows from $\|\Theta\|_{\mathrm{L\text{-}nuc}} = \left\|\Theta L^{1/2}\right\|_{\mathrm{nuc}} \geq \left\|\Theta L^{1/2}\right\|_{\mathrm{F}} = \|\Theta\|_{\mathrm{L}}$.

Lemma 23 *Let* $\mathcal{B}(D) := \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} | \|\Theta\|_\infty \leq \alpha, \ \|\Theta\|_{\mathrm{L}} \leq D, \ \|\Theta\|_{\mathrm{L\text{-}nuc}} \leq \frac{D^2}{16\alpha d_1 R}\right\}$, *and,*
$Z_D := \sup_{\Theta \in \mathcal{B}(D)} \left(-\frac{1}{n}\sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 + \frac{2}{d_1}\|\Theta\|_{\mathrm{L}}^2\right)$, *then,*

$$\mathbb{P}\left\{Z_D \geq \frac{3}{2d_1}D^2\right\} \leq \exp\left(-\frac{nD^4}{32\alpha^4 d_1^2}\right). \tag{71}$$

Above lemma is proved in Section A.2. Let $\beta = \sqrt{\frac{10}{9}}$, then the sets,

$$\mathcal{S}_\ell = \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} | \|\Theta\|_\infty \leq \alpha, \beta^{\ell-1}\mu \leq \|\Theta\|_{\mathrm{L}} \leq \beta^\ell \mu, \ \|\Theta\|_{\mathrm{L\text{-}nuc}} \leq \frac{(\beta^\ell \mu)^2}{16\alpha d_1 R}\right\}, \ \ell = 1,2,3,\ldots, \tag{72}$$

cover the set $\mathcal{A}$, that is $\mathcal{A} \subset \cup_{\ell=1}^\infty \mathcal{S}_\ell$ and $\mathcal{S}_\ell \subseteq \mathcal{B}(\beta^\ell \mu)$. This gives us,

$$\mathbb{P}\left\{\exists\, \Theta \in \mathcal{A} \text{ s.t. } \frac{1}{n}\sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}\|\Theta\|_{\mathrm{L}}^2\right\}$$

$$\leq \sum_{\ell=1}^\infty \mathbb{P}\left\{\exists\, \Theta \in \mathcal{S}_\ell \text{ s.t. } \frac{1}{n}\sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}\|\Theta\|_{\mathrm{L}}^2\right\}$$

$$\leq \sum_{\ell=1}^\infty \mathbb{P}\left\{\exists\, \Theta \in \mathcal{B}(\beta^\ell \mu) \text{ s.t. } \frac{1}{n}\sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}\|\Theta\|_{\mathrm{L}}^2\right\} \tag{73}$$

If there exists a $\Theta \in \mathcal{B}(\beta^\ell \mu)$ such that $\frac{1}{n}\sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}\|\Theta\|_{\mathrm{L}}^2$ then,

$$Z_{\beta^\ell \mu} \geq -\frac{1}{n}\sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 + \frac{2}{d_1}\|\Theta\|_{\mathrm{L}}^2 > \frac{5}{3d_1}\|\Theta\|_{\mathrm{L}}^2 \geq \frac{5}{3d_1}\beta^{2\ell-2}\mu^2 = \frac{3}{2d_1}(\beta^\ell \mu)^2,$$

which gives us,

$$\mathbb{P}\left\{\exists\, \Theta \in \mathcal{A} \ni \frac{1}{n}\sum_{i=1}^n \left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}\|\Theta\|_{\mathrm{L}}^2\right\} \leq \sum_{\ell=1}^\infty \mathbb{P}\left\{Z_{\beta^\ell \mu} > \frac{3}{2d_1}(\beta^\ell \mu)^2\right\}$$

$$\overset{(a)}{\leq} \sum_{\ell=1}^\infty \exp\left(-\frac{n(\beta^\ell \mu)^4}{32\alpha^4 d_1^2}\right)$$

$$\overset{(b)}{\leq} \sum_{\ell=1}^\infty \exp\left(-\frac{4\ell(\beta-1)n\mu^4}{32\alpha^4 d_1^2}\right)$$

$$\overset{(c)}{\leq} 2\exp\left(-\frac{4(\beta-1)n\mu^4}{32\alpha^4 d_1^2}\right) \tag{74}$$

where (a) is from Lemma 23, (b) is true since $\beta^{4\ell} \geq 4\ell(\beta-1)$ when $\beta \geq 1$ and (c) is obtained by summing the geometric series, with common ratio less than $1/2$, in previous inequality. Finally if we assume that $n \leq 2^6 d_1^2 \sigma^2 \log(2d)$ and $n \leq 2^2 (d_1 \sigma_{\min}(L)^{-1})^{2/3} \log(2d)$, then we have $2^2 \log(2d) \leq 4(\beta-1)n\mu^4/32\alpha^4 d_1^2$ as follows

$$
\frac{4(\beta-1)n\mu^4}{32\alpha^4 d_1^2} = \frac{4(\beta-1)n(16\alpha d_1 R)^4}{32\alpha^4 d_1^2} = 2^{13}(\beta-1)d_1^2 \max\left\{\frac{\sigma^2 \log^2(2d)}{n}, \frac{\sigma_{\min}(L)^{-2}\log^4(2d)}{n^3}\right\}
$$
$$
\geq 2^2 \log(2d) \tag{75}
$$

A.2 Proof of Lemma 23

Notice that the $\frac{2}{d_1}\||\Theta\||_{\mathrm{L}}^2$ is the mean of $\frac{1}{n}\sum_{i=1}^n \langle\!\langle \Theta, X^{(i)}\rangle\!\rangle^2$,

$$
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \langle\!\langle \Theta, X^{(i)}\rangle\!\rangle^2\right] = \frac{1}{d_1}\sum_{j\in[d1]}\sum_{k,l\in[d2]}(\Theta_{j,k}-\Theta_{j,l})^2 P_{k,l}
$$
$$
= \frac{2}{d_1}\sum_j\sum_k \Theta_{j,k}^2 \sum_l P_{k,l} - 2\sum_{k,l}\Theta_{j,k}\Theta_{j,l}P_{k,l}
$$
$$
\overset{(a)}{=} \frac{2}{d_1}\sum_j \langle\!\langle \Theta_j\Theta_j^T, \mathrm{diag}(P_k)\rangle\!\rangle - 2\langle\!\langle \Theta_j\Theta_j^T, P\rangle\!\rangle
$$
$$
= \frac{2}{d_1}\sum_j \langle\!\langle \Theta_j\Theta_j^T, L\rangle\!\rangle = \frac{2}{d_1}\left\|\!\left|\Theta L^{1/2}\right|\!\right\|_{\mathrm{F}}^2
$$

where, in (a) $P_k = \sum_{l\in[d2]} P_{k,l}$ and Θ_j is the j-th row of Θ. Therefore we use the following standard technique to get a handle on supremum of deviation from mean.

First, we use bounded differences property to prove that Z_D concentrates around its mean. We write $Z_D(X^{(1)}, \ldots, X^{(n)})$ to represent Z_D as a function of n independent random variables. Now, let $X^{(i)}$ and $\tilde{X}^{(i)}$ be two realization of the i-th $(1 \leq i \leq n)$ random parameter of Z_D, then,

$$
\left|Z_D(X^{(1)}, \ldots, X^{(i)}, \ldots, X^{(n)}) - Z_D(X^{(1)}, \ldots, \tilde{X}^{(i)}, \ldots, X^{(n)})\right|
$$
$$
= \left|\sup_{\Theta\in\mathcal{B}(D)}\left(-\frac{1}{n}\sum_{i=1}^n \langle\!\langle \Theta, X^{(i)}\rangle\!\rangle^2 + \frac{2}{d_1}\||\Theta\||_{\mathrm{L}}^2\right) - \right.
$$
$$
\left. \sup_{\Theta'\in\mathcal{B}(D)}\left(-\frac{1}{n}\left(\sum_{i=1,i\neq i'}^n \langle\!\langle \Theta', X^{(i)}\rangle\!\rangle^2 + \langle\!\langle \Theta', \tilde{X}^{(i')}\rangle\!\rangle^2\right) + \frac{2}{d_1}\||\Theta'\||_{\mathrm{L}}^2\right)\right| \tag{76}
$$

Now WLOG assume that $Z_D(X^{(1)}, \ldots, X^{(i)}, \ldots, X^{(n)}) \geq Z_D(X^{(1)}, \ldots, \tilde{X}^{(i)}, \ldots, X^{(n)})$ and the first supremum is achieved at $\bar{\Theta}$, which gives us.

$$
= \sup_{\Theta \in \mathcal{B}(D)} \left(-\frac{1}{n} \sum_{i=1}^{n} \langle\!\langle \Theta, X^{(i)} \rangle\!\rangle^2 + \frac{2}{d_1} \|\!|\Theta|\!\|_{\mathrm{L}}^2 \right) -
$$

$$
\sup_{\Theta' \in \mathcal{B}(D)} \left(-\frac{1}{n} \left(\sum_{\substack{i=1 \\ i \neq i'}}^{n} \langle\!\langle \Theta', X^{(i)} \rangle\!\rangle^2 + \langle\!\langle \Theta', \tilde{X}^{(i')} \rangle\!\rangle^2 \right) + \frac{2}{d_1} \|\!|\Theta'|\!\|_{\mathrm{L}}^2 \right)
$$

$$
\leq \left(-\frac{1}{n} \sum_{i=1}^{n} \langle\!\langle \bar{\Theta}, X^{(i)} \rangle\!\rangle^2 + \frac{2}{d_1} \|\!|\bar{\Theta}|\!\|_{\mathrm{L}}^2 \right) - \left(-\frac{1}{n} \left(\sum_{\substack{i=1 \\ i \neq i'}}^{n} \langle\!\langle \bar{\Theta}, X^{(i)} \rangle\!\rangle^2 + \langle\!\langle \bar{\Theta}, \tilde{X}^{(i')} \rangle\!\rangle^2 \right) + \frac{2}{d_1} \|\!|\bar{\Theta}|\!\|_{\mathrm{L}}^2 \right)
$$

$$
\leq \sup_{\Theta \in \mathcal{B}(D)} \frac{1}{n} \left| \langle\!\langle \Theta, X^{(i)} \rangle\!\rangle^2 - \langle\!\langle \Theta, \tilde{X}^{(i)} \rangle\!\rangle^2 \right|
$$

$$
\leq \frac{4\alpha^2}{n}, \tag{77}
$$

where the last inequality is true since, $\|\!|\Theta|\!\|_{\infty} \leq \alpha$ for any $\Theta \in \mathcal{B}(D) \subseteq \Omega_{\alpha}$. Now we upper bound $\mathbb{E}[Z_D]$ as follows,

$$
\mathbb{E}[Z_D] \overset{(a)}{\leq} 2\mathbb{E} \left[\sup_{\Theta \in \mathcal{B}(D)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle\!\langle \Theta, X^{(i)} \rangle\!\rangle^2 \right]
$$

$$
\overset{(b)}{\leq} 4\alpha \mathbb{E} \left[\sup_{\Theta \in \mathcal{B}(D)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle\!\langle \Theta L^{1/2}, X^{(i)} L^{-1/2} \rangle\!\rangle \right]
$$

$$
\leq 4\alpha \mathbb{E} \left[\sup_{\Theta \in \mathcal{B}(D)} \|\!|\Theta|\!\|_{\mathrm{L\text{-}nuc}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X^{(i)} L^{-1/2} \right\|_2 \right]
$$

$$
\leq 4\alpha \sup_{\Theta \in \mathcal{B}(D)} \|\!|\Theta|\!\|_{\mathrm{L\text{-}nuc}} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X^{(i)} L^{-`1/2} \right\|_2 \right],
$$

where (a) is standard symmetrization argument using i.i.d. Rademacher variables $\{\varepsilon_i\}_{i=1}^n$ and since $|\langle\!\langle \Theta, X^{(i)} \rangle\!\rangle| \leq 2\alpha$ we use Ledoux-Talagrand contraction inequality Ledoux and Talagrand (2013) to obtain (b).

Lemma 24 *Let* $R = \max\left\{ \sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}(L)^{-1/2} \log(2d)}{n} \right\}$. *For* $\{X^{(i)}\}_{i=1}^n$ *as defined in the graph sampling and for a binary random variable* ε_i *such that* $\mathbb{E}[\varepsilon_i | X^{(i)}] = 0$ *and* $|\varepsilon_i| \leq 1$, *we have,*

$$
\mathbb{P}\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X^{(i)} L^{-1/2} \right\|_2 \geq \sqrt{32} R \right\} \leq \frac{1}{(d_1 + d_2)^3}, \quad \text{and}, \quad \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X^{(i)} L^{-1/2} \right\|_2 \right] \leq 4R. \tag{78}
$$

Proof of the lemma is in Section A.3. Now using Lemma 24 we have $\mathbb{E}\left[Z_D\right]$ $\leq 16 R\alpha \sup_{\Theta \in \mathcal{B}(D)} \|\Theta\|_{\text{L-nuc}} \leq \frac{D^2}{d_1}$. Then using the bounded differences property and the upper bound on the mean, we get the McDiarmid's concentration,

$$\mathbb{P}\left\{Z_D - D^2/d_1 \geq t\right\} \leq \mathbb{P}\left\{Z_D - \mathbb{E}\left[Z_D\right] \geq t\right\}$$

$$\leq \exp\left(-\frac{nt^2}{8\alpha^4}\right) \tag{79}$$

and putting $t = D^2/2d_1$ gives us the theorem.

A.3 Proof of Lemma 24

Let $W_i := \frac{1}{n}\varepsilon_i X^{(i)} L^{-1/2} = \frac{1}{n}\varepsilon_i e_{j(i)}\left(e_{k(i)} - e_{l(i)}\right)^T L^{-1/2}$ and pseudo-inverse of L be $L^\dagger$, then, $\|W_i\|_2 \leq \sigma_{\min}(L)^{-1/2}\sqrt{2}/n$,

$$\mathbb{E}\left[W_i W_i^T\right] = \mathbb{E}\left[\frac{1}{n^2}e_{j(i)}\left(e_{k(i)} - e_{l(i)}\right)^T L^{-1/2}L^{-1/2}\left(e_{k(i)} - e_{l(i)}\right)e_{j(i)}^T\right]$$

$$= \mathbb{E}\left[\frac{1}{n^2}e_{j(i)}e_{j(i)}^T\right]\mathbb{E}\left[\left(e_{k(i)} - e_{l(i)}\right)^T L^\dagger\left(e_{k(i)} - e_{l(i)}\right)\right]$$

$$= \frac{1}{n^2 d_1}\mathbf{I}_{d_1 \times d_1} \times 2\left(\mathbb{E}\left[e_{k(i)}^T L^\dagger e_{k(i)}\right] - \mathbb{E}\left[e_{k(i)}^T L^\dagger e_{l(i)}\right]\right)$$

$$= \frac{2}{n^2 d_1}\left(\sum_{u \in [d_1]} P_u L_{u,u}^\dagger - \sum_{u,v \in [d_1]} P_{u,v} L_{u,v}^\dagger\right)\mathbf{I}_{d_1 \times d_1}$$

$$= \frac{2}{n^2 d_1}\langle\!\langle L, L^\dagger\rangle\!\rangle\mathbf{I}_{d_1 \times d_1}$$

$$= \frac{2(d_2 - G)}{n^2 d_1}\mathbf{I}_{d_1 \times d_1}, \tag{80}$$

$$\mathbb{E}\left[W_i^T W_i\right] = L^{-1/2}\mathbb{E}\left[\frac{1}{n^2}\left(e_{k(i)} - e_{l(i)}\right)\left(e_{k(i)} - e_{l(i)}\right)^T\right]L^{-1/2}$$

$$= \frac{1}{n^2}L^{-1/2}\left(\sum_{u,v=1}^{d_2}\left(e_u - e_v\right)\left(e_u - e_v\right)^T P_{u,v}\right)L^{-1/2}$$

$$= \frac{1}{n^2}L^{-1/2}\left(2L\right)L^{-1/2}$$

$$= \frac{2}{n^2}\left(\mathbf{I}_{d_2 \times d_2} - \sum_{k \in [G]} g_k g_k^T / \|g_k\|_2^2\right), \text{ and,} \tag{81}$$

$$\max\left\{\left\|\mathbb{E}\left[\sum_{i=1}^n W_i W_i^T\right]\right\|_2, \left\|\mathbb{E}\left[\sum_{i=1}^n W_i^T W_i\right]\right\|_2\right\} \leq \sum_{i=1}^n \max\left\{\left\|\mathbb{E}\left[W_i W_i^T\right]\right\|_2, \left\|\mathbb{E}\left[W_i^T W_i\right]\right\|_2\right\} \tag{82}$$

$$\leq \frac{2}{n}\sigma. \tag{83}$$

44

where $\sigma = \max\left\{\frac{d_2-G}{d_1}, 1\right\}$. Now by Matrix Bernstein concentration theorem Tropp (2015) we have,

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X^{(i)} L^{-1/2}\right\|_2 \geq t\right\} \leq \exp\left(\frac{-nt^2/2}{2\sigma + \sqrt{2}\sigma_{\min}(L)^{-1/2}t/3}\right), \text{ and,} \tag{84}$$

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X^{(i)} L^{-1/2}\right\|_2\right] \leq \sqrt{\frac{4\sigma\log(d_1+d_2)}{n}} + \frac{\sqrt{2\sigma_{\min}(L)^{-1}}}{3n}\log(d_1+d_2). \tag{85}$$

Choosing $t = \max\left\{\sqrt{\frac{24\sigma\log(d_1+d_2)}{n}}, \frac{16\sqrt{2\sigma_{\min}(L)^{-1}}\log(d_1+d_2)}{n}\right\}$ produces the desired result.

A.4 Proof of Lemma 21

The gradient can be written down as,

$$\nabla\mathcal{L}(\Theta^*) = \frac{1}{n}\sum_{i=1}^n \left(y_i - \frac{\exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}{1 + \exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}\right) X^{(i)}. \tag{86}$$

Then Lemma 24 directly gives the result because,

$$\mathbb{E}\left[y_i - \frac{\exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}{1 + \exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}\Bigg| X^{(i)}\right] = 0 \quad \text{and} \quad \left|y_i - \frac{\exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}{1 + \exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}\right| \leq 1.$$

A.5 Proof of Lemma 22

Denote the singular value decomposition of $\Theta^* L^{1/2}$ by $\Theta^* L^{1/2} = U\Sigma V^T$, where $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_2 \times d_2}$ are orthogonal matrices. For a given $r \in [\min\{d_1, d_2 - G\}]$, let $U_r = [u_1, \ldots, u_r]$ and $V_r = [v_1, \ldots, v_r]$, where $u_i \in \mathbb{R}^{d_1 \times 1}$ and $v_i \in \mathbb{R}^{d_2 \times 1}$ are the left and right singular vectors corresponding to the i-th largest singular value, respectively. Define T to be the subspace spanned by all matrices in $\mathbb{R}^{d_1 \times d_2}$ of the form $U_r A^T$ or $B V_r^T$ for any $A \in \mathbb{R}^{d_2 \times r}$ or $B \in \mathbb{R}^{d_1 \times r}$, respectively. The orthogonal projection of any matrix $M \in \mathbb{R}^{d_1 \times d_2}$ onto the space T is given by $\mathcal{P}_T(M) = U_r U_r^T M + M V_r V_r^T - U_r U_r^T M V_r V_r^T$. The projection of M onto the complement space $T^\perp$ is $\mathcal{P}_{T^\perp}(M) = (I - U_r U_r^T)M(I - V_r V_r^T)$. The subspace T and the respective projections onto T and $T^\perp$ play crucial a role in the analysis of nuclear norm minimization, since they define the sub-gradient of the nuclear norm at Θ^*. We refer to Candès and Recht (2009) for more detailed treatment of this topic.

Let $\Delta' = \mathcal{P}_T(\Delta L^{1/2})$ and $\Delta'' = \mathcal{P}_{T^\perp}(\Delta L^{1/2})$. Notice that $\mathcal{P}_T(\Theta^* L^{1/2}) = U_r \Sigma_r V_r^T$, where $\Sigma_r \in \mathbb{R}^{r \times r}$ is the diagonal matrix formed by the top r singular values. Since $\mathcal{P}_T(\Theta^* L^{1/2})$ and Δ'' have row and column spaces that are orthogonal, it follows from Lemma 2.3 in Recht et al. (2010) that

$$\left\|\!\left\|\mathcal{P}_T(\Theta^* L^{1/2}) - \Delta''\right\|\!\right\|_{\text{nuc}} = \left\|\!\left\|\mathcal{P}_T(\Theta^* L^{1/2})\right\|\!\right\|_{\text{nuc}} + \left\|\!\left\|\Delta''\right\|\!\right\|_{\text{nuc}}.$$

Hence, in view of the triangle inequality,

$$
\begin{aligned}
\left\|\left|\widehat{\Theta} L^{1/2}\right|\right\|_{\text{nuc}} &= \left\|\left|\mathcal{P}_T(\Theta^* L^{1/2}) + \mathcal{P}_{T^\perp}(\Theta^* L^{1/2}) - \Delta' - \Delta''\right|\right\|_{\text{nuc}} \\
&\geq \left\|\left|\mathcal{P}_T(\Theta^* L^{1/2}) - \Delta''\right|\right\|_{\text{nuc}} - \left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2}) - \Delta'\right|\right\|_{\text{nuc}} \\
&= \left\|\left|\mathcal{P}_T(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}} + \left\|\left|\Delta''\right|\right\|_{\text{nuc}} - \left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2}) - \Delta'\right|\right\|_{\text{nuc}} \\
&\geq \left\|\left|\mathcal{P}_T(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}} + \left\|\left|\Delta''\right|\right\|_{\text{nuc}} - \left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}} - \left\|\left|\Delta'\right|\right\|_{\text{nuc}} \\
&= \left\|\left|\Theta^* L^{1/2}\right|\right\|_{\text{nuc}} + \left\|\left|\Delta''\right|\right\|_{\text{nuc}} - 2\left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}} - \left\|\left|\Delta'\right|\right\|_{\text{nuc}}. \quad (87)
\end{aligned}
$$

Because $\widehat{\Theta}$ is an optimal solution, we have

$$
\begin{aligned}
\lambda\left(\left\|\left|\widehat{\Theta} L^{1/2}\right|\right\|_{\text{nuc}} - \left\|\left|\Theta^* L^{1/2}\right|\right\|_{\text{nuc}}\right) &\leq -\mathcal{L}(\Theta^*) + \mathcal{L}(\widehat{\Theta}) \\
&\overset{(a)}{\leq} \langle\!\langle \Delta L^{1/2}, \nabla\mathcal{L}(\Theta^*) L^{-1/2}\rangle\!\rangle \\
&\overset{(b)}{\leq} \left\|\left|\Delta\right|\right\|_{\text{L-nuc}}\left\|\left|\nabla\mathcal{L}(\Theta^*) L^{-1/2}\right|\right\|_2 \leq \frac{\lambda}{2}\left\|\left|\Delta\right|\right\|_{\text{L-nuc}}, \quad (88)
\end{aligned}
$$

where (a) holds due to the convexity of $-\mathcal{L}$; (b) follows from the Cauchy-Schwarz inequality; the last inequality holds due to the assumption that $\lambda \geq 2\|\!|\nabla\mathcal{L}(\Theta^*)|\!\|_2$. Combining (87) and (88) yields

$$
2\left(\left\|\left|\Delta''\right|\right\|_{\text{nuc}} - 2\left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}} - \left\|\left|\Delta'\right|\right\|_{\text{nuc}}\right) \leq \left\|\left|\Delta\right|\right\|_{\text{L-nuc}} \leq \left\|\left|\Delta'\right|\right\|_{\text{nuc}} + \left\|\left|\Delta''\right|\right\|_{\text{nuc}}.
$$

Thus $\|\!|\Delta''|\!\|_{\text{nuc}} \leq 3\|\!|\Delta'|\!\|_{\text{L-nuc}} + 4\left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}}$. By triangle inequality,

$$
\left\|\left|\Delta\right|\right\|_{\text{L-nuc}} \leq 4\left\|\left|\Delta'\right|\right\|_{\text{nuc}} + 4\left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}}.
$$

Notice that $\Delta' = U_r U_r^T \Delta L^{1/2} + (I - U_r U_r^T)\Delta L^{1/2} V_r V_r^T$. Both $U_r U_r^T \Delta L^{1/2}$ and $(I - U_r U_r^T)\Delta L^{1/2} V_r V_r^T$ have rank at most r. Thus Δ' has rank at most $2r$. Hence, $\|\!|\Delta'|\!\|_{\text{nuc}} \leq \sqrt{2r}\|\!|\Delta'|\!\|_{\text{F}} \leq \sqrt{2r}\|\!|\Delta L^{1/2}|\!\|_{\text{F}} \leq \sqrt{2r}\|\!|\Delta|\!\|_{\text{L}}$. Then the theorem follows because $\left\|\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\right\|_{\text{nuc}} = \sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^* L^{1/2})$.

Appendix B. Proof of the Information-theoretic Graph Sampling Lower Bound, Theorem 4

The proof uses Fano Inequality based packing set argument to get an lower bound on the error of any (measurable) estimator. We will construct a packing set in Ω_α with a minimum distance of δ between any pair of elements in the packing.

Let $\{\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(M)}\}$ be a set of M matrices within the set Ω_α, satisfying $\|\!|\Theta^{(\ell_1)} - \Theta^{(\ell_1)}|\!\|_{\text{L}} \geq \delta$ for all $\ell_1, \ell_2 \in [M]$. Now, $\Theta^{(N)}$ is uniformly drawn from this set and then the comparison results (according to MNL model) of n randomly chosen pairs of items,

each drawn according to the probability matrix P and each compared by uniformly chosen user. Let $\widehat{N}$ be the best estimator of N from the observations. Then we can show that,

$$\sup_{\Theta^* \in \Omega_\alpha} \mathbb{P}\left\{\left\|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right\|_{\mathrm{L}}^2 \geq \frac{\delta^2}{2}\right\} \;\geq\; \mathbb{P}\left\{\widehat{N} \neq N\right\}, \tag{89}$$

Now we have converted the problem of finding the minimum estimation error, into finding the minimum probability error of a M-ary hypothesis testing problem. If we can prove that the above RHS is lower bounded by $1/2$, we are done.

The generalized Fanos inequality along with data processing inequality gives us,

$$\mathbb{P}\left\{\widehat{N} \neq N\right\} \;\geq\; 1 - \frac{\mathbb{E}[\,I(\widehat{N}; N)\,] + \log 2}{\log M} \tag{90}$$

$$\geq\; 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1, \ell_2 \in [M]} D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) + \log 2}{\log M}, \tag{91}$$

where $D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$ denotes the *expected* Kullback-Leibler divergence between the probability distributions of the comparison results of the observed nd_1 pairs, for $N = \ell_1$ and $N = \ell_2$. The expectation is taken over different choices for the selected pairs for comparison.

$$D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) = \frac{n}{d_1} \sum_{\substack{i \in [d_1] \\ \{j,j'\} \subset [d_2]}} 2P_{j,j'} \left[\frac{e^{\Theta_{ij}^{(\ell_1)}}}{e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}}} \log \left(\frac{e^{\Theta_{ij}^{(\ell_1)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}}\right)}{e^{\Theta_{ij}^{(\ell_2)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_2)}} + e^{\Theta_{ij'}^{(\ell_2)}}\right)} \right) \right.$$
$$\tag{92}$$
$$\left. + \frac{e^{\Theta_{ij'}^{(\ell_1)}}}{e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}}} \log \left(\frac{e^{\Theta_{ij'}^{(\ell_1)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}}\right)}{e^{\Theta_{ij'}^{(\ell_2)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_2)}} + e^{\Theta_{ij'}^{(\ell_2)}}\right)} \right) \right]$$
$$\tag{93}$$

where n is the number of pairs of items selected and compared by one random user each, $P_{j,j'}$ is half the probability with which item pair $\{j, j'\}$ is selected and the observation probabilities come from the standard MNL model. Let $x_{ijj'} \equiv e^{\Theta_{ij'}^{(\ell_1)}} / (e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}})$ and $y_{ijj'} \equiv e^{\Theta_{ij'}^{(\ell_2)}} / (e^{\Theta_{ij}^{(\ell_2)}} + e^{\Theta_{ij'}^{(\ell_2)}})$.

$$D_{\mathrm{KL}}(\Theta^{(\ell_1)}\|\Theta^{(\ell_2)}) \overset{(a)}{=} n \sum_{i\in[d_1]} \frac{1}{d_1} \sum_{\{j,j'\}\subset[d_2]} 2P_{j,j'}\left[x_{ijj'}\log\frac{x_{ijj'}}{y_{ijj'}} + (1-x_{ijj'})\log\frac{1-x_{ijj'}}{1-y_{ijj'}}\right] \tag{94}$$

$$\overset{(a)}{\leq} n \sum_{i\in[d_1]} \frac{1}{d_1} \sum_{\{j,j'\}\subset[d_2]} 2P_{j,j'}\left[x_{ijj'}\frac{x_{ijj'}-y_{ijj'}}{y_{ijj'}} + (1-x_{ijj'})\frac{y_{ijj'}-x_{ijj'}}{1-y_{ijj'}}\right] \tag{95}$$

$$= 2n \sum_{i\in[d_1]} \frac{1}{d_1} \sum_{\{j,j'\}\subset[d_2]} \frac{(x_{ijj'}-y_{ijj'})P_{j,j'}(x_{ijj'}-y_{ijj'})}{y_{ijj'}(1-y_{ijj'})} \tag{96}$$

$$\overset{(b)}{\leq} 8ne^{2\alpha} \sum_{i\in[d_1]} \frac{1}{d_1} \sum_{\{j,j'\}\subset[d_2]} (x_{ijj'}-y_{ijj'})P_{j,j'}(x_{ijj'}-y_{ijj'}), \tag{97}$$

where (a) is due to the fact that $\log(x/y) \leq (x-y)/y$ for $x/y \geq 0$ and (b) is true because $|\Theta_{ij}^{(\ell_2)}| \leq \alpha$ implies, $y_{ijj'} = e^{\Theta_{ij}^{(\ell_2)}}/(e^{\Theta_{ij}^{(\ell_2)}} + e^{\Theta_{ij'}^{(\ell_2)}}) \geq e^{-2\alpha}/2$ which in turn implies, $y_{ijj'}(1-y_{ijj'}) \geq e^{-2\alpha}(2-e^{-2\alpha})/4 \geq e^{-2\alpha}/4$. Let $f(z) = 1/(1+e^{-z})$, a 1-Lipschitz function, it can be seen that $(x_{ijj'}-y_{ijj'})^2 = (f(\Theta_{ij}^{(\ell_1)}-\Theta_{ij'}^{(\ell_1)}) - f(\Theta_{ij}^{(\ell_2)}-\Theta_{ij'}^{(\ell_2)}))^2 \leq ((\Theta_{ij}^{(\ell_1)}-\Theta_{ij'}^{(\ell_1)}) - (\Theta_{ij}^{(\ell_2)}-\Theta_{ij'}^{(\ell_2)}))^2$. This gives us,

$$D_{\mathrm{KL}}(\Theta^{(\ell_1)}\|\Theta^{(\ell_2)}) \leq \frac{8ne^{2\alpha}}{d_1} \sum_{i\in[d_1]} \sum_{\{j,j'\}\subset[d_2]} P_{u,v}((\Theta_{ij}^{(\ell_1)}-\Theta_{ij}^{(\ell_2)}) - (\Theta_{ij'}^{(\ell_1)}-\Theta_{ij'}^{(\ell_2)}))^2, \tag{98}$$

$$\overset{(a)}{\leq} \frac{8ne^{2\alpha}}{d_1} \sum_{i\in[d_1]} (\Theta^{(\ell_1)}-\Theta^{(\ell_2)})_i L(\Theta^{(\ell_1)}-\Theta^{(\ell_2)})_i, \tag{99}$$

$$= \frac{8ne^{2\alpha}}{d_1} \sum_{i\in[d_1]} (\Theta^{(\ell_1)}-\Theta^{(\ell_2)})_i L(\Theta^{(\ell_1)}-\Theta^{(\ell_2)})_i, \tag{100}$$

$$= \frac{8ne^{2\alpha}}{d_1} \left\|\left|(\Theta^{(\ell_1)}-\Theta^{(\ell_2)})L^{1/2}\right|\right\|_{\mathrm{F}}^2 \tag{101}$$

$$= \frac{8ne^{2\alpha}}{d_1} \left\|\left|\Theta^{(\ell_1)}-\Theta^{(\ell_2)}\right|\right\|_{\mathrm{L}}^2 \tag{102}$$

$$\tag{103}$$

where (a) is due to the fact that $L = \mathrm{diag}(P_u) - P$ is the Laplacian of the probability matrix P, and Θ_i denotes the i-th row of matrix Θ. Combining the above with (91), we get,

$$\mathbb{P}\left\{\widehat{N} \neq N\right\} \geq 1 - \frac{\binom{M}{2}^{-1}\sum_{\ell_1,\ell_2\in[M]}(8ne^{2\alpha}/d_1)\left\|\left|\Theta^{(\ell_1)}-\Theta^{(\ell_2)}\right|\right\|_{\mathrm{L}}^2 + \log 2}{\log M}. \tag{104}$$

The remainder of the proof relies on the following probabilistic packing.

Lemma 25 *For each $r \in \{1, \ldots, d_1\}$, and for any positive $\delta > 0$ there exists a family of $d_1 \times d_2$ dimensional matrices $\{\Theta^{(1)}, \ldots, \Theta^{(M(\delta))}\}$ with cardinality $M(\delta) = \lfloor \exp(rd_1/256) \rfloor$ such that each matrix is rank r and the following bounds hold:*

$$\left\| \left\| \Theta^{(\ell)} \right\| \right\|_{\mathrm{L}} \leq \delta \,, \text{ for all } \ell \in [M] \tag{105}$$

$$\left\| \left\| \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \right\| \right\|_{\mathrm{L}} \geq \delta \,, \text{ for all } \ell_1, \ell_2 \in [M] \tag{106}$$

$$\Theta^{(\ell)} \in \Omega_{\tilde{\alpha}} \,, \text{ for all } \ell \in [M] \,, \tag{107}$$

with $\tilde{\alpha} = (8\delta/d_2)\sqrt{2 \log d}$ for $d = (d_1 + d_2)/2$.

Now if we assume $\delta \leq \alpha d_2 / 8\sqrt{2 \log d}$, we get $\Theta^{(\ell)} \in \Omega_\alpha$ for $\ell \in [M]$. The above lemma also implies that $\left\| \left\| \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \right\| \right\|_{\mathrm{F}}^2 \leq 4\delta^2$ which implies,

$$\mathbb{P}\left\{ \widehat{N} \neq N \right\} \geq 1 - \frac{32 n e^{2\alpha} \delta^2 / d_1 + \log 2}{rd_1/256} \geq \frac{1}{2} \,, \tag{108}$$

where the last inequality holds when $\delta \leq (e^{-\alpha}/128)\sqrt{rd_1^2/n}$. Along with (89), this proves that,

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\left[\left\| \left\| \widehat{\Theta} - \Theta^* \right\| \right\|_{\mathrm{L}} \right] \geq \frac{\delta}{2} \,, \tag{109}$$

for all $\delta \leq \min\{\alpha d_2/8\sqrt{2 \log d}, (e^{-\alpha}/128)\sqrt{rd_1^2/n}\}$.

Similarly using the following lemma, Lemma 26, we can prove that,

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\left[\left\| \left\| \widehat{\Theta} - \Theta^* \right\| \right\|_{\mathrm{L}} \right] \geq \frac{\delta}{2} \,, \tag{110}$$

for all $\delta \leq \min\{\alpha\sqrt{rd_1}/\mathrm{tr}\left(\sqrt{(L_r)^\dagger}\right), (e^{-\alpha}/128)\sqrt{rd_1^2/n}\}$.

Lemma 26 *For each $r \in \{1, \ldots, d_1\}$, and for any positive $\delta > 0$ there exists a family of $d_1 \times d_2$ dimensional matrices $\{\Theta^{(1)}, \ldots, \Theta^{(M(\delta))}\}$ with cardinality $M(\delta) = \lfloor \exp(rd_1/256) \rfloor$ such that each matrix is rank r and the following bounds hold:*

$$\left\| \left\| \Theta^{(\ell)} \right\| \right\|_{\mathrm{L}} \leq \delta \,, \text{ for all } \ell \in [M] \tag{111}$$

$$\left\| \left\| \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \right\| \right\|_{\mathrm{L}} \geq \delta \,, \text{ for all } \ell_1, \ell_2 \in [M] \tag{112}$$

$$\Theta^{(\ell)} \in \Omega_{\tilde{\alpha}} \,, \text{ for all } \ell \in [M] \,, \tag{113}$$

with $\tilde{\alpha} = \delta\sqrt{\mathrm{tr}\left((L_r)^\dagger\right)}/\sqrt{rd_1}$, where L_r is the (best) rank r approximation of L.

Now combining (109) and (110), and maximizing the RHS proves the theorem.

B.1 Proof of Lemma 25

Following the construction in Negahban and Wainwright (2012), we use probabilistic method to prove the existence of the desired family. We will show that the following procedure succeeds in producing the desired family with probability at least half, which proves its existence. Let $d = (d_1 + d_2)/2$, and suppose $d_2 \geq d_1$ without loss of generality. For the choice of $M' = e^{rd_2/576}$, and for each $\ell \in [M']$, generate a rank-r matrix $\Theta^{(\ell)} \in \mathbb{R}^{d_1 \times d_2}$ as follows:

$$\Theta^{(\ell)} \; = \; \frac{\delta}{\sqrt{rd_2}} U(V^{(\ell)})^T \left(\mathbf{I}_{d_2 \times d_2} - \sum_{i \in [G]} \frac{g_i g_i^T}{g_i^T g_i} \right), \tag{114}$$

where $U \in \mathbb{R}^{d_1 \times r}$ is a random orthogonal basis such that $U^T U = \mathbf{I}_{r \times r}$ and $V^{(\ell)} \in \mathbb{R}^{d_2 \times r}$ is a random matrix with each entry $V_{ij}^{(\ell)} \in \{-1, +1\}$ chosen independently and uniformly at random. By construction, notice that,

$$\left\vert\!\left\vert\!\left\vert \Theta^{(\ell)} \right\vert\!\right\vert\!\right\vert_{\mathrm{L}}^2 = \frac{\delta^2}{r\, d_2} \left\vert\!\left\vert\!\left\vert (V^{(\ell)})^T L^{1/2} \right\vert\!\right\vert\!\right\vert_{\mathrm{F}}^2 = \frac{\delta^2}{r\, d_2} \sum_{i \in [d_2]} (V_i^{(\ell)})^T L V_i^{(\ell)} \leq \frac{\delta^2}{r\, d_2} \|L\|_2 \sum_{i \in [d_2]} \left\vert\!\left\vert\!\left\vert V_i^{(\ell)} \right\vert\!\right\vert\!\right\vert_{\mathrm{F}}^2 = \delta^2, \tag{115}$$

where $V_i^{(\ell)}$ is the i-th column of $V^{(\ell)}$, since $g_{i=1}^G$ span the null space of the Laplacian L, $\|L\|_2 \leq 1$, and $\left\vert\!\left\vert\!\left\vert V^{(\ell)} \right\vert\!\right\vert\!\right\vert_{\mathrm{F}} = \sqrt{rd_2}$. Now, consider $\left\vert\!\left\vert\!\left\vert \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \right\vert\!\right\vert\!\right\vert_{\mathrm{F}}^2 = (\delta^2/(rd_2)) \left\vert\!\left\vert\!\left\vert (V^{(\ell_1)} - V^{(\ell_2)})^T \right\vert\!\right\vert\!\right\vert_{\mathrm{L}}^2 \equiv f(V^{(\ell_1)}, V^{(\ell_2)})$ which is a function over $2rd_2$ i.i.d. random Rademacher variables $V^{(\ell_1)}$ and $V^{(\ell_2)}$ which define $\Theta^{(\ell_1)}$ and $\Theta^{(\ell_2)}$ respectively. Since f is Lipschitz in the following sense, we can apply McDiarmid's concentration inequality. For all $(V^{(\ell_1)}, V^{(\ell_2)})$ and $(\widetilde{V}^{(\ell_1)}, \widetilde{V}^{(\ell_2)})$ that differ in only one variable, say $\widetilde{V}^{(\ell_1)} = V^{(\ell_1)} + 2e_{ij}$, for some standard basis matrix e_{ij}, we have

$$\left\vert f(V^{(\ell_1)}, V^{(\ell_2)}) - f(\widetilde{V}^{(\ell_1)}, \widetilde{V}^{(\ell_2)}) \right\vert =$$
$$\left\vert \frac{\delta^2}{r\, d_2} \left\vert\!\left\vert\!\left\vert (V^{(\ell_1)} - V^{(\ell_2)})^T \right\vert\!\right\vert\!\right\vert_{\mathrm{L}}^2 - \frac{\delta^2}{r\, d_2} \left\vert\!\left\vert\!\left\vert (V^{(\ell_1)} - V^{(\ell_2)} + 2e_{ij})^T \right\vert\!\right\vert\!\right\vert_{\mathrm{L}}^2 \right\vert \tag{116}$$

$$= \frac{\delta^2}{r\, d_2} \left\vert 2e_i^T L(V_j^{(\ell_1)} - V_j^{(\ell_2)}) + 4e_i^T L e_i \right\vert \tag{117}$$

$$= \frac{\delta^2}{r\, d_2} \left(2\|L_i\|_1 \left\vert\!\left\vert\!\left\vert V_j^{(\ell_1)} - V_j^{(\ell_2)} \right\vert\!\right\vert\!\right\vert_{\infty} + 4P_i \right) \tag{118}$$

$$\leq \frac{12\, \delta^2}{r\, d_2}, \tag{119}$$

where we used the fact that $\|L_i\|_1 = 2P_i \leq 2$ and $V^{(\ell_1)} - V^{(\ell_2)}$ is entry-wise bounded by 2. The expectation $\mathbb{E}[f(V^{(\ell_1)}, V^{(\ell_2)})]$ is

$$\frac{\delta^2}{r\, d_2} \mathbb{E}\left[\left\vert\!\left\vert\!\left\vert (V^{(\ell_1)} - V^{(\ell_2)})^T \right\vert\!\right\vert\!\right\vert_{\mathrm{L}}^2 \right] = \frac{2\delta^2}{r\, d_2} \mathbb{E}\left[\left\vert\!\left\vert\!\left\vert (V^{(\ell_1)})^T \right\vert\!\right\vert\!\right\vert_{\mathrm{L}}^2 \right] \leq 2\delta^2, \tag{120}$$

where we use equation (115). Applying McDiarmid's inequality with bounded difference $12\delta^2/(rd_2)$, we get that

$$\mathbb{P}\left\{ f(V^{(\ell_1)}, V^{(\ell_2)}) \leq 2\delta^2 - t \right\} \; \leq \; \exp\left\{ -\frac{t^2\, r\, d_2}{144\, \delta^4} \right\}, \tag{121}$$

Since there are less than $(M')^2$ pairs of (ℓ_1, ℓ_2), setting $t = \delta^2$ and applying the union bound gives

$$\mathbb{P}\left\{ \min_{\ell_1, \ell_2 \in [M']} \left\|\!\left\|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right\|\!\right\|_{\mathrm{F}}^2 \geq \delta^2 \right\} \geq 1 - \exp\left\{ -\frac{r\,d_2}{144} + 2\log M' \right\} \geq \frac{7}{8}, \quad (122)$$

where we used $M' = \exp\{rd_2/576\}$ and $d_2 \geq 607$.

We are left to prove that $\Theta^{(\ell)}$'s are in $\Omega_{(8\delta/d_2)\sqrt{2\log d_2}}$ as defined in (10). Since we removed the mean for each connected component, such that $\Theta^{(\ell)}g_i = 0$, $\forall i \in [G]$ by construction, we only need to show that the maximum entry is bounded by $(8\delta/d_2)\sqrt{2\log d_2}$. We first prove an upper bound in (124) for a fixed $\ell \in [M']$, and use this to show that there exists a large enough subset of matrices satisfying this bound. From (114), consider $(UV^T)_{ij} = \langle\!\langle u_i, v_j \rangle\!\rangle$, where $u_i \in \mathbb{R}^r$ is the first r entries of a random vector drawn uniformly from the d_2-dimensional sphere, and $v_j \in \mathbb{R}^r$ is drawn uniformly at random from $\{-1, +1\}^r$ with $\|v_j\| = \sqrt{r}$. Using Levy's theorem for concentration on the sphere Ledoux (2005), we have

$$\mathbb{P}\left\{|\langle\!\langle u_i, v_j \rangle\!\rangle| \geq t\right\} \leq 2\exp\left\{ -\frac{d_2\,t^2}{8\,r} \right\}. \quad (123)$$

Notice that by the definition (114), $\max_{i,j} |\Theta_{ij}^{(\ell)}| \leq (2\delta/\sqrt{rd_2}) \max_{i,j} |\langle\!\langle u_i, v_j \rangle\!\rangle|$. Setting $t = \sqrt{(32r/d_2)\log d_2}$ and taking the union bound over all $d_1 d_2$ indices, we get

$$\mathbb{P}\left\{ \max_{i,j} |\Theta_{ij}^{(\ell)}| \leq \frac{2\delta\sqrt{32\log d_2}}{d_2} \right\} \geq 1 - 2d_1 d_2 \exp\left\{ -4\log d_2 \right\} \geq \frac{1}{2}, \quad (124)$$

for a fixed $\ell \in [M']$. Consider the event that there exists a subset $S \subset [M']$ of cardinality $M = (1/4)M'$ with the same bound on maximum entry, then from (124) we get

$$\mathbb{P}\left\{ \exists S \subset [M'] \text{ such that } \left\|\!\left\|\Theta^{(\ell)}\right\|\!\right\|_\infty \leq \frac{2\delta\sqrt{32\log d_2}}{d_2} \text{ for all } \ell \in S \right\} \geq \sum_{m=M}^{M'} \binom{M'}{m}\left(\frac{1}{2}\right)^m, \quad (125)$$

which is larger than half for our choice of $M < M'/2$.

B.2 Proof of Lemma 26

Inspired from the construction in Negahban and Wainwright (2012), we furnish the following probabilistic argument for the existence of the desired family. For the choice of $M = \lfloor e^{rd_1/256} \rfloor$, and for each $\ell \in [M]$, generate a rank-r matrix $\Theta^{(\ell)} \in \mathbb{R}^{d_1 \times d_2}$ as follows:

$$\Theta^{(\ell)} = \frac{\delta}{\sqrt{rd_1}} V^{(\ell)} \sqrt{\Lambda_r^\dagger} U_r^T, \quad (126)$$

where the columns of $U_r \in \mathbb{R}^{d_2 \times r}$ are the top r singular vectors of $L = U\Lambda U^T$, Λ_r is a diagonal matrix in $\mathbb{R}^{r \times r}$ and its diagonal elements are the top r singular values of L corresponding to columns of U_r, $\dagger$ represents the Moore-Penrose pseudo inverse, and $V^{(\ell)}$ is a random matrix with each entry $V_{ij}^{(\ell)} \in \{-1, +1\}$ chosen independently and uniformly at random. First by definition, $\left\|\!\left\|\Theta^{(\ell)}\right\|\!\right\|_{\mathrm{L}} = (\delta/\sqrt{rd_1})\left\|\!\left\|V^{(\ell)}\right\|\!\right\|_{\mathrm{F}} \leq \delta$, since $\left\|\!\left\|V^{(\ell)}\right\|\!\right\|_{\mathrm{F}} = \sqrt{rd_1}$.

Define f as $f(V^{(\ell_1)}, V^{(\ell_2)}) \equiv \left\vert\left\vert\left\vert \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \right\vert\right\vert\right\vert_{\mathrm{L}}^2 = (\delta^2/(rd_1)) \left\vert\left\vert\left\vert V^{(\ell_1)} - V^{(\ell_2)} \right\vert\right\vert\right\vert_{\mathrm{F}}^2$ which is a function of $2rd_1$ i.i.d. random Rademacher variables. Now we can apply McDiarmid's concentration inequality since f is Lipschitz as folows. For all $(V^{(\ell_1)}, V^{(\ell_2)})$ and $(\widetilde{V}^{(\ell_1)}, \widetilde{V}^{(\ell_2)})$ that differ in only one variable, say $\widetilde{V}^{(\ell_1)} = V^{(\ell_1)} + 2e_{ij}$, for some standard basis matrix e_{ij}, we have

$$
\begin{aligned}
& \left\vert f(V^{(\ell_1)}, V^{(\ell_2)}) - f(\widetilde{V}^{(\ell_1)}, \widetilde{V}^{(\ell_2)}) \right\vert \\
&= \left\vert \frac{\delta^2}{r\,d_2} \left\vert\left\vert\left\vert V^{(\ell_1)} - V^{(\ell_2)} \right\vert\right\vert\right\vert_{\mathrm{F}}^2 - \frac{\delta^2}{r\,d_2} \left\vert\left\vert\left\vert V^{(\ell_1)} - V^{(\ell_2)} + 2e_{ij} \right\vert\right\vert\right\vert_{\mathrm{F}}^2 \right\vert \\
&= \left\vert \frac{\delta^2}{r\,d_2} \left\vert\left\vert\left\vert 2e_{ij} \right\vert\right\vert\right\vert_{\mathrm{F}}^2 + \frac{\delta^2}{r\,d_2} \langle\!\langle (V^{(\ell_1)} - V^{(\ell_2)}), 2e_{ij} \rangle\!\rangle \right\vert \\
&\leq \frac{4\,\delta^2}{r\,d_1} + \frac{\delta^2}{r\,d_1} \left\vert\left\vert\left\vert V^{(\ell_1)} - V^{(\ell_2)} \right\vert\right\vert\right\vert_\infty \left\vert\left\vert\left\vert 2e_{ij} \right\vert\right\vert\right\vert_1 \\
&\leq \frac{8\,\delta^2}{r\,d_1} \; ,
\end{aligned}
\tag{127}
$$

where the penultimate step is true since $(V^{(\ell_1)} - V^{(\ell_2)})$ is entry-wise bounded by 2. The expectation $\mathbb{E}[f(V^{(\ell_1)}, V^{(\ell_2)})]$ is

$$
\begin{aligned}
\frac{\delta^2}{r\,d_1} \mathbb{E} \left[\left\vert\left\vert\left\vert (V^{(\ell_1)} - V^{(\ell_2)}) \right\vert\right\vert\right\vert_{\mathrm{F}}^2 \right] &= \frac{2\delta^2}{r\,d_1} \mathbb{E} \left[\left\vert\left\vert\left\vert V^{(\ell_1)} \right\vert\right\vert\right\vert_{\mathrm{F}}^2 \right] \\
&= 2\,\delta^2 \; .
\end{aligned}
\tag{128}
$$

Now applying McDiarmid's inequality on the function f, we get that

$$
\mathbb{P} \left\{ f(V^{(\ell_1)}, V^{(\ell_2)}) \leq 2\delta^2 - t \right\} \leq \exp \left\{ -\frac{t^2\,r\,d_1}{64\,\delta^4} \right\} ,
\tag{129}
$$

Setting $t = \delta^2$ and applying the union bound gives us,

$$
\mathbb{P} \left\{ \min_{\ell_1, \ell_2 \in [M]} \left\vert\left\vert\left\vert \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \right\vert\right\vert\right\vert_{\mathrm{F}}^2 \geq \delta^2 \right\} \geq 1 - \exp \left\{ -\frac{r\,d_1}{64} + 2\log M \right\} > 0 \; .
\tag{130}
$$

In the last step, we used $M = \lfloor \exp\{rd_1/\,256\} \rfloor$. At last we prove that $\Theta^{(\ell)}$'s are in $\Omega_{\delta\sqrt{\mathrm{tr}((L_r)^\dagger)}/rd_1}$ as defined in (10). Since we know that g_i belongs to the kernel of L for all $i \in [G]$, $\Theta^{(\ell)} g = 0$ by construction (6). From (126), consider $(V\sqrt{\Lambda_r^\dagger} U_r^T)_{ij} = \langle\!\langle v_i, \sqrt{\Lambda_r^\dagger}(u_r)_j \rangle\!\rangle$, where $(u_r)_j \in \mathbb{R}^r$ is the vector of i-th entries of the top r singular vectors of L, and $v_i \in \mathbb{R}^r$ is drawn uniformly at random from $\{-1, +1\}^r$.

$$
\left\vert \langle\!\langle v_i, \sqrt{\Lambda_r^\dagger}(u_r)_j \rangle\!\rangle \right\vert \leq \left\vert\left\vert\left\vert v_i \right\vert\right\vert\right\vert_\infty \left\vert\left\vert\left\vert \sqrt{\Lambda_r^\dagger}(u_r)_j \right\vert\right\vert\right\vert_1 \leq \sqrt{\mathrm{tr}\left(\Lambda_r^\dagger\right)} = \sqrt{\mathrm{tr}\left((L_r)^\dagger\right)} \; .
\tag{131}
$$

The above inequality proves that $\left\vert\left\vert\left\vert \Theta^{(\ell)} \right\vert\right\vert\right\vert_\infty$ is upper bounded as desired.

Appendix C. Proof of Theorem 7

We first introduce some additional notations used in the proof. Recall that $\mathcal{L}(\Theta)$ is the log likelihood function. Let $\nabla\mathcal{L}(\Theta) \in \mathbb{R}^{d_1 \times d_2}$ denote its gradient such that $\nabla_{ij}\mathcal{L}(\Theta) = \frac{\partial\mathcal{L}(\Theta)}{\partial\Theta_{ij}}$. Let $\nabla^2\mathcal{L}(\Theta) \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ denote its Hessian matrix such that $\nabla^2_{ij,i'j'}\mathcal{L}(\Theta) = \frac{\partial^2\mathcal{L}(\Theta)}{\partial\Theta_{ij}\partial\Theta_{i'j'}}$. By the definition of $\mathcal{L}(\Theta)$ in (32), we have

$$\nabla\mathcal{L}(\Theta^*) = -\frac{1}{k\,d_1}\sum_{i=1}^{d_1}\sum_{\ell=1}^{k} e_i(e_{v_{i,\ell}} - p_{i,\ell})^T \,, \tag{132}$$

where $p_{i,\ell}$ denotes the conditional choice probability at ℓ-th position. Precisely, $p_{i,\ell} = \sum_{j\in S_{i,\ell}} p_{j|(i,\ell)}e_j$ where $p_{j|(i,\ell)}$ is the probability that item j is chosen at ℓ-th position from the top by the user i conditioned on the top $\ell - 1$ choices such that $p_{j|(i,\ell)} \equiv \mathbb{P}\{v_{i,\ell} = j | v_{i,1}, \ldots, v_{i,\ell-1}, S_i\} = e^{\Theta^*_{ij}}/(\sum_{j'\in S_{i,\ell}} e^{\Theta_{ij'}})$ and $S_{i,\ell} \equiv S_i \setminus \{v_{i,1}, \ldots, v_{i,\ell-1}\}$, where S_i is the set of alternatives presented to the i-th user and $v_{i,\ell}$ is the item ranked at the ℓ-th position by the user i. Notice that for $i \neq i'$, $\frac{\partial^2\mathcal{L}(\Theta)}{\partial\Theta_{ij}\partial\Theta_{i'j'}} = 0$ and the Hessian is

$$\frac{\partial^2\mathcal{L}(\Theta)}{\partial\Theta_{ij}\partial\Theta_{ij'}} = \frac{1}{k\,d_1}\sum_{\ell=1}^{k}\mathbb{I}(j\in S_{i,\ell})\frac{\partial p_{j|(i,\ell)}}{\partial\Theta_{ij'}}$$

$$= \frac{1}{k\,d_1}\sum_{\ell=1}^{k}\mathbb{I}(j,j'\in S_{i,\ell})\left(p_{j|(i,\ell)}\mathbb{I}(j=j') - p_{j|(i,\ell)}p_{j'|(i,\ell)}\right). \tag{133}$$

This Hessian matrix is a block-diagonal matrix $\nabla^2\mathcal{L}(\Theta) = \mathrm{diag}(H^{(1)}(\Theta), \ldots, H^{(d_1)}(\Theta))$ with

$$H^{(i)}(\Theta) = \frac{1}{k\,d_1}\sum_{\ell=1}^{k}\left(\mathrm{diag}(p_{i,\ell}) - p_{i,\ell}p_{i,\ell}^T\right). \tag{134}$$

Let $\Delta = \Theta^* - \widehat{\Theta}$ where $\widehat{\Theta}$ is the optimal solution of the convex program in (31). We first introduce three key technical lemmas. The first lemma follows from Lemma 1 of Negahban and Wainwright (2012), and shows that Δ is approximately low-rank.

Lemma 27 *If $\lambda \geq 2\|\|\nabla\mathcal{L}(\Theta^*)\|\|_2$, then we have*

$$\|\Delta\|_{\mathrm{nuc}} \leq 4\sqrt{2r}\|\Delta\|_{\mathrm{F}} + 4\sum_{j=r+1}^{\min\{d_1,d_2\}}\sigma_j(\Theta^*)\,, \tag{135}$$

for all $r \in [\min\{d_1, d_2\}]$.

Proof of the above lemma is omitted because of its similarity to that of Lemma 22. The following lemma provides a bound on the gradient using the concentration in measure of sum of independent random matrices Tropp (2011).

Lemma 28 *For any positive constant* $c \geq 1$ *and* $k \leq (1/e)\ d_2(4\log d_2 + \log d_1)$, *with probability at least* $1 - 2d^{-c} - d_2^{-3}$,

$$\|\|\nabla \mathcal{L}(\Theta^*)\|\|_2$$

$$\leq \sqrt{\frac{4(1+c)\ \log d}{k\ d_1^2}} \max \left\{ \sqrt{d_1/d_2},\ e^{2\alpha}\sqrt{4(1+c)\log(d)}(8\log d_2 + 2\log d_1)\log k \right\} . \quad (136)$$

Since we are typically interested in the regime where the number of samples is much smaller than the dimension $d_1 \times d_2$ of the problem, the Hessian is typically not positive definite. However, when we restrict our attention to the vectorized Δ with relatively small nuclear norm, then we can prove restricted strong convexity, which gives the following bound.

Lemma 29 (Restricted Strong Convexity for collaborative ranking) *Fix any* $\Theta \in \Omega_\alpha$ *and assume* $24 \leq k \leq \min\{d_1^2, (d_1^2+d_2^2)/(2d_1)\}\log d$. *Under the random sampling model of the alternatives* $\{j_{i\ell}\}_{i\in[d_1],\ell\in[k]}$ *and the random outcome of the comparisons described in section 1, with probability larger than* $1 - 2d^{-2^{18}}$,

$$\text{Vec}(\Delta)^T\ \nabla^2\mathcal{L}(\Theta)\ \text{Vec}(\Delta) \ \geq\ \frac{e^{-4\alpha}}{24\ d_1 d_2}\|\|\Delta\|\|_F^2 , \quad (137)$$

for all Δ *in* $\mathcal{A}$ *where*

$$\mathcal{A} = \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2} \,\middle|\, \|\|\Delta\|\|_\infty \leq 2\alpha,\ \sum_{j\in[d_2]} \Delta_{ij} = 0 \text{ for all } i \in [d_1] \text{ and } \|\|\Delta\|\|_F^2 \geq \mu\|\|\Delta\|\|_{\text{nuc}} \right\} . \quad (138)$$

with

$$\mu \ \equiv\ 2^{10}\ e^{2\alpha}\ \alpha\ d_2 \sqrt{\frac{d_1\ \log d}{k\ \min\{d_1, d_2\}}} . \quad (139)$$

Building on these lemmas, the proof of Theorem 7 is divided into the following two cases. In both cases, we will show that

$$\|\|\Delta\|\|_F^2 \ \leq\ 72\ e^{4\alpha}c_0\lambda_0\ d_1 d_2\ \|\|\Delta\|\|_{\text{nuc}} , \quad (140)$$

with high probability. Applying Lemma 27 proves the desired theorem. We are left to show Eq. (140) holds.

Case 1: Suppose $\|\|\Delta\|\|_F^2 \geq \mu \|\|\Delta\|\|_{\text{nuc}}$. With $\Delta = \Theta^* - \widehat{\Theta}$, the Taylor expansion yields

$$\mathcal{L}(\widehat{\Theta}) = \mathcal{L}(\Theta^*) - \langle\!\langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle\!\rangle + \frac{1}{2}\text{Vec}(\Delta)\nabla^2\mathcal{L}(\Theta)\text{Vec}^T(\Delta), \quad (141)$$

where $\Theta = a\widehat{\Theta}+(1-a)\Theta^*$ for some $a \in [0,1]$. It follows from Lemma 29 that with probability at least $1 - 2d^{-2^{18}}$,

$$\mathcal{L}(\widehat{\Theta}) - \mathcal{L}(\Theta^*) \ \geq\ -\langle\!\langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle\!\rangle + \frac{e^{-4\alpha}}{48\ d_1\ d_2}\|\|\Delta\|\|_F^2$$

$$\geq\ -\|\|\nabla \mathcal{L}(\Theta^*)\|\|_2\|\|\Delta\|\|_{\text{nuc}} + \frac{e^{-4\alpha}}{48\ d_1\ d_2}\|\|\Delta\|\|_F^2 .$$

54

From the definition of $\widehat{\Theta}$ as an optimal solution of the minimization, we have

$$\mathcal{L}(\widehat{\Theta}) - \mathcal{L}(\Theta^*) \leq \lambda \left(\left\|\left\|\Theta^*\right\|\right\|_{\text{nuc}} - \left\|\left\|\widehat{\Theta}\right\|\right\|_{\text{nuc}} \right) \leq \lambda \left\|\left\|\Delta\right\|\right\|_{\text{nuc}} .$$

By the assumption, we choose $\lambda \geq 480\lambda_0$. In view of Lemma 28, this implies that $\lambda \geq 2\left\|\left\|\nabla\mathcal{L}(\Theta^*)\right\|\right\|_2$ with probability at least $1 - 2d^{-3}$. It follows that with probability at least $1 - 2d^{-3} - 2d^{-2^{18}}$,

$$\frac{e^{-4\alpha}}{48d_1d_2}\left\|\left\|\Delta\right\|\right\|_{\text{F}}^2 \leq \left(\lambda + \left\|\left\|\nabla\mathcal{L}(\Theta^*)\right\|\right\|_2\right)\left\|\left\|\Delta\right\|\right\|_{\text{nuc}} \leq \frac{3\lambda}{2}\left\|\left\|\Delta\right\|\right\|_{\text{nuc}} .$$

By our assumption on $\lambda \leq c_0\lambda_0$, this proves the desired bound in Eq. (140)
Case 2: Suppose $\left\|\left\|\Delta\right\|\right\|_{\text{F}}^2 \leq \mu \left\|\left\|\Delta\right\|\right\|_{\text{nuc}}$. By the definition of μ and the fact that $c_0 \geq 480$, it follows that $\mu \leq 72 \, e^{4\alpha}c_0\lambda_0 \, d_1d_2$, and we get the same bound as in Eq. (140).

C.1 Proof of Lemma 28

Define $X_i = -e_i \sum_{\ell=1}^k (e_{v_{i,\ell}} - p_{i,\ell})^T$ such that $\nabla\mathcal{L}(\Theta^*) = \frac{1}{k\,d_1}\sum_{i=1}^{d_1} X_i$, which is a sum of d_1 independent random matrices. Although $\left\|\left\|X_i\right\|\right\|_2$ can be as large as $O(k)$, this occurs with very low probability. We make this precise in the following lemma and focus on the case where $\left\|\left\|X_i\right\|\right\|_2 = O(\sqrt{k})$ for all $i \in [d_1]$.

Lemma 30 *For a fixed $i \in [d_1]$ and $j \in [d_2]$, if $k \leq (1/e) \, d_2 \, (4\log d_2 + \log d_1)$, then the number of times the item j is observed by the user i is at most $8(\log d_2) + 2(\log d_1)$ with probability larger than $1 - 1/(d_2^4 d_1)$.*

Proof is given in the end of this Section. Applying union bound over the d_1 items and d_2 users, we have the multiplicity in sampling for any item for all users is bounded by $8(\log d_2) + 2(\log d_1)$ with probability at least $1 - d_2^{-3}$. We denote this event by $\mathcal{A}$ and let $\mathbb{I}(\mathcal{A})$ be the indicator function that all the multiplicities in sampling are bounded. We first upper bound $\left\|\left\|(\sum_i X_i)\mathbb{I}(\mathcal{A})\right\|\right\|_2$ using the Matrix Bernstein inequality Tropp (2011).

$$\left\|\left\|X_i\mathbb{I}(\mathcal{A})\right\|\right\|_2 = \left\|\mathbb{I}(\mathcal{A}) \sum_{\ell=1}^k \left(e_{v_{i,\ell}} - p_{i,\ell}\right)\right\|$$

$$\overset{(a)}{\leq} \left\|\mathbb{I}(\mathcal{A}) \sum_{\ell=1}^k e_{v_{i,l}}\right\| + \left\|\mathbb{I}(\mathcal{A}) \sum_{\ell=1}^k p_{i,\ell}\right\|$$

$$\overset{(b)}{\leq} (8(\log d_2) + 2(\log d_1))\sqrt{\min\{k, d_2\}}\left(1 + \left(\sum_{\ell=1}^k \frac{e^{2\alpha}}{\ell}\right)\right)$$

$$\overset{(c)}{\leq} \sqrt{k}(8(\log d_2) + 2(\log d_1))\left(1 + 2e^{2\alpha}\,\log k\right)$$

$$\leq 3\sqrt{k}(8(\log d_2) + 2(\log d_1))e^{2\alpha}\,\log k , \qquad (142)$$

where (a) is by triangle inequality, (b) is because under the given event $\mathcal{A}$ each term in $\sum_\ell e_{v_{i,\ell}}$ and $\sum_l p_{i,\ell}$ are upper bounded by $\log d_2$ and $\left(\sum_{\ell=1}^k \frac{e^{2\alpha}}{\ell}\right)\log d_2$ respectively and because there can be at most $\min\{d_2, k\}$ non-zero entries in the two vectors $\sum_\ell e_{v_{i,\ell}}$ and

$\sum_\ell p_{i,\ell}$ and, (c) is due to the fact that k-th harmonic number $\sum_{\ell=1}^k \frac{1}{\ell}$ is upper bounded by $\log k$. We also have,

$$\left\|\left|\sum_i \mathbb{E}\left[X_i X_i^T \mathbb{I}\left(\mathcal{A}\right)\right]\right|\right\|_2 \leq \left\|\left|\sum_i \mathbb{E}\left[X_i X_i^T\right]\right|\right\|_2$$

$$\leq \left\|\left|\sum_{i=1}^{d_1} e_i e_i^T \mathbb{E}\left[\sum_{\ell,\ell'=1}^k \left(e_{v_{i,\ell}} - p_{i,\ell}\right)^T \left(e_{v_{i,\ell'}} - p_{i,\ell'}\right)\right]\right|\right\|_2$$

$$= \left\|\left|\sum_{i=1}^{d_1} e_i e_i^T \mathbb{E}\left[\sum_{\ell=1}^k \left(e_{v_{i,\ell}} - p_{i,\ell}\right)^T \left(e_{v_{i,\ell}} - p_{i,\ell}\right)\right]\right|\right\|_2$$

$$= \left\|\left|\sum_{i=1}^{d_1} e_i e_i^T \mathbb{E}\left[\sum_{\ell=1}^k e_{v_{i,\ell}}^T e_{v_{i,\ell}} - p_{i,\ell}^T p_{i,\ell}\right]\right|\right\|_2$$

$$\leq \left\|\left|\sum_{i=1}^{d_1} e_i e_i^T \mathbb{E}\left[\sum_{\ell=1}^k e_{v_{i,\ell}}^T e_{v_{i,\ell}}\right]\right|\right\|_2$$

$$= k\||\mathbf{I}_{d_1 \times d_1}|\|_2 = k, \tag{143}$$

and

$$\left\|\left|\sum_{i=1}^{d_1} \mathbb{E}\left[X_i^T X_i \mathbb{I}\left(\mathcal{A}\right)\right]\right|\right\|_2 \leq \left\|\left|\sum_{i=1}^{d_1} \mathbb{E}\left[X_i^T X_i\right]\right|\right\|_2$$

$$\leq \left\|\left|\sum_{i=1}^{d_1} \mathbb{E}\left[\sum_{\ell,\ell'=1}^k (e_{v_{i,\ell}} - p_{i,\ell})(e_{v_{i,\ell'}} - p_{i,\ell'})^T\right]\right|\right\|_2$$

$$= \left\|\left|\sum_{i=1}^{d_1} \mathbb{E}\left[\sum_{\ell=1}^k (e_{v_{i,\ell}} - p_{i,\ell})(e_{v_{i,\ell}} - p_{i,\ell})^T\right]\right|\right\|_2 \tag{144}$$

$$= \left\|\left|\sum_{i=1}^{d_1} \mathbb{E}\left[\sum_{\ell=1}^k e_{v_{i,\ell}} e_{v_{i,\ell}}^T - p_{i,\ell} p_{i,\ell}^T\right]\right|\right\|_2$$

$$\leq \left\|\left|\sum_{i=1}^{d_1} \mathbb{E}\left[\sum_{\ell=1}^k e_{v_{i,\ell}} e_{v_{i,\ell}}^T\right]\right|\right\|_2$$

$$= \left\|\left|\sum_{i=1}^{d_1} \frac{k}{d_2}\mathbf{I}_{d_2 \times d_2}\right|\right\|_2 = \frac{kd_1}{d_2} . \tag{145}$$

By matrix Bernstein inequality Tropp (2011),

$$\mathbb{P}\left(\||\nabla\mathcal{L}(\Theta^*)\mathbb{I}\left(\mathcal{A}\right)|\|_2 > t\right)$$

$$\leq (d_1 + d_2)\exp\left(\frac{-k^2 d_1^2 t^2/2}{(d_1 k/\min\{d_2, d_1\}) + (3e^{2\alpha}k^{3/2}d_1(8(\log d_2) + 2(\log d_1))\log k\, t/3)}\right),$$

56

which gives the tail probability of $2d^{-c}$ for the choice of

$$
\begin{aligned}
t &= \max\left\{\sqrt{\frac{4(1+c)\log d}{k\,d_1\,\min\{d_2,d_1\}}}, \; \frac{4(1+c)e^{2\alpha}\log(d)\,(8(\log d_2)+2(\log d_1))\log k}{k^{1/2}\,d_1}\right\} \\
&= \frac{\sqrt{4(1+c)\log d}}{k^{1/2}\,d_1}\max\left\{\sqrt{d_1/d_2}, \; e^{2\alpha}\sqrt{4(1+c)\log(d)}\,(8(\log d_2)+2(\log d_1))\log k\right\}.
\end{aligned}
$$

Now with a high probability of $1-\frac{2}{d^c}-\frac{1}{d_2^3}$ the desired bound is true.

C.2 Proof of Lemma 30

In a classical balls-in-bins setting, we consider k as the number of balls and d_2 as the number of bins. We can consider the number of balls in a particular bin as the number of times the user i observes item j. Let the event that this number is at least δ be denoted by the event A_δ^j. Then, $\mathbb{P}\left\{A_\delta^j\right\} \leq \binom{k}{\delta}\frac{1}{d_2^\delta} \leq \left(\frac{ke}{d_2\delta}\right)^\delta$. Using the fact that $(1/x)^x \leq a$ for any $x \geq (2\log(1/a))/(\log\log(1/a))$, we let $x = d_2\delta/(ke)$ to get

$$
\left(\frac{ke}{d_2\delta}\right)^\delta \leq a^{\frac{ke}{d_2}},
$$

for $\delta \geq (ke/d_2)(2\log(1/a))/(\log\log(1/a))$. Choosing $a = (1/d_2^4 d_1)^{d_2/ke}$, we have $\mathbb{P}\left\{A_\delta^j\right\} \leq 1/(d_1 d_2^4)$, for a choice of $\delta = 2\,\log(d_2^4 d_1) \geq 2\log(d_2^4 d_1)/(\log((d_2/ke)\log(d_2^4 d_1)))$.

C.3 Proof of Lemma 29

Recall that the Hessian matrix is a block-diagonal matrix with the i-th block $H^{(i)}(\Theta)$ given by (134). We use the following remark from Hajek et al. (2014) to bound the Hessian.

Remark 31 *(Hajek et al., 2014, Claim 1) Given $\theta \in \mathbb{R}^r$, let p be the column probability vector with $p_i = e^{\theta_i}/(e^{\theta_1} + \cdots + e^{\theta_\rho})$ for each $i \in [\rho]$ and for any positive integer ρ. If $|\theta_i| \leq \alpha$, for all $i \in [\rho]$, then*

$$
e^{2\alpha}\left(\operatorname{diag}(p) - pp^T\right) \succeq \frac{1}{\rho}\operatorname{diag}(\mathbb{1}) - \frac{1}{\rho^2}\mathbb{1}\mathbb{1}^T.
$$

By letting $\mathbb{1}_{S_{i,\ell}} = \sum_{j\in S_{i,\ell}} e_j$ and applying the above claim, we have

$$
\begin{aligned}
e^{2\alpha}H^{(i)}(\Theta) &\succeq \frac{1}{k\,d_1}\sum_{\ell=1}^{k}\left(\frac{1}{k-\ell+1}\operatorname{diag}(\mathbb{1}_{S_{i,\ell}}) - \frac{1}{(k-\ell+1)^2}\mathbb{1}_{S_{i,\ell}}\mathbb{1}_{S_{i,\ell}}^T\right) \\
&= \frac{1}{2\,k\,d_1}\sum_{\ell=1}^{k}\frac{1}{(k-\ell+1)^2}\sum_{j,j'\in S_{i,\ell}}(e_j - e_{j'})(e_j - e_{j'})^T \\
&\succeq \frac{1}{2\,k^3\,d_1}\sum_{\ell=1}^{k}\sum_{j,j'\in S_{i,\ell}}(e_j - e_{j'})(e_j - e_{j'})^T.
\end{aligned}
$$

Hence,

$$\text{Vec}(\Delta)\nabla^2\mathcal{L}(\Theta)\text{Vec}^T(\Delta) = \sum_{i=1}^{d_1}(\Delta^T e_i)^T H^{(i)}(\Theta)(\Delta^T e_i)$$

$$\geq \frac{e^{-2\alpha}}{2\,k^3\,d_1}\sum_{i=1}^{d_1}\sum_{\ell=1}^{k}\sum_{j,j'\in S_{i,\ell}}\big\||e_i^T\Delta(e_j-e_{j'})|\big\|_2^2.$$

By changing the order of the summation, we get that

$$\sum_{\ell=1}^{k}\sum_{j,j'\in S_{i,\ell}}\big\||e_i^T\Delta(e_j-e_{j'})|\big\|_2^2 = \sum_{\ell,\ell'=1}^{k}\langle\!\langle\Delta,e_{i,j_{i,\ell}}-e_{i,j_{i,\ell'}}\rangle\!\rangle^2\sum_{\ell''=1}^{k}\mathbb{I}\big(\sigma_i(j_{i,\ell''})$$

$$\leq \min\{\sigma_i(j_{i,\ell}),\sigma_i(j_{i,\ell'})\}\big).$$

Define

$$\chi_{i,\ell,\ell',\ell''} \equiv \mathbb{I}\big(\sigma_i(j_{i,\ell''})\leq\min\{\sigma_i(j_{i,\ell}),\sigma_i(j_{i,\ell'})\}\big), \tag{146}$$

and let

$$H(\Delta) \equiv \frac{e^{-2\alpha}}{2\,k^3\,d_1}\sum_{i=1}^{d_1}\sum_{\ell,\ell'=1}^{k}\langle\!\langle\Delta,e_{i,j_{i,\ell}}-e_{i,j_{i,\ell'}}\rangle\!\rangle^2\sum_{\ell''=1}^{k}\chi_{i,\ell,\ell',\ell''}.$$

Then we have $\text{Vec}^T(\Delta)\nabla^2\mathcal{L}(\Theta)\text{Vec}(\Delta)\geq H(\Delta)$. To prove the theorem, it suffices to bound $H(\Delta)$ from the below. First, we prove a lower bound on the expectation $\mathbb{E}[H(\Delta)]$. Notice that for $\ell\neq\ell'$, the conditional expectation of $\chi_{i,\ell,\ell',\ell''}$'s, given the set of alternatives presented to user i is

$$\mathbb{E}\Big[\sum_{\ell''=1}^{k}\chi_{i,\ell,\ell',\ell''}\,\big|\,j_{i,1},\ldots,j_{i,k}\Big] = 1 + \sum_{\ell''\neq\ell,\ell'}\frac{\exp(\theta_{i,j_{i,\ell''}})}{\exp(\theta_{i,j_{i,\ell''}})+\exp(\theta_{i,j_{i,\ell'}})+\exp(\theta_{i,j_{i,\ell}})}$$

$$\geq 1 + \frac{k-2}{1+2e^{2\alpha}} \geq \frac{k}{3e^{2\alpha}}.$$

Then,

$$\mathbb{E}[H(\Delta)] = \frac{e^{-2\alpha}}{2\,k^3\,d_1}\sum_{i,\ell,\ell'}\mathbb{E}\Big[\langle\!\langle\Delta,e_{i,j_{i,\ell}}-e_{i,j_{i,\ell'}}\rangle\!\rangle^2\mathbb{E}\big[\sum_{\ell''=1}^{k}\chi_{i,\ell,\ell',\ell''}\,\big|\,j_{i,1},\ldots,j_{i,k}\big]\Big]$$

$$\geq \frac{e^{-4\alpha}}{6\,k^2\,d_1}\sum_{i=1}^{d_1}\sum_{\ell,\ell'\in[k]}\mathbb{E}\Big[\langle\!\langle\Delta,e_{i,j_{i,\ell}}-e_{i,j_{i,\ell'}}\rangle\!\rangle^2\Big]$$

$$= \frac{e^{-4\alpha}}{6\,k^2\,d_1}\sum_{i=1}^{d_1}\sum_{\ell\neq\ell'\in[k]}\left(\frac{2}{d_2}\sum_{j=1}^{d_2}\Delta_{ij}^2 - \frac{2}{d_2^2}\sum_{j,j'=1}^{d_2}\Delta_{ij}\Delta_{ij'}\right)$$

$$= \frac{e^{-4\alpha}(k-1)}{3\,k\,d_1\,d_2}\|\Delta\|_{\text{F}}^2\,, \tag{147}$$

58

where the last equality holds because $\sum_{j\in[d_2]}\Delta_{ij}=0$ for $\Delta\in\Omega_{2\alpha}$ and for all $i\in[d_1]$.

We are left to prove that $H(\Delta)$ cannot deviate from its mean too much. Suppose there exists a $\Delta\in\mathcal{A}$ such that Eq. (137) is violated, i.e. $H(\Delta)<(e^{-4\alpha}/(24\,d_1d_2))\|\!|\Delta\|\!|_{\mathrm{F}}^2$. We will show this happens with a small probability. From Eq. (147), we get that for $k\geq 24$,

$$
\begin{aligned}
\mathbb{E}[H(\Delta)]-H(\Delta) &\geq \frac{(7k-8)}{24k}\frac{e^{-4\alpha}}{d_1\,d_2}\|\!|\Delta\|\!|_{\mathrm{F}}^2 \\
&\geq \frac{(20/3)\,e^{-4\alpha}}{24\,d_1d_2}\|\!|\Delta\|\!|_{\mathrm{F}}^2 .
\end{aligned}
\tag{148}
$$

We use a peeling argument as in (Negahban and Wainwright, 2012, Lemma 3), Van De Geer (2000) to upper bound the probability that Eq. (148) is true. We first construct the following family of subsets to cover $\mathcal{A}$ such that $\mathcal{A}\subseteq\bigcup_{\ell=1}^{\infty}\mathcal{S}_\ell$. Recall $\mu=2^{10}e^{2\alpha}\alpha d_2\sqrt{(d_1\log d)/(k\min\{d_1,d_2\})}$, define in (139). Notice that since for any $\Delta\in\mathcal{A}$, $\|\!|\Delta\|\!|_{\mathrm{F}}^2\geq\mu\|\!|\Delta\|\!|_{\mathrm{nuc}}\geq\mu\|\!|\Delta\|\!|_{\mathrm{F}}$, it follows that $\|\!|\Delta\|\!|_{\mathrm{F}}\geq\mu$. Then, we can cover $\mathcal{A}$ with the family of sets

$$
\mathcal{S}_\ell=\Big\{\Delta\in\mathbb{R}^{d_1\times d_2}\,\Big|\,\|\!|\Delta\|\!|_\infty\leq 2\alpha,\,\beta^{\ell-1}\mu\leq\|\!|\Delta\|\!|_{\mathrm{F}}\leq\beta^\ell\mu,
$$
$$
\sum_{j\in[d_2]}\Delta_{ij}=0\text{ for all }i\in[d_1],\text{ and }\|\!|\Delta\|\!|_{\mathrm{nuc}}\leq\beta^{2\ell}\mu\Big\},
$$

where $\beta=\sqrt{10/9}$ and for $\ell\in\{1,2,3,\ldots\}$. This implies that when there exists a $\Delta\in\mathcal{A}$ such that (148) holds, then there exists an $\ell\in\mathbb{Z}_+$ such that $\Delta\in\mathcal{S}_\ell$ and

$$
\begin{aligned}
\mathbb{E}[H(\Delta)]-H(\Delta) &\geq \frac{(20/3)\,e^{-4\alpha}}{24\,d_1d_2}\beta^{2(\ell-1)}\mu^2 \\
&\geq \frac{e^{-4\alpha}}{4\,d_1d_2}\beta^{2\ell}\mu^2 .
\end{aligned}
\tag{149}
$$

Applying the union bound over $\ell\in\mathbb{Z}_+$, we get from (148) and (149) that

$$
\mathbb{P}\left\{\exists\Delta\in\mathcal{A}\,,\,H(\Delta)<\frac{e^{-4\alpha}}{24\,d_1d_2}\|\!|\Delta\|\!|_{\mathrm{F}}^2\right\}
$$
$$
\leq\sum_{\ell=1}^{\infty}\mathbb{P}\left\{\sup_{\Delta\in\mathcal{S}_\ell}\big(\,\mathbb{E}[H(\Delta)]-H(\Delta)\,\big)>\frac{e^{-4\alpha}}{4\,d_1d_2}(\beta^\ell\mu)^2\right\}
$$
$$
\leq\sum_{\ell=1}^{\infty}\mathbb{P}\left\{\sup_{\Delta\in\mathcal{B}(\beta^\ell\mu)}\big(\,\mathbb{E}[H(\Delta)]-H(\Delta)\,\big)>\frac{e^{-4\alpha}}{4\,d_1d_2}(\beta^\ell\mu)^2\right\},
\tag{150}
$$

where we define a new set $\mathcal{B}(D)$ such that $\mathcal{S}_\ell\subseteq\mathcal{B}(\beta^\ell\mu)$:

$$
\mathcal{B}(D)=\Big\{\Delta\in\mathbb{R}^{d_1\times d_2}\,\big|\,\|\!|\Delta\|\!|_\infty\leq 2\alpha,\|\!|\Delta\|\!|_{\mathrm{F}}\leq D,
$$
$$
\sum_{j\in[d_2]}\Delta_{ij}=0\text{ for all }i\in[d_1],\mu\|\!|\Delta\|\!|_{\mathrm{nuc}}\leq D^2\Big\}.
\tag{151}
$$

The following key lemma provides the upper bound on this probability.

Lemma 32 *For $(16 \min\{d_1, d_2\} \log d)/(3d_1) \leq k \leq d_1^2 \log d$,*

$$\mathbb{P}\left\{\sup_{\Delta \in \mathcal{B}(D)} \left(\mathbb{E}[H(\Delta)] - H(\Delta) \right) \geq \frac{e^{-4\alpha}}{4d_1 d_2} D^2 \right\} \leq \exp\left\{ -\frac{e^{-4\alpha} k D^4}{2^{19} \alpha^4 d_1 d_2^2} \right\}. \quad (152)$$

Let $\eta = \exp\left(-\frac{e^{-4\alpha} 4k(\beta - 1.002)\mu^4}{2^{19}\alpha^4 d_1 d_2^2} \right)$. Applying the tail bound to (150), we get

$$\mathbb{P}\left\{ \exists \Delta \in \mathcal{A} , \ H(\Delta) < \frac{e^{-4\alpha}}{24\,d_1 d_2} \|\|\Delta\|\|_{\mathrm{F}}^2 \right\} \leq \sum_{\ell=1}^{\infty} \exp\left\{ -\frac{e^{-4\alpha} k(\beta^\ell \mu)^4}{2^{19}\alpha^4 d_1 d_2^2} \right\}$$

$$\overset{(a)}{\leq} \sum_{\ell=1}^{\infty} \exp\left\{ -\frac{e^{-4\alpha} 4k\ell(\beta - 1.002)\mu^4}{2^{19}\alpha^4 d_1 d_2^2} \right\}$$

$$\leq \frac{\eta}{1-\eta},$$

where (a) holds because $\beta^x \geq x \log \beta \geq x(\beta - 1.002)$ for the choice of $\beta = \sqrt{10/9}$. By the definition of μ,

$$\eta = \exp\left\{ -\frac{2^{23} e^{4\alpha} d_2^2 d_1 (\log d)^2 (\beta - 1.002)}{k(\min\{d_1, d_2\})^2} \right\} \leq \exp\{ -2^{18} \log d \},$$

where the last inequality follows from the assumption that $k \leq \max\{d_1, d_2^2/d_1\} \log d = (d_2^2 d_1 \log d)/(\min\{d_1, d_2\})^2$, and $\beta - 1.002 \geq 2^{-5}$. Since for $d \geq 2$, $\exp\{-2^{18} \log d\} \leq 1/2$ and thus $\eta \leq 1/2$, the lemma follows by assembling the last two displayed inequalities.

C.4 Proof of Lemma 32

Recall that

$$H(\Delta) = \frac{e^{-2\alpha}}{2\,k^3\,d_1} \sum_{i=1}^{d_1} \sum_{\ell,\ell'=1}^{k} \langle\!\langle \Delta, e_{i,j_{i,\ell}} - e_{i,j_{i,\ell'}} \rangle\!\rangle^2 \sum_{\ell''=1}^{k} \chi_{i,\ell,\ell',\ell''} ,$$

with $\chi_{i,\ell,\ell',\ell''} = \mathbb{I}\left(\sigma_i(j_{i,\ell''}) \leq \min\{\sigma_i(j_{i,\ell}), \sigma_i(j_{i,\ell'})\} \right)$. Let $Z = \sup_{\Delta \in \mathcal{B}(D)} \mathbb{E}[H(\Delta)] - H(\Delta)$ be the worst-case random deviation of $H(\Delta)$ form its mean. We prove an upper bound on Z by showing that $Z - \mathbb{E}[Z] \leq e^{-4\alpha} D^2/(64d_1 d_2)$ with high probability, and $\mathbb{E}[Z] \leq 9e^{-4\alpha} D^2/(40d_1 d_2)$. This proves the desired claim in Lemma 32.

To prove the concentration of Z, we utilize the random utility model (RUM) theoretic interpretation of the MNL model. The random variable Z depends on the random choice of alternatives $\{j_{i,\ell}\}_{i \in [d_1], \ell \in [k]}$ and the random k-wise ranking outcomes $\{\sigma_i\}_{i \in [d_1]}$. The random utility theory, pioneered by Thurstone (1927); Marschak (1960); Luce (1959), tells us that the k-wise ranking from the MNL model has the same distribution as first drawing independent (unobserved) utilities $u_{i,\ell}$'s of the item $j_{i,\ell}$ for user i according to the standard Gumbel Cumulative Distribution Function (CDF) $F(c - \Theta_{i,j_{i,\ell}})$ with $F(c) = e^{-e^{-c}}$, and then ranking the k items for user i according to their respective utilities. Given this definition of the MNL model, we have $\chi_{i,\ell,\ell',\ell''} = \mathbb{I}\left(u_{i,\ell''} \geq \max\{u_{i,\ell}, u_{i,\ell'}\} \right)$. Thus Z is a function of independent

choices of the items and their (unobserved) utilities, i.e. $Z = f(\{(j_{i,\ell}, u_{i,\ell})\}_{i\in[d_1],\ell\in[k]})$. Let $x_{i,\ell} = (j_{i,\ell}, u_{i,\ell})$ and write $H(\Delta)$ as $H(\Delta, \{x_{i,\ell}\}_{i\in[d_1],\ell\in[k]})$. This allows us to bound the difference and apply McDiarmid's tail bound. Note that for any $i \in [d_1]$, $\ell \in [k]$, $x_{1,1}, \ldots, x_{d_1,k}$, and $x'_{i,\ell}$,

$$
\begin{aligned}
&\left| f\left(x_{1,1}, \ldots, x_{i,\ell}, \ldots, x_{d_1,k}\right) - f\left(x_{1,1}, \ldots, x'_{i,\ell}, \ldots, x_{d_1,k}\right) \right| \\
&= \Big| \sup_{\Delta \in \mathcal{B}(D)} \left(\mathbb{E}\left[H(\Delta)\right] - H(\Delta, x_{1,1}, \ldots, x_{i,\ell}, \ldots, x_{d_1,k})\right) - \\
&\quad \sup_{\Delta \in \mathcal{B}(D)} \left(\mathbb{E}\left[H(\Delta)\right] - H(\Delta, x_{1,1}, \ldots, x'_{i,\ell}, \ldots, x_{d_1,k})\right) \Big| \\
&\leq \sup_{\Delta \in \mathcal{B}(D)} \left| H(\Delta, x_{1,1}, \ldots, x_{i,\ell}, \ldots, x_{d_1,k}) - H(\Delta, x_{1,1}, \ldots, x'_{i,\ell}, \ldots, x_{d_1,k}) \right| \\
&\overset{(a)}{\leq} \frac{e^{-2\alpha}}{2\,k^3\,d_1} \sup_{\Delta \in \mathcal{B}(D)} \Big\{ 2 \sum_{\ell' \in [k]} \langle\!\langle \Delta, e_{i,j_{i,\ell}} - e_{i,j_{i,\ell'}} \rangle\!\rangle^2 \sum_{\ell''=1}^{k} \chi_{i,\ell,\ell',\ell''} + \\
&\quad \sum_{\ell',\ell'' \in [k]} \langle\!\langle \Delta, e_{i,j_{i,\ell'}} - e_{i,j_{i,\ell''}} \rangle\!\rangle^2 \chi_{i,\ell',\ell'',\ell} \Big\} \\
&\overset{(b)}{\leq} \frac{8\alpha^2 e^{-2\alpha}}{k^3\,d_1} \Big\{ 2 \sum_{\ell' \in [k]\setminus\{\ell\}} \sum_{\ell''=1}^{k} \chi_{i,\ell,\ell',\ell''} + \sum_{\ell',\ell'' \in [k], \ell' \neq \ell''} \chi_{i,\ell',\ell'',\ell} \Big\} \\
&\leq \frac{16\alpha^2 e^{-2\alpha}}{k\,d_1} \;,
\end{aligned}
$$

where (a) follows because for a fixed i and ℓ, the random variable $x_{i,\ell} = (j_{i,\ell}, u_{i,\ell})$ can appear in three terms, i.e. $\sum_{\ell',\ell''} \langle\!\langle \Delta, e_{i,j_{i,\ell}} - e_{i,j_{i,\ell'}} \rangle\!\rangle^2 \chi_{i,\ell,\ell',\ell''} + \sum_{\ell',\ell''} \langle\!\langle \Delta, e_{i,j_{i,\ell'}} - e_{i,j_{i,\ell}} \rangle\!\rangle^2 \chi_{i,\ell',\ell,\ell''} + \sum_{\ell',\ell''} \langle\!\langle \Delta, e_{i,j_{i,\ell'}} - e_{i,j_{i,\ell''}} \rangle\!\rangle^2 \chi_{i,\ell',\ell'',\ell}$, and (b) follows because $|\Delta_{ij}| \leq 2\alpha$ for all i, j since $\Delta \in \mathcal{B}(D)$. The last inequality follows because in the worst case, $\sum_{\ell' \in [k]\setminus\{\ell\}} \sum_{\ell''=1}^{k} \chi_{i,\ell,\ell',\ell''} \leq k(k-1)/2$ and $\sum_{\ell',\ell'' \in [k], \ell' \neq \ell''} \chi_{i,\ell',\ell'',\ell} \leq k(k-1)$. This holds with equality if $\sigma_i(j_{i,\ell}) = k$ and $\sigma_i(j_{i,\ell}) = 1$, respectively. By bounded differences inequality, we have

$$
\mathbb{P}\left\{ Z - \mathbb{E}[Z] \geq t \right\} \leq \exp\left(-\frac{k^2\,d_1^2\,t^2}{2^7\,\alpha^4 e^{-4\alpha} d_1 k} \right),
$$

It follows that for the choice of $t = e^{-4\alpha} D^2/(64 d_1 d_2)$,

$$
\mathbb{P}\left\{ Z - \mathbb{E}[Z] \geq \frac{e^{-4\alpha} D^2}{64 d_1 d_2} \right\} \leq \exp\left(-\frac{e^{-4\alpha} k D^4}{2^{19} \alpha^4 d_1 d_2^2} \right).
$$

We are left to prove the upper bound on $\mathbb{E}[Z]$ using symmetrization and contraction. Define random variables

$$
Y_{i,\ell,\ell',\ell''}(\Delta) \equiv (\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2 \chi_{i,\ell,\ell',\ell''} \;, \tag{153}
$$

where the randomness is in the choice of alternatives $j_{i,\ell}, j_{i,\ell'}$, and $j_{i,\ell''}$, and the outcome of the comparisons of those three alternatives.

The main challenge in applying the symmetrization to $\sum_{\ell,\ell',\ell''\in[k]} Y_{i,\ell,\ell',\ell''}(\Delta)$ is that we need to partition the summation over the set $[k] \times [k] \times [k]$ into subsets of independent random variables, such that we can apply the standard symmetrization argument. To this end, we prove in the following lemma a generalization of the well-known problem of scheduling a round robin tournament to a tournament of matches involving three teams each. No teams are present in more than one triple in a single round, and we want to minimize the number of rounds to cover all combination of triples are matched. For example, when there are $k = 6$ teams, there is a simple construction of such a tournament: $T_1 = \{(1,2,3),(4,5,6)\}$, $T_2 = \{1,2,4),(3,5,6)\}$, $T_3 = \{(1,2,5),(3,4,6)\}$, $T_4 = \{(1,2,6),(3,4,5)\}$, $T_5 = \{(1,3,4),(2,5,6)\}$, $T_6 = \{(1,3,5),(2,4,6)\}$, $T_7 = \{(1,3,6),(2,4,5)\}$, $T_8 = \{(1,4,5),(2,3,6)\}$, $T_9 = \{(1,4,6),(2,3,5)\}$, $T_{10} = \{(1,5,6),(2,3,4)\}$. This is a perfect scheduling of a tournament with three teams in each match. For a general k, the following lemma provides a construction with $O(k^2)$ rounds.

Lemma 33 *There exists a partition $(T_1, \ldots, T_N)$ of $[k] \times [k] \times [k]$ for some $N \leq 24k^2$ such that T_a's are disjoint subsets of $[k] \times [k] \times [k]$, $\bigcup_{a\in[N]} T_a = [k] \times [k] \times [k]$, $|T_a| \leq \lfloor k/3 \rfloor$ and for any $a \in [N]$ the set of random variables in T_a satisfy*

$$\{Y_{i,\ell,\ell',\ell''}\}_{i\in[d_1],(\ell,\ell',\ell'')\in T_a} \text{ are mutually independent } .$$

Now, we are ready to partition the summation.

$$
\begin{aligned}
\mathbb{E}[Z] &= \frac{e^{-2\alpha}}{2\,k^3\,d_1}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i\in[d_1]} \sum_{\ell,\ell',\ell''\in[k]} \big\{\mathbb{E}[Y_{i,\ell,\ell',\ell''}(\Delta)] - Y_{i,\ell,\ell',\ell''}(\Delta)\big\}\Big] \\
&= \frac{e^{-2\alpha}}{2\,k^3\,d_1}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i\in[d_1]} \sum_{a\in[N]} \sum_{(\ell,\ell',\ell'')\in T_a} \big\{\mathbb{E}[Y_{i,\ell,\ell',\ell''}(\Delta)] - Y_{i,\ell,\ell',\ell''}(\Delta)\big\}\Big] \\
&\leq \frac{e^{-2\alpha}}{2\,k^3\,d_1}\sum_{a\in[N]}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i\in[d_1]} \sum_{(\ell,\ell',\ell'')\in T_a} \big\{\mathbb{E}[Y_{i,\ell,\ell',\ell''}(\Delta)] - Y_{i,\ell,\ell',\ell''}(\Delta)\big\}\Big] \\
&\leq \frac{e^{-2\alpha}}{k^3\,d_1}\sum_{a\in[N]}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i\in[d_1]} \sum_{(\ell,\ell',\ell'')\in T_a} \xi_{i,\ell,\ell',\ell''} Y_{i,\ell,\ell',\ell''}(\Delta)\Big] \\
&= \frac{e^{-2\alpha}}{k^3\,d_1}\sum_{a\in[N]}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i\in[d_1]} \sum_{(\ell,\ell',\ell'')\in T_a} \xi_{i,\ell,\ell',\ell''}(\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2 \chi_{i,\ell,\ell',\ell''}\Big] ,(154)
\end{aligned}
$$

where the first inequality follows from the fact that sum of the supremum is no less than the supremum of the sum, and the second inequality follows from standard symmetrization argument applied to independent random variables $\{Y_{i,\ell,\ell',\ell''}(\Delta)\}_{i\in[d_1],(\ell,\ell',\ell'')\in T_a}$ with i.i.d. Rademacher random variables $\xi_{i,\ell,\ell',\ell''}$'s. Since $(\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2 \chi_{i,\ell,\ell',\ell''} \leq 4\alpha|\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}}|\chi_{i,\ell,\ell',\ell''}$, we have by the Ledoux-Talagrand contraction inequality Ledoux and Tala-

grand (2013) that

$$
\mathbb{E}\Big[\sup_{\Delta \in \mathcal{B}(D)} \sum_{i \in [d_1]} \sum_{(\ell,\ell',\ell'') \in T_a} \xi_{i,\ell,\ell',\ell''} (\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2 \chi_{i,\ell,\ell',\ell''} \Big]
$$

$$
\leq \ 8\alpha \mathbb{E}\Big[\sup_{\Delta \in \mathcal{B}(D)} \sum_{i \in [d_1]} \sum_{(\ell,\ell',\ell'') \in T_a} \xi_{i,\ell,\ell',\ell''}\, \chi_{i,\ell,\ell',\ell''} \, \langle\!\langle \Delta, e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T \rangle\!\rangle \Big] \qquad (155)
$$

Applying Hölder's inequality, we get that

$$
\Big| \sum_{i \in [d_1]} \sum_{(\ell,\ell',\ell'') \in T_a} \xi_{i,\ell,\ell',\ell''}\, \chi_{i,\ell,\ell',\ell''} \, \langle\!\langle \Delta, e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T \rangle\!\rangle \Big|
$$

$$
\leq \ \|\!\|\Delta\|\!\|_{\mathrm{nuc}} \Big\|\!\Big\|\!\Big\| \sum_{i \in [d_1]} \sum_{(\ell,\ell',\ell'') \in T_a} \xi_{i,\ell,\ell',\ell''}\, \chi_{i,\ell,\ell',\ell''} \, \big(e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\big) \Big\|\!\Big\|\!\Big\|_2 . \qquad (156)
$$

We are left to prove that the expected value of the right-hand side of the above inequality is bounded by $C\|\!\|\Delta\|\!\|_{\mathrm{nuc}} \sqrt{kd_1 \log d / \min\{d_1, d_2\}}$ for some numerical constant C. For $i \in [d_1]$ and $(\ell, \ell', \ell'') \in T_a$, let $W_{i,\ell,\ell',\ell''} = \xi_{i,\ell,\ell',\ell''}\, \chi_{i,\ell,\ell',\ell''} \, \big(e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\big)$ be independent zero-mean random matrices, such that

$$
\|\!\|W_{i,\ell,\ell',\ell''}\|\!\|_2 = \Big\|\!\Big\|\!\Big\| \xi_{i,\ell,\ell',\ell''}\, \chi_{i,\ell,\ell',\ell''} \, \big(e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\big) \Big\|\!\Big\|\!\Big\|_2 \leq \sqrt{2}\,,
$$

almost surely, and

$$
\begin{aligned}
\mathbb{E}[W_{i,\ell,\ell',\ell''} W_{i,\ell,\ell',\ell''}^T] &= \mathbb{E}\big[\big(e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T(e_{j_{i,\ell}} - e_{j_{i,\ell'}})e_i^T\big)\chi_{i,\ell,\ell',\ell''}\big] \\
&= 2\mathbb{E}\big[\chi_{i,\ell,\ell',\ell''}\big] e_i e_i^T \\
&\preceq 2 e_i e_i^T \,,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[W_{i,\ell,\ell',\ell''}^T W_{i,\ell,\ell',\ell''}] &= \mathbb{E}\big[\big((e_{j_{i,\ell}} - e_{j_{i,\ell'}})e_i^T e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\big)\chi_{i,\ell,\ell',\ell''}\big] \\
&\preceq \mathbb{E}\big[(e_{j_{i,\ell}} - e_{j_{i,\ell'}})e_i^T e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\big] \\
&= \frac{2}{d_2}\mathbf{I}_{d_2 \times d_2} - \frac{2}{d_2^2}\mathbb{1}\mathbb{1}^T \,.
\end{aligned}
$$

This gives

$$
\begin{aligned}
\sigma^2 &= \max \left\{ \Big\|\!\Big\|\!\Big\| \sum_{\substack{i \in [d_1] \\ (\ell,\ell',\ell'') \in T_a}} \mathbb{E}[W_{i,\ell,\ell',\ell''} W_{i,\ell,\ell',\ell''}^T] \Big\|\!\Big\|\!\Big\|_2 , \Big\|\!\Big\|\!\Big\| \sum_{\substack{i \in [d_1] \\ (\ell,\ell',\ell'') \in T_a}} \mathbb{E}[W_{i,\ell,\ell',\ell''}^T W_{i,\ell,\ell',\ell''}] \Big\|\!\Big\|\!\Big\|_2 \right\} \\
&\leq \max \left\{ 2|T_a|, \frac{2d_1|T_a|}{d_2} \right\} = \frac{2d_1|T_a|}{\min\{d_1,d_2\}} \leq \frac{2d_1 k}{3\min\{d_1,d_2\}} \,,
\end{aligned}
$$

since we have designed T_a's such that $|T_a| \leq k/3$. Applying matrix Bernstein inequality Tropp (2011) yields the tail bound

$$\mathbb{P}\left\{\left|\!\left|\!\left|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a} W_{i,\ell,\ell',\ell''}\right|\!\right|\!\right|_2 \geq t\right\} \leq (d_1+d_2)\exp\left(\frac{-t^2/2}{\sigma^2+\sqrt{2}t/3}\right).$$

Choosing $t = \max\left\{\sqrt{32kd_1\log d/(3\min\{d_1,d_2\})}, (16\sqrt{2}/3)\log d\right\}$, we obtain with probability at least $1-2d^{-3}$,

$$\left|\!\left|\!\left|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a} W_{i,\ell,\ell',\ell''}\right|\!\right|\!\right|_2 \leq \max\left\{\sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}}, \frac{16\sqrt{2}\log d}{3}\right\}.$$

It follows from the fact $\left|\!\left|\!\left|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a} W_{i,\ell,\ell',\ell''}\right|\!\right|\!\right|_2 \leq \sum_{i,(\ell,\ell',\ell'')}\left|\!\left|\!\left|W_{i,\ell,\ell',\ell''}\right|\!\right|\!\right|_2 \leq \frac{\sqrt{2}d_1 k}{3}$ that

$$\mathbb{E}\left[\left|\!\left|\!\left|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a} W_{i,\ell,\ell',\ell''}\right|\!\right|\!\right|_2\right] \leq \max\left\{\sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}}, \frac{16\sqrt{2}\log d}{3}\right\} + \frac{2\sqrt{2}d_1 k}{3d^3}$$

$$\leq 2\sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}},$$

where the last inequality follows from the assumption that $(16\min\{d_1,d_2\}\log d)/(3d_1) \leq k \leq d_1^2\log d$. Substituting this in the RHS of Eq. (156), and then together with Eqs. (155) and (154), this gives the following desired bound:

$$\mathbb{E}[Z] \leq \sum_{a\in[N]}\sup_{\Delta\in\mathcal{B}(D)}\frac{16\alpha e^{-2\alpha}}{k^3 d_1}\sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}}\|\!|\Delta\|\!|_{\mathrm{nuc}}$$

$$\leq \sum_{a\in[N]}\frac{e^{-4\alpha}\sqrt{2}}{16\sqrt{3}k^2 d_1 d_2}\underbrace{\left(2^{10}e^{2\alpha}\alpha d_2\sqrt{\frac{d_1\log d}{k\min\{d_1,d_2\}}}\right)}_{=\mu}\|\!|\Delta\|\!|_{\mathrm{nuc}}$$

$$\leq \frac{9e^{-4\alpha}D^2}{40d_1 d_2},$$

where the last inequality holds because $N \leq 4k^2$ and $\mu\|\!|\Delta\|\!|_{\mathrm{nuc}} \leq D^2$.

C.5 Proof of Lemma 33

Recall that $Y_{i,\ell,\ell',\ell''}(\Delta) = (\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2\chi_{i,\ell,\ell',\ell''}$, as defined in (153). From the random utility model (RUM) interpretation of the MNL model presented in Section 1, it is not difficult to show that $Y_{i,\ell,\ell',\ell''}$ and $Y_{i,\tilde{\ell},\tilde{\ell}',\tilde{\ell}''}$ are mutually independent if the two triples (ℓ,ℓ',ℓ'') and $(\tilde{\ell},\tilde{\ell}',\tilde{\ell}'')$ do not overlap, i.e., no index is present in both triples.

Now, borrowing the terminologies from round robin tournaments, we construct a schedule for a tournament with k teams where each match involve three teams. Let $T_{a,b}$ denote

a set of triples playing at the same round, indexed by two integers $a \in \{3, \ldots, 2k-3\}$ and $b \in \{5, \ldots, 2k-1\}$. Hence, there are total $N = (2k-5)^2$ rounds.

Each round (a, b) consists of disjoint triples and is defined as

$$T_{a,b} \equiv \left\{ (\ell, \ell', \ell'') \in [k] \times [k] \times [k] \mid \ell < \ell' < \ell'', \ell + \ell' = a, \text{ and } \ell' + \ell'' = b \right\}.$$

We need to prove that (a) there is no missing triple; and (b) no team plays twice in a single round. First, for any ordered triple (ℓ, ℓ', ℓ''), there exists $a \in \{3, \ldots, 2k-3\}$ and $b \in \{5, \ldots, 2k-1\}$ such that $\ell + \ell' = a$ and $\ell' + \ell'' = b$. This proves that all ordered triples are covered by the above construction. Next, given a pair (a, b), no two triples in $T_{a,b}$ can share the same team. Suppose there exists two distinct ordered triples (ℓ, ℓ', ℓ'') and $(\tilde{\ell}, \tilde{\ell}', \tilde{\ell}'')$ both in $T_{a,b}$, and one of the triples are shared. Then, from the two equations $\ell + \ell' = \tilde{\ell} + \tilde{\ell}' = a$ and $\ell' + \ell'' = \tilde{\ell}' + \tilde{\ell}'' = b$, it follows that all three indices must be the same, which is a contradiction. This proves the desired claim for ordered triples.

One caveat is that we wanted to cover the whole $[k] \times [k] \times [k]$, and not just the ordered triples. In the above construction, for example, a triple $(3, 2, 1)$ does not appear. This can be resolved by simply taking all $T_{a,b}$'s from the above construction, and make 6 copies of each round, and permuting all the triples in each copy according to the same permutation over $\{1, 2, 3\}$. This increases the total rounds to $N = 6(2k-5)^2 \leq 24k^2$. Note that $|T_{a,b}| \leq \lfloor k/3 \rfloor$ since no item can be in more than one triple.

Appendix D. Proof of Estimating Approximate Low-rank Matrices in Corollary 9

We follow closely the proof of a similar corollary in Negahban and Wainwright (2012). First fix a threshold $\tau > 0$, and set $r = \max\{j | \sigma_j(\Theta^*) > \tau\}$. With this choice of r, we have

$$\sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) = \tau \sum_{j=r+1}^{\min\{d_1, d_2\}} \frac{\sigma_j(\Theta^*)}{\tau} \leq \tau \sum_{j=r+1}^{\min\{d_1, d_2\}} \left(\frac{\sigma_j(\Theta^*)}{\tau} \right)^q \leq \tau^{1-q} \rho_q.$$

Also, since $r\tau^q \leq \sum_{j=1}^r \sigma_j(\Theta^*)^q \leq \rho_q$, it follows that $\sqrt{r} \leq \sqrt{\rho_q}\tau^{-q/2}$. Using these bounds, Eq. (35) is now

$$\left\|\widehat{\Theta} - \Theta\right\|_{\mathrm{F}}^2 \leq \underbrace{288\sqrt{2} c_0 e^{4\alpha} d_1 d_2 \lambda_0}_{=A} \left(\sqrt{\rho_q}\tau^{-q/2} \left\|\widehat{\Theta} - \Theta\right\|_{\mathrm{F}} + \tau^{1-q} \rho_q \right).$$

With the choice of $\tau = A$ and due to the fact that $x^2 \leq bx + c$ implies $x \leq (b + \sqrt{b^2 + 4c})/2$ we have,

$$\left\|\widehat{\Theta} - \Theta\right\|_{\mathrm{F}} \leq 2\sqrt{\rho_q} A^{(2-q)/2}.$$

Appendix E. Proof of the Information-theoretic Lower Bound in Theorem 10

The proof uses information-theoretic methods which reduces the estimation problem to a multiway hypothesis testing problem. To prove a lower bound on the expected error, it

suffices to prove that,

$$\sup_{\Theta^* \in \Omega_\alpha} \mathbb{P}\left\{\left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}}^2 \geq \frac{\delta^2}{4}\right\} \geq \frac{1}{2}. \tag{157}$$

To prove the above claim, we follow the standard recipe of constructing a packing in Ω_α. Consider a family $\{\Theta^{(1)}, \ldots, \Theta^{(M(\delta))}\}$ of $d_1 \times d_2$ dimensional matrices contained in Ω_α satisfying $\left\|\!\left\|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right\|\!\right\|_{\mathrm{F}} \geq \delta$ for all $\ell_1, \ell_2, \in [M(\delta)]$. We will use M to refer to $M(\delta)$ for simplify the notation. Suppose we draw an index $L \in [M(\delta)]$ uniformly at random, and we are given direct observations σ_i as per MNL model with $\Theta^* = \Theta^{(L)}$ on a randomly chosen set of k items S_i for each user $i \in [d_1]$. It follows from triangular inequality that

$$\sup_{\Theta^* \in \Omega_\alpha} \mathbb{P}\left\{\left\|\!\left\|\widehat{\Theta} - \Theta^*\right\|\!\right\|_{\mathrm{F}}^2 \geq \frac{\delta^2}{4}\right\} \geq \mathbb{P}\left\{\widehat{L} \neq L\right\}, \tag{158}$$

where $\widehat{L}$ is the resulting best estimate of the multiway hypothesis testing on L. The generalized Fano's inequality gives

$$\mathbb{P}\left\{\widehat{L} \neq L | S(1), \ldots, S(d_1)\right\} \geq 1 - \frac{I(\widehat{L}; L) + \log 2}{\log M} \tag{159}$$

$$\geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1, \ell_2 \in [M]} D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) + \log 2}{\log M}, \tag{160}$$

where $D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$ denotes the Kullback-Leibler divergence between the distributions of the partial rankings $\mathbb{P}\left\{\sigma_1, \ldots, \sigma_{d_1} | \Theta^{(\ell_1)}, S(1), \ldots, S(d_1)\right\}$ and $\mathbb{P}\left\{\sigma_1, \ldots, \sigma_{d_1} | \Theta^{(\ell_2)}, S(1), \ldots, S(d_1)\right\}$. The second inequality follows from a standard technique, which we repeat here for completeness. Let $\Sigma = \{\sigma_1, \ldots, \sigma_{d_1}\}$ denote the observed outcome of comparisons. Since L–$\Theta^{(L)}$–Σ–$\widehat{L}$ form a Markov chain, the data processing inequality gives $I(\widehat{L}; L) \leq I(\Sigma; L)$. For simplicity, we drop the conditioning on the set of alternatives $\{S(1), \ldots, S(d_1)\}$, and and let $p(\cdot)$ denotes joint, marginal, and conditional distribution of respective random variables. It follows that

$$
\begin{aligned}
I(\Sigma; L) &= \sum_{\ell \in [M], \Sigma} p(\Sigma | \ell) \frac{1}{M} \log \frac{p(\ell, \Sigma)}{p(\ell) p(\Sigma)} \\
&= \frac{1}{M} \sum_{\ell \in [M]} \sum_\Sigma p(\Sigma | \ell) \log \frac{p(\Sigma | \ell)}{\frac{1}{M} \sum_{\ell'} p(\Sigma | \ell')} \\
&\leq \frac{1}{M^2} \sum_{\ell, \ell' \in [M]} \sum_\Sigma p(\Sigma | \ell) \log \frac{p(\Sigma | \ell)}{p(\Sigma | \ell')} \\
&= \frac{1}{M^2} \sum_{\ell, \ell' \in [M]} D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}), \tag{161}
\end{aligned}
$$

where the first inequality follows from Jensen's inequality. To compute the KL-divergence, recall that from the RUM interpretation of the MNL model (see Section 1), one can generate sample rankings Σ by drawing random variables with exponential distributions with mean

$e^{\Theta_{ij}^*}$'s. Precisely, let $X^{(\ell)} = [X_{ij}^{(\ell)}]_{i\in[d_1], j\in S_i}$ denote the set of random variables, where $X_{ij}^{(\ell)}$ is drawn from the exponential distribution with mean $e^{-\Theta_{ij}^{(\ell)}}$. The MNL ranking follows by ordering the alternatives in each S_i according to this $\{X_{ij}^{(\ell)}\}_{j\in S_i}$ by ranking the smaller ones on the top. This forms a Markov chain L–$X^{(L)}$–Σ, and the standard data processing inequality gives

$$
\begin{aligned}
D_{\mathrm{KL}}(\Theta^{(\ell_1)}\|\Theta^{(\ell_2)}) &\leq D_{\mathrm{KL}}(X^{(\ell_1)}\|X^{(\ell_2)}) &(162)\\
&= \sum_{i\in[d_1]}\sum_{j\in S_i}\left\{ e^{\Theta_{ij}^{(\ell_1)}-\Theta_{ij}^{(\ell_2)}} - (\Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)}) - 1 \right\} &(163)\\
&\leq \frac{e^{2\alpha}}{4\alpha^2}\sum_{i\in[d_1]}\sum_{j\in S_i}(\Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)})^2 , &(164)
\end{aligned}
$$

where the last inequality follows from the fact that $e^x - x - 1 \leq (e^{2\alpha}/(4\alpha^2))x^2$ for any $x \in [-2\alpha, 2\alpha]$. Taking expectation over the randomly chosen set of alternatives,

$$
\mathbb{E}_{S(1),\ldots,S(d_1)}[D_{\mathrm{KL}}(\Theta^{(\ell_1)}\|\Theta^{(\ell_2)})] \leq \frac{e^{2\alpha}\,k}{4\,\alpha^2\,d_2}\left\|\!\left\|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right\|\!\right\|_{\mathrm{F}}^2 . \tag{165}
$$

Combined with (160), we get that

$$
\begin{aligned}
\mathbb{P}\left\{\widehat{L}\neq L\right\} &= \mathbb{E}_{S(1),\ldots,S(d_1)}[\mathbb{P}\left\{\widehat{L}\neq L | S(1),\ldots,S(d_1)\right\}] &(166)\\
&\geq 1 - \frac{\binom{M}{2}^{-1}\sum_{\ell_1,\ell_2\in[M]}(e^{2\alpha}k/(4\alpha^2 d_2))\left\|\!\left\|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right\|\!\right\|_{\mathrm{F}}^2 + \log 2}{\log M} , &(167)
\end{aligned}
$$

The remainder of the proof relies on the following probabilistic packing.

Lemma 34 *Let $d_2 \geq d_1 \geq 607$ be positive integers. Then for each $r \in \{1,\ldots,d_1\}$, and for any positive $\delta > 0$ there exists a family of $d_1 \times d_2$ dimensional matrices $\{\Theta^{(1)},\ldots,\Theta^{(M(\delta))}\}$ with cardinality $M(\delta) = \lfloor (1/4)\exp(rd_2/576)\rfloor$ such that each matrix is rank r and the following bounds hold:*

$$
\begin{aligned}
\left\|\!\left\|\Theta^{(\ell)}\right\|\!\right\|_{\mathrm{F}} &\leq \delta , \text{ for all } \ell \in [M] &(168)\\
\left\|\!\left\|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right\|\!\right\|_{\mathrm{F}} &\geq \delta , \text{ for all } \ell_1, \ell_2 \in [M] &(169)\\
\Theta^{(\ell)} &\in \Omega_{\tilde{\alpha}} , \text{ for all } \ell \in [M] , &(170)
\end{aligned}
$$

with $\tilde{\alpha} = (8\delta/d_2)\sqrt{2\log d}$ for $d = (d_1 + d_2)/2$.

We omit the proof of the above lemma since it is similar to that of Lemma 25. Suppose $\delta \leq \alpha d_2/(8\sqrt{2\log d})$ such that the matrices in the packing set are entry-wise bounded by α, then the above lemma implies that $\left\|\!\left\|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right\|\!\right\|_{\mathrm{F}}^2 \leq 4\delta^2$, which gives

$$
\mathbb{P}\left\{\widehat{L}\neq L\right\} \geq 1 - \frac{\frac{e^{2\alpha}k\delta^2}{\alpha^2 d_2} + \log 2}{\frac{rd}{576} - 2\log 2} \geq \frac{1}{2} ,
$$

where the last inequality holds for $\delta^2 \leq (\alpha^2 d_2/(e^{2\alpha}k))((rd/1152) - 2\log 2)$. If we assume $rd \geq 3195$ for simplicity, this bound on δ can be simplified to $\delta \leq \alpha e^{-\alpha}\sqrt{r\,d_2\,d/(2304\,k)}$. Together with (157) and (158), this proves that for all $\delta \leq \min\{\alpha d_2/(8\sqrt{2\log d})$, $\alpha e^{-\alpha}\sqrt{r\,d_2\,d/(2304\,k)}\}$,

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\Big[\big\|\big\|\widehat{\Theta} - \Theta^*\big\|\big\|_{\mathrm{F}}\Big] \geq \frac{\delta}{4} \,.$$

Choosing δ appropriately to maximize the right-hand side finishes the proof of the desired claim.

Appendix F. Proof of Pairwise Rank Breaking in Theorem 11

Analogous to Section C, we define the gradient $\nabla\mathcal{L}(\Theta)$ as $\nabla_{ij}\mathcal{L} = \frac{\partial \mathcal{L}(\Theta)}{\partial \Theta_{ij}}$ and $\Delta \equiv \widehat{\Theta} - \Theta^*$, and provide two main technical lemmas.

Lemma 35 *If $\lambda \geq 2\|\|\nabla\mathcal{L}(\Theta^*)\|\|_2$, then we have,*

$$\|\|\Delta\|\|_{\mathrm{nuc}} \leq 4\sqrt{2r}\|\|\Delta\|\|_{\mathrm{F}} + 4 \sum_{j=\rho+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^*) \,, \tag{171}$$

for all $\rho \in [\min\{d_1, d_2\}]$.

Proof This follows from the proof of Lemma 27, which only depends on the convexity of $\mathcal{L}(\Theta)$. ∎

Lemma 36 *For any positive constant $c \geq 1$, if $k \leq \max\{d_1, d_2^2/d_1\}\log d$ and $d_1 \geq 4$ then with probability at least $1 - 2d^{-c}$,*

$$\|\|\nabla\mathcal{L}(\Theta^*)\|\|_2 \leq \sqrt{\frac{16(c+4)\log d}{k\,d_1^2}} \max\left\{\sqrt{\max\left\{\frac{1}{4}, \frac{d_1}{d_2}\right\}}, \frac{2}{3}\sqrt{\frac{2(c+4)\log d}{k}}\right\} \,. \tag{172}$$

The proof of this lemma is provided in Section F.1. We will simplify the above lemma by assuming, $2(c+4)\log d \leq k$ which implies the last term in RHS is less than equal to first term,

$$\frac{2}{3}\sqrt{\frac{2(4+c)\log d}{k}} \leq \sqrt{\frac{1}{4}} \,. \tag{173}$$

(173) simplifies (172) as,

$$\begin{aligned}
\|\|\nabla\mathcal{L}(\Theta^*)\|\|_2 &\leq \sqrt{\frac{16(c+4)\log d}{k\,d_1^2} \max\left\{\frac{1}{4}, \frac{d_1}{d_2}\right\}} \\
&\leq \sqrt{\frac{32d\,(c+4)\log d}{k\,d_1^2\,d_2}} \\
&\overset{(a)}{\leq} \sqrt{32(c+4)\lambda} \,,
\end{aligned} \tag{174}$$

68

where (a) is due to (41) .

For Lemma 35 and further proof of Theorem11 we want $\lambda \geq 2\||\nabla \mathcal{L}(\Theta)\||_2$, therefore we assume that,

$$\lambda \in [2\sqrt{32(c+4)}\lambda, c_p\lambda], \text{ for some } c_p \geq 2\sqrt{32(c+4)} \tag{175}$$

Similar to the k-wise ranking,we will divide the proof into two cases and each part we will prove that $\||\Delta\||_{\mathrm{F}}^2 \leq 36e^{2\alpha} c \lambda d_1 d_2 \||\Delta\||_{\mathrm{nuc}}$ with probability at least $1 - 2/d^c - 2/d^{2^{13}}$. We define a new constant μ as,

$$\mu = 16\alpha\sqrt{\frac{48 \, d_1 d_2^2 \log d}{k \, \min\{d_1, d_2\}}} . \tag{176}$$

Case 1: Assume $\mu\||\Delta\||_{\mathrm{nuc}} \leq \||\Delta\||_{\mathrm{F}}^2$.
Since $\mathcal{L}$ is a sum of a linear function of Θ and log-sum-exponential functions, which are convex, we know that $\mathcal{L}$ is a convex function of Θ. Therefore, by convexity and Taylor expansion we get,

$$\mathcal{L}(\hat{\Theta}) = \mathcal{L}(\Theta^*) - \langle\langle\nabla\mathcal{L}(\Theta^*), \Delta\rangle\rangle + \tag{177}$$

$$\frac{1}{2! \, d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{\Theta_{i,u_{i,m_1}}} e^{\Theta_{i,u_{i,m_2}}}}{\left(e^{\Theta_{i,u_{i,m_1}}} + e^{\Theta_{i,u_{i,m_2}}}\right)^2} \left(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\right)^2,$$

where $\Theta = a\Theta^* + (1-a)\hat{\Theta}$ for some $a \in [0,1]$ and $\mathcal{P}_0 = \{(i,j)|\ 1 \leq i < j \leq k\}$. We lower bound the final term in (177) as,

$$\frac{1}{2! \, d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{\Theta_{i,u_{i,m_1}}} e^{\Theta_{i,u_{i,m_2}}}}{\left(e^{\Theta_{i,u_{i,m_1}}} + e^{\Theta_{i,u_{i,m_2}}}\right)^2} \left(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\right)^2$$

$$\overset{(a)}{\geq} \frac{1}{2 \, d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{-\alpha}e^{\alpha}}{(e^{-\alpha} + e^{\alpha})^2} \left(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\right)^2$$

$$\geq \frac{1}{2 \, d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{-2\alpha}}{4} \left(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\right)^2, \tag{178}$$

where (a) is due to the fact that Δ_{ij}'s are upper and lower bounded by α and $-\alpha$ respectively. We can bound this term further according to the following Lemma.

Lemma 37 *For* $(4\log d)/9 \leq k \leq \max\{d_1, d_2^2/d_1\}\log d$, *with probability at least* $1 - 2d^{-2^{13}}$,

$$\frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \left(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\right)^2 \geq \frac{1}{3d_1 d_2}\||\Delta\||_{\mathrm{F}}^2, \tag{179}$$

for all $\Delta \in \mathcal{A}_p$ *where,*

$$\mathcal{A} = \left\{\Delta \in \mathbb{R}^{d_1 \times d_2} \ \Big| \ \||\Delta\||_\infty \leq 2\alpha, \ \sum_{j\in[d_2]} \Delta_{ij} = 0 \ , \text{ for all } i \in [d_2], \text{ and } \mu\||\Delta\||_{\mathrm{nuc}} \leq \||\Delta\||_{\mathrm{F}}^2\right\}. \tag{180}$$

The proof is given in Section F.2. Now using Lemma 37 and (178) with high probability we get,

$$\frac{1}{2!\,d_1\binom{k}{2}}\sum_{i=1}^{d_1}\sum_{(m_1,m_2)\in\mathcal{P}_0}\frac{e^{\Theta_{i,u_{i,m_1}}}e^{\Theta_{i,u_{i,m_2}}}}{\left(e^{\Theta_{i,u_{i,m_1}}}+e^{\Theta_{i,u_{i,m_2}}}\right)^2}\left(\Delta_{i,u_{i,m_1}}-\Delta_{i,u_{i,m_1}}\right)^2\geq\frac{e^{-2\alpha}}{24\,d_1\,d_2}\|\!|\Delta\|\!|_{\mathrm{F}}^2\,.$$

(181)

Incorporating the above inequality in (177) we obtain,

$$\frac{e^{-2\alpha}}{24\,d_1\,d_2}\|\!|\Delta\|\!|_{\mathrm{F}}^2\leq\mathcal{L}(\hat{\Theta})-\mathcal{L}(\Theta^*)+\langle\!\langle\nabla\mathcal{L}(\Theta^*),\Delta\rangle\!\rangle\,.$$

(182)

From the definition of $\hat{\Theta}$ we have $\mathcal{L}(\hat{\Theta})-\mathcal{L}(\Theta^*)\leq\lambda\left(\|\!|\Theta^*\|\!|_{\mathrm{nuc}}-\left\|\!\left|\hat{\Theta}\right|\!\right\|_{\mathrm{nuc}}\right)\leq\lambda\|\!|\Delta\|\!|_{\mathrm{nuc}}$, and we assume that $\lambda\geq2\sqrt{32(c+1)}\,\lambda$, so that $\lambda\geq2\|\!|\nabla\mathcal{L}(\Theta^*)\|\!|_2$ is true with a probability of at least $1-2d^{-c}$ from Lemma 36 . These give us the following with at least probability $1-2d^{-c}-2d^{-2^{13}}$.

$$\frac{e^{-2\alpha}}{24\,d_1\,d_2}\|\!|\Delta\|\!|_{\mathrm{F}}^2\leq\lambda\|\!|\Delta\|\!|_{\mathrm{nuc}}+\|\!|\nabla\mathcal{L}(\Theta^*)\|\!|_2\|\!|\Delta\|\!|_{\mathrm{nuc}}$$

$$\leq\frac{3\lambda}{2}\|\!|\Delta\|\!|_{\mathrm{nuc}}$$

(183)

which gives us,

$$\|\!|\Delta\|\!|_{\mathrm{F}}^2\leq36e^{2\alpha}\,\lambda\,d_1d_2\,\|\!|\Delta\|\!|_{\mathrm{nuc}}$$

$$\overset{(a)}{\leq}36e^{2\alpha}\,c_p\,\lambda\,d_1d_2\,\|\!|\Delta\|\!|_{\mathrm{nuc}}\,,$$

(184)

where (a) is due to the fact that $\lambda\leq c_p\lambda$.

Case 2: Assume $\|\!|\Delta\|\!|_{\mathrm{F}}^2\leq\mu\|\!|\Delta\|\!|_{\mathrm{nuc}}$.
Here we prove that $\mu\leq36\,e^{2\alpha}\,c_p\lambda\,d_1d_2$.

$$\frac{\mu}{36\,e^{2\alpha}\,c_p\lambda\,d_1d_2}\overset{(a)}{\leq}\frac{\alpha}{e^{2\alpha}}\times\frac{16\sqrt{48}}{72\sqrt{32(c+4)}}\times\sqrt{\frac{d_1d_2}{min\{d_1,d_2\}d}}$$

$$\overset{(b)}{\leq}1\times\frac{16\sqrt{48}}{72\sqrt{32\times4}}\times\sqrt{\frac{\max\{d_1,d_2\}}{d}}$$

$$\overset{(c)}{\leq}\sqrt{\frac{\max\{d_1,d_2\}}{2d}}$$

$$\overset{(d)}{\leq}1\,,$$

(185)

where (a) is by substituting μ, λ and c_p from (176), (41) and (175) respectively, (b) is because $x\leq e^x$ (c) is because $d=(\max\{d_1,d_2\}+\min\{d_1,d_2\})/2$.

Now combining the above result with (171) we get with probability at least $1-2d^{-c}-2d^{-2^{13}}$,

$$\frac{1}{d_1d_2}\|\!|\Delta\|\!|_{\mathrm{F}}^2\leq144\sqrt{2}e^{2\alpha}c_p\lambda\sqrt{r}\|\!|\Delta\|\!|_{\mathrm{F}}+144e^{2\alpha}c_p\lambda\sum_{j=\rho+1}^{\min\{d_1,d_2\}}\sigma_j(\Theta^*)\,.$$

(186)

F.1 Proof of Lemma 36

From definition of $\mathcal{L}(\Theta)$ in (39) we get,

$$\nabla\mathcal{L}_p(\Theta^*) = \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e_i \left(e_{l_i(m_1,m_2)} - e_{h_i(m_1,m_2)}\right)^T}{\exp\left(\Theta^*_{i,h_i(m_1,m_2)} - \Theta^*_{i,l_i(m_1,m_2)}\right) + 1} , \tag{187}$$

where $\mathcal{P}_0 = \{(i,j)| \quad 1 \leq i < j \leq k\}$. We use the matrix Bernstein inequality Tropp (2011) for the sum of independent matrices. Similar to Lemma 42, we can partition the set of all pairs $\mathcal{P}_0$ into $(k-1)$ sets $\{\mathcal{P}_a\}_{a\in[k-1]}$ of $k/2$ disjoint pairs each. Define $Y_a \equiv \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_a} \tilde{X}_{i,m_1,m_2}$, and

$$\tilde{X}_{i,m_1,m_2} \equiv \frac{\exp\left(\Theta^*_{i,l_i(m_1,m_2)}\right)}{\exp\left(\Theta^*_{i,h_i(m_1,m_2)}\right) + \exp\left(\Theta^*_{i,l_i(m_1,m_2)}\right)} e_i \left(e_{l_i(m_1,m_2)} - e_{h_i(m_1,m_2)}\right)^T ,$$

such that

$$\nabla\mathcal{L}_p(\Theta^*) = \frac{1}{d_1\binom{k}{2}} \sum_{a=1}^{k-1} \tilde{Y}_a . \tag{188}$$

For a fixed value of a, it is easy to see that $\tilde{X}_{i,m_1,m_2}$'s are independent. Further, we can easily show that $\mathbb{E}\left[\tilde{X}_{i,m_1,m_2}\right] = 0$, and $\|\tilde{X}_{i,m_1,m_2}\|_2 \leq \sqrt{2}$. We also have,

$$\mathbb{E}\left[\tilde{X}_{i,m_1,m_2}\tilde{X}^T_{i,m_1,m_2}\right]$$

$$\preceq 2\, e_i e_i^T\, \mathbb{E}\left[\mathbb{E}\left[\frac{\exp\left(\Theta^*_{i,l_i(m_1,m_2)}\right)^2}{\left(\exp\left(\Theta^*_{i,u_{i,m_1}}\right) + \exp\left(\Theta^*_{i,u_{i,m_2}}\right)\right)^2}\,\Bigg|\, u_{i,m_1}, u_{i,m_1}\right]\right]$$

$$\overset{(a)}{=} 2\, e_i e_i^T\, \mathbb{E}\left[\frac{\exp\left(\Theta^*_{iu_{i,m_1}}\right)\exp\left(\Theta^*_{iu_{i,m_2}}\right)}{\left(\exp\left(\Theta^*_{i,u_{i,m_1}}\right) + \exp\left(\Theta^*_{i,u_{i,m_2}}\right)\right)^2}\right]$$

$$\overset{(b)}{\preceq} \frac{1}{2} e_i e_i^T , \tag{189}$$

where we get (a) from the MNL model for the random choice of $l_i(m_1,m_2)$, (b) is due to the fact that $xy/(x+y)^2 \leq 1/4$ for all $x,y > 0$.

Let $p_{i,m_1,m_2} \equiv \dfrac{\left(\exp(\Theta^*_{i,u_{i,m_1}})e_{u_{i,m_1}} + \exp(\Theta^*_{i,u_{i,m_2}})e_{u_{i,m_2}} \right)}{\left(\exp(\Theta^*_{i,u_{i,m_1}}) + \exp(\Theta^*_{i,u_{i,m_2}}) \right)}$, then we have,

$$
\begin{aligned}
\mathbb{E}\left[\tilde{X}^T_{i,m_1,m_2} \tilde{X}_{i,m_1,m_2} \right] &= \mathbb{E}\left[(e_{h_i(m_1,m_2)} - p_{i,m_1,m_2})(e_{h_i(m_1,m_2)} - p_{i,m_1,m_2})^T \right] \\
&= \mathbb{E}\left[e_{h_i(m_1,m_2)} e^T_{h_i(m_1,m_2)} \right] - \mathbb{E}\left[p_{i,m_1,m_2} p^T_{i,m_1,m_2} \right] \\
&\overset{(a)}{\preceq} \mathbb{E}\left[e_{u_{i,m_1}} e^T_{u_{i,m_1}} + e_{u_{i,m_2}} e^T_{u_{i,m_2}} \right] \\
&= \frac{2}{d_2} \mathbf{I}_{d_2 \times d_2} ,
\end{aligned}
\tag{190}
$$

where (a) comes from the fact that $p_{i,m_1,m_2} p^T_{i,m_1,m_2}$ is a positive semi-definite matrix. Therefore using (189) and (190), we get

$$
\begin{aligned}
\sigma^2 &\equiv \left\{ \left\| \sum_{\substack{i \in [d_1] \\ (m_1,m_2) \in \mathcal{P}_a}} \mathbb{E}\left[\tilde{X}_{i,m_1,m_2} \tilde{X}^T_{i,m_1,m_2} \right] \right\|_2 , \left\| \sum_{\substack{i \in [d_1] \\ (m_1,m_2) \in \mathcal{P}_a}} \mathbb{E}\left[\tilde{X}^T_{i,m_1,m_2} \tilde{X}_{i,m_1,m_2} \right] \right\|_2 \right\} \\
&\leq k \max\left\{ \frac{1}{4}, \frac{d_1}{d_2} \right\} .
\end{aligned}
\tag{191}
$$

Define $\rho \equiv \max\{1/4, d_1/d_2\}$, then by the matrix Bernstein inequality Tropp (2011), $\forall \ a \in [k-1]$,

$$
\mathbb{P}\left(\left\| \tilde{Y}_a \right\|_2 > t \right) \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{k\rho + \sqrt{2}t/3} \right) ,
$$

which gives a tail probability of $2d^{-c}/(k-1)$ for the choice of

$$
t = \max\left\{ \sqrt{4k\rho\left((1+c)\log d + \log(k-1)\right)}, \frac{4\sqrt{2}((1+c)\log d + \log(k-1))}{3} \right\} .
\tag{192}
$$

For this choice of t, using union bound we can get the probabilistic bound on the derivative of log likelihood as,

$$
\begin{aligned}
\mathbb{P}\left(\|\nabla \mathcal{L}_p(\Theta^*)\|_2 \geq \frac{k-1}{d_1 \binom{k}{2}} t \right) &\leq \mathbb{P}\left(\sum_{a=1}^{k-1} \left\| \tilde{Y}_a \right\|_2 \geq (k-1)t \right) \\
&\overset{(a)}{\leq} \mathbb{P}\left(\max_{a \in [k-1]} \left\| \tilde{Y}_a \right\|_2 \geq t \right) \\
&\overset{(b)}{\leq} \sum_{a=1}^{k-1} \mathbb{P}\left(\left\| \tilde{Y}_a \right\|_2 \geq t \right) \\
&= 2 d^{-c} ,
\end{aligned}
\tag{193}
$$

where we obtain (a) by pigeon-hole principle which implies that among a set of numbers, there should be, at the very least one number greater or equal to the average of the set of numbers and (b) by union-bound. Assuming $k \leq \max\{d_1, d_2^2/d_1\} \log d$ and $d_1 \geq 4$, we have,

$$(c+1)\log d + \log(k-1) \leq (c+4)\log d , \tag{194}$$

from $\log(k-1) \leq \log\left(\max\{d_1, d_2^2/d_1\} \log d\right) \leq \log(((d_1^2 + d_2^2)\log d)/d_1) \leq \log\left((4\,d^2 \log d)/d_1\right)$ $\leq 3 \log d$. This proves the desired lemma.

F.2 Proof of Lemma 37

With a slight abuse of notation, we define $\tilde{H}$ as

$$\tilde{H}(\Delta) \;\; \equiv \;\; \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \left(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\right)^2 , \tag{195}$$

and provide a lower bound. The mean is easily computed as

$$\mathbb{E}\left[\tilde{H}(\Delta)\right] = \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \left[\frac{2}{d_2}\sum_{j\in[d_2]}\Delta_{ij}^2 - \frac{2}{d_2^2}\sum_{j\in[d_2]}\Delta_{ij}\sum_{j'\in[d_2]}\Delta_{ij'}\right]$$
$$= \frac{2}{d_1 d_2}\|\|\Delta\|\|_{\mathrm{F}}^2 , \tag{196}$$

where we used the fact that $\sum_j \Delta_{ij} = 0$. We want to upper bound the probability that $\tilde{H}(\Delta) \leq \frac{1}{3d_1 d_2}\|\|\Delta\|\|_{\mathrm{F}}^2$ for some $\Delta \in \mathcal{A}$. As in the case of k-wise ranking we using the following peeling argument used in (Negahban and Wainwright, 2012, Lemma 3), Van De Geer (2000). The strategy is to split this above event as union of many event events as follows. We construct the following family of subsets $\{\tilde{\mathcal{S}}_\ell\}$ such that $\mathcal{A} \subseteq \cup_{\ell=1}^{\infty}\tilde{\mathcal{S}}_\ell$ and,

$$\tilde{\mathcal{S}}_\ell = \left\{\Delta \in \mathbb{R}^{d_1\times d_2} \;\middle|\; \|\|\Delta\|\|_{\infty} \leq 2\alpha, \; \beta^{\ell-1}\mu \leq \|\|\Delta\|\|_{\mathrm{F}} \leq \beta^\ell\mu, \right.$$
$$\left. \sum_{j\in[d_2]} \Delta_{ij} = 0 \text{ for all } i \in [d_2], \text{ and } \|\|\Delta\|\|_{\mathrm{nuc}} \leq \beta^{2\ell}\mu\right\} , \tag{197}$$

where $\beta = \sqrt{10/9}$ and $\ell \in \{1, 2, 3, \ldots\}$. This is true since, for any $\Delta \in \mathcal{A}$, $\|\|\Delta\|\|_{\mathrm{F}}^2 \geq \mu\|\|\Delta\|\|_{\mathrm{nuc}}$ and this implies $\|\|\Delta\|\|_{\mathrm{F}}^2 \geq \mu\|\|\Delta\|\|_{\mathrm{F}}$ (or, $\|\|\Delta\|\|_{\mathrm{F}} \geq \mu$). Also note that,

$$\tilde{H}(\Delta) \leq \frac{1}{3d_1 d_2}\|\|\Delta\|\|_{\mathrm{F}}^2 \implies \frac{2}{d_1 d_2}\|\|\Delta\|\|_{\mathrm{F}}^2 - \tilde{H}(\Delta) \geq \frac{5}{3d_1 d_2}\|\|\Delta\|\|_{\mathrm{F}}^2$$
$$\implies \left(\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)\right) \geq \frac{5}{3d_1 d_2}\|\|\Delta\|\|_{\mathrm{F}}^2 . \tag{198}$$

Therefore using union bound we get,

$$
\begin{aligned}
&\mathbb{P}\Big(\exists \, \Delta \in \mathcal{A} \,\, s.t. \,\, \tilde{H}(\Delta) \leq \frac{1}{3d_1 d_2}\|\!|\Delta|\!\|_{\mathrm{F}}^2\Big) \\
&\leq \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{\Delta \in \tilde{\mathcal{S}}_\ell} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{5}{3d_1 d_2}\|\!|\Delta|\!\|_{\mathrm{F}}^2\right) \\
&\overset{(a)}{\leq} \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{\Delta \in \tilde{\mathcal{S}}_\ell} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{3}{2d_1 d_2}(\beta^\ell \mu)^2\right) \\
&\overset{(b)}{\leq} \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{\Delta \in \tilde{\mathcal{B}}(\beta^\ell \mu)} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{3}{2d_1 d_2}(\beta^\ell \mu)^2\right) ,
\end{aligned}
\tag{199}
$$

where $\tilde{\mathcal{B}}(\mathcal{D})$ is defined as,

$$
\tilde{\mathcal{B}}(D) = \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2} \,\middle|\, \|\!|\Delta|\!\|_\infty \leq 2\alpha, \,\, \|\!|\Delta|\!\|_{\mathrm{F}} \leq D, \right.
$$
$$
\left. \sum_{j \in [d_2]} \Delta_{ij} = 0 \text{ for all } i \in [d_2], \text{ and } \mu\|\!|\Delta|\!\|_{\mathrm{nuc}} \leq D^2 \right\} ,
\tag{200}
$$

and $(a)$is true because for $\Delta \in \tilde{\mathcal{S}}_l$,

$$
\frac{5}{3d_1 d_2}\|\!|\Delta|\!\|_{\mathrm{F}}^2 \geq \frac{5}{3d_1 d_2}(\beta^{\ell-1}\mu)^2 = \frac{3}{2d_1 d_2}(\beta^\ell \mu)^2 ,
\tag{201}
$$

and (b) is true because $\tilde{\mathcal{S}}_\ell \subset \tilde{\mathcal{B}}(\beta^\ell \mu)$. Now we use following lemma to upper bound (199).

Lemma 38 *For $4(\log d)/3 \leq k \leq d^2 \log d$,*

$$
\mathbb{P}\left(\sup_{\Delta \in \tilde{\mathcal{B}}(D)} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{3}{2d_1 d_2}D^2\right) \leq \exp\left(\frac{-kD^4}{2048 \, \alpha^4 \, d_1 d_2^2}\right)
\tag{202}
$$

74

Proof has been relegated to Section F.3. Now by (199) and Lemma 38 we get,

$$
\begin{aligned}
\mathbb{P}\left(\exists \, \Delta \in \mathcal{A} \text{ s.t. } \tilde{H}(\Delta) \le \frac{1}{3d_1 d_2} \|\!|\Delta|\!\|_{\mathrm{F}}^2\right) &\le \sum_{\ell=1}^{\infty} \exp\left(\frac{-k \left(\beta^\ell \, \mu\right)^4}{2048 \, \alpha^4 \, d_1 d_2^2}\right) \\
&\overset{(a)}{\le} \sum_{\ell=1}^{\infty} \exp\left(\frac{-2^{13} \, 9 \, \beta^{4\ell} \, d_1 d_2^2 \log^2 d}{k \, \min^2\{d_1, d_2\}}\right) \\
&\overset{(b)}{\le} \sum_{\ell=1}^{\infty} \exp\left(\frac{-2^{13} \, 9 \, 4\ell \times \frac{1}{36} \, d_1 d_2^2 \log^2 d}{k \, \min^2\{d_1, d_2\}}\right) \\
&\overset{(c)}{\le} \sum_{\ell=1}^{\infty} \exp\left(-2^{13} \, \ell \, \log d\right) \\
&= \sum_{\ell=1}^{\infty} \left(\frac{1}{d^{2^{13}}}\right)^\ell \\
&\overset{(d)}{=} \frac{1/d^{2^{13}}}{1 - 1/d^{2^{13}}} \\
&\overset{(e)}{\le} \frac{2}{d^{2^{13}}} \, ,
\end{aligned}
\tag{203}
$$

where we get (a) by substituting μ from (176), (b) by the fact that for $\beta = \sqrt{10/9}$ and $x \ge 1$, $\beta^x \ge x \log \beta \ge x(\beta - 1) \ge x/32$, (c) is obtained by assuming $k \le \max\{d_1, d_2^2/d_1\} \log d$, we get (d) because we are summing an infinite geometric sequence with common ratio of $1/d^{2^{13}}$ and (e) is because for $d \ge 2$, $1/d^{2^{13}}$ is less than $1/2$.

F.3 Proof of Lemma 38

With a slight abuse of notations, let $\tilde{Z} \equiv \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \left(\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)\right)$. Notice that $\tilde{Z}$ is a function of $d_1 k$ random variables, $\{u_{i,\ell}\}_{i \in [d_1], \ell \in [k]}$. We apply the McDiarmid's bounded differences inequality. Let $\tilde{Z}_1$ and $\tilde{Z}_2$ be two realizations of $\tilde{Z}$ where value of only one random variable $u_{i',\ell'}$ is changed to $u'_{i',\ell'}$. Also with a little more abuse of notation the two realizations of $\tilde{H}(\Delta)$ are written as $\tilde{H}(\Delta', u_{1,1}, \ldots, u_{i',\ell'}, \ldots, u_{d_1,k})$ and $\tilde{H}(\Delta', u_{1,1}, \ldots, u'_{i',\ell'}, \ldots, u_{d_1,k})$. We let Δ^* be the maximizer of $\max\{\tilde{Z}_1, \tilde{Z}_2\}$. Maximum

absolute difference between them is upper bounded as follows,

$$
\begin{aligned}
&|\tilde{Z}_1 - \tilde{Z}_2| \\
&= \left| \max_{\Delta \in \tilde{\mathcal{B}}(D)} \left(\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta, u_{1,1}, \ldots, u_{i',\ell'}, \ldots, u_{d_1,k}) \right) - \right. \\
&\qquad \left. \sup_{\Delta' \in \tilde{\mathcal{B}}(D)} \left(\mathbb{E}\left[\tilde{H}(\Delta')\right] - \tilde{H}(\Delta', u_{1,1}, \ldots, u'_{i',\ell'}, \ldots, u_{d_1,k}) \right) \right| \\
&\overset{(a)}{\leq} \left| \left(\mathbb{E}\left[\tilde{H}(\Delta^*)\right] - \tilde{H}(\Delta^*, u_{1,1}, \ldots, u_{i',\ell'}, \ldots, u_{d_1,k}) \right) - \right. \\
&\qquad \left. \left(\mathbb{E}\left[\tilde{H}(\Delta^*)\right] - \tilde{H}(\Delta^*, u_{1,1}, \ldots, u'_{i',\ell'}, \ldots, u_{d_1,k}) \right) \right| \\
&\leq \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \left| \tilde{H}(\Delta, u_{1,1}, \ldots, u_{i',\ell'}, \ldots, u_{d_1,k}) - \tilde{H}(\Delta, u_{1,1}, \ldots, u'_{i',\ell'}, \ldots, u_{d_1,k}) \right| \\
&\overset{(b)}{\leq} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \left| \frac{1}{d_1 \binom{k}{2}} \sum_{\ell \neq \ell'} \left(\Delta_{i', u_{i',\ell}} - \Delta_{i', u_{i',\ell'}} \right)^2 - \left(\Delta_{i', u_{i',\ell}} - \Delta_{i', u'_{i',\ell'}} \right)^2 \right| \\
&\overset{(c)}{\leq} \frac{1}{d_1 \binom{k}{2}} (k-1)(4\alpha)^2 = \frac{32\alpha^2}{d_1 k}.
\end{aligned}
\tag{204}
$$

where (a) follows from the fact that Δ^* is maximizer of $\max\{\tilde{Z}_1, \tilde{Z}_2\}$, (b) is due to the fact that the terms which change because of $u'_{i',\ell'}$ are the $k-1$ difference square terms between $\Delta_{i u_{i',\ell \neq \ell'}}$ and $\Delta_{i, u_{i',\ell'}}$ and (c) is because maximum and minimum value of difference square terms are $(4\alpha)^2$ and 0 respectively. Using McDiarmid's bounded differences inequality we get,

$$
\mathbb{P}\{\tilde{Z} - \mathbb{E}\left[\tilde{Z}\right] \geq \epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{d_1 k \left(\frac{32\alpha^2}{d_1 k}\right)^2} \right),
\tag{205}
$$

because of (204) and the fact that there are $d_1 k$ random variables. We upper bound $\mathbb{E}\left[\tilde{Z}\right]$ as follows.

$$
\begin{aligned}
\mathbb{E}\left[\tilde{Z}\right] &= \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2) \in \mathcal{P}_0} \mathbb{E}\left[\left(\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}\right)^2\right] - \left(\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}\right)^2 \\
&\overset{(a)}{\leq} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2) \in \mathcal{P}_0} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}\right)^2 \\
&\overset{(b)}{\leq} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{a=1}^{k-1} \sum_{(m_1,m_2) \in \mathcal{P}_a} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}\right)^2 \\
&\overset{(c)}{\leq} \sum_{a=1}^{k-1} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2) \in \mathcal{P}_a} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}\right)^2,
\end{aligned}
\tag{206}
$$

where (a) is by standard symmetrization technique as used in k-wise ranking and $\{\xi_{i,m_1,m_2}\}_{i \in [d_1],\, m_1,m_2 \in [k]}$ are i.i.d. Rademacher variables, (b) is due to the fact that we can partition set of all pairs into $k - 1$ independent sets as in (188) and (c) is because of fact that supremum of sum is less than or equal to sum of supremum and the linearity of expectation. Since $|\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}| \leq 4\alpha$, we can use Ledoux-Talagrand contraction inequality Ledoux and Talagrand (2013) on (206) to get,

$$
\begin{aligned}
E[\tilde{Z}] &\leq \sum_{a=1}^{k-1} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2) \in \mathcal{P}_a} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}\right)^2 \\
&\leq \sum_{a=1}^{k-1} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2) \in \mathcal{P}_a} 4\alpha\, 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\,u_{i,m_1}} - \Delta_{i,\,u_{i,m_2}}\right) \\
&\overset{(a)}{\leq} \sum_{a=1}^{k-1} \frac{8\alpha}{d_1 \binom{k}{2}} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \langle\!\langle \sum_{i=1}^{d_1} \sum_{(m_1,m_2) \in \mathcal{P}_a} \tilde{W}_{i,m_1,m_2}, \Delta \rangle\!\rangle \\
&\overset{(b)}{\leq} \sum_{a=1}^{k-1} \frac{8\alpha}{d_1 \binom{k}{2}} \mathbb{E}\left[\left\|\!\left\|\!\left\| \sum_{i=1}^{d_1} \sum_{(m_1,m_2) \in \mathcal{P}_a} \tilde{W}_{i,m_1,m_2} \right\|\!\right\|\!\right\|_2\right] \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \|\!\|\!\|\Delta\|\!\|\!\|_{\mathrm{nuc}},
\end{aligned}
\tag{207}
$$

where we get (a) by putting $\tilde{W}_{i,m_1,m_2} = \tilde{\xi}_{i,m_1,m_2} e_i (e_{u_{i,m_1}} - e_{u_{i,m_2}})^T$ and (b) is due to Hölder's inequality $(\langle\!\langle x, y \rangle\!\rangle \leq \|\!\|\!\|x\|\!\|\!\|_2 \|\!\|\!\|y\|\!\|\!\|_{\mathrm{nuc}})$. Now we use Bernstein's inequality Tropp (2011) to up-perbound the above expectation terms. First fix a to value in $[k-1]$. We can easily show that $\tilde{W}_{i,m_1,m_2}$ is zero mean and,

$$
\left\|\!\left\|\!\left\| \tilde{W}_{i,m_1,m_2} \right\|\!\right\|\!\right\|_2 \leq \sqrt{2}.
\tag{208}
$$

We also get,

$$\mathbb{E}\left[\tilde{W}_{i,m_1,m_2}\tilde{W}_{i,m_1,m_2}^T\right] = 2e_i e_i^T \mathbb{E}\left[1 - e_{u_{i,m_1}}^T e_{u_{i,m_2}}\right]$$

$$\preceq e_i e_i^T \left(2 - \frac{2}{d_2}\right)$$

$$\preceq 2e_i e_i^T , \tag{209}$$

and,

$$\mathbb{E}\left[\tilde{W}_{i,m_1,m_2}^T \tilde{W}_{i,m_1,m_2}\right] = \mathbb{E}\left[2e_{u_{i,m_1}}e_{u_{i,m_1}}^T - 2e_{u_{i,m_1}}e_{u_{i,m_2}}^T\right]$$

$$\preceq \frac{2}{d_2}\mathbf{I}_{d_2 \times d_2} - \frac{2}{d_2^2}\mathbf{1}\mathbf{1}_{d_2 \times d_2}$$

$$\preceq \frac{2}{d_2}\mathbf{I}_{d_2 \times d_2} . \tag{210}$$

Therefore, using (209) and (210), the standard deviation of $\sum_{(i,m_1,m_2)} Z_{i,m_2,m_2}$ is,

$$\sigma^2 = \max\left\{\left\|\left\|\left\|\sum_{\substack{i\in[d_1]\\(m_1,m_2)\in\mathcal{P}_a}} \mathbb{E}\left[\tilde{W}_{i,m_2,m_2}\tilde{W}_{i,m_2,m_2}^T\right]\right\|\right\|\right\|_2, \left\|\left\|\left\|\sum_{\substack{i\in[d_1]\\(m_1,m_2)\in\mathcal{P}_a}} \mathbb{E}\left[\tilde{W}_{i,m_2,m_2}^T\tilde{W}_{i,m_2,m_2}\right]\right\|\right\|\right\|_2\right\}$$

$$\leq \max\left\{\frac{d_1 k}{2}\frac{2}{d_1}\|\|\mathbf{I}\|\|_2, \frac{d_1 k}{2}\frac{2}{d_2}\|\|\mathbf{I}\|\|_2\right\}$$

$$= \frac{kd_1}{\min\{d_1, d_2\}} . \tag{211}$$

By matrix Bernstein inequality Tropp (2011), $\forall \ a \in [k-1]$,

$$\mathbb{P}\left(\left\|\left\|\left\|\sum_{i\in[d_1]}\sum_{(m_1,m_2)\in\mathcal{P}_a} \tilde{W}_{i,m_2,m_2}\right\|\right\|\right\|_2 > t\right) \leq (d_1 + d_2)\exp\left(\frac{-t^2/2}{2kd_1/\min\{d_1, d_2\} + \sqrt{2}t/3}\right) ,$$

which gives a tail probability of $2d^{-c_1}$ for the choice of

$$t = \max\left\{\sqrt{\frac{8kd_1\left((1+c_1)\log d\right)}{\min\{d_1, d_2\}}}, \frac{4\sqrt{2}\left((1+c_1)\log d\right)}{3}\right\}$$

$$= \sqrt{\frac{8kd_1\left((1+c_1)\log d\right)}{\min\{d_1, d_2\}}}, \text{when } k \geq 4(c_1 + 1)\log d/9 . \tag{212}$$

Therefore $\forall \ a \in [k-1]$,

$$\mathbb{E}\left[\left\|\left\|\left\|\sum_{i=1}^{d_1}\sum_{(m_1,m_2)\in\mathcal{P}_a} \tilde{W}_{i,m_2,m_2}\right\|\right\|\right\|_2\right] \leq \sqrt{\frac{8kd_1\left((1+c_1)\log d\right)}{\min\{d_1, d_2\}}} + \frac{2}{d^{c_1}}\frac{\sqrt{2}d_1 k}{2} , \tag{213}$$

because from (208) we get $\left\lVert\left\lvert\sum_{\substack{i\in[d_1]\\(m_1,m_2)\in\mathcal{P}_a}}\tilde{W}_{i,m_2,m_2}\right\rvert\right\rVert_2 \leq \sum_{\substack{i\in[d_1]\\(m_1,m_2)\in\mathcal{P}_a}}\left\lVert\left\lvert\tilde{W}_{i,m_2,m_2}\right\rvert\right\rVert_2 \leq \frac{d_1k}{2(\sqrt{2})}$.
From (207) and (213), putting $c_1 = 2$, we get,

$$
\begin{aligned}
\mathbb{E}\left[\tilde{Z}\right] &\leq \sum_{a=1}^{k-1}\frac{8\alpha}{d_1\binom{k}{2}}\left(\sqrt{\frac{24\,kd_1\log d}{\min\{d_1,d_2\}}} \;+\; \frac{\sqrt{2}d_1k}{d^2}\right)\sup_{\Delta\in\tilde{\mathcal{B}}(D)}\lVert\lvert\Delta\rvert\rVert_{\text{nuc}}\\
&\overset{(a)}{\leq} 8\alpha\left(2\sqrt{\frac{24\,\log d}{k\,d_1\min\{d_1,d_2\}}} \;+\; \frac{2\sqrt{2}}{d^2}\right)\frac{D^2}{\mu}\\
&\overset{(b)}{\leq} 16\alpha\sqrt{\frac{48\log d}{k\,d_1\min\{d_1,d_2\}}}D^2\frac{1}{16\alpha}\sqrt{\frac{k\,\min\{d_1,d_2\}}{48d_1d_2^2\log d}}\\
&= \frac{D^2}{d_1d_2}\;,
\end{aligned}
\tag{214}
$$

where (a) is obtained because of (200) which gives $\sup_{D\in\mathcal{B}(D)}\lVert\lvert\Delta\rvert\rVert_{\text{nuc}} \leq D^2/\mu$ and (b) can be got by assuming $k \leq d^2\log d$. Using the above bound in (205) we get,

$$
\mathbb{P}\{\tilde{Z} - D^2/(d_1d_2) \geq \epsilon\} \leq \mathbb{P}\{\tilde{Z} - \mathbb{E}\left[\tilde{Z}\right] \geq \epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{d_1k\left(\frac{32\alpha^2}{d_1k}\right)^2}\right)\;,
\tag{215}
$$

and using $\epsilon = D^2/(2d_1d_2)$ will get us the required bound.

Appendix G. Proof of Bundled Choices Theorem 13

We use similar notations and techniques as the proof of Theorem 7 in Appendix C. From the definition of $\mathcal{L}(\Theta)$ in Eq. (51), we have for the true parameter Θ^*, the gradient evaluated at the true parameter is

$$
\nabla\mathcal{L}(\Theta^*) = -\frac{1}{n}\sum_{i=1}^{n}(e_{u_i}e_{v_i}^T - p_i)\;,
\tag{216}
$$

where p_i denotes the conditional probability of the MNL choice for the i-th sample. Precisely, $p_i = \sum_{j_1\in S_i}\sum_{j_2\in T_i} p_{j_1,j_2|S_i,T_i}e_{j_1}e_{j_2}^T$ where $p_{j_1,j_2|S_i,T_i}$ is the probability that the pair of items (j_1,j_2) is chosen at the i-th sample such that $p_{j_1,j_2|S_i,T_i} \equiv \mathbb{P}\{(u_i,v_i) = (j_1,j_2)|S_i,T_i\} = e^{\Theta^*_{j_1,j_2}}/(\sum_{j_1'\in S_i,j_2'\in T_i}e^{\Theta^*_{j_1',j_2'}})$, where (u_i,v_i) is the pair of items selected by the i-th user among the set of pairs of alternatives $S_i \times T_i$. The Hessian can be computed as

$$
\frac{\partial^2\mathcal{L}(\Theta)}{\partial\Theta_{j_1,j_2}\partial\Theta_{j_1',j_2'}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\big((j_1,j_2)\in S_i\times T_i\big)\frac{\partial p_{j_1,j_2|S_i,T_i}}{\partial\Theta_{j_1',j_2'}}
\tag{217}
$$

$$
= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\big((j_1,j_2),(j_1',j_2')\in S_i\times T_i\big)\Big(p_{j_1,j_2|S_i,T_i}\mathbb{I}((j_1,j_2)=(j_1',j_2')) - p_{j_1,j_2|S_i,T_i}p_{j_1',j_2'|S_i,T_i}\Big)\;,
\tag{218}
$$

We use $\nabla^2 \mathcal{L}(\Theta) \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ to denote this Hessian. Let $\Delta = \Theta^* - \widehat{\Theta}$ where $\widehat{\Theta}$ is an optimal solution to the convex optimization in (49). We introduce the following key technical lemmas. The following lemma provides a bound on the gradient using the concentration of measure for sum of independent random matrices Tropp (2011).

Lemma 39 *For any positive constant $c \geq 1$ and $n \geq (4(1+c)e^{2\alpha}d_1 d_2 \log d)/\max\{d_1, d_2\}$, with probability at least $1 - 2d^{-c}$,*

$$\|\!|\!|\nabla \mathcal{L}(\Theta^*)|\!|\!\|_2 \ \leq \ \sqrt{\frac{4(1+c)e^{2\alpha}\max\{d_1, d_2\} \log d}{d_1\, d_2\, n}} \ . \tag{219}$$

Since we are typically interested in the regime where the number of samples is much smaller than the dimension $d_1 \times d_2$ of the problem, the Hessian is typically not positive definite. However, when we restrict our attention to the vectorized Δ with relatively small nuclear norm, then we can prove restricted strong convexity, which gives the following bound.

Lemma 40 (Restricted Strong Convexity for bundled choice modeling) *Fix any $\Theta \in \Omega_\alpha$ and assume $(\min\{d_1, d_2\}/\min\{k_1, k_2\}) \log d \leq n \leq \min\{d^5 \log d, k_1 k_2 \max\{d_1^2, d_2^2\} \log d\}$. Under the random sampling model of the alternatives $\{j_{ia}\}_{i\in[n],a\in[k_1]}$ from the first set of items $[d_1]$, $\{j_{ib}\}_{i\in[n],b\in[k_1]}$ from the second set of items $[d_2]$ and the random outcome of the comparisons described in section 1, with probability larger than $1 - 2d^{-2^{25}}$,*

$$\mathrm{Vec}(\Delta)^T \nabla^2 \mathcal{L}(\Theta) \, \mathrm{Vec}(\Delta) \ \geq \ \frac{e^{-2\alpha}}{8\, d_1\, d_2}\|\!|\!|\Delta|\!|\!\|_F^2 \, , \tag{220}$$

for all Δ in $\mathcal{A}$ where

$$\mathcal{A} = \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2} \,\big|\, \|\!|\!|\Delta|\!|\!\|_\infty \leq 2\alpha \, , \sum_{j_1 \in [d_1], j_2 \in [d_2]} \Delta_{j_1 j_2} = 0 \ \text{and} \ \|\!|\!|\Delta|\!|\!\|_F^2 \geq \mu' \|\!|\!|\Delta|\!|\!\|_{\mathrm{nuc}} \right\} . \tag{221}$$

with

$$\mu' \ \equiv \ 2^{10}\, \alpha\, d_1 d_2 \sqrt{\frac{\log d}{n\, \min\{d_1, d_2\}\, \min\{k_1, k_2\}}} \ . \tag{222}$$

Building on these lemmas, the proof of Theorem 13 is divided into the following two cases. In both cases, we will show that

$$\|\!|\!|\Delta|\!|\!\|_F^2 \ \leq \ 12\, e^{2\alpha} c_1 \lambda\, d_1 d_2 \|\!|\!|\Delta|\!|\!\|_{\mathrm{nuc}} \, , \tag{223}$$

with high probability. Finally, applying an omitted result similar to Lemma 27 proves the desired theorem. We are left to show Eq. (223) holds.

Case 1: Suppose $\|\!|\!|\Delta|\!|\!\|_F^2 \geq \mu' \|\!|\!|\Delta|\!|\!\|_{\mathrm{nuc}}$**.** With $\Delta = \Theta^* - \widehat{\Theta}$, the Taylor expansion yields

$$\mathcal{L}(\widehat{\Theta}) = \mathcal{L}(\Theta^*) - \langle\!\langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle\!\rangle + \frac{1}{2}\mathrm{Vec}(\Delta)\nabla^2 \mathcal{L}(\Theta)\mathrm{Vec}^T(\Delta), \tag{224}$$

where $\Theta = a\widehat{\Theta} + (1-a)\Theta^*$ for some $a \in [0,1]$. It follows from Lemma 40 that with probability at least $1 - 2d^{-2^{25}}$,

$$\mathcal{L}(\widehat{\Theta}) - \mathcal{L}(\Theta^*) \geq -\|\|\nabla\mathcal{L}(\Theta^*)\|\|_2\|\|\Delta\|\|_{\mathrm{nuc}} + \frac{e^{-2\alpha}}{8\,d_1\,d_2}\|\|\Delta\|\|_{\mathrm{F}}^2 .$$

From the definition of $\widehat{\Theta}$ as an optimal solution of the minimization, we have

$$\mathcal{L}(\widehat{\Theta}) - \mathcal{L}(\Theta^*) \leq \lambda\left(\|\|\Theta^*\|\|_{\mathrm{nuc}} - \left\|\left\|\widehat{\Theta}\right\|\right\|_{\mathrm{nuc}}\right) \leq \lambda\|\|\Delta\|\|_{\mathrm{nuc}} .$$

By the assumption, we choose $\lambda \geq 8\lambda_0$. In view of Lemma 39, this implies that $\lambda \geq 2\|\|\nabla\mathcal{L}(\Theta^*)\|\|_2$ with probability at least $1 - 2d^{-3}$. It follows that with probability at least $1 - 2d^{-3} - 2d^{-2^{25}}$,

$$\frac{e^{-2\alpha}}{8d_1d_2}\|\|\Delta\|\|_{\mathrm{F}}^2 \leq \left(\lambda + \|\|\nabla\mathcal{L}(\Theta^*)\|\|_2\right)\|\|\Delta\|\|_{\mathrm{nuc}} \leq \frac{3\lambda}{2}\|\|\Delta\|\|_{\mathrm{nuc}} .$$

By our assumption on $\lambda \leq c_1\lambda_0$, this proves the desired bound in Eq. (223)

Case 2: Suppose $\|\|\Delta\|\|_{\mathrm{F}}^2 \leq \mu'\|\|\Delta\|\|_{\mathrm{nuc}}$. By the definition of μ and the fact that $c_1 \geq 128/\sqrt{\min\{k_1, k_2\}}$, it follows that $\mu' \leq 12\,e^{2\alpha}c_1\lambda\,d_1d_2$, and we get the same bound as in Eq. (223).

G.1 Proof of Lemma 39

Define $X_i = -(e_{u_i}e_{v_i}^T - p_i)$ such that $\nabla\mathcal{L}(\Theta^*) = (1/n)\sum_{i=1}^n X_i$, which is a sum of n independent random matrices. Note that since p_i is entry-wise bounded by $e^{2\alpha}/(k_1 k_2)$,

$$\|\|X_i\|\|_2 \leq 1 + \frac{e^{2\alpha}}{\sqrt{k_1 k_2}} ,$$

and

$$\sum_{i=1}^n \mathbb{E}[X_i X_i^T] = \sum_{i=1}^n (\mathbb{E}[e_{u_i}e_{u_i}^T] - p_i p_i^T) \tag{225}$$

$$\preceq \sum_{i=1}^n \mathbb{E}[e_{u_i}e_{u_i}^T] \tag{226}$$

$$\preceq \frac{e^{2\alpha}\,n}{d_1}\mathbf{I}_{d_1 \times d_1} , \tag{227}$$

where the last inequality follows from the fact that for any given S_i, u_i will be chosen with probability at most $e^{2\alpha}/k_1$, if it is in the set S_i which happens with probability k_1/d_1. Therefore,

$$\left\|\left\|\sum_{i=1}^n \mathbb{E}[X_i X_i^T]\right\|\right\|_2 \leq \frac{e^{2\alpha}\,n}{d_1} . \tag{228}$$

Similarly,

$$\left\|\left\|\sum_{i=1}^n \mathbb{E}[X_i^T X_i]\right\|\right\|_2 \leq \frac{e^{2\alpha}\,n}{d_2} . \tag{229}$$

Applying matrix Bernstein inequality Tropp (2011), we get

$$\mathbb{P}\left\{\||\nabla\mathcal{L}(\Theta^*)|\|_2 > t\right\}$$

$$\leq (d_1 + d_2)\exp\left\{\frac{-n^2 t^2/2}{(e^{2\alpha}n\max\{d_1,d_2\}/(d_1 d_2)) + ((1 + (e^{2\alpha}/\sqrt{k_1 k_2}))nt/3)}\right\}, \quad (230)$$

which gives the desired tail probability of $2d^{-c}$ for the choice of

$$
\begin{aligned}
t &= \max\left\{\sqrt{\frac{4(1+c)e^{2\alpha}\max\{d_1,d_2\}\log d}{d_1 d_2 n}}, \frac{4(1+c)(1 + \frac{e^{2\alpha}}{\sqrt{k_1 k_2}})\log d}{3n}\right\} \\
&= \sqrt{\frac{4(1+c)e^{2\alpha}\max\{d_1,d_2\}\log d}{d_1 d_2 n}},
\end{aligned}
$$

where the last equality follows from the assumption, $n \geq (4(1+c)e^{2\alpha}d_1 d_2 \log d)/\max\{d_1, d_2\}$.

G.2 Proof of Lemma 40

Thee quadratic form of the Hessian defined in (218) can be lower bounded by

$$\mathrm{Vec}(\Delta)^T \nabla^2\mathcal{L}(\Theta)\,\mathrm{Vec}(\Delta) \geq \underbrace{\frac{e^{-2\alpha}}{2\,k_1^2\,k_2^2\,n}\sum_{i=1}^{n}\sum_{j_1,j_1'\in S_i}\sum_{j_2,j_2'\in T_i}\left(\Delta_{j_1,j_2} - \Delta_{j_1',j_2'}\right)^2}_{\equiv H(\Delta)}, \quad (231)$$

which follows from Remark 31. To lower bound $H(\Delta)$, we first compute the mean:

$$\mathbb{E}[H(\Delta)] = \frac{e^{-2\alpha}}{2\,k_1^2\,k_2^2\,n}\sum_{i=1}^{n}\mathbb{E}\Big[\sum_{j_1,j_1'\in S_i}\sum_{j_2,j_2'\in T_i}\left(\Delta_{j_1,j_2} - \Delta_{j_1',j_2'}\right)^2\Big] \quad (232)$$

$$= \frac{e^{-2\alpha}}{d_1\,d_2}\||\Delta|\|_{\mathrm{F}}^2, \quad (233)$$

where we used the fact that $\mathbb{E}[\sum_{j_1\in S_i, j_2\in T_i}\Delta_{j_1,j_2}] = (k_1 k_2/(d_1 d_2))\sum_{j_1'\in[d_1],j_2'\in[d_2]}\Delta_{j_1',j_2'} = 0$ for $\Delta \in \Omega_{2\alpha}$ in (51).

We now prove that $H(\Delta)$ does not deviate from its mean too much. Suppose there exists a $\Delta \in \mathcal{A}$ defined in (221) such that Eq. (220) is violated, i.e. $H(\Delta) < (e^{-2\alpha}/(8d_1 d_2))\||\Delta|\|_{\mathrm{F}}^2$. In this case,

$$\mathbb{E}[H(\Delta)] - H(\Delta) \geq \frac{7\,e^{-2\alpha}}{8d_1 d_2}\||\Delta|\|_{\mathrm{F}}^2. \quad (234)$$

We will show that this happens with a small probability. We use the same peeling argument as in Appendix C with

$$
\begin{aligned}
\mathcal{S}_\ell = \Big\{\Delta \in \mathbb{R}^{d_1\times d_2} \mid \||\Delta|\|_\infty \leq 2\alpha, \beta^{\ell-1}\mu' \leq \||\Delta|\|_{\mathrm{F}} \leq \beta^\ell \mu', \\
\sum_{j_1\in[d_1],j_2\in[d_2]}\Delta_{j_1,j_2} = 0, \text{ and } \||\Delta|\|_{\mathrm{nuc}} \leq \beta^{2\ell}\mu'\Big\},
\end{aligned}
$$

where $\beta = \sqrt{10/9}$ and for $\ell \in \{1, 2, 3, \ldots\}$, and μ' is defined in (222). By the peeling argument, there exists an $\ell \in \mathbb{Z}_+$ such that $\Delta \in \mathcal{S}_\ell$ and

$$\mathbb{E}[H(\Delta)] - H(\Delta) \geq \frac{7\, e^{-2\alpha}}{8d_1 d_2}\beta^{2\ell-2}(\mu')^2 \geq \frac{7\, e^{-2\alpha}}{9\, d_1 d_2}\beta^{2\ell}(\mu')^2 \, . \tag{235}$$

Applying the union bound over $\ell \in \mathbb{Z}_+$,

$$\mathbb{P}\left\{\exists \Delta \in \mathcal{A}\,,\ H(\Delta) < \frac{e^{-2\alpha}}{8\, d_1\, d_2}|\!|\!|\Delta|\!|\!|_{\mathrm{F}}^2\right\}$$

$$\leq \sum_{\ell=1}^{\infty} \mathbb{P}\left\{\sup_{\Delta \in \mathcal{S}_\ell}\left(\, \mathbb{E}[H(\Delta)] - H(\Delta)\,\right) > \frac{7\, e^{-2\alpha}}{9 d_1 d_2}(\beta^\ell \mu')^2\right\}$$

$$\leq \sum_{\ell=1}^{\infty} \mathbb{P}\left\{\sup_{\Delta \in \mathcal{B}'(\beta^\ell \mu')}\left(\, \mathbb{E}[H(\Delta)] - H(\Delta)\,\right) > \frac{7 e^{-2\alpha}}{9 d_1 d_2}(\beta^\ell \mu')^2\right\}\,, \tag{236}$$

where we define the set $\mathcal{B}'(D)$ such that $\mathcal{S}_\ell \subseteq \mathcal{B}'(\beta^\ell \mu')$:

$$\mathcal{B}'(D) = \left\{\Delta \in \mathbb{R}^{d_1 \times d_2} \,\big|\, \|\Delta\|_\infty \leq 2\alpha, |\!|\!|\Delta|\!|\!|_{\mathrm{F}} \leq D, \sum_{j_1 \in [d_1], j_2 \in [d_2]} \Delta_{j_1 j_2} = 0, \mu'|\!|\!|\Delta|\!|\!|_{\mathrm{nuc}} \leq D^2\right\}\,. \tag{237}$$

The following key lemma provides the upper bound on this probability.

Lemma 41 *For* $(\min\{d_1, d_2\}/\min\{k_1, k_2\})\log d \leq n \leq d^5 \log d$,

$$\mathbb{P}\left\{\sup_{\Delta \in \mathcal{B}'(D)}\left(\, \mathbb{E}[H(\Delta)] - H(\Delta)\,\right) \geq \frac{e^{-2\alpha}D^2}{2 d_1 d_2}\right\} \leq \exp\left\{-\frac{n \min\{k_1^2, k_2^2\}\, k_1 k_2\, D^4}{2^{10}\alpha^4 d_1^2 d_2^2}\right\}\,. \tag{238}$$

Let $\eta = \exp\left(-\frac{n k_1 k_2 \min\{k_1^2, k_2^2\}(\beta-1.002)(\mu')^4}{2^{10}\alpha^4 d_1^2 d_2^2}\right)$. Applying the tail bound to (236), we get

$$\mathbb{P}\left\{\exists \Delta \in \mathcal{A}\,,\ H(\Delta) < \frac{e^{-2\alpha}}{8\, d_1 d_2}|\!|\!|\Delta|\!|\!|_{\mathrm{F}}^2\right\} \leq \sum_{\ell=1}^{\infty} \exp\left\{-\frac{n\, k_1 k_2\, \min\{k_1^2, k_2^2\}\,(\beta^\ell \mu')^4}{2^{10}\alpha^4 d_1^2 d_2^2}\right\}$$

$$\overset{(a)}{\leq} \sum_{\ell=1}^{\infty} \exp\left\{-\frac{n k_1 k_2 \min\{k_1^2, k_2^2\}\ell(\beta - 1.002)(\mu')^4}{2^{10}\alpha^4 d_1^2 d_2^2}\right\}$$

$$\leq \frac{\eta}{1-\eta}\,,$$

where (a) holds because $\beta^x \geq x \log \beta \geq x(\beta - 1.002)$ for the choice of $\beta = \sqrt{10/9}$. By the definition of μ',

$$\eta = \exp\left\{-\frac{2^{30}\, k_1 k_2\, \max\{d_2^2, d_1^2\}(\log d)^2(\beta - 1.002)}{n}\right\} \leq \exp\{-2^{25}\log d\}\,,$$

where the last inequality follows from the assumption that $n \leq k_1 k_2 \max\{d_1^2, d_2^2\}\log d$, and $\beta - 1.002 \geq 2^{-5}$. Since for $d \geq 2$, $\exp\{-2^{25}\log d\} \leq 1/2$ and thus $\eta \leq 1/2$, the lemma follows by assembling the last two displayed inequalities.

G.3 Proof of Lemma 41

Let $Z \equiv \sup_{\Delta \in \mathcal{B}'(D)} \mathbb{E}[H(\Delta)] - H(\Delta)$ and consider the tail bound using McDiarmid's inequality. Note that Z has a bounded difference of $(8\alpha^2 e^{-2\alpha} \max\{k_1, k_2\})/(k_1^2 k_2^2 n)$ when one of the $k_1 k_2 n$ independent random variables are changed, which gives

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geq t\} \leq \exp\left(-\frac{k_1^4 k_2^4 n^2 t^2}{64\alpha^4 e^{-4\alpha} \max\{k_1^2, k_2^2\} k_1 k_2 n}\right). \tag{239}$$

With the choice of $t = D^2/(4e^{2\alpha} d_1 d_2)$, this gives

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq \frac{e^{-2\alpha}}{4 d_1 d_2} D^2\right\} \leq \exp\left(-\frac{k_1^3 k_2^3 n D^4}{2^{10} \alpha^4 d_1^2 d_2^2 \max\{k_1^2, k_2^2\}}\right). \tag{240}$$

We first construct a partition of the space similar to Lemma 33. Let

$$\tilde{k} \equiv \min\{k_1, k_2\}. \tag{241}$$

Lemma 42 *There exists a partition $(\mathcal{T}_1, \ldots, \mathcal{T}_N)$ of $\{[k_1] \times [k_2]\} \times \{[k_1] \times [k_2]\}$ for some $N \leq 2k_1^2 k_2^2/\tilde{k}$ such that $\mathcal{T}_\ell$'s are disjoint subsets, $\bigcup_{\ell \in [N]} \mathcal{T}_\ell = \{[k_1] \times [k_2]\} \times \{[k_1] \times [k_2]\}$, $|\mathcal{T}_\ell| \leq \tilde{k}$ and for any $\ell \in [N]$ the set of random variables in $\mathcal{T}_\ell$ satisfy*

$$\{(\Delta_{j_{i,a}, j_{i,b}} - \Delta_{j_{i,a'}, j_{i,b'}})^2\}_{i \in [n], ((a,b), (a',b')) \in \mathcal{T}_\ell} \text{ are mutually independent }.$$

where $j_{i,a}$ for $i \in [n]$ and $a \in [k_1]$ denote the a-th chosen item to be included in the set S_i.

Now we prove an upper bound on $\mathbb{E}[Z]$ using the symmetrization technique. Recall that $j_{i,a}$ is independently and uniformly chosen from $[d_1]$ for $i \in [n]$ and $a \in [k_1]$. Similarly, $j_{i,b}$ is independently and uniformly chosen from $[d_1]$ for $i \in [n]$ and $b \in [k_2]$.

$$\mathbb{E}[Z]$$

$$= \frac{e^{-2\alpha}}{2 k_1^2 k_2^2 n} \mathbb{E}\left[\sup_{\Delta \in \mathcal{B}'(D)} \sum_{i=1}^n \sum_{\substack{a,a' \in [k_1] \\ b,b' \in [k_2]}} \mathbb{E}\left[(\Delta_{j_{i,a}, j_{i,b}} - \Delta_{j_{i,a'}, j_{i,b'}})^2\right] - (\Delta_{j_{i,a}, j_{i,b}} - \Delta_{j_{i,a'}, j_{i,b'}})^2\right]$$

$$\leq \frac{e^{-2\alpha}}{2 k_1^2 k_2^2 n} \sum_{\ell \in [N]} \mathbb{E}\left[\sup_{\Delta \in \mathcal{B}'(D)} \sum_{i=1}^n \sum_{(j_1, j_2, j_1', j_2') \in \mathcal{T}_\ell} \mathbb{E}\left[(\Delta_{j_1, j_2} - \Delta_{j_1', j_2'})^2\right] - (\Delta_{j_1, j_2} - \Delta_{j_1', j_2'})^2\right]$$

$$\leq \frac{e^{-2\alpha}}{k_1^2 k_2^2 n} \sum_{\ell \in [N]} \mathbb{E}\left[\sup_{\Delta \in \mathcal{B}'(D)} \sum_{i=1}^n \sum_{(j_1, j_2, j_1', j_2') \in \mathcal{T}_\ell} \xi_{i, j_1, j_2, j_1', j_2'} (\Delta_{j_1, j_2} - \Delta_{j_1', j_2'})^2\right], \tag{242}$$

where the first inequality follows for the fact that the supremum of the sum is smaller than the sum of supremum, and the second inequality follows from standard symmetrization

84

with i.i.d. Rademacher random variables $\xi_{i,j_1,j_2,j_1',j_2'}$'s. It follows from Ledoux-Talagrand contraction inequality Ledoux and Talagrand (2013) that

$$\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}'(D)}\sum_{i=1}^n\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\left(\Delta_{j_1,j_2}-\Delta_{j_1',j_2'}\right)^2\right] \tag{243}$$

$$\leq 8\alpha\,\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}'(D)}\sum_{i=1}^n\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\left(\Delta_{j_1,j_2}-\Delta_{j_1',j_2'}\right)\right] \tag{244}$$

$$\leq 8\alpha\,\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}'(D)}\|\!|\Delta|\!\|_{\mathrm{nuc}}\left|\!\left|\!\left|\sum_{i=1}^n\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\left(e_{j_1,j_2}-e_{j_1',j_2'}\right)\right|\!\right|\!\right|_2\right] \tag{245}$$

$$\leq \frac{8\alpha D^2}{\mu'}\mathbb{E}\left[\left|\!\left|\!\left|\sum_{i=1}^n\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\left(e_{j_1,j_2}-e_{j_1',j_2'}\right)\right|\!\right|\!\right|_2\right], \tag{246}$$

where the second inequality follows for the Hölder's inequality and the last inequality follows from $\mu'\|\!|\Delta|\!\|_{\mathrm{nuc}}\leq D^2$ for all $\Delta\in\mathcal{B}'(D)$. To bound the expected spectral norm of the random matrix, we use matrix Bernstein's inequality. Note that $\left|\!\left|\!\left|\xi_{i,j_1,j_2,j_1',j_2'}c\right|\!\right|\!\right|_2\leq\sqrt{2}$ almost surely, $\mathbb{E}[(e_{j_1,j_2}-e_{j_1',j_2'})(e_{j_1,j_2}-e_{j_1',j_2'})^T]\preceq(2/d_1)\mathbf{I}_{d_1\times d_1}$, and $\mathbb{E}[(e_{j_1,j_2}-e_{j_1',j_2'})^T(e_{j_1,j_2}-e_{j_1',j_2'})]\preceq(2/d_2)\mathbf{I}_{d_2\times d_2}$. It follows that $\sigma^2=2n|\mathcal{T}_\ell|/\min\{d_1,d_2\}$, where $|\mathcal{T}_\ell|\leq\min\{k_1,k_2\}$. It follows that

$$\mathbb{P}\left\{\left|\!\left|\!\left|\sum_{i=1}^n\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\left(e_{j_1,j_2}-e_{j_1',j_2'}\right)\right|\!\right|\!\right|_2>t\right\}$$

$$\leq(d_1+d_2)\exp\left\{\frac{-t^2/2}{2n\min\{k_1,k_2\}/\min\{d_1,d_2\}+\sqrt{2}t/3}\right\},$$

Choosing $t=\max\{\sqrt{64n(\min\{k_1,k_2\}/\min\{d_1,d_2\})\log d},(16\sqrt{2}/3)\log d\}$, we obtain a bound on the spectral norm of t with probability at least $1-2d^{-7}$. From the fact that $\left|\!\left|\!\left|\sum_{i=1}^n\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\left(e_{j_1,j_2}-e_{j_1',j_2'}\right)\right|\!\right|\!\right|_2\leq(n/\sqrt{2})\min\{k_1,k_2\}$, it follows that

$$\mathbb{E}\left[\left|\!\left|\!\left|\sum_{i=1}^n\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\left(e_{j_1,j_2}-e_{j_1',j_2'}\right)\right|\!\right|\!\right|_2\right] \tag{247}$$

$$\leq \max\left\{\sqrt{\frac{64\,n\,\min\{k_1,k_2\}\log d}{\min\{d_1,d_2\}}},(16\sqrt{2}/3)\log d\right\}+\frac{2n\min\{k_1,k_2\}}{\sqrt{2}d^7} \tag{248}$$

$$\leq \sqrt{\frac{66\,n\,\min\{k_1,k_2\}\log d}{\min\{d_1,d_2\}}} \tag{249}$$

which follows form the assumption that $n \min\{k_1, k_2\} \geq \min\{d_1, d_2\} \log d$ and $n \leq d^5 \log d$. Substituting this bound in (242), and (246), we get that

$$
\begin{aligned}
\mathbb{E}[Z] &\leq \frac{16 e^{-2\alpha} \alpha D^2}{\mu'} \sqrt{\frac{66 \log d}{n \min\{k_1, k_2\} \min\{d_1, d_2\}}} && (250) \\
&\leq \frac{e^{-2\alpha} D^2}{4 \, d_1 d_2} \,. && (251)
\end{aligned}
$$

Appendix H. Proof of the Information-theoretic Lower Bound in Theorem 16

This proof follow closely the proof of Theorem 10 in Appendix E. We apply the generalized Fano's inequality in the same way to get Eq. (160)

$$
\mathbb{P}\left\{\widehat{L} \neq L\right\} \geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1, \ell_2 \in [M]} D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) + \log 2}{\log M} \,, \tag{252}
$$

The main challenge in this case is that we can no longer directly apply the RUM interpretation to compete $D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$. This will result in over estimating the KL-divergence, because this approach does not take into account that we only take the top winner, out of those $k_1 k_2$ alternatives. Instead, we compute the divergence directly, and provide an appropriate bound. Let the set of k_1 rows and k_2 columns chosen in one of the n samples

be $S \subset [d_1]$ and $T \subset [d_2]$ respectively. Then,

$$D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$$

$$\overset{(a)}{=} \frac{n}{\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \sum_{\substack{i \in S \\ j \in T}} \frac{e^{\Theta_{ij}^{(\ell_1)}}}{\sum_{\substack{i' \in S \\ j' \in T}} e^{\Theta_{i'j'}^{(\ell_1)}}} \log \left(\frac{e^{\Theta_{ij}^{(\ell_1)}} \sum_{\substack{i' \in S \\ j' \in T}} e^{\Theta_{i'j'}^{(\ell_2)}}}{e^{\Theta_{ij}^{(\ell_2)}} \sum_{\substack{i' \in S \\ j' \in T}} e^{\Theta_{i'j'}^{(\ell_1)}}} \right)$$

$$\overset{(b)}{\leq} \frac{n}{\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \left(\sum_{i,j} \frac{e^{2\Theta_{ij}^{(\ell_1)}} \sum_{i',j'} e^{\Theta_{i'j'}^{(\ell_2)}} - e^{\Theta_{ij}^{(\ell_1)} + \Theta_{ij}^{(\ell_2)}} \sum_{i',j'} e^{\Theta_{i'j'}^{(\ell_1)}}}{e^{\Theta_{ij}^{(\ell_2)}} \left(\sum_{i',j'} e^{\Theta_{i'j'}^{(\ell_1)}} \right)^2} \right)$$

$$\overset{(c)}{\leq} \frac{ne^{2\alpha}}{k_1^2 k_2^2 \binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \sum_{i,j} \left(e^{2\Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)}} \sum_{i',j'} e^{\Theta_{i'j'}^{(\ell_2)}} - e^{\Theta_{ij}^{(\ell_1)}} \sum_{i',j'} e^{\Theta_{i'j'}^{(\ell_1)}} \right)$$

$$= \frac{ne^{2\alpha}}{k_1^2 k_2^2 \binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \left(\sum_{i',j'} e^{\Theta_{i'j'}^{(\ell_2)}} \sum_{i,j} \frac{\left(e^{\Theta_{ij}^{(\ell_1)}} - e^{\Theta_{ij}^{(\ell_2)}} \right)^2}{e^{\Theta_{ij}^{(\ell_2)}}} - \left(\sum_{i,j} (e^{\Theta_{ij}^{(\ell_1)}} - e^{\Theta_{ij}^{(\ell_2)}}) \right)^2 \right)$$

$$\overset{(d)}{\leq} \frac{ne^{4\alpha}}{k_1 k_2 \binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \sum_{i,j} \left(e^{\Theta_{ij}^{(\ell_1)}} - e^{\Theta_{ij}^{(\ell_2)}} \right)^2$$

$$\overset{(e)}{\leq} \frac{ne^{5\alpha}}{k_1 k_2 \binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \sum_{i,j} \left(\Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)} \right)^2$$

$$\overset{(f)}{=} \frac{ne^{5\alpha}}{d_1 d_2} \left\| \Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)} \right\|_{\mathrm{F}}^2$$

Here (a) is by definition of KL-distance and the fact that S, T are chosen uniformly from all possible such sets and (b) is due to the fact that $\log(x) \leq x - 1$ with $x = (e^{\Theta_{ij}^{(\ell_1)}} \sum_{i' \in S, j' \in T} e^{\Theta_{i'j'}^{(\ell_2)}}) / (e^{\Theta_{ij}^{(\ell_2)}} \sum_{i' \in S, j' \in T} e^{\Theta_{i'j'}^{(\ell_1)}})$. The constants at (c) is due to the fact that each element of $\Theta^{(\ell_1)}$ is upper bounded by α and lower bounded by $-\alpha$. We can get (d) by removing the second term which is always negative, and using the bond of α. (e) is obtained because e^x where $-\alpha \leq x \leq \alpha$ is Lipschitz continuous with Lipschitz constant e^α. At last (f) is obtained by simple counting of the occurrences of each ij. Thus we have,

$$\mathbb{P}\left\{ \widehat{L} \neq L \right\} \geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1, \ell_2 \in [M]} \frac{ne^{5\alpha}}{d_1 d_2} \left\| \Theta_{ij}^{(\ell_2)} - \Theta_{ij}^{(\ell_2)} \right\|_{\mathrm{F}}^2 + \log 2}{\log M}, \qquad (253)$$

The remainder of the proof relies on the following probabilistic packing.

Lemma 43 *Let $d_2 \geq d_1$ be sufficiently large positive integers. Then for each $r \in \{1, \ldots, d_1\}$, and for any positive $\delta > 0$ there exists a family of $d_1 \times d_2$ dimensional matrices $\{\Theta^{(1)}, \ldots, \Theta^{(M(\delta))}\}$ with cardinality $M(\delta) = \lfloor (1/4) \exp(rd_2/576) \rfloor$ such that each matrix is rank r and*

the following bounds hold:

$$\left|\!\left|\!\left|\Theta^{(\ell)}\right|\!\right|\!\right|_{\mathrm{F}} \quad \leq \quad \delta \ , \ \textit{for all } \ell \in [M] \tag{254}$$

$$\left|\!\left|\!\left|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right|\!\right|\!\right|_{\mathrm{F}} \quad \geq \quad \frac{1}{2}\delta \ , \ \textit{for all } \ell_1, \ell_2 \in [M] \tag{255}$$

$$\Theta^{(\ell)} \quad \in \quad \Omega_{\tilde{\alpha}} \ , \ \textit{for all } \ell \in [M] \ , \tag{256}$$

with $\tilde{\alpha} = (8\delta/d_2)\sqrt{2\log d}$ *for* $d = (d_1 + d_2)/2$.

Suppose $\delta \leq \alpha d_2/(8\sqrt{2\log d})$ such that the matrices in the packing set are entry-wise bounded by α, then the above lemma 43 implies that $\left|\!\left|\!\left|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right|\!\right|\!\right|_{\mathrm{F}}^2 \leq 4\delta^2$, which gives

$$\mathbb{P}\left\{\widehat{L} \neq L\right\} \quad \geq \quad 1 - \frac{\frac{e^{5\alpha}n4\delta^2}{d_1 d_2} + \log 2}{\frac{rd_2}{576} - 2\log 2} \quad \geq \quad \frac{1}{2} \ , \tag{257}$$

where the last inequality holds for $\delta^2 \leq (rd_1 d_2^2/(1152 e^{5\alpha}n))$ and assuming $rd_2 \geq 1600$. Together with (257) and (255), this proves that for all $\delta \leq \min\{\alpha d_2/(8\sqrt{2\log d}),$ $\sqrt{rd_1 d_2^2/(9216\, e^{5\alpha}n)}\}$,

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\left[\left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\mathrm{F}}\right] \quad \geq \quad \delta/4 \ .$$

Choosing δ appropriately to maximize the right-hand side finishes the proof of the desired claim. Also by symmetry, we can apply the same argument to get similar bound with d_1 and d_2 interchanged.

H.1 Proof of Lemma 43

We show that the following procedure succeeds in producing the desired family with probability at least half, which proves its existence. Let $d = (d_1 + d_2)/2$, and suppose $d_2 \geq d_1$ without loss of generality. For the choice of $M' = e^{rd_2/576}$, and for each $\ell \in [M']$, generate a rank-r matrix $\Theta^{(\ell)} \in \mathbb{R}^{d_1 \times d_2}$ as follows:

$$\Theta^{(\ell)} \quad = \quad \frac{\delta}{\sqrt{rd_2}}U(V^{(\ell)})^T - \frac{\delta}{\sqrt{rd_2}}\frac{\mathbb{1}^T U(V^{(\ell)})^T\mathbb{1}}{d_1 d_2}\mathbb{1}\mathbb{1}^T \ , \tag{258}$$

where $U \in \mathbb{R}^{d_1 \times r}$ is a random orthogonal basis such that $U^T U = \mathbf{I}_{r \times r}$ and $V^{(\ell)} \in \mathbb{R}^{d_2 \times r}$ is a random matrix with each entry $V_{ij}^{(\ell)} \in \{-1, +1\}$ chosen independently and uniformly at random. By construction, notice that $\left|\!\left|\!\left|\Theta^{(\ell)}\right|\!\right|\!\right|_{\mathrm{F}} \leq (\delta/\sqrt{rd_2})\left|\!\left|\!\left|U(V^{(\ell)})^T\right|\!\right|\!\right|_{\mathrm{F}} = \delta$.

Now, by triangular inequality, we have

$$\left|\!\left|\!\left|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right|\!\right|\!\right|_{\mathrm{F}}$$

$$\geq \quad \frac{\delta}{\sqrt{rd_2}}\left|\!\left|\!\left|U(V^{(\ell_1)} - V^{(\ell_2)})^T\right|\!\right|\!\right|_{\mathrm{F}} - \frac{\delta\,|\mathbb{1}^T U(V^{(\ell_1)} - V^{(\ell_2)})^T\mathbb{1}|}{d_1 d_2 \sqrt{rd_2}}\left|\!\left|\!\left|\mathbb{1}\mathbb{1}^T\right|\!\right|\!\right|_{\mathrm{F}}$$

$$\geq \quad \frac{\delta}{\sqrt{rd_2}}\underbrace{\left|\!\left|\!\left|V^{(\ell_1)} - V^{(\ell_2)}\right|\!\right|\!\right|_{\mathrm{F}}}_{A} - \frac{\delta}{\sqrt{r\,d_1 d_2^2}}\big(\underbrace{|\mathbb{1}^T U(V^{(\ell_1)})^T\mathbb{1}|}_{B} + |\mathbb{1}^T U(V^{(\ell_2)})^T\mathbb{1}|\big) \ .$$

88

We will prove that the first term is bounded by $A \geq \sqrt{rd_2}$ with probability at least $7/8$ for all M' matrices, and we will show that we can find M matrices such that the second term is bounded by $B \leq 8\sqrt{2rd_2 \log(32r) \log(32d)}$ with probability at least $7/8$. Together, this proves that with probability at least $3/4$, there exists M matrices such that

$$\left\|\!\left\|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right\|\!\right\|_{\mathrm{F}} \;\geq\; \delta\Big(1 - \sqrt{\frac{2^7 \log(32r)\,\log(32d)}{d_1 d_2}}\Big) \;\geq\; \frac{1}{2}\delta\,,$$

for all $\ell_1,\,\ell_2 \in [M]$ and for sufficiently large d_1 and d_2.

Applying similar McDiarmid's inequality as Eq. (122) in Appendix E, it follows that $A^2 \geq rd_2$ with probability at least $7/8$ for $M' = e^{rd_2/576}$ and a sufficiently large d_2.

To prove a bound on B, we will show that for a given ℓ,

$$\mathbb{P}\left\{|\mathbb{1}^T U (V^{(\ell)})^T \mathbb{1}| \leq 8\sqrt{2rd_2 \log(32r)\log(32d)}\right\} \;\geq\; \frac{7}{8}\,. \tag{259}$$

Then using the similar technique as in (125), it follows that we can find $M = (1/4)M'$ matrices all satisfying this bound and also the bound on the max-entry in (260). We are left to prove (259). We apply a series of concentration inequalities. Let H_1 be the event that $\{|\langle\!\langle V_i^{(\ell)}, \mathbb{1}\rangle\!\rangle| \leq \sqrt{2d_2 \log(32r)}$ for all $i \in [r]\}$. Then, applying the standard Hoeffding's inequality, we get that $\mathbb{P}\{H_1\} \geq 15/16$, where $V_i^{(\ell)}$ is the i-th column of $V^{(\ell)}$. We next change the variables and represent $\mathbb{1}^T U$ as $\sqrt{d_1}u^T \tilde{U}$, where u is drawn uniformly at random from the unit sphere and $\tilde{U}$ is a r dimensional subspace drawn uniformly at random. By symmetry, $\sqrt{d_1}u^T \tilde{U}$ have the same distribution as $\mathbb{1}^T U$. Let H_2 be the event that $\{|\langle\!\langle \tilde{U}_i, (V^{(\ell)})^T \mathbb{1}\rangle\!\rangle| \leq \sqrt{16r(d_2/d_1) \log(32r)\log(32d)}$ for all $i \in [d_1]\}$, where $\tilde{U}_i$ is the i-th row of $\tilde{U}$. Then, applying Levy's theorem for concentration on the sphere Ledoux (2005), we have $\mathbb{P}\{H_2|H_1\} \geq 15/16$. Finally, let H_3 be the event that $\{|\sqrt{d_1}\langle\!\langle u, \tilde{U}(V^{(\ell)})^T\rangle\!\rangle \mathbb{1}| \leq 8\sqrt{2rd_2 \log(32r)\log(32d)}\}$. Then, again applying Levy's concentration, we get $\mathbb{P}\{H_3|H_1, H_2\} \geq 15/16$. Collecting all three concentration inequalities, we get that with probability at least $13/16$, $|\mathbb{1}^T U (V^{(\ell)})^T \mathbb{1}| \leq 8\sqrt{2rd_2 \log(32r)\log(32d)}$, which proves Eq. (259).

We are left to prove that $\Theta^{(\ell)}$'s are in $\Omega_{(8\delta/d_2)\sqrt{2\log d_2}}$ as defined in (51). Similar to Eq. (124), applying Levy's concentration gives

$$\mathbb{P}\left\{ \max_{i,j} |\Theta_{ij}^{(\ell)}| \leq \frac{2\delta\sqrt{32\log d_2}}{d_2} \right\} \;\geq\; 1 - 2\exp\left\{-2\log d_2\right\} \;\geq\; \frac{1}{2}\,, \tag{260}$$

for a fixed $\ell \in [M']$. Then using the similar technique as in (125), it follows that there exists $M = (1/4)M'$ matrices all satisfying this bound and also the bound on B in Eq. (259).

References

A. Agarwal, S. Negahban, and M. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *In NIPS*, pages 37–45, 2010.

A. Agarwal, P. Patil, and S. Agarwal. Accelerated spectral ranking. In *International Conference on Machine Learning*, pages 70–79, 2018.

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.

J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.

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

A. R. Benson, R. Kumar, and A. Tomkins. A discrete choice model for subset selection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 37–45. ACM, 2018.

S. Bhojanapalli and P. Jain. Universal matrix completion. *arXiv preprint arXiv:1402.2324*, 2014.

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

C. Borgs, J. Chayes, C. E. Lee, and D. Shah. Iterative collaborative filtering for sparse matrix estimation. *arXiv preprint arXiv:1712.00710*, 2017.

V. S. Borkar, N. Karamchandani, and S. Mirani. Randomized Kaczmarz for rank aggregation from pairwise comparisons. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 389–393. IEEE, 2016.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1955.

J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202. ACM, 2013.

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

Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized mle are both optimal for top-k ranking. *arXiv preprint arXiv:1707.09971*, 2017.

F. Chierichetti, R. Kumar, and A. Tomkins. Learning a mixture of two multinomial logits. In *International Conference on Machine Learning*, pages 960–968, 2018.

C. Chu, P. Leslie, and A. Sorensen. Bundle-size pricing as an approximation to mixed bundling. *The American Economic Review*, pages 263–303, 2011.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In *International Conference on Computational Learning Theory*, pages 1–15. Springer, 2005.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

P. Diaconis and R. L. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 262–268, 1977.

M. Falahatgar, A. Jain, A. Orlitsky, V. Pichapati, and V. Ravindrakumar. The limits of maxing, ranking, and preference learning. In *International Conference on Machine Learning*, pages 1426–1435, 2018.

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

R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

D. F. Gleich and L.-h. Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 60–68. ACM, 2011.

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

I. C. Gormley and T. B. Murphy. A grade of membership model for rank data. *Bayesian Analysis*, 4(2):265–295, 2009.

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

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

R. Heckel, M. Simchowitz, K. Ramchandran, and M. J. Wainwright. Approximate ranking from pairwise comparisons. *arXiv preprint arXiv:1801.01253*, 2018.

D. R. Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, pages 384–406, 2004.

P. Jain and S. Oh. Provable tensor factorization with missing data. preprint arXiv:1406.2784, 2014.

P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, pages 665–674, 2013.

M. Jang, S. Kim, C. Suh, and S. Oh. Top-k ranking from pairwise comparisons: When spectral ranking is optimal. *arXiv preprint arXiv:1603.04153*, 2016.

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

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

S. Katariya, L. Jain, N. Sengupta, J. Evans, and R. Nowak. Adaptive sampling for coarse ranking. *arXiv preprint arXiv:1802.07176*, 2018.

J. Katz-Samuels and C. Scott. Nonparametric preference completion. *arXiv preprint arXiv:1705.08621*, 2017.

E. Kazemi, L. Chen, S. Dasgupta, and A. Karbasi. Comparison based learning from weak oracles. *arXiv preprint arXiv:1802.06942*, 2018.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010a.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(2057-2078):1, 2010b.

A. Khetan and S. Oh. Data-driven rank breaking for efficient rank aggregation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, 2016a.

A. Khetan and S. Oh. Computational and statistical tradeoffs in learning to rank. In *Advances in Neural Information Processing Systems 29*, pages 739–747, 2016b.

M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Soc., 2005.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

D. R. Luce. *Individual Choice Behavior*. Wiley, New York, 1959.

L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

J. Marschak. Binary-choice constraints and random utility indicators. In *Proceedings of a symposium on mathematical methods in the social sciences*, volume 7, pages 19–38, 1960.

A. K. Massimino and M. A. Davenport. As you like it: Localization via paired comparisons. *arXiv preprint arXiv:1802.10489*, 2018.

L. Maystre and M. Grossglauser. Fast and accurate inference of Plackett–Luce models. In *Advances in Neural Information Processing Systems*, pages 172–180, 2015.

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

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

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

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

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 2012.

S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

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

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

G. Ongie, L. Balzano, D. Pimentel-Alarcón, R. Willett, and R. D. Nowak. Tensor methods for nonlinear matrix completion. *arXiv preprint arXiv:1804.10266*, 2018.

A. Pananjady, C. Mao, V. Muthukumar, M. J. Wainwright, and T. A. Courtade. Worst-case vs average-case design for estimation from fixed pairwise comparisons. *arXiv preprint arXiv:1707.06217*, 2017.

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

R. L. Plackett. The analysis of permutations. *Applied Statistics*, pages 193–202, 1975.

S. Ragain and J. Ugander. Pairwise choice Markov chains. In *Advances in Neural Information Processing Systems*, pages 3198–3206, 2016.

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

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

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

W. Rui, J. Xu, S. Rayadurgam, M. Lelarge, L. Massoulié, and B. Hajek. Clustering and inference from pairwise comparisons. In *SIGMETRICS'15 Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, volume 43, page 2, 2015.

N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright. When is it better to compare than to score? *arXiv preprint arXiv:1406.6618*, 2014.

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. *The Journal of Machine Learning Research*, 17(1):2049–2095, 2016a.

N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, 2016b.

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

O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

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

K. Train. *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*, volume 10. MIT press, 1986.

J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Comput. Math.*, 2011.

J. A. Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

A. Tsokos, S. Narayanan, I. Kosmidis, G. Baio, M. Cucuringu, G. Whitaker, and F. J. Király. Modeling outcomes of soccer matches. *arXiv preprint arXiv:1807.01623*, 2018.

S. Van De Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

M. Vojnovic and S. Yun. Parameter estimation for generalized thurstone choice models. In *International Conference on Machine Learning*, pages 498–506, 2016.

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

M. Wilber, S. Kwak, and S. Belongie. Cost-effective HITs for relative similarity comparisons. In *Human Computation and Crowdsourcing (HCOMP)*, Pittsburgh, 2014. URL `http://arxiv.org/abs/1404.3291`.

M. Yuan and C. Zhang. On tensor completion via nuclear norm minimization. *arXiv preprint arXiv:1405.1773*, 2014.

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