COLLABORATIVE RANKING FROM ORDINAL DATA

BY

KIRAN KOSHY THEKUMPARAMPIL

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Adviser:

Assistant Professor Sewoong Oh

ABSTRACT

Personalized recommendation systems have to predict preferences of a user for items that have not seen by the user. For cardinal (ratings) data, personalized preference prediction has been efficiently solved over the past few years using matrix factorization related techniques. Recent studies have shown that ordinal (comparison) data can outperform cardinal data in learning preferences, but there has not been much study on learning personalized preferences from ordinal data. This thesis presents a matrix factorization inspired, convex relaxation algorithm to collaboratively learn hidden preferences of users through the multinomial logit (MNL) model, a discrete choice model. It also shows that the algorithm is efficient in terms of the number of observations needed.

To my parents, for their love, support, and faith in me

ACKNOWLEDGMENTS

I would like to thank my adviser Prof. Sewoong Oh for his guidance and support. I also would like to thank my collaborators Prof. Sahand Negahban and Prof. Jiaming Xu, whose contributions and valuable insights helped this project greatly. I would like to express my gratitude to my academic advisor Prof. Bruce Hajek for his help and support throughout my master's program. I would also like to thank my colleagues at the Coordinated Science Laboratory for all the stimulating discussions we had. Finally, I would like to thank my friends and my family. This work was partially supported by NSF awards CNS-1527754 and CCF-1553452.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

We are in an information revolution now. Humans are accumulating a wealth of information driven by the enormous reach and scale of the Internet and information technology. Institutions are collecting large amounts of data everyday regarding their client base.

In particular, a large amount of individual preference data on different sets of choices such as Netflix movie ratings [1], Foursquare restaurant check-ins, the videos watched from a group of suggested ones on YouTube, etc., are collected. They are collected in the hopes of better understanding the needs and interests of individuals and population as a whole so as to provide personalized products and services better catered to them.

These preference data are primarily used in recommendation systems and for revenue management. In both the cases, it is important to predict hidden preferences of individuals for items which are not yet "seen" or compared by the individuals. In revenue management, we need to predict the preference trends in the population and various demographic groups from partial preference data.

But aiming to achieve these without any further assumptions on the structure of the problem is futile. A reasonable assumption we can make would be that, individuals with similar preferences of compared or scored items would have similar preferences of "unseen" items too. We will call this *correlation assumption*. For example, if users A and B are both die-hard fans of the movies "Die Hard" and "The Matrix" and user B likes "The Terminator", then it is very likely that user A also likes "The Terminator". This assumption points to a collaborative learning algorithm for estimation preferences from partial observations. But first we will look at the type of preference data collected.

1.2 Cardinal versus ordinal data

There can be two kinds of preference data.

1. *Cardinal data*: It consists of ratings of different items, by individuals, according to a scoring system. For example, in Netflix [1], users are asked to rate the movies they have watched with integer scores on a scale of 0 to 5.

2. *Ordinal data*: Ordinal data contains comparison results or choices made by different individuals. It can be the choice made by the individual, given a set of alternatives or a full-blown ordering or ranking of the given set. For example, in single transferable voting system the voters are asked to order their choice of candidates in decreasing order of preference [2].

In the case of cardinal data, many fast and efficient algorithms were proposed in the past few years, which exploit matrix factorization techniques and assume a low-rank score matrix structure to collaboratively predict scores of unrated items for individual users [3].

On superficial analysis, one may say that using comparison data (ordinal data) would decrease the accuracy of learning, by arguing that ratings provide more information. Ratings data can always be converted to comparison data and not vice versa. But a recent study has shown that there can be a lot of noise in the cardinal, data and inferences made from ordinal data can be superior [4]. This may be so, because the scoring systems are often arbitrarily chosen, and different individuals may interpret them differently. That study also empirically found that the time (and in turn cost) associated with procuring comparison data is considerably less than the time associated with scoring. Thus the results of the above mentioned study justifies collaboratively learning preferences from ordinal data.

1.3 Problem statement

The problem statement is as follows. There are a set of users and a set of items. Each user has given partial comparison data of a subset of the items. This comparison data can be of different forms:

(a) *graph-sampled pairwise ranking*, where each user compares (graph-sampled) pairs of items and reports the result of these comparisons. We will see more about graph sampling in Section 2.1.

(b) *k-wise ranking*, where each user ranks a subset of k items according to her order of preference.

(b) *n choices*, where each user is given n subsets of items of size k_2 each, and she picks her foremost preferred item from each of the n subsets. Note that pairwise comparison is a special case of this with $k_2 = 2$.

We will assume that data satisfies the *correlation assumption* described in the previous section. The goal is to find the overall ranking of all the items for each user and also predict the outcome of any future comparisons made by any user. Results for graph-sampled pairwise ranking were first published online in [5], and k-wise ranking and n choices cases were published in [6].

1.4 Multinomial logit (MNL) model

The origins of theory of choices can be traced to the 18th century mathematician Condorcet [7], who tried to combine partial ranking of candidates by voters in an election into a full ranking. But foundations of the current theoretical models for preferences or choices were developed by economists and psychologists in the 20th century for studying consumer preferences and population behavior. Among the various models developed, a popular one is the random utility model (RUM) or Thurstone model [8]. In RUM, given a set of alternative choices S, the decision maker assigns a utility score for each choice x as

$$U(S,x) = V(S,x) + \varepsilon(S,x), \tag{1.1}$$

where $V(S,x)$ is a deterministic function which depends on the ground truth "taste" of the decision maker and $\varepsilon(S,x)$ is the i.i.d. (among different choices) random noise or "confusion" generated by sampling from cumulative distribution function (CDF) $F(\varepsilon)$. The randomness is crucial in justifying seemingly contradictory observations in real data, such as in-transitivity in ranking of triplets (example: $a < b$, $b < c$, and $c < a$) or different ordering of items given by the same person at different times (example: $a < b < c$ and $a < c < b$). RUMs are proven to be useful in various applications, like consumer market analyses [9].

In this thesis a special case of the RUM called the multinomial logit (MNL) model [8] is used. In the MNL model we have a set of d_1 users and a set of d_2 items and a "quality" parameter matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ which is believed to have a low rank of r. The rows correspond to the users and the columns correspond to the items. Each user i is presented with set $S_i \subseteq [d_2]$ of k alternatives. 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

$$\mathbb{P}\left\{v_{i,1}, \ldots, v_{i,k}\right\} \equiv \prod_{\ell=1}^{k} \frac{e^{\Theta^*_{i,v_{i,\ell}}}}{\sum_{j \in S_{i,\ell}} e^{\Theta^*_{i,j}}} \,, \tag{1.2}$$

where $S_{i,\ell} \equiv S_i \backslash \{v_{i,1}, \ldots, v_{i,\ell-1}\}$ and $S_{i,1} \equiv S_i$. The rank r criterion naturally captures the *correlation assumption* from Section 1.1, by capturing the similarities among users and items by representing them on a low-dimensional space. Thus the rank r criterion implies the preference parameter vector of every user is a mixture of a standard set of r basis parameter vectors.

The log-likelihood function of MNL is concave; thus we can efficiently use maximum likelihood estimator (MLE) for finding ground truth. The MNL model is extensively applied in choice modelling applications such as marketing and estimating airline travel demands due to its tractability and accuracy [10]. It can be shown that the MNL is a special case of RUM with the noise ε following the CDF $F(\varepsilon) = e^{-e^{-\varepsilon}}$ of the standard Gumbel distribution. This interpretation is crucial in our theoretical analysis.

1.5 Low-rank regularization using nuclear norm minimization

We use the following basic meta-algorithm strategy for solving our collaborative learning problems. Since Θ^* is approximately low rank, a natural strategy is to maximize the log-likelihood $\mathcal{L}(\Theta)$ under a fixed rank constraint, as

$$\widehat{\Theta} \in \arg \min_{\text{rank}(\Theta)=r} -\mathcal{L}(\Theta)\,. \tag{1.3}$$

Even if we know the rank of Θ^*, this is a difficult and non-convex optimization problem.

We know that the nuclear norm ball is the convex hull of rank-1 matrices. Analogous to l_1-norm in the case of sparse solutions, the nuclear norm is a tight convex surrogate for low-rank solutions. Therefore, our algorithms will solve the following nuclear norm regularized optimization problem,

$$\widehat{\Theta} \in \arg\min_{\Theta\in\Omega} -\mathcal{L}(\Theta) \,+\, \lambda|\!|\!|\Theta|\!|\!|_{\text{nuc}}, \tag{1.4}$$

where Ω is a convex constraint which takes care of identifiability and Lipschitz smoothness conditions. Nuclear norm regularization has been widely used [11] in similar settings; however, provable guarantees typically exist only for the quadratic loss function $\mathcal{L}(\Theta)$ [12, 13]. Our analyses extends these results to by first proving that $-\mathcal{L}(\cdot)$ satisfies the restricted strong convexity property with high probability.

1.6 Related work

Preliminary inference algorithms on ordinal data were done with the Plackett-Luce (PL) model, which is a special case of MNL where there is only one row in Θ^* and every person in the population is assumed to have the same preference. The PL model has been studied extensively in the past couple of years, and many efficient algorithms for estimating the population preferences have been developed [14, 15, 16], and sample complexity was characterized for the MLE [17]. However, the PL model can only capture a single overall ranking.

To overcome the lack of personalized preferences in the PL model, recently a few algorithms were proposed which try to learn a mixed PL model with only r classes of users by clustering [18] and a tensor-based approaches [19]. MNL can be thought of as a generalization of r class models, where the preference of each user is a mixture of r basis preferences. But existing work on MNL model is restricted to only pairwise comparisons [20, 21]. [21] proposes an algorithm minimizing a convex loss function with nuclear norm minimization. It is shown that this approach achieves a statistically optimal generalization error rate. The approach in this thesis is inspired by [20], which a uses similar convex relaxation as does this thesis, but for pairwise comparisons of independent and identically (i.i.d.) sampled pairs of items. We generalize the results of [5] by analyzing more general sampling models beyond i.i.d. sampled pairwise comparisons.

1.7 Notation

$|||A|||_{\mathrm{F}} = \sqrt{\sum_{i,j} A_{ij}^2}$ and $|||A|||_\infty$denote the Frobenius norm, and the ℓ_∞ norm. $|||A|||_{\mathrm{nuc}} = \sum_i \sigma_i(A)$ denotes the nuclear norm, which is the sum of singular values $\{\sigma_i(A)\}_i$ of matrix A. $\langle\langle u, v\rangle\rangle \equiv \sum_i u_i v_i$. The set of the first N positive integers is denoted by $[N] = \{1, \ldots, N\}$.

1.8 Organization of chapters

In this thesis, we organize the results into three chapters, one each for graph sampled pairwise comparison, k-wise ranking, and bundle choices, which is a generalized version of the n choices model described in Section 1.3. In each chapter we will provide an algorithm for solving the problem at hand, an upper bound on the error in the Frobenius norm of parameter matrix in terms of the number of samples and dimensions of the problem and an information theoretic lower bound for the performance of the best possible estimator for the problem.

CHAPTER 2

COMPARISON OF GRAPH SAMPLED PAIRS

2.1 Learning from pairwise comparisons of graph sampled pairs

Here, we analyze a generalized sampling of items for pairwise comparison. Consider a weighted complete graph of d_2 nodes, which represent items. The edge weight $2P_{i,j}$ between nodes i and j represents the probability with which the pair (i, j) is chosen for comparison by any user. We capture this probability in a real-symmetric adjacency matrix P such that $P_{i,i} = 0 \ \forall i \in [d_2]$, $P_{i,j} = P_{j,i}$, and $\sum_{i,j \in [d_2]} P_{i,j} = 1$.

Outcomes of these comparisons follow the MNL preference model from Section 1.4, which is parameterized by a rank-r matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, and where the probability with which user i prefers item j over item k is $1/\left(1 + \exp\left(\Theta^*_{ik} - \exp \Theta^*_{ij}\right)\right)$. We also assume that $|||\Theta|||_\infty \leq \alpha$, so we have a handle on the dynamic range of the feasible probabilities. While in practice we do not require the ℓ_∞ norm constraint, we need it for the analysis. For a related problem of matrix completion, where the loss $\mathcal{L}(\theta)$ is quadratic, either a similar condition on the ℓ_∞ norm is required or a different condition on incoherence is required.

In the most general setting of graph based sampling, some items may never be compared with each other; that is, $P_{i,j}$ can be zero 0 for some (i, j). Moreover items may be partitioned into groups of items where inter-group comparisons never occur. For example, it does not make sense to compare a television and a radio system. Given this, users might still assess these items using similar metrics like quality or price. Therefore it makes sense to solve the preference learning problem of different groups together rather than separately. To capture this intrinsic relation between groups, we define the notion of disjoint components in the graph. Let $g_i \in \{0, 1\}^{d_2}$ such that

$g_{i,j} = 1$ if item j is in group i, else $g_{i,j} = 0$. We also assume that no item can be present in more than one group; that is, $\sum_{i=1}^{G} g_i = \mathbb{1}$, where G is the number of groups.

Since probabilities of the pairwise comparison are oblivious to shifts of each row by a constant, for each distribution, we have an equivalence class of Θ^* which are consistent:

$$[\Theta^*] = \{\Theta^* + \sum_{i=1}^{G} u_i g_i^T \mid u_i \in \mathbb{R}^{d_1}\}\,. \tag{2.1}$$

To overcome this identifiability issue, we assume that the mean of each group is zero,

$$\Theta^* g_i = 0, \quad \forall\, i \in \{1, 2, \ldots, G\}\,. \tag{2.2}$$

Graph Laplacian: We define the graph Laplacian of matrix P as

$$L = \operatorname{diag}(P_u) - P\,, \tag{2.3}$$

where $P_u = \sum_v P_{u,v}$. Notice that L is singular and zero eigenvalues are in the direction of vectors $\{g_i\}_{i=1}^{G}$. 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 the 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 $L^x := U\Sigma^x U^T$ for all $x \in \mathbb{R}$ and $L^\dagger := L^{-1} = U\Sigma^{-1}U^T$. We also define the Laplacian induced norms of matrices,

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

We use these Laplacian induced norms because they 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. Now the log-likelihood of the comparison outcomes w.r.t. a parameter matrix Θ is

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} y_i\langle\langle\Theta, X^{(i)}\rangle\rangle - \log\left(1+\exp\left(\langle\langle\Theta, X^{(i)}\rangle\rangle\right)\right) \ . \tag{2.4}$$

Now we propose and analyze the following convex optimization problem,

$$\hat{\Theta} \in \operatorname*{argmin}_{\Theta\in\Omega_\alpha} - \mathcal{L}(\Theta) + \lambda|||\Theta|||_{\text{L-nuc}}, \tag{2.5}$$

where

$$\Omega_\alpha = \left\{\Theta \in \mathbb{R}^{d_1\times d_2} | |||\Theta|||_\infty \le \alpha, \Theta g_i = 0, \ \forall\, i \in [G]\right\}, \tag{2.6}$$

and $\lambda = 2\sqrt{32}\max\left\{\sqrt{\frac{\sigma\log(2d)}{n}}, \frac{\sigma_{\min}^{-1/2}\log(2d)}{n}\right\}$ with $\sigma = \max\{(d_2-G)/d_1, 1\}$.

2.2 Performance guarantee

Following is the main result for the comparison of graph sampled pairs.

Theorem 1. *Let $d = (d_1 + d_2)/2$ and suppose $n \le \min\{2^6 d_1^2\sigma^2, 2^3\left(d_1\sigma_{\min}^{-1}\right)^{2/3}\}\log(2d)$. Then under the described graph sampling and MNL preference model and solving the optimization problem in* (2.5) *gives*

$$\frac{\left|\left|\left|\left(\Theta^* - \widehat{\Theta}\right)L^{1/2}\right|\right|\right|_{\mathrm{F}}^2}{\sqrt{d_1}}$$
$$\le 36\lambda\sqrt{d_1}\left(\alpha + \frac{1}{\psi(2\alpha)}\right)\left(\sqrt{2r}\left|\left|\left|\left(\Theta^* - \widehat{\Theta}\right)L^{1/2}\right|\right|\right|_{\mathrm{F}} + \sum_{j=r+1}^{\min\{d_1,d_2-G\}}\sigma_j(\Theta^* L^{1/2})\right) \tag{2.7}$$

with probability greater than $1 - 2/(2d)^3$, where $\sigma = \max\{(d_2-G)/d_1, 1\}$.

Proof. The proof of the theorem relies on the following two lemmas. The first lemma shows that the negative of the log-likelihood satisfies restricted strong convexity with high probability.

Lemma 2.2.1 (Restricted strong convexity). *Let* $\mathcal{A}(\alpha) = \left\{\Theta \in \mathbf{R}^{d_1 \times d_2}, |||\Theta|||_\infty \leq \alpha, |||\Theta|||_{\text{L-nuc}} \leq \frac{|||\Theta L^{1/2}|||_{\text{F}}^2}{16\alpha d_1 R}\right\}$ *and* $R = \max\left\{\sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}^{-1/2}\log(2d)}{n}\right\}$. *When* $n \leq 2^6 d_1^2 \sigma^2 \log(2d)$ *and* $n \leq 2^2 (d_1 \sigma_{\min}^{-1})^{2/3} \log(2d)$,

$$\frac{1}{n}\sum_{i=1}^{n} (\langle\langle \Theta, X_i \rangle\rangle)^2 \geq \frac{1}{3d_1} |||\Theta|||_{\text{L}}^2, \quad \forall\, \Theta \in \mathcal{A}(\alpha), \tag{2.8}$$

with probability at least $1 - 2(2d)^{-4}$.

Here the upper bound on n may not be necessary; 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. The proof is in Section A.1.1. The next lemma says that the gradient of the log-likelihood at the actual parameter matrix Θ^* is controllably small.

Lemma 2.2.2 (**Bounded gradient**). *Let* $R = \max\left\{\sqrt{\frac{\sigma \log(2d)}{n}}, \frac{\sigma_{\min}^{-1/2}\log(2d)}{n}\right\}$. *The spectral norm of the gradient of the log-likelihood at the actual parameter matrix,* $\nabla\mathcal{L}(\Theta^*)$, *can be upper-bounded with high probability as follows,*

$$\mathbb{P}\left\{\left\|\nabla\mathcal{L}(\Theta^*)L^{-1/2}\right\|_2 \geq \sqrt{32}R\right\} \leq \frac{1}{(d_1+d_2)^3}\,. \tag{2.9}$$

Proof of the above lemma is in Section A.1.4. Let $\Delta = \hat{\Theta} - \Theta^*$.

Case 1: $\Delta \notin \mathcal{A}(2\alpha)$ Then,

$$|||\Delta|||_{\text{L}} \leq 32\alpha d_1 R |||\Delta|||_{\text{L-nuc}}\,.$$

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

$$\begin{aligned}-\mathcal{L}(\hat{\Theta}) = -\mathcal{L}(\Theta^*) &+ \langle\langle -\nabla\mathcal{L}(\Theta^*), \Delta\rangle\rangle \\ &+ \frac{1}{2n}\sum_{i=1}^{n} \psi\left(\langle\langle \Theta^*, X^{(i)}\rangle\rangle + s\langle\langle \Delta, X^{(i)}\rangle\rangle\right)\langle\langle \Delta, X^{(i)}\rangle\rangle^2\,,\end{aligned} \tag{2.10}$$

where $\psi(x) = e^x/(1+e^x)^2$, $x \in [-2\alpha, 2\alpha]$, and $s \in [0,1]$. Next using Lemma

2.2.1 and the fact that $\psi(x)$ attains its minimum at $x = 2\alpha$, we get

$$-\mathcal{L}(\hat{\Theta}) + \mathcal{L}(\Theta^*) + \langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle \geq \frac{1}{2n}\sum_{i=1}^{n} \psi(2\alpha) \langle\!\langle \Delta, X^{(i)} \rangle\!\rangle^2 \geq \frac{\psi(2\alpha)}{6d_1} |\!|\!| \Delta |\!|\!|_{\mathrm{L}}^2, \tag{2.11}$$

with probability at least $1 - 1/(d_1 + d_2)^3$. Since $\hat{\Theta}$ is the minimizer for the objective function (2.5), 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

$$\begin{aligned} \frac{\psi(2\alpha)}{6d_1} |\!|\!| \Delta |\!|\!|_{\mathrm{L}}^2 &\leq -\mathcal{L}(\hat{\Theta}) + \mathcal{L}(\Theta^*) + \langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle \\ &\leq \lambda \left(|\!|\!| \Theta^* |\!|\!|_{\text{L-nuc}} - \left|\!\left|\!\left| \hat{\Theta} \right|\!\right|\!\right|_{\text{L-nuc}} \right) + \langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle \\ &\leq \lambda \left(|\!|\!| \Delta |\!|\!|_{\text{L-nuc}} \right) + \langle\!\langle \nabla\mathcal{L}(\Theta^*) L^{-1/2}, \Delta L^{1/2} \rangle\!\rangle \qquad (2.12) \\ &\leq \lambda \left(|\!|\!| \Delta |\!|\!|_{\text{L-nuc}} \right) + \left\| \nabla\mathcal{L}(\Theta^*) L^{-1/2} \right\|_2 |\!|\!| \Delta |\!|\!|_{\text{L-nuc}}, \qquad (2.13) \end{aligned}$$

where the last two inequalities follow from the triangle inequality for the nuclear norm and the generalized Hölder's inequality. Now we put $\lambda = 2\sqrt{32}R$ and use Lemma 2.2.2 to get

$$|\!|\!| \Delta |\!|\!|_{\mathrm{L}}^2 \leq \frac{6d_1}{\psi(2\alpha)} \left(\lambda + \frac{\lambda}{2} \right) |\!|\!| \Delta |\!|\!|_{\text{L-nuc}} \leq \frac{9d_1\lambda}{\psi(2\alpha)} |\!|\!| \Delta |\!|\!|_{\text{L-nuc}}, \tag{2.14}$$

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

$$|\!|\!| \Delta |\!|\!|_{\mathrm{L}}^2 \leq 9 \left(\alpha + \frac{1}{\psi(2\alpha)} \right) d_1 \lambda |\!|\!| \Delta |\!|\!|_{\text{L-nuc}}. \tag{2.15}$$

Lemma 2.2.3. *If $\lambda \geq 2|\!|\!| \nabla\mathcal{L}(\Theta^*) |\!|\!|_2$, then we have*

$$|\!|\!| \Delta |\!|\!|_{\text{L-nuc}} \;\; \leq \;\; 4\sqrt{2r} |\!|\!| \Delta |\!|\!|_{\mathrm{L}} + 4 \sum_{j=\rho+1}^{\min\{d_1, d_2 - G\}} \sigma_j(\Theta^* L^{1/2}), \tag{2.16}$$

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

Proof of the above the lemma is in Section A.1.5. Finally, utilizing the

Lemma 2.2.3, we get

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

□

Since $\left|\left|\left|\Theta^*-\widehat{\Theta}\right|\right|\right|_{\mathrm{L}} \geq \sigma_{min}^{1/2}\left|\left|\left|\Theta^*-\widehat{\Theta}\right|\right|\right|_{\mathrm{F}}$, this theorem automatically gives us the error bound of $\left|\left|\left|\Theta^*-\widehat{\Theta}\right|\right|\right|_{\mathrm{F}}$. The above bound shows a natural splitting of the error into two terms, one corresponding to the *estimation error* for the rank-r component and the second one corresponding to the *approximation error* for how well one can approximate Θ^* with a rank-r matrix. We also give the following corollary for the exact low-rank case.

Corollary 2.2.4 (Exact rank-r matrix). *Under the same hypothesis as in Theorem 1, if Θ^* is exactly rank r, we get*

$$\frac{1}{\sqrt{d_1}}\left|\left|\left|\left(\Theta^*-\widehat{\Theta}\right)L^{1/2}\right|\right|\right|_{\mathrm{F}} \leq 576\left(\alpha+\frac{1}{\psi(2\alpha)}\right)\sqrt{r}\max\left\{\sqrt{\frac{\sigma\ d_1\log(2d)}{n}},\right.$$
$$\left.\frac{\sqrt{\left(\sigma_{\min}^{-1}d_1\right)}\ \log(2d)}{n}\right\}, \tag{2.17}$$

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

We conjecture that the second term in the maximization is only an artifact of the analysis and does not reflect the actual error. We have corroborated the conjecture using simulation results on graphs with very small spectral gap.

2.3 Information-theoretic lower bound

Now we also provide the information theoretic lower bound for the performance for the best estimator.

Theorem 2. *Suppose Θ^* has rank r. Under the graph based sampling model as described in Section 2.1, there is a universal numerical constant $c>0$*

such that

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\Big[\frac{1}{\sqrt{d_1}} \left|\!\left|\!\left| \widehat{\Theta} - \Theta^* \right|\!\right|\!\right|_{\mathrm{L}}\Big] \geq \quad c \min \left\{ e^{-\alpha}\sqrt{\frac{r\, d_1}{n}}, \frac{\alpha\sqrt{r}}{\sqrt{\mathrm{tr}\left(L_r^{\dagger}\right)}} \right\} , \tag{2.18}$$

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, $L = U\Lambda U^T$.

Proof for this lower bound has been relegated to Appendix A.2. It can be easily seen that when d_1 and d_2 are comparable, the lower bound matches the first term in the upper-bound given earlier, except for a polylog factor.

2.4 Corollaries for i.i.d. sampled pairs (complete graph)

It can be easily checked that when P is the uniform sampling matrix, the error bound we get here matches past results [20]. As a corollary to the graph sampling, we provide the following upper and lower bounds for the case of complete uniform graph G.

Corollary 2.4.1 (Complete graph G upper bound). *Under the same hypothesis as in Theorem 1, if G is a complete graph, we get*

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

with probability greater than $1 - 2/(2d)^3$, where $\sigma = \max\{(d_2-1)/d_1, 1\}$.

Corollary 2.4.2 (Complete graph G lower bound). *Suppose Θ^* has rank r. Under the previously described graph-based sampling model with the graph being a complete graph, there is a universal numerical constant $c > 0$*

such that

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\Big[\frac{1}{\sqrt{d_1\,(d_2-1)}} \big|\!\big|\!\big|\widehat{\Theta} - \Theta^*\big|\!\big|\!\big|_{\mathrm{F}}\Big] \geq \; c \, \min\left\{ e^{-\alpha}\sqrt{\frac{r\,d_1}{n}}\, , \, \frac{\alpha}{\sqrt{(d_2-1)}} \right\} \,, \tag{2.19}$$

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

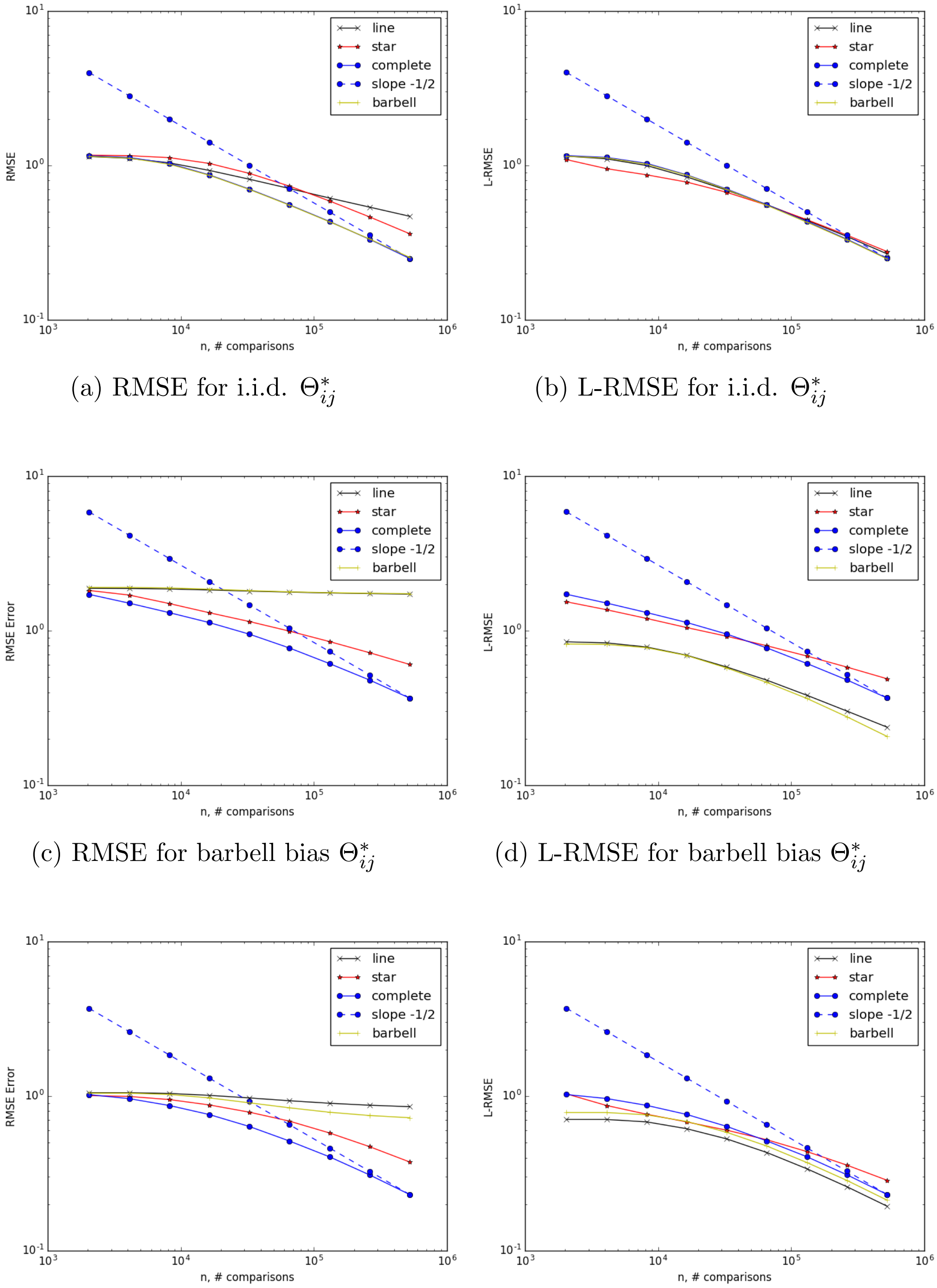
2.5 Simulation results

We present two experiments. One of the major challenges while solving the convex optimization problem (2.5) is the non-differentiable nuclear norm regularizer. We solve this issue by following the proximal gradient method as given in [22]. Another constraint, of zero row sum, is forced by adding a Frobenious norm regularizer to the objective function. We will not worry about the α constraint, as it would be automatically sorted out by the algorithm. Another issue was that convergence rates for some of the graph structures were slow; in particular, the star graph (which will be described in the following section) had an extremely slow convergence rate, which was 10-20 times slower than the slowest of the other graphs. To overcome this, we implemented a modified Barzilai-Borwein (BB) rule-based algorithm for accelerating the proximal gradient descent [23], which we found to be an extremely useful step-size free algorithm.

2.5.1 Error versus number of samples for different graphs

In Figure 2.1, we plot the error of our nuclear norm minimization-based algorithm versus number of samples (in log-scale), n, for $d_1 = d_2 = 300$, $r = 4$, $\alpha = 5.0$, and $G = 1$. We consider two errors here: root mean squared error (RMSE) $= \frac{|\!|\!|\Theta - \widehat{\Theta}|\!|\!|_{\mathrm{F}}}{\sqrt{d_1 d_2}}$ and Laplacian induced RMSE (L-RMSE)$=$ $\frac{|\!|\!|(\Theta-\widehat{\Theta})L^{1/2}|\!|\!|_{\mathrm{F}}}{\sqrt{d_1}}$. We plot these errors for four shapes of graph: 1) line graph, 2) star-shaped graph, 3) complete graph, and 4) barbell-shaped graph.

First, in Figure 2.1a and 2.1b, we plot RMSE and L-RMSE errors for different graphs using i.i.d. generated Θ^*_{ij}. We see that L-RMSE curves for



(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 2.1: Error versus number of samples for pairwise comparison with different graph shapes

different graphs are the same (and slopes in log-scale are, as expected, close to $-1/2$), whereas the RMSE barbell and complete graphs do slightly better. But we will see that these are not the worst distributions of Θ^*_{ij} for barbell and line graphs.

In Figure 2.1c and 2.1d, we again plot the errors for different graphs but for different non-i.i.d. generated Θ^*_{ij}. Here, the items are divided into two sets (corresponding to each side of the barbell graph), such that corresponding Θ^*_{ij}'s are i.i.d. inside a set but have similar but shifted means across the sets. We call this type of preference data *barbell biased*. As expected, L-RMSE behave similar to the i.i.d. case, with the complete and star graphs doing worse than the others. But the RMSE error blows up in the case of the line and barbell-shaped graphs because of the shifts in the mean.

Finally, in Figure 2.1e and 2.1f, we plot the errors for yet another type of non-i.i.d. generated Θ^*_{ij}'s. Here 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 we move in the descending order. We call this type of preference data *line biased*. As expected, the L-RMSE error behaves similar to that of the i.i.d. case. But again the RMSE error blows up in the case of the line and barbell-shaped graphs because of the shifting means.

2.5.2 Error versus number of groups

Next, we present the error versus number of components in a graph G, where each component is a complete graph and $d_1 = d_2 = 360$, $r = 4$, $\alpha = 5.0$, and $n = 2^{14}$. Figure 2.2 plots the errors versus G, when the components are solved together and separately using our algorithm. We see that solving the components together keeps the errors more or less the same as the number of groups increase, but if we are solving the groups separately, the error increases with the number of groups.

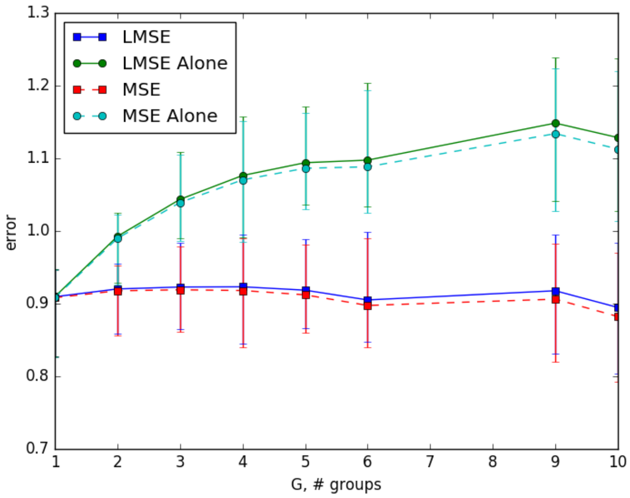


Figure 2.2: Error versus G, number of groups for comparison of graph sampled pairs

CHAPTER 3

K-WISE RANKING

3.1 Collaborative ranking from k-wise comparisons

Similar to the comparison of the graph-sampled pairs case, let Θ^* be the $d_1 \times d_2$ dimensional approximately r-rank 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 set-up, when a user i is presented with a set of k alternatives, $S_i \subseteq [d_2]$, she reveals her preferences as a ranked list over those items. To simplify the notations, we assume all users compare the same number k of items, but all the analysis generalizes to the case when the size might differ from user to user. According to the MNL model from Section 1.4, if $v_{i,\ell} \in S_i$ denotes the ℓ-th best choice of user i, then the probability of the ranking $\{v_{i,1}, v_{i,2}, \dots, v_{i,k}\}$ is

$$\mathbb{P}\{v_{i,1}, \dots, v_{i,k}\} = \prod_{\ell=1}^{k} \frac{e^{\Theta^*_{i,v_{i,\ell}}}}{\sum_{j \in S_{i,\ell}} e^{\Theta^*_{i,j}}} \,, \tag{3.1}$$

where $S_{i,\ell} \equiv S_i \setminus \{v_{i,1}, \dots, v_{i,\ell-1}\}$ and $S_{i,1} \equiv S_i$.

Similar to the graph sampling case, ranking distribution (3.1) 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 $i \in [d_1]$. Let $\alpha \equiv \max_{i,j_1,j_2} |\Theta^*_{ij_1} - \Theta^*_{ij_2}|$ denote the dynamic range of the underlying Θ^*, so that the probability of every ranking is bounded away from zero, by quantity of the order of $O(e^{-\alpha})$, which we assume to be a constant w.r.t. d_1 and d_2. Given this definition, we solve the following optimization

$$\widehat{\Theta} \in \arg\min_{\Theta \in \Omega} -\mathcal{L}(\Theta) + \lambda |||\Theta|||_{\text{nuc}}, \tag{3.2}$$

where

$$\mathcal{L}(\Theta) \;=\; \frac{1}{k\,d_1}\sum_{i=1}^{d_1}\sum_{\ell=1}^{k}\left(\langle\!\langle\Theta, e_i e_{v_{i,\ell}}^T\rangle\!\rangle - \log\left(\sum_{j\in S_{i,\ell}}\exp\left(\langle\!\langle\Theta, e_i e_j^T\rangle\!\rangle\right)\right)\right)\;, \tag{3.3}$$

and

$$\Omega_\alpha = \Big\{A \in \mathbb{R}^{d_1\times d_2} \,\big|\, |\!|\!|A|\!|\!|_\infty \le \alpha \text{ and } \sum_{j\in[d_2]} A_{ij} = 0,\;\; \forall i \in [d_1]\Big\}\,. \tag{3.4}$$

3.2 Performance guarantee

We provide an upper bound on the resulting error of our convex relaxation, when a *multi-set* of items S_i, drawn uniformly at random with replacement, is presented to user i. That is, for a given k, $S_i = \{j_{i,1},\dots,j_{i,k}\}$ where $j_{i,\ell}$'s are independently drawn uniformly at random over the d_2 items. If an item is sampled more than once, then we assume that the user treats these two instances of the item as if they were two distinct items with the same MNL weights. 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. [24]) (cf. Appendix B.1.4). Similar assumptions have been made in existing analyses on learning low-rank models from noisy observations, e.g. [13]. Let $d \equiv (d_1+d_2)/2$, and let $\sigma_j(\Theta^*)$ denote the j-th singular value of the matrix Θ^*. Define

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

Theorem 3. *Under the described sampling model, assume* $24\;\le k \le \min\{d_1^2\log d, (d_1^2+d_2^2)/(2d_1)\log d,\;(1/e)\,d_2(4\log d_2+2\log d_1)\}$*, and* $\lambda \in [480\lambda_0, c_0\lambda_0]$ *with any constant* $c_0 = O(1)$ *larger than 480. Then, solving the optimization* (3.2) *achieves*

$$\frac{1}{d_1 d_2}\Big|\!\Big|\!\Big|\widehat{\Theta}-\Theta^*\Big|\!\Big|\!\Big|_{\mathrm{F}}^2 \;\le\; 288\,e^{4\alpha}c_0\lambda_0\left(\sqrt{2r}\,\Big|\!\Big|\!\Big|\widehat{\Theta}-\Theta^*\Big|\!\Big|\!\Big|_{\mathrm{F}} + \sum_{j=r+1}^{\min\{d_1,d_2\}}\sigma_j(\Theta^*)\right) \tag{3.5}$$

for any $r \in \{1,\dots,\min\{d_1,d_2\}\}$ *with probability at least* $1-2d^{-3}-d_2^{-3}$ *where*

$d = (d_1 + d_2)/2$.

Proof. Let $\nabla \mathcal{L}(\Theta) \in \mathbb{R}^{d_1 \times d_2}$ be the gradient of the log-likelihood $\mathcal{L}(\Theta)$ (3.3) such that $\nabla_{ij}\mathcal{L}(\Theta) = \frac{\partial \mathcal{L}(\Theta)}{\partial \Theta_{ij}}$, and $\nabla^2 \mathcal{L}(\Theta) \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ be the its Hessian such that $\nabla^2_{ij,i'j'}\mathcal{L}(\Theta) = \frac{\partial^2 \mathcal{L}(\Theta)}{\partial \Theta_{ij} \partial \Theta_{i'j'}}$. By definition of $\mathcal{L}(\Theta)$ (3.3), we have

$$\nabla \mathcal{L}(\Theta^*) \;\; = \;\; -\frac{1}{k\, d_1} \sum_{i=1}^{d_1} \sum_{\ell=1}^{k} e_i (e_{v_{i,\ell}} - p_{i,\ell})^T \;, \tag{3.6}$$

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 $p_{j|(i,\ell)} \equiv \mathbb{P}\left\{v_{i,\ell} = j | v_{i,1}, \ldots, v_{i,\ell-1}, S_i\right\} = e^{\Theta^*_{ij}} / (\sum_{j' \in S_{i,\ell}} e^{\Theta_{ij'}})$ 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

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\Theta)}{\partial \Theta_{ij} \partial \Theta_{ij'}} &= \frac{1}{k\, d_1} \sum_{\ell=1}^{k} \mathbb{I}\big(j \in S_{i,\ell}\big) \frac{\partial p_{j|(i,\ell)}}{\partial \Theta_{ij'}} \\ &= \frac{1}{k\, d_1} \sum_{\ell=1}^{k} \mathbb{I}\big(j, j' \in S_{i,\ell}\big) \left(p_{j|(i,\ell)} \mathbb{I}(j = j') - p_{j|(i,\ell)} p_{j'|(i,\ell)}\right) . \end{aligned} \tag{3.7}$$

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

$$H^{(i)}(\Theta) = \frac{1}{k\, d_1} \sum_{\ell=1}^{k} \left(\text{diag}(p_{i,\ell}) - p_{i,\ell} p_{i,\ell}^T\right) \,. \tag{3.8}$$

Let $\Delta = \Theta^* - \widehat{\Theta}$ where $\widehat{\Theta}$ is the optimal solution of the convex program in (3.2). Now we first introduce three key technical lemmas. The first lemma shows that when Θ^* is approximately low rank, Δ is also approximately low-rank.

Lemma 3.2.1. *If $\lambda \geq 2 |||\nabla \mathcal{L}(\Theta^*)|||_2$, then we have*

$$|||\Delta|||_{\text{nuc}} \;\; \leq \;\; 4\sqrt{2r} |||\Delta|||_{\text{F}} + 4 \sum_{j=\rho+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) \;, \tag{3.9}$$

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

Proof of the above lemma is omitted since it is similar to that of Lemma 2.2.3. The next lemma proves that the actual parameter matrix Θ^* is close to the optimum, $\widehat{\Theta}$, of the optimization problem (3.2), in terms of the gradient at Θ^*.

Lemma 3.2.2. *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* $1 - 2d^{-c} - d_2^{-3}$*,*

$$|||\nabla\mathcal{L}(\Theta^*)|||_2 \leq \sqrt{\frac{4(1+c)\,\log d}{k\, d_1^2}} \max\Big\{\sqrt{d_1/d_2}, \\ e^{2\alpha}\sqrt{4(1+c)\log(d)}(8\log d_2 + 2\log d_1)\log k\Big\}\;. \tag{3.10}$$

The final lemma proves that $\mathcal{L}(\Theta)$ satisfies restricted strong convexity when Δ is small enough.

Lemma 3.2.3 (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 1.1, with probability larger than* $1 - 2d^{-2^{18}}$*,*

$$\text{Vec}(\Delta)^T\, \nabla^2\mathcal{L}(\Theta)\,\text{Vec}(\Delta) \;\;\geq\;\; \frac{e^{-4\alpha}}{24\, d_1 d_2}|||\Delta|||_{\text{F}}^2\;, \tag{3.11}$$

for all Δ *in* $\mathcal{A}$ *where*

$$\mathcal{A} = \Big\{\Delta \in \mathbb{R}^{d_1\times d_2}\;\big|\;|||\Delta|||_\infty \leq 2\alpha\,, \\ \sum_{j\in[d_2]}\Delta_{ij} = 0 \textit{ for all } i \in [d_1] \textit{ and } |||\Delta|||_{\text{F}}^2 \geq \mu|||\Delta|||_{\text{nuc}}\Big\}\;, \tag{3.12}$$

with

$$\mu \;\;\equiv\;\; 2^{10}\, e^{2\alpha}\,\alpha\, d_2\sqrt{\frac{d_1\,\log d}{k\,\min\{d_1, d_2\}}}\;. \tag{3.13}$$

Building on these lemmas, the proof of Theorem 3 is divided into the following two cases. In both cases, we will show that

$$|||\Delta|||_{\mathrm{F}}^2 \ \le\ 72\, e^{4\alpha} c_0 \lambda_0\, d_1 d_2\, |||\Delta|||_{\mathrm{nuc}}\ , \tag{3.14}$$

with high probability. Applying Lemma 3.2.1 proves the desired theorem. We are left to show that (3.14) holds.

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

$$\mathcal{L}(\widehat{\Theta}) = \mathcal{L}(\Theta^*) - \langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle + \frac{1}{2}\mathrm{Vec}(\Delta)\nabla^2\mathcal{L}(\Theta)\mathrm{Vec}^T(\Delta), \tag{3.15}$$

where $\Theta = a\widehat{\Theta} + (1-a)\Theta^*$ for some $a \in [0,1]$. It follows from Lemma 3.2.3 that with probability at least $1 - 2d^{-2^{18}}$,

$$\begin{aligned}\mathcal{L}(\widehat{\Theta}) - \mathcal{L}(\Theta^*) \ &\ge\ -\langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle + \frac{e^{-4\alpha}}{48\, d_1\, d_2}|||\Delta|||_{\mathrm{F}}^2 \\ &\ge\ -|||\nabla\mathcal{L}(\Theta^*)|||_2 |||\Delta|||_{\mathrm{nuc}} + \frac{e^{-4\alpha}}{48\, d_1\, d_2}|||\Delta|||_{\mathrm{F}}^2\ .\end{aligned}$$

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

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

By the assumption, we choose $\lambda \ge 480\lambda_0$. In view of Lemma 3.2.2, this implies that $\lambda \ge 2|||\nabla\mathcal{L}(\Theta^*)|||_2$ with probability at least $1 - 2d^{-3}$. It follows that with probability at least $1 - 2d^{-3} - 2d^{-2^{18}}$,

$$\frac{e^{-4\alpha}}{48 d_1 d_2}|||\Delta|||_{\mathrm{F}}^2 \ \le\ \left(\lambda + |||\nabla\mathcal{L}(\Theta^*)|||_2\right) |||\Delta|||_{\mathrm{nuc}} \ \le\ \frac{3\lambda}{2}|||\Delta|||_{\mathrm{nuc}}\ .$$

By our assumption of $\lambda \le c_0\lambda_0$, this proves the desired bound in (3.14).

Case 2: Suppose $|||\Delta|||_{\mathrm{F}}^2 \le \mu\, |||\Delta|||_{\mathrm{nuc}}$. By the definition of μ and the fact that $c_0 \ge 480$, it follows that $\mu \le 72\, e^{4\alpha} c_0 \lambda_0\, d_1 d_2$, and we get the same bound as in Eq. (3.14). □

Proofs of the lemmas are provided in Appendix B.1. Optimizing over the choices of r, we get the following corollaries.

Corollary 3.2.4 (**Exact low-rank matrices**). *Suppose* Θ^* *has rank at most* r. *Under the hypotheses of Theorem 3, solving the optimization* (3.2) *with the choice of the regularization parameter* $\lambda \in [480\lambda_0, c_0\lambda_0]$ *achieves with probability at least* $1 - 2d^{-3} - d_2^{-3}$,

$$\frac{1}{\sqrt{d_1 d_2}} \left|\!\left|\!\left| \widehat{\Theta} - \Theta^* \right|\!\right|\!\right|_{\mathrm{F}} \;\leq\; 288\sqrt{2}e^{6\alpha}c_0\sqrt{\frac{r(d_1 \log d + d_2\,(\log d)^2(\log 2d)^4)}{k\,d_1}} . \tag{3.16}$$

When Θ^* is a rank-r matrix, then the number of degrees of freedom in representing Θ^* is $r(d_1+d_2) - r^2 = O(r(d_1+d_2))$. The above corollary shows that for achieving an arbitrarily small error, the number of samples, $(k\,d_1)$, needs to scale as $O(rd_1(\log d) + rd_2\,(\log d)^2(\log 2d)^4)$, which is only a poly-logarithmic factor larger than the degrees of freedom of the matrix Θ^*. In Section 3.4, we directly provide a lower bound on the error using information theoretic method.

Now we relax the exact low-rank condition of the underlying matrix Θ^* and consider the more realistic scenario when it is only approximately low-rank. Following [13] we formalize this notion with "ℓ_q-ball" of matrices defined as

$$\mathbb{B}_q(\rho_q) \;\equiv\; \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j \in [\min\{d_1,d_2\}]} |\sigma_j(\Theta^*)|^q \leq \rho_q\} \,. \tag{3.17}$$

When $q = 0$, this is a set of rank-ρ_0 matrices, and when $q \in (0,1]$, this is a set of matrices whose singular values decay. Optimizing the choice of r in Theorem 3, we get the following result.

Corollary 3.2.5 (**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 3, solving the optimization* (3.2) *with the choice of the regularization parameter* $\lambda \in [480\lambda_0, c_0\lambda_0]$ *achieves with probability at least* $1 - 2d^{-3}$,

$$\begin{aligned}&\frac{1}{\sqrt{d_1 d_2}} \left|\!\left|\!\left| \widehat{\Theta} - \Theta^* \right|\!\right|\!\right|_{\mathrm{F}} \\ &\qquad \leq\; \frac{2\sqrt{\rho_q}}{\sqrt{d_1 d_2}} \left(288\sqrt{2}c_0 e^{6\alpha} \sqrt{\frac{d_1 d_2 (d_1 \log d + d_2\,(\log d)^2(\log 2d)^2)}{k\,d_1}} \right)^{\frac{2-q}{2}} .\end{aligned} \tag{3.18}$$

A proof of this Corollary is provided in Appendix B.2.

3.3 Experiments

3.3.1 Simulation results

The left panel of Figure 3.1 confirms the scaling of the error rate as predicted by Corollary 3.2.4. In the inset, we can see that the lines merge to a single line when the sample size is rescaled appropriately. We make a choice of $\lambda = (1/2)\sqrt{(\log d)/(kd^2)}$, since experimental results suggest that this provides small error (right panel). This choice is almost independent of α and is smaller than proposed in Theorem 3. We generate random rank-r matrices of dimension $d \times d$, where $\Theta^* = UV^T$ 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 row-mean is subtracted from each row, 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, since de-meaning can be written as $\Theta - \Theta\mathbb{1}\mathbb{1}^T/d_2$, and both terms in the operation are in the same column space as Θ, which is of rank r. The root mean squared error (RMSE) is plotted where RMSE $= (1/d)|||\Theta^* - \widehat{\Theta}|||_{\mathrm{F}}$. We implement and solve the convex optimization (3.2) using the proximal gradient descent method as analyzed in [22]. The right panel in Figure 3.1 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.

3.3.2 Jester dataset

The Jester dataset has 73×10^3 users who rate subsets of 100 jokes on a continuous scale of $[-10, 10]$. Since the scale is continuous, we can directly generate ordinal data from the scores. Only the users who rated all the jokes were used. For each user, k jokes were randomly selected in a biased manner, such that some jokes are more likely to get selected than others. Then our convex relaxation algorithm and the Borda count, a simple rank aggregator

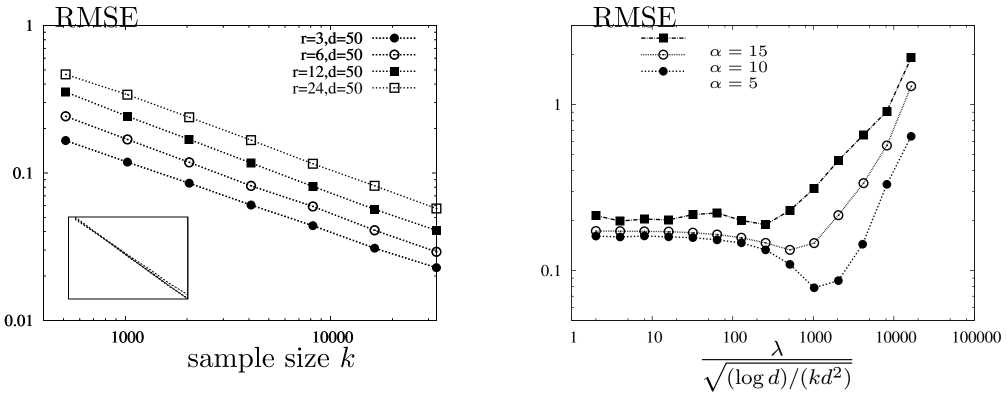


Figure 3.1: The (rescaled) RMSE scales as $\sqrt{r(\log d)/k}$ as expected from Corollary 3.2.4 for fixed $d = 50$ (left). In the inset, the same data is plotted versus rescaled sample size $k/(r \log d)$. The (rescaled) RMSE is stable for a broad range of λ and α for fixed $d = 50$ and $r = 3$ (right).

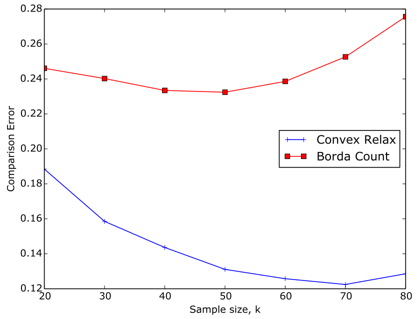


Figure 3.2: Average prediction error versus sample size for convex relaxation and Borda count

for learning a single ranking of the population, were used to predict outcomes of comparison among the remaining $100 - k$ jokes. Average error rates of the predictions for both methods are plotted for different values of k in Figure 3.2. The convex relaxation algorithm performs better, as expected, since it can predict personalized preference for each user.

3.4 Information-theoretic lower bound for k-wise ranking

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). We construct an appropriate packing over the set of low-rank matrices with bounded entries in Ω_α defined as (3.4), 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.

Theorem 4. *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*

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\Big[\frac{1}{\sqrt{d_1 d_2}} \left|\!\left|\!\left| \widehat{\Theta} - \Theta^* \right|\!\right|\!\right|_{\mathrm{F}}\Big] \geq c \min \left\{ \alpha e^{-\alpha} \sqrt{\frac{r\, d_2}{k\, d_1}} \,,\, \frac{\alpha d_2}{\sqrt{d_1 d_2 \log d}} \right\} , \tag{3.19}$$

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 B.3. 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 (3.5). It is the dominant term in the bound whenever the number of samples is larger than the number of 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 4 establishes that the upper bound in Theorem 3 is minimax-optimal up to a logarithmic factor in the dimension d.

3.5 Pairwise ranking breaking of k-wise ranking

In this section we consider a different algorithm for solving the k-wise ranking case. The new algorithm, called *rank breaking*, breaks up the k-wise rank into

$\binom{k}{2}$ pairwise comparisons and then solves an optimization problem which tries to maximize the likelihood of these newly generated pairwise comparisons, by considering them as independent events.

Assume the same observation model as in k-wise ranking. Assume that $u_{i,m}$, $i \in [k]$, $m \in [k]$, denotes the m-th element observed by the i-th user. The difference from the k-wise case is that here we convert the k-wise ranking data into pairwise ranking data, and then we solve the optimization problem as mentioned in k-wise ranking with a modified pairwise likelihood function,

$$\mathcal{L}(\Theta) = \frac{1}{d_1\binom{k}{2}} \sum_{i\in[d_1]} \sum_{(m_1,m_2)\in\mathcal{P}_0} \Theta_{i,\, h_i(m_1,m_2)} - \log\Big(\exp\big(\Theta_{i,\, u_{i,m_1}}\big) + \exp\big(\Theta_{i,\, u_{i,m_2}}\big)\Big) \ , \tag{3.20}$$

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

$$\hat{\Theta} \in \arg\min_{\Theta\in\Omega_\alpha} -\mathcal{L}(\Theta) + \lambda|||\Theta|||_{\text{nuc}} \, . \tag{3.21}$$

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

$$\lambda \ \equiv \ \sqrt{\frac{d \log d}{k\, d_1^2\, d_2}} \ . \tag{3.22}$$

Theorem 5. *Under the described sampling model, assume* $2(c + 4) \log d \le k \le \max\{d_1, d_2^2/d_1\} \log d$, $d_1 \ge 4$, *and* $\lambda \in [2\sqrt{32(c + 4)}\lambda, c_p\lambda]$ *with any constant* $c = O(1)$ *larger than* $2\sqrt{32(c + 4)}$. *Then, solving the optimization* (3.21) *achieves*

$$\frac{1}{d_1 d_2}\Big|\Big|\Big|\widehat{\Theta} - \Theta^*\Big|\Big|\Big|_{\text{F}}^2 \ \le \ 144\sqrt{2}\, e^{2\alpha} c\lambda\sqrt{r}\, \Big|\Big|\Big|\widehat{\Theta} - \Theta^*\Big|\Big|\Big|_{\text{F}} + 144 e^{2\alpha} c\lambda \sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^*) \ , \tag{3.23}$$

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 is provided in Appendix B.4.

Corollary 3.5.1 (Exact low-rank matrices). *Suppose* Θ^* *has rank at most* r. *Under the assumptions of Theorem 5, solving the optimization* (3.21) *with the choice of the regularization parameter* $\lambda \in [2\sqrt{32(c+4)}\lambda, c\lambda]$ *achieves with probability at least* $1 - 2d^{-c} - 2d^{-2^{13}}$,

$$\frac{1}{\sqrt{d_1 d_2}} \left|\left|\left| \widehat{\Theta} - \Theta^* \right|\right|\right|_{\mathrm{F}} \leq 144\sqrt{2} e^{2\alpha} c_p \sqrt{\frac{r d \log d}{k\, d_1}} \; . \tag{3.24}$$

From the above results we see that, error wise, the *rank breaking* algorithm does as well as the direct k-wise algorithm provided in Section 3.1. But due to the breaking of k-wise ranking into $O(k^2)$ pairwise comparisons, we lose a factor of $O(k)$ in the per iteration time complexity of any gradient based optimization methods.

CHAPTER 4

BUNDLED CHOICES

In this chapter we will look at a mathematically generalized version of n choices (cf. Section 1.3) called bundled choices.

4.1 Choice modeling for bundled purchase history

In this section, we use the MNL model to study another scenario of practical interest: choice modeling from bundled purchase history. For this scenario, 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. For example, there are d_1 toothpastes to choose from and d_2 toothbrushes 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. Given these sets of alternatives, each user makes a "bundled" purchase and we use (u_i, v_i) to denote the bundled pair of alternatives (e.g. a toothbrush and a toothpaste) 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

$$\mathbb{P}\left\{(u_i, v_i) = (j_1, j_2)\right\} \;=\; \frac{e^{\Theta^*_{j_1,j_2}}}{\sum_{j_1' \in S_i, j_2' \in T_i} e^{\Theta^*_{j_1',j_2'}}} \;, \tag{4.1}$$

for all $j_1 \in S_i$ and $j_2 \in T_i$. The distribution (4.1) 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: $[\Theta^*] \equiv \{A \in \mathbb{R}^{d_1 \times d_2} \mid A = \Theta^* + c\mathbb{1}\mathbb{1}^T \text{ for some } c \in \mathbb{R}\}$. Since we can only estimate Θ^* up to this equivalent class, we search for the ones that sum to zero, i.e. $\sum_{j_1\in[d_1],j_2\in[d_2]}\Theta^*_{j_1,j_2}=0$. Similar to k-wise ranking, 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 the probability of any choice is bounded away from zero. Assuming Θ^* is well approximated by a low-rank matrix, we solve the following convex relaxation, given the observed bundled purchase history $\{(u_i,v_i,S_i,T_i)\}_{i\in[n]}$:

$$\widehat{\Theta}\in\arg\min_{\Theta\in\Omega'_\alpha}\mathcal{L}(\Theta)+\lambda|||\Theta|||_{\rm nuc}\;, \tag{4.2}$$

where the (negative) log-likelihood function according to (4.1) is

$$\mathcal{L}(\Theta)=-\frac{1}{n}\sum_{i=1}^{n}\left(\langle\!\langle\Theta,e_{u_i}e_{v_i}^T\rangle\!\rangle-\log\left(\sum_{j_1\in S_i,j_2\in T_i}\exp\left(\langle\!\langle\Theta,e_{j_1}e_{j_2}^T\rangle\!\rangle\right)\right)\right)\,,\ \text{and} \tag{4.3}$$

$$\Omega_\alpha\equiv\left\{A\in\mathbb{R}^{d_1\times d_2}\;\middle|\;|||A|||_\infty\leq\alpha,\text{ and }\sum_{j_1\in[d_1],j_2\in[d_2]}A_{j_1,j_2}=0\right\}\,. \tag{4.4}$$

Notice that in this case we do not model individual preferences, but the preference of the whole population. Compared to collaborative ranking, rows and columns of Θ^* correspond to an alternative from the first and second category, respectively; each sample corresponds to the purchase choice of a user which follow the MNL model with Θ^*; each person is presented subsets S_i and T_i of items from each category; and each choice represents the most preferred bundled pair of alternatives from the set of alternatives presented to the user.

4.2 Performance guarantee

We provide an upper bound on the error achieved by our convex relaxation, when the *multi-set* 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. As in the previous chapters, sampling with replacement is necessary for the analysis. Define

$$\lambda = \sqrt{\frac{e^{2\alpha}\max\{d_1,d_2\}\log d}{n\, d_1\, d_2}}\ . \tag{4.5}$$

Theorem 6. *Under the described sampling model, assume* $16e^{2\alpha}\min\{d_1,d_2\}\log d \leq n \leq \min\{d^5, k_1k_2\max\{d_1^2,d_2^2\}\}\log d$*, and* $\lambda \in [8\lambda, c_1\lambda]$ *with any constant* $c_1 = O(1)$ *larger than* $\max\{8, 128/\sqrt{\min\{k_1,k_2\}}\}$*. Then, solving the optimization* (4.2) *achieves*

$$\frac{1}{d_1d_2}\left|\left|\left|\widehat{\Theta}-\Theta^*\right|\right|\right|_{\mathrm{F}}^2 \;\leq\; 48\sqrt{2}\,e^{2\alpha}c_1\lambda\sqrt{r}\,\left|\left|\left|\widehat{\Theta}-\Theta^*\right|\right|\right|_{\mathrm{F}} + 48e^{2\alpha}c_1\lambda\sum_{j=r+1}^{\min\{d_1,d_2\}}\sigma_j(\Theta^*)\ , \tag{4.6}$$

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 C.1. Optimizing over r gives the following corollaries.

Corollary 4.2.1 (Exact low-rank matrices). *Suppose* Θ^* *has rank at most* r*. Under the assumptions of Theorem 6, solving the optimization* (4.2) *with the choice of the regularization parameter* $\lambda \in [8\lambda, c_1\lambda]$ *achieves with probability at least* $1-2d^{-3}$*,*

$$\frac{1}{\sqrt{d_1d_2}}\left|\left|\left|\widehat{\Theta}-\Theta^*\right|\right|\right|_{\mathrm{F}} \;\leq\; 48\sqrt{2}e^{3\alpha}c_1\sqrt{\frac{r(d_1+d_2)\log d}{n}}\ . \tag{4.7}$$

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. For approximately low-rank matrices in an ℓ_1-ball as defined in (3.17), we show an upper bound on the error, whose error exponent reduces from 1 to $(2-q)/2$.

Corollary 4.2.2 (Approximately low-rank matrices). *Suppose* $\Theta^* \in \mathbb{B}_q(\rho_q)$ *for some* $q \in (0,1]$ *and* $\rho_q > 0$*. Under the assumptions of Theorem 6, solving the optimization* (4.2) *with the choice of the regularization parameter*

$\lambda \in [8\lambda, c_1\lambda]$ *achieves with probability at least* $1 - 2d^{-3}$,

$$\frac{1}{\sqrt{d_1 d_2}} \left|\!\left|\!\left| \widehat{\Theta} - \Theta^* \right|\!\right|\!\right|_{\mathrm{F}} \leq \frac{2\sqrt{\rho_q}}{\sqrt{d_1 d_2}} \left(48\sqrt{2} c_1 e^{3\alpha} \sqrt{\frac{d_1 d_2 (d_1 + d_2) \log d}{n}} \right)^{\frac{2-q}{2}} . \quad (4.8)$$

Since the proof is almost identical to the proof of Corollary 3.2.5 in Appendix B.2, we omit it.

4.3 Information-theoretic lower bound

As in previous chapters, we provide the lower bound on the worst-case error of the best possible estimator.

Theorem 7. *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*

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\Big[\frac{1}{\sqrt{d_1 d_2}} \left|\!\left|\!\left| \widehat{\Theta} - \Theta^* \right|\!\right|\!\right|_{\mathrm{F}} \Big] \geq c \min \left\{ \sqrt{\frac{e^{-5\alpha}\, r\, (d_1 + d_2)}{n}} \,,\, \frac{\alpha (d_1 + d_2)}{\sqrt{d_1 d_2 \log d}} \right\} . \quad (4.9)$$

See Appendix C.2 for the proof. The first term is dominant when the sample size is comparable to the latent dimension of the problem. This shows that Theorem 6 is minimax optimal up to a logarithmic factor.

4.4 Conclusion and future work

We presented measurement-efficient convex programs to learn MNL parameters from ordinal data, motivated by two scenarios: recommendation systems and bundled purchases. We gave algorithms to learn preferences from three different kinds of ordinal data: comparison of graph-sampled pairwise data, k-wise ranking, and bundles choices. We take the first-principles approach of identifying the fundamental limits and also developing efficient algorithms matching those fundamental trade-offs. There are several remaining challenges. First, the nuclear norm minimization, while polynomial-time, is still slow, due to the computation of the singular value decomposition (SVD) at every iteration. We want first-order methods that are efficient with provable guarantees. The main challenge is providing a good initialization to start such non-convex approaches. Second, we could extend the graph-sampling to both k-wise ranking and bundled choices. Finally, practical use of the model and the algorithm needs to be tested on real datasets of purchase history and recommendations.

One way to speed up the algorithm would be to optimize directly over the low rank decomposition of the matrix U, V where $\Theta = UV^T$, but this makes the optimization problem non-convex with possible local minima and saddle points. Then the challenge would be finding a good initialization point close to the global minimum for the iterative algorithm. Another way to reduce the runtime is by parallelizing the algorithm. There has been a recent work which does such parallelization [21].

Although our convex relaxation based algorithm achieves near optimal error with the available information, there is always the possibility of providing contextual information, such as the features of the item and users to improve the accuracy of the output. We could represent the parameter matrix as a function of the contextual information and then try to learn these functions instead of the parameters.

APPENDIX A

PROOFS OF GRAPH-SAMPLED PAIRS

A.1 Proof of Theorem 1: performance guarantee for comparison of graph-sampled pairs

A.1.1 Proof of Lemma 2.2.1

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 \geq \frac{1}{3d_1}\|\!|\Theta|\!\|_{\mathrm{L}}^2,\ \forall\, \Theta \in \mathcal{A}\right\} \\
=&1-\mathbb{P}\left\{\exists\, \Theta \in \mathcal{A} \ni \frac{1}{n}\sum_{i=1}^{n}\left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}\|\!|\Theta|\!\|_{\mathrm{L}}^2\right\} \qquad \text{(A.1)}
\end{aligned}$$

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

$$\|\!|\Theta|\!\|_{\mathrm{L}}^2 \geq 16\alpha d_1 R \|\!|\Theta|\!\|_{\text{L-nuc}} \geq 16\alpha d_1 R\|\!|\Theta|\!\|_{\mathrm{L}} \implies \|\!|\Theta|\!\|_{\mathrm{L}} \geq 16\alpha d_1 R := \mu. \qquad \text{(A.2)}$$

Lemma A.1.1. *Let* $\mathcal{B}(D) := \left\{\Theta \in \mathbb{R}^{d_1\times d_2} | \|\!|\Theta|\!\|_\infty \leq \alpha,\ \|\!|\Theta|\!\|_{\mathrm{L}} \leq D,\ \|\!|\Theta|\!\|_{\text{L-nuc}} \leq \frac{D^2}{16\alpha d_1 R}\right\}$ *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*

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

The above lemma is proved in Section A.1.2. Let $\beta = \sqrt{\frac{10}{9}}$, then the sets

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

for $\ell = 1, 2, 3, \ldots$ 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

$$\begin{aligned}
&\mathbb{P}\left\{\exists\, \Theta \in \mathcal{A} \ni \frac{1}{n}\sum_{i=1}^{n}\left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}|\!|\!|\Theta|\!|\!|_{\mathrm{L}}^2\right\} \\
&\leq \sum_{\ell=1}^{\infty}\mathbb{P}\left\{\exists\, \Theta \in \mathcal{S}_\ell \ni \frac{1}{n}\sum_{i=1}^{n}\left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}|\!|\!|\Theta|\!|\!|_{\mathrm{L}}^2\right\} \\
&\leq \sum_{\ell=1}^{\infty}\mathbb{P}\left\{\exists\, \Theta \in \mathcal{B}(\beta^\ell\mu) \ni \frac{1}{n}\sum_{i=1}^{n}\left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}|\!|\!|\Theta|\!|\!|_{\mathrm{L}}^2\right\} \qquad \text{(A.4)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}\left\{\exists\, \Theta \in \mathcal{A} \ni \frac{1}{n}\sum_{i=1}^{n}\left(\langle\!\langle \Theta, X^{(i)}\rangle\!\rangle\right)^2 < \frac{1}{3d_1}|\!|\!|\Theta|\!|\!|_{\mathrm{L}}^2\right\} &\leq \sum_{\ell=1}^{\infty}\mathbb{P}\left\{Z_{\beta^\ell\mu} > \frac{3}{2d_1}(\beta^\ell\mu)^2\right\} \\
&\overset{(a)}{\leq} \sum_{\ell=1}^{\infty}\exp\left(-\frac{n(\beta^\ell\mu)^4}{32\alpha^4 d_1^2}\right) \\
&\overset{(b)}{\leq} \sum_{\ell=1}^{\infty}\exp\left(-\frac{4\ell(\beta-1)n\mu^4}{32\alpha^4 d_1^2}\right) \\
&\overset{(c)}{\leq} 2\exp\left(-\frac{4(\beta-1)n\mu^4}{32\alpha^4 d_1^2}\right),
\end{aligned}$$

where (a) is from Lemma A.1.1, (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 in the previous inequality. Finally we get the desired result, when we have $2^2\log(2d) \leq 4(\beta-1)n\mu^4/32\alpha^4 d_1^2$, which follows from $n \leq 2^6 d_1^2\sigma^2\log(2d)$ and $n \leq 2^2(d_1\sigma_{\min}^{-1})^{2/3}\log(2d)$, because

$$\begin{aligned}
2^2\log(2d) \leq \frac{4(\beta-1)n\mu^4}{32\alpha^4 d_1^2} &= \frac{4(\beta-1)n(16\alpha d_1 R)^4}{32\alpha^4 d_1^2} \\
&= 2^{13}(\beta-1)d_1^2\max\left\{\frac{\sigma^2\log^2(2d)}{n}, \frac{\sigma_{\min}^{-2}\log^4(2d)}{n^3}\right\}. \qquad \text{(A.5)}
\end{aligned}$$

A.1.2 Proof of Lemma A.1.1

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

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

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 the supremum of deviation from the mean.

First, we use the bounded differences property of differences 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 realizations of the i-th $(1\le i\le n)$ random parameter of Z_D. Then

$$\begin{aligned}
&\left|Z_D(X^{(1)},\ldots,X^{(i)},\ldots,X^{(n)}) - Z_D(X^{(1)},\ldots,\tilde{X}^{(i)},\ldots,X^{(n)})\right| \\
=&\left|\sup_{\Theta\in\mathcal{B}(D)}\left(-\frac{1}{n}\sum_{i=1}^{n}\langle\langle\Theta, X^{(i)}\rangle\rangle^2 + \frac{2}{d_1}|||\Theta|||_{\mathrm{L}}^2\right)\right. \\
&\left.-\sup_{\Theta'\in\mathcal{B}(D)}\left(-\frac{1}{n}\left(\sum_{\substack{i=1\\i\ne i'}}^{n}\langle\langle\Theta', X^{(i)}\rangle\rangle^2 + \langle\langle\Theta', \tilde{X}^{(i')}\rangle\rangle^2\right) + \frac{2}{d_1}|||\Theta'|||_{\mathrm{L}}^2\right)\right|. \quad (A.6)
\end{aligned}$$

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

$$\begin{aligned}
&= \sup_{\Theta \in \mathcal{B}(D)} \left(-\frac{1}{n} \sum_{i=1}^{n} \langle\!\langle \Theta, X^{(i)} \rangle\!\rangle^2 + \frac{2}{d_1} |\!|\!|\Theta|\!|\!|_{\mathrm{L}}^2 \right) \\
&- \sup_{\Theta' \in \mathcal{B}(D)} \left(-\frac{1}{n} \left(\sum_{\substack{i=1 \\ i \neq i'}}^{n} \langle\!\langle \Theta', X^{(i)} \rangle\!\rangle^2 + \langle\!\langle \Theta', \tilde{X}^{(i')} \rangle\!\rangle^2 \right) + \frac{2}{d_1} |\!|\!|\Theta'|\!|\!|_{\mathrm{L}}^2 \right) \\
&\leq \left(-\frac{1}{n} \sum_{i=1}^{n} \langle\!\langle \bar{\Theta}, X^{(i)} \rangle\!\rangle^2 + \frac{2}{d_1} |\!|\!|\bar{\Theta}|\!|\!|_{\mathrm{L}}^2 \right) \\
&- \left(-\frac{1}{n} \left(\sum_{\substack{i=1 \\ i \neq i'}}^{n} \langle\!\langle \bar{\Theta}, X^{(i)} \rangle\!\rangle^2 + \langle\!\langle \bar{\Theta}, \tilde{X}^{(i')} \rangle\!\rangle^2 \right) + \frac{2}{d_1} |\!|\!|\bar{\Theta}|\!|\!|_{\mathrm{L}}^2 \right) \\
&\leq \sup_{\Theta \in \mathcal{B}(D)} \frac{1}{n} \left| \langle\!\langle \Theta, X^{(i)} \rangle\!\rangle^2 - \langle\!\langle \Theta, \tilde{X}^{(i)} \rangle\!\rangle^2 \right| \\
&\leq \frac{4\alpha^2}{n}, \qquad \text{(A.7)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[Z_D] &\overset{(a)}{\leq} 2\mathbb{E}\left[\sup_{\Theta \in \mathcal{B}(D)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle\!\langle \Theta, X^{(i)} \rangle\!\rangle^2 \right] \\
&\overset{(b)}{\leq} 4\alpha\mathbb{E}\left[\sup_{\Theta \in \mathcal{B}(D)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle\!\langle \Theta L^{1/2}, X^{(i)} L^{-1/2} \rangle\!\rangle \right] \\
&\leq 4\alpha\mathbb{E}\left[\sup_{\Theta \in \mathcal{B}(D)} |\!|\!|\Theta|\!|\!|_{\text{L-nuc}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X^{(i)} L^{-1/2} \right\|_2 \right] \\
&\leq 4\alpha \sup_{\Theta \in \mathcal{B}(D)} |\!|\!|\Theta|\!|\!|_{\text{L-nuc}} \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X^{(i)} L^{-`1/2} \right\|_2 \right],
\end{aligned}$$

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 the Ledoux-Talagrand contraction to obtain(b).

Lemma A.1.2. *For $\{X^{(i)}\}_{i=1}^n$ as defined in the graph sampling and for a* *binary random variable* ε_i *such that* $\mathbb{E}\left[\varepsilon_i | X^{(i)}\right] = 0$ *and* $|\varepsilon_i| \leq 1$*, we have*

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X^{(i)} L^{-1/2}\right\|_2 \geq \sqrt{32}\max\left\{\sqrt{\frac{\sigma\log(d_1+d_2)}{n}}, \frac{\sigma_{\min}^{-1/2}\log(d_1+d_2)}{n}\right\}\right\}$$

$$\leq \frac{1}{(d_1+d_2)^3} \quad \textit{and} \tag{A.8}$$

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X^{(i)} L^{-1/2}\right\|_2\right] \leq 4\max\left\{\sqrt{\frac{\sigma\log(d_1+d_2)}{n}}, \frac{\sigma_{\min}^{-1/2}\log(d_1+d_2)}{n}\right\}. \tag{A.9}$$

Proof of the lemma is in Section A.1.3. Now using Lemma A.1.2 we have $\mathbb{E}\left[Z_D\right] \leq 16R\alpha\sup_{\Theta\in\mathcal{B}(D)} |||\Theta|||_{\text{L-nuc}} \leq \frac{D^2}{d_1}$. Now using the bounded differences property and the upper bound on the mean, we get the McDiarmid's concentration,

$$\begin{aligned}\mathbb{P}\left\{Z_D - D^2/d_1 \geq t\right\} &\leq \mathbb{P}\left\{Z_D - \mathbb{E}\left[Z_D\right] \geq t\right\}\\ &\leq \exp\left(-\frac{nt^2}{8\alpha^4}\right),\end{aligned} \tag{A.10}$$

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

A.1.3 Proof of Lemma A.1.2

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 $L^{\dagger} = L^{-1}$, then, $\|W_i\|_2 \le \sigma_{\min}^{-1/2}\sqrt{2}/n$,

$$\begin{aligned}
\mathbb{E}\left[W_i W_i^T\right] &\preceq \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} e_{j(i)}\left(e_{k(i)} - e_{l(i)}\right)^T L^{-1/2} L^{-1/2}\left(e_{k(i)} - e_{l(i)}\right) e_{j(i)}^T\right] \\
&= \mathbb{E}\left[\frac{1}{n^2} e_{j(i)} e_{j(i)}^T\right] \mathbb{E}\left[\left(e_{k(i)} - e_{l(i)}\right)^T L^{\dagger}\left(e_{k(i)} - e_{l(i)}\right)\right] \\
&= \frac{1}{n^2 d_1}\mathbf{I}_{d_1\times d_1} \times 2\left(\mathbb{E}\left[e_{k(i)}^T L^{\dagger} e_{k(i)}\right] - \mathbb{E}\left[e_{k(i)}^T L^{\dagger} e_{l(i)}\right]\right) \\
&= \frac{2}{n^2 d_1}\left(\sum_{u\in[d_1]} P_u L^{\dagger}_{u,u} - \sum_{u,v\in[d_1]} P_{u,v} L^{\dagger}_{u,v}\right)\mathbf{I}_{d_1\times d_1} \\
&= \frac{2}{n^2 d_1}\langle\!\langle L, L^{\dagger}\rangle\!\rangle \mathbf{I}_{d_1\times d_1} \\
&\le \frac{2d_2}{n^2 d_1}\mathbf{I}_{d_1\times d_1}, \qquad \text{(A.11)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[W_i^T W_i\right] &\preceq L^{-1/2}\mathbb{E}\left[\frac{1}{n^2}\left(e_{k(i)} - e_{l(i)}\right)\left(e_{k(i)} - e_{l(i)}\right)^T\right] L^{-1/2} \\
&= \frac{1}{n^2} L^{-1/2}\left(\sum_{u,v=1}^{d_2}\left(e_u - e_v\right)\left(e_u - e_v\right)^T P_{u,v}\right) L^{-1/2} \\
&= \frac{1}{n^2} L^{-1/2}\left(2L\right) L^{-1/2} \\
&= \frac{2}{n^2} U U^T, \text{ and} \qquad \text{(A.12)}
\end{aligned}$$

$$\begin{aligned}
&\max\left\{\left\|\mathbb{E}\left[\sum_{i=1}^{n} W_i W_i^T\right]\right\|_2, \left\|\mathbb{E}\left[\sum_{i=1}^{n} W_i^T W_i\right]\right\|_2\right\} \\
&\quad \le \sum_{i=1}^{n}\max\left\{\left\|\mathbb{E}\left[W_i W_i^T\right]\right\|_2, \left\|\mathbb{E}\left[W_i^T W_i\right]\right\|_2\right\} \le \frac{2}{n}\sigma, \qquad \text{(A.13)}
\end{aligned}$$

where $\sigma = \max\left\{\frac{d_2 - G}{d_1}, 1\right\}$.

Now by matrix Bernstein concentration theorem [25], we have

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X^{(i)}\right\|_2 \geq t\right\} \leq \exp\left(\frac{-nt^2/2}{2\sigma + \sqrt{2}\sigma_{\min}^{-1/2}t/3}\right) \text{ and} \tag{A.14}$$

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X^{(i)}\right\|_2\right] \leq \sqrt{\frac{4\sigma\log(d_1+d_2)}{n}} + \frac{\sqrt{2\sigma_{\min}^{-1}}}{3n}\log(d_1+d_2)\,. \tag{A.15}$$

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

A.1.4 Proof of Lemma 2.2.2

The gradient can be written as

$$\nabla\mathcal{L}(\Theta^*) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{\exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}{1+\exp(\langle\!\langle\Theta^*, X^{(i)}\rangle\!\rangle)}\right)X^{(i)}. \tag{A.16}$$

Then Lemma A.1.2 directly gives the result because

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

A.1.5 Proof of Lemma 2.2.3

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 BV_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 $\mathcal{P}_T(M) = U_r U_r^T M + MV_rV_r^T - U_rU_r^T MV_rV_r^T$. The projection of M onto the complement space $T^\perp$ is $\mathcal{P}_{T^\perp}(M) = (I - U_rU_r^T)M(I - V_rV_r^T)$. 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 [12] 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 [11] that

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

Hence, in view of the triangle inequality,

$$\begin{aligned}\left|\!\left|\!\left|\widehat{\Theta} L^{1/2}\right|\!\right|\!\right|_{\text{nuc}} &= \left|\!\left|\!\left|\mathcal{P}_T(\Theta^* L^{1/2}) + \mathcal{P}_{T^\perp}(\Theta^* L^{1/2}) - \Delta' - \Delta''\right|\!\right|\!\right|_{\text{nuc}} \\ &\geq \left|\!\left|\!\left|\mathcal{P}_T(\Theta^* L^{1/2}) - \Delta''\right|\!\right|\!\right|_{\text{nuc}} - \left|\!\left|\!\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2}) - \Delta'\right|\!\right|\!\right|_{\text{nuc}} \\ &= \left|\!\left|\!\left|\mathcal{P}_T(\Theta^* L^{1/2})\right|\!\right|\!\right|_{\text{nuc}} + |\!|\!|\Delta''|\!|\!|_{\text{nuc}} - \left|\!\left|\!\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2}) - \Delta'\right|\!\right|\!\right|_{\text{nuc}} \\ &\geq \left|\!\left|\!\left|\mathcal{P}_T(\Theta^* L^{1/2})\right|\!\right|\!\right|_{\text{nuc}} + |\!|\!|\Delta''|\!|\!|_{\text{nuc}} - \left|\!\left|\!\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\!\right|\!\right|_{\text{nuc}} - |\!|\!|\Delta'|\!|\!|_{\text{nuc}} \\ &= \left|\!\left|\!\left|\Theta^* L^{1/2}\right|\!\right|\!\right|_{\text{nuc}} + |\!|\!|\Delta''|\!|\!|_{\text{nuc}} - 2\left|\!\left|\!\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\!\right|\!\right|_{\text{nuc}} - |\!|\!|\Delta'|\!|\!|_{\text{nuc}}.\end{aligned} \tag{A.17}$$

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

$$\begin{aligned}\lambda\left(\left|\!\left|\!\left|\widehat{\Theta} L^{1/2}\right|\!\right|\!\right|_{\text{nuc}} - |\!|\!|\Theta^* L^{1/2}|\!|\!|_{\text{nuc}}\right) &\leq -\mathcal{L}(\Theta^*) + \mathcal{L}(\widehat{\Theta}) \\ &\overset{(a)}{\leq} \langle\!\langle \Delta L^{1/2}, \nabla\mathcal{L}(\Theta^*) L^{-1/2}\rangle\!\rangle \\ &\overset{(b)}{\leq} |\!|\!|\Delta|\!|\!|_{\text{L-nuc}} \left|\!\left|\!\left|\nabla\mathcal{L}(\Theta^*) L^{-1/2}\right|\!\right|\!\right|_2 \leq \frac{\lambda}{2}|\!|\!|\Delta|\!|\!|_{\text{L-nuc}},\end{aligned} \tag{A.18}$$

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 (A.17) and (A.18) yields

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

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

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

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, $|\!|\!|\Delta'|\!|\!|_{\rm nuc} \leq \sqrt{2r}|\!|\!|\Delta'|\!|\!|_{\rm F} \leq \sqrt{2r}\left|\!\left|\!\left|\Delta L^{1/2}\right|\!\right|\!\right|_{\rm F} \leq \sqrt{2r}|\!|\!|\Delta|\!|\!|_{\rm L}$. Then the theorem follows because $\left|\!\left|\!\left|\mathcal{P}_{T^\perp}(\Theta^* L^{1/2})\right|\!\right|\!\right|_{\rm nuc} = \sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^* L^{1/2})$.

A.2 Proof of Theorem 2: information-theoretic graph sampling lower bound

The proof uses the Fano's 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 $\left|\!\left|\!\left|\Theta^{(\ell_1)} - \Theta^{(\ell_1)}\right|\!\right|\!\right|_{\rm L} \geq \delta$ for all $\ell_1, \ell_2 \in [M]$. Now, $\Theta^{(N)}$ is uniformly drawn from this set and then comparison results of n randomly chosen pairs of items, each drawn according to the probability matrix P and each compared by uniformly chosen user according to MNL model parameterized by $\Theta^{(N)}$, are generated. Let $\widehat{N}$ be the best estimator of N from the observations. Then we can show that

$$\sup_{\Theta^* \in \Omega_\alpha} \mathbb{P}\left\{ \left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\rm L}^2 \geq \frac{\delta^2}{2} \right\} \;\geq\; \mathbb{P}\left\{ \widehat{N} \neq N \right\}. \tag{A.19}$$

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

The generalized Fano's inequality along with data processing inequality gives us

$$\begin{aligned} \mathbb{P}\left\{\widehat{N} \neq N\right\} &\geq 1 - \frac{\mathbb{E}[\, I(\widehat{N}; N)\,] + \log 2}{\log M} && \text{(A.20)} \\ &\geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1,\ell_2 \in [M]} D_{\rm KL}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) + \log 2}{\log M}\,, && \text{(A.21)} \end{aligned}$$

where $D_{\rm KL}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$ 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.

$$D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$$

$$= n \sum_{i \in [d_1]} \frac{1}{d_1} \sum_{\{j,j'\} \subset [d_2]} 2P_{u,v} \Bigg[\frac{e^{\Theta_{ij}^{(\ell_1)}}}{e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}}} \log \left(\frac{e^{\Theta_{ij}^{(\ell_1)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}} \right)}{e^{\Theta_{ij}^{(\ell_2)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_2)}} + e^{\Theta_{ij'}^{(\ell_2)}} \right)} \right) \tag{A.22}$$

$$+ \frac{e^{\Theta_{ij'}^{(\ell_1)}}}{e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}}} \log \left(\frac{e^{\Theta_{ij'}^{(\ell_1)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_1)}} + e^{\Theta_{ij'}^{(\ell_1)}} \right)}{e^{\Theta_{ij'}^{(\ell_2)}} \Big/ \left(e^{\Theta_{ij}^{(\ell_2)}} + e^{\Theta_{ij'}^{(\ell_2)}} \right)} \right) \Bigg], \tag{A.23}$$

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 $y_{ijj'} \equiv e^{\Theta_{ij'}^{(\ell_2)}} / (e^{\Theta_{ij}^{(\ell_2)}} + e^{\Theta_{ij'}^{(\ell_2)}})$.

$$D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$$

$$\overset{(a)}{=} n \sum_{i \in [d_1]} \frac{1}{d_1} \sum_{\{j,j'\} \subset [d_2]} 2P_{u,v} \left[x_{ijj'} \log \frac{x_{ijj'}}{y_{ijj'}} + (1 - x_{ijj'}) \log \frac{1 - x_{ijj'}}{1 - y_{ijj'}} \right] \tag{A.24}$$

$$\overset{(b)}{\leq} n \sum_{i \in [d_1]} \frac{1}{d_1} \sum_{\{j,j'\} \subset [d_2]} 2P_{u,v} \left[x_{ijj'} \frac{x_{ijj'} - y_{ijj'}}{y_{ijj'}} + (1 - x_{ijj'}) \frac{y_{ijj'} - x_{ijj'}}{1 - y_{ijj'}} \right] \tag{A.25}$$

$$= 2n \sum_{i \in [d_1]} \frac{1}{d_1} \sum_{\{j,j'\} \subset [d_2]} \frac{(x_{ijj'} - y_{ijj'}) P_{u,v} (x_{ijj'} - y_{ijj'})}{y_{ijj'}(1 - y_{ijj'})} \tag{A.26}$$

$$\overset{(b)}{\leq} 8ne^{2\alpha} \sum_{i \in [d_1]} \frac{1}{d_1} \sum_{\{j,j'\} \subset [d_2]} (x_{ijj'} - y_{ijj'}) P_{u,v} (x_{ijj'} - y_{ijj'}), \tag{A.27}$$

where (a) is due to the fact that $\log(x/y) \leq (x-y)/y \leq (x-y)/y$ for $x/y \geq 0$ and (b) is true be-cause $|\Theta_{ij}^{(\ell_2)}| \leq \alpha$ 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 $(x_{ijj'} - y_{ijj'})^2 \le (f(\Theta^{(\ell_1)}_{ij} - \Theta^{(\ell_1)}_{ij'}) - f(\Theta^{(\ell_2)}_{ij} - \Theta^{(\ell_2)}_{ij'}))^2 \le ((\Theta^{(\ell_1)}_{ij} - \Theta^{(\ell_1)}_{ij'}) - (\Theta^{(\ell_2)}_{ij} - \Theta^{(\ell_2)}_{ij'}))^2$. This gives us

$$D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) \le \frac{8ne^{2\alpha}}{d_1} \sum_{i \in [d_1]} \sum_{\{j,j'\} \subset [d_2]} P_{u,v}((\Theta^{(\ell_1)}_{ij} - \Theta^{(\ell_2)}_{ij}) - (\Theta^{(\ell_1)}_{ij'} - \Theta^{(\ell_2)}_{ij'}))^2, \tag{A.28}$$

$$\begin{aligned}
&\overset{(a)}{\le} \frac{8ne^{2\alpha}}{d_1} \sum_{i \in [d_1]} (\Theta^{(\ell_1)} - \Theta^{(\ell_2)})_i L (\Theta^{(\ell_1)} - \Theta^{(\ell_2)})_i && \text{(A.29)}\\
&= \frac{8ne^{2\alpha}}{d_1} \sum_{i \in [d_1]} (\Theta^{(\ell_1)} - \Theta^{(\ell_2)})_i L (\Theta^{(\ell_1)} - \Theta^{(\ell_2)})_i && \text{(A.30)}\\
&= \frac{8ne^{2\alpha}}{d_1} |||(\Theta^{(\ell_1)} - \Theta^{(\ell_2)}) L^{1/2}|||_{\mathrm{F}}^2 && \text{(A.31)}\\
&= \frac{8ne^{2\alpha}}{d_1} |||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{L}}^2 , && \text{(A.32)}
\end{aligned}$$

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 A.21, we get

$$\mathbb{P}\left\{\widehat{N} \neq N\right\} \ge 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1, \ell_2 \in [M]} (8ne^{2\alpha}/d_1) |||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{L}}^2 + \log 2}{\log M} . \tag{A.33}$$

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

Lemma A.2.1. *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:*

$$\begin{aligned}
|||\Theta^{(\ell)}|||_{\mathrm{L}} &\le \delta \ , \ \textit{for all } \ell \in [M] && \text{(A.34)}\\
|||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{L}} &\ge \delta \ , \ \textit{for all } \ell_1, \ell_2 \in [M] && \text{(A.35)}\\
\Theta^{(\ell)} &\in \Omega_{\tilde{\alpha}} \ , \ \textit{for all } \ell \in [M] \ , && \text{(A.36)}
\end{aligned}$$

with $\tilde{\alpha} = \delta \sqrt{\mathrm{tr}\left(\Lambda_r^{\dagger}\right)} / \sqrt{rd_1}$.

Now if we assume $\delta \le \alpha \sqrt{rd_1} / \mathrm{tr}\left(\sqrt{\Lambda_r^{\dagger}}\right)$, we get $|||\Theta^{(\ell)}|||_{\infty}$ for $\ell \in [M]$.

The above lemma also implies that $\left|\!\left|\!\left|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right|\!\right|\!\right|_{\mathrm{F}}^2 \leq 4\delta^2$, which implies

$$\mathbb{P}\left\{\widehat{N} \neq N\right\} \geq 1 - \frac{32ne^{2\alpha}\delta^2/d_1 + \log 2}{rd_1/256} \geq \frac{1}{2}\,, \tag{A.37}$$

where the last inequality holds when $\delta \leq (e^{-\alpha}/128)\sqrt{rd_1^2/n}$. Along with (A.19), this proves that

$$\inf_{\widehat{\Theta}} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\left[\left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\mathrm{L}}\right] \geq \frac{\delta}{2}\,, \tag{A.38}$$

for all $\delta \leq \min\{\alpha\sqrt{rd_1}/\mathrm{tr}\left(\sqrt{\Lambda_r^\dagger}\right), (e^{-\alpha}/128)\sqrt{rd_1^2/n}\}$. Now maximizing the RHS proves the theorem.

A.2.1 Proof of Lemma

Inspired by the previous work that has been done, we furnish a 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:

$$\Theta^{(\ell)} \;=\; \frac{\delta}{\sqrt{rd_1}} V^{(\ell)} \sqrt{\Lambda_r^\dagger} U_r^T\,, \tag{A.39}$$

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, $|||\Theta^{(\ell)}|||_{\mathrm{L}} = (\delta/\sqrt{rd_1})|||V^{(\ell)}|||_{\mathrm{F}} \leq \delta$, since $|||V^{(\ell)}|||_{\mathrm{F}} = \sqrt{rd_1}$.

Define f as $f(V^{(\ell_1)}, V^{(\ell_2)}) \equiv \left|\!\left|\!\left|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}\right|\!\right|\!\right|_{\mathrm{L}}^2 = (\delta^2/(rd_1))\left|\!\left|\!\left|V^{(\ell_1)} - V^{(\ell_2)}\right|\!\right|\!\right|_{\mathrm{F}}^2$, 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

$$\begin{aligned}
\big|f(V^{(\ell_1)},&V^{(\ell_2)}) - f(\widetilde{V}^{(\ell_1)},\widetilde{V}^{(\ell_2)})\big| \\
&= \left| \frac{\delta^2}{r\,d_2} \left|\!\left|\!\left| V^{(\ell_1)} - V^{(\ell_2)} \right|\!\right|\!\right|_{\rm F}^2 - \frac{\delta^2}{r\,d_2} \left|\!\left|\!\left| V^{(\ell_1)} - V^{(\ell_2)} + 2e_{ij} \right|\!\right|\!\right|_{\rm F}^2 \right| \\
&= \left| \frac{\delta^2}{r\,d_2} |\!|\!| 2e_{ij} |\!|\!|_{\rm F}^2 + \frac{\delta^2}{r\,d_2} \langle\!\langle (V^{(\ell_1)} - V^{(\ell_2)}), 2e_{ij} \rangle\!\rangle \right| \\
&\leq \frac{4\,\delta^2}{r\,d_1} + \frac{\delta^2}{r\,d_1} \left|\!\left|\!\left| V^{(\ell_1)} - V^{(\ell_2)} \right|\!\right|\!\right|_\infty |\!|\!| 2e_{ij} |\!|\!|_1 \\
&\leq \frac{8\,\delta^2}{r\,d_1}\,, \qquad (A.40)
\end{aligned}$$

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

$$\begin{aligned}
\frac{\delta^2}{r\,d_1} \mathbb{E}\left[\left|\!\left|\!\left| (V^{(\ell_1)} - V^{(\ell_2)}) \right|\!\right|\!\right|_{\rm F}^2 \right] &= \frac{2\delta^2}{r\,d_1} \mathbb{E}\left[\left|\!\left|\!\left| V^{(\ell_1)} \right|\!\right|\!\right|_{\rm F}^2 \right] \\
&= 2\,\delta^2\,. \qquad (A.41)
\end{aligned}$$

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

$$\mathbb{P}\left\{ f(V^{(\ell_1)}, V^{(\ell_2)}) \leq 2\delta^2 - t \right\} \leq \exp\left\{ -\frac{t^2\,r\,d_1}{64\,\delta^4} \right\}. \qquad (A.42)$$

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

$$\mathbb{P}\left\{ \min_{\ell_1,\ell_2 \in [M]} \left|\!\left|\!\left| \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \right|\!\right|\!\right|_{\rm F}^2 \geq \delta^2 \right\} \geq 1 - \exp\left\{ -\frac{r\,d_1}{64} + 2\log M \right\} > 0\,. \qquad (A.43)$$

In the last step, we used $M = \lfloor \exp\{rd_1/\,256\} \rfloor$. At last we prove that $\Theta^{(\ell)}$'s are in $\Omega_{\delta\sqrt{{\rm tr}(\Lambda_r^\dagger)/rd_1}}$ as defined in (2.6). Since we know that g belongs to the kernel of L for all $g \in \mathcal{G}$, $\Theta^{(\ell)} g = 0$ by construction. From (A.39), 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 $\{-1,+1\}^r$.

$$\left| \langle\!\langle v_i, \sqrt{\Lambda_r^\dagger}(u_r)_j \rangle\!\rangle \right| \leq |\!|\!| v_i |\!|\!|_\infty \left|\!\left|\!\left| \sqrt{\Lambda_r^\dagger}(u_r)_j \right|\!\right|\!\right|_1 \leq \sqrt{{\rm tr}\left(\Lambda_r^\dagger\right)}\,. \qquad (A.44)$$

The above proves that $\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \Theta^{(\ell)} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_{\infty}$ is upper bounded as desired.

APPENDIX B

PROOFS OF K-WISE RANKING

B.1 Proof of Theorem 3: performance guarantee for k-wise ranking

B.1.1 Proof of Lemma 3.2.2

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 $\lvert\lvert\lvert X_i \rvert\rvert\rvert_2$ 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 $\lvert\lvert\lvert X_i \rvert\rvert\rvert_2 = O(\sqrt{k})$ for all $i \in [d_1]$.

Lemma B.1.1. *For a fixed $i \in [d_1]$ and $j \in [d_2]$, if $k \le (1/e)\, d_2\, (4 \log d_2 + \log d_1)$, 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

$|||\left(\sum_i X_i\right) \mathbb{I}\left(\mathcal{A}\right)|||_2$ using the matrix Bernstein inequality [26].

$$\begin{aligned}
|||X_i \mathbb{I}\left(\mathcal{A}\right)|||_2 &= \left\| \mathbb{I}\left(\mathcal{A}\right) \sum_{\ell=1}^{k} \left(e_{v_{i,\ell}} - p_{i,\ell}\right) \right\| \\
&\overset{(a)}{\leq} \left\| \mathbb{I}\left(\mathcal{A}\right) \sum_{\ell=1}^{k} e_{v_{i,l}} \right\| + \left\| \mathbb{I}\left(\mathcal{A}\right) \sum_{\ell=1}^{k} p_{i,\ell} \right\| \\
&\overset{(b)}{\leq} (8(\log d_2) + 2(\log d_1)) \sqrt{\min\{k, d_2\}} \left(1 + \ \left(\sum_{\ell=1}^{k} \frac{e^{2\alpha}}{\ell} \right) \right) \\
&\overset{(c)}{\leq} \sqrt{k}(8(\log d_2) + 2(\log d_1)) \big(1 + 2e^{2\alpha} \ \log k\big) \\
&\leq 3\sqrt{k}(8(\log d_2) + 2(\log d_1)) e^{2\alpha} \ \log k \ , \qquad \text{(B.1)}
\end{aligned}$$

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 $\min\{\sqrt{d_2, k}\}$ 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

$$\begin{aligned}
\left|\left|\left| \sum_i \mathbb{E}\left[X_i X_i^T \mathbb{I}\left(\mathcal{A}\right)\right] \right|\right|\right|_2 &\leq \left|\left|\left| \sum_i \mathbb{E}\left[X_i X_i^T\right] \right|\right|\right|_2 \\
&\leq \left|\left|\left| \sum_{i=1}^{d_1} e_i e_i^T \mathbb{E} \left[\sum_{\ell,\ell'=1}^{k} \left(e_{v_{i,\ell}} - p_{i,\ell}\right)^T \left(e_{v_{i,\ell'}} - p_{i,\ell'}\right) \right] \right|\right|\right|_2 \\
&= \left|\left|\left| \sum_{i=1}^{d_1} e_i e_i^T \mathbb{E} \left[\sum_{\ell=1}^{k} \left(e_{v_{i,\ell}} - p_{i,\ell}\right)^T \left(e_{v_{i,\ell}} - p_{i,\ell}\right) \right] \right|\right|\right|_2 \\
&= \left|\left|\left| \sum_{i=1}^{d_1} e_i e_i^T \mathbb{E} \left[\sum_{\ell=1}^{k} e_{v_{i,\ell}}^T e_{v_{i,\ell}} - p_{i,\ell}^T p_{i,\ell} \right] \right|\right|\right|_2 \\
&\leq \left|\left|\left| \sum_{i=1}^{d_1} e_i e_i^T \mathbb{E} \left[\sum_{\ell=1}^{k} e_{v_{i,\ell}}^T e_{v_{i,\ell}} \right] \right|\right|\right|_2 \\
&= \ k |||\mathbf{I}_{d_1 \times d_1}|||_2 = k \qquad \text{(B.2)}
\end{aligned}$$

and

$$\begin{aligned}
\left|\!\left|\!\left|\sum_{i=1}^{d_1}\mathbb{E}\left[X_i^T X_i \mathbb{I}\left(\mathcal{A}\right)\right]\right|\!\right|\!\right|_2 &\leq \left|\!\left|\!\left|\sum_{i=1}^{d_1}\mathbb{E}\left[X_i^T X_i\right]\right|\!\right|\!\right|_2 \\
&\leq \left|\!\left|\!\left|\sum_{i=1}^{d_1}\mathbb{E}\left[\sum_{\ell,\ell'=1}^{k}(e_{v_{i,\ell}} - p_{i,\ell})(e_{v_{i,\ell'}} - p_{i,\ell'})^T\right]\right|\!\right|\!\right|_2 \\
&= \left|\!\left|\!\left|\sum_{i=1}^{d_1}\mathbb{E}\left[\sum_{\ell=1}^{k}(e_{v_{i,\ell}} - p_{i,\ell})(e_{v_{i,\ell}} - p_{i,\ell})^T\right]\right|\!\right|\!\right|_2 \qquad \text{(B.3)} \\
&= \left|\!\left|\!\left|\sum_{i=1}^{d_1}\mathbb{E}\left[\sum_{\ell=1}^{k} e_{v_{i,\ell}} e_{v_{i,\ell}}^T - p_{i,\ell}p_{i,\ell}^T\right]\right|\!\right|\!\right|_2 \\
&\leq \left|\!\left|\!\left|\sum_{i=1}^{d_1}\mathbb{E}\left[\sum_{\ell=1}^{k} e_{v_{i,\ell}} e_{v_{i,\ell}}^T\right]\right|\!\right|\!\right|_2 \\
&= \left|\!\left|\!\left|\sum_{i=1}^{d_1}\frac{k}{d_2}\mathbf{I}_{d_2\times d_2}\right|\!\right|\!\right|_2 = \frac{kd_1}{d_2}\ . \qquad \text{(B.4)}
\end{aligned}$$

By matrix Bernstein inequality [26],

$$\begin{aligned}
&\mathbb{P}\Big(|\!|\!|\nabla\mathcal{L}(\Theta^*)\mathbb{I}\left(\mathcal{A}\right)|\!|\!|_2 > t\Big) \\
&\quad \leq (d_1 + d_2)\exp\Big(\frac{-k^2\, d_1^2\, t^2/2}{(d_1 k/\min\{d_2, d_1\}) + (3e^{2\alpha}k^{3/2}d_1(\log d_2^8 d_1^2)\log k\ t/3)}\Big)\ , \qquad \text{(B.5)}
\end{aligned}$$

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

$$\begin{aligned}
t &= \max\left\{\sqrt{\frac{4(1+c)\ \log d}{k\, d_1\ \min\{d_2, d_1\}}}\,,\ \frac{4(1+c)e^{2\alpha}\log(d)\ (8(\log d_2) + 2(\log d_1))\log k}{k^{1/2}\, d_1}\right\} \qquad \text{(B.6)} \\
&= \frac{\sqrt{4(1+c)\ \log d}}{k^{1/2}\, d_1}\max\left\{\sqrt{\frac{d_1}{d_2}}\,,\ \ e^{2\alpha}\sqrt{4(1+c)\log(d)}\ (\log d_2^8 d_1^2)\log k\right\}. \qquad \text{(B.7)}
\end{aligned}$$

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

B.1.2 Proof of Lemma B.1.1

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^j_δ. Then, $\mathbb{P}\left\{A^j_\delta\right\} \le \binom{k}{\delta}\frac{1}{d_2^\delta} \le \left(\frac{ke}{d_2\delta}\right)^\delta$. Using the fact that $(1/x)^x \le a$ for any $x \ge (2\log(1/a))/(\log\log(1/a))$, we let $x = d_2\delta/(ke)$ to get

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

for $\delta \ge (ke/d_2)(2\log(1/a))/(\log\log(1/a))$. Choosing $a = (1/d_2^4 d_1)^{d_2/ke}$, we have $\mathbb{P}\left\{A^j_\delta\right\} \le 1/(d_1 d_2^4)$, for a choice of
$\delta = 2\ \log(d_2^4 d_1) \ge 2\log(d_2^4 d_1)/(\log((d_2/ke)\log(d_2^4 d_1)))$.

B.1.3 Proof of Lemma 3.2.3

Recall that the Hessian matrix is a block-diagonal matrix with the i-th block $H^{(i)}(\Theta)$ given by (3.8). We use the following remark from [17] to bound the Hessian.

Remark B.1.2 ([17, 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 $|\theta_i| \le \alpha$, for all $i \in [\rho]$, then*

$$e^{2\alpha}\Big(\mathrm{diag}(p) - pp^T\Big) \;\succeq\; \frac{1}{\rho}\mathrm{diag}(\mathbb{1}) - \frac{1}{\rho^2}\mathbb{1}\mathbb{1}^T \; .$$

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

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

Hence,

$$\begin{aligned}
\mathrm{Vec}(\Delta) \nabla^2 \mathcal{L}(\Theta) \mathrm{Vec}^T(\Delta) &= \sum_{i=1}^{d_1} (\Delta^T e_i)^T H^{(i)}(\Theta) (\Delta^T e_i) \\
&\geq \frac{e^{-2\alpha}}{2\, k^3\, d_1} \sum_{i=1}^{d_1} \sum_{\ell=1}^{k} \sum_{j,j' \in S_{i,\ell}} \left|\!\left|\!\left| e_i^T \Delta (e_j - e_{j'}) \right|\!\right|\!\right|_2^2 .
\end{aligned}$$

By changing the order of the summation, we get that

$$\begin{aligned}
&\sum_{\ell=1}^{k} \sum_{j,j' \in S_{i,\ell}} \left|\!\left|\!\left| e_i^T \Delta (e_j - e_{j'}) \right|\!\right|\!\right|_2^2 \\
&\quad = \sum_{\ell,\ell'=1}^{k} \langle\!\langle \Delta, e_{i,j_{i,\ell}} - e_{i,j_{i,\ell'}} \rangle\!\rangle^2 \sum_{\ell''=1}^{k} \mathbb{I}\big(\, \sigma_i(j_{i,\ell''}) \leq \min\{\sigma_i(j_{i,\ell}), \sigma_i(j_{i,\ell'})\} \,\big).
\end{aligned} \tag{B.8}$$

Define

$$\chi_{i,\ell,\ell',\ell''} \;\equiv\; \mathbb{I}\big(\, \sigma_i(j_{i,\ell''}) \leq \min\{\sigma_i(j_{i,\ell}), \sigma_i(j_{i,\ell'})\} \,\big) \;, \tag{B.9}$$

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 $\mathrm{Vec}^T(\Delta) \nabla^2 \mathcal{L}(\Theta) \mathrm{Vec}(\Delta) \geq H(\Delta)$. 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

$$\begin{aligned}\mathbb{E}\Big[\sum_{\ell''=1}^{k}\chi_{i,\ell,\ell',\ell''}\,\big|\,j_{i,1},\ldots,j_{i,k}\Big] &= 1+\sum_{\ell''\neq\ell,\ell'}\frac{\exp(\theta_{i,j_{i,\ell''}})}{\exp(\theta_{i,j_{i,\ell''}})+\exp(\theta_{i,j_{i,\ell'}})+\exp(\theta_{i,j_{i,\ell}})}\\ &\geq 1+\frac{k-2}{1+2e^{2\alpha}}\geq\frac{k}{3e^{2\alpha}}.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}[H(\Delta)] &= \frac{e^{-2\alpha}}{2\,k^3\,d_1}\sum_{i,\ell,\ell'}\mathbb{E}\Big[\langle\!\langle\Delta,e_{i,j_{i,\ell}}-e_{i,j_{i,\ell'}}\rangle\!\rangle^2\mathbb{E}\big[\sum_{\ell''=1}^{k}\chi_{i,\ell,\ell',\ell''}\,\big|\,j_{i,1},\ldots,j_{i,k}\big]\Big]\\ &\geq \frac{e^{-4\alpha}}{6\,k^2\,d_1}\sum_{i=1}^{d_1}\sum_{\ell,\ell'\in[k]}\mathbb{E}\left[\langle\!\langle\Delta,e_{i,j_{i,\ell}}-e_{i,j_{i,\ell'}}\rangle\!\rangle^2\right]\\ &= \frac{e^{-4\alpha}}{6\,k^2\,d_1}\sum_{i=1}^{d_1}\sum_{\ell\neq\ell'\in[k]}\left(\frac{2}{d_2}\sum_{j=1}^{d_2}\Delta_{ij}^2-\frac{2}{d_2^2}\sum_{j,j'=1}^{d_2}\Delta_{ij}\Delta_{ij'}\right)\\ &= \frac{e^{-4\alpha}(k-1)}{3\,k\,d_1\,d_2}|\!|\!|\Delta|\!|\!|_{\rm F}^2\,, \qquad\qquad (B.10)\end{aligned}$$

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 (3.11) is violated, i.e. $H(\Delta)<(e^{-4\alpha}/(24\,d_1d_2))|\!|\!|\Delta|\!|\!|_{\rm F}^2$. We will show this happens with a small probability. From (B.10), we get that for $k\geq 24$,

$$\begin{aligned}\mathbb{E}[H(\Delta)]-H(\Delta) &\geq \frac{(7k-8)}{24k}\frac{e^{-4\alpha}}{d_1\,d_2}|\!|\!|\Delta|\!|\!|_{\rm F}^2\\ &\geq \frac{(20/3)\,e^{-4\alpha}}{24\,d_1d_2}|\!|\!|\Delta|\!|\!|_{\rm F}^2\,. \qquad\qquad (B.11)\end{aligned}$$

We use a peeling argument as in [13, Lemma 3], [27] to upper bound the probability that(B.11) 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 (3.13). Notice that since for any $\Delta\in\mathcal{A}$, $|\!|\!|\Delta|\!|\!|_{\rm F}^2\geq\mu|\!|\!|\Delta|\!|\!|_{\rm nuc}\geq\mu|\!|\!|\Delta|\!|\!|_{\rm F}$, it follows that $|\!|\!|\Delta|\!|\!|_{\rm F}\geq\mu$.

Then, we can cover $\mathcal{A}$ with the family of sets

$$\mathcal{S}_\ell = \Big\{ \Delta \in \mathbb{R}^{d_1\times d_2} \,\Big|\, |||\Delta|||_\infty \le 2\alpha \,,\, \beta^{\ell-1}\mu \le |||\Delta|||_{\mathrm{F}} \le \beta^\ell \mu \,, \\ \sum_{j\in[d_2]} \Delta_{ij} = 0 \text{ for all } i \in [d_1], \text{ and } |||\Delta|||_{\mathrm{nuc}} \le \beta^{2\ell}\mu \Big\} \,, \tag{B.12}$$

where $\beta = \sqrt{10/9}$ and for $\ell \in \{1,2,3,\ldots\}$, which implies that when there exists a $\Delta \in \mathcal{A}$ such that (B.11) holds, then there exists an $\ell \in \mathbb{Z}_+$ such that $\Delta \in \mathcal{S}_\ell$ and

$$\begin{aligned} \mathbb{E}[H(\Delta)] - H(\Delta) &\ge \frac{(20/3)\, e^{-4\alpha}}{24\, d_1 d_2} \beta^{2(\ell-1)}\mu^2 \\ &\ge \frac{e^{-4\alpha}}{4\, d_1 d_2} \beta^{2\ell}\mu^2 \,. \end{aligned} \tag{B.13}$$

Applying the union bound over $\ell \in \mathbb{Z}_+$, we get from (B.11) and (B.13) that

$$\begin{aligned} &\mathbb{P}\Big\{ \exists \Delta \in \mathcal{A} \,,\; H(\Delta) < \frac{e^{-4\alpha}}{24\, d_1 d_2} |||\Delta|||_{\mathrm{F}}^2 \Big\} \\ &\le \sum_{\ell=1}^{\infty} \mathbb{P}\Big\{ \sup_{\Delta\in\mathcal{S}_\ell} \big(\, \mathbb{E}[H(\Delta)] - H(\Delta)\, \big) > \frac{e^{-4\alpha}}{4\, d_1 d_2} (\beta^\ell \mu)^2 \Big\} \\ &\le \sum_{\ell=1}^{\infty} \mathbb{P}\Bigg\{ \sup_{\Delta\in\mathcal{B}(\beta^\ell\mu)} \big(\, \mathbb{E}[H(\Delta)] - H(\Delta)\, \big) > \frac{e^{-4\alpha}}{4\, d_1 d_2} (\beta^\ell \mu)^2 \Bigg\} \,, \end{aligned} \tag{B.14}$$

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

$$\mathcal{B}(D) = \big\{ \Delta \in \mathbb{R}^{d_1\times d_2} \,\big|\, \|\Delta\|_\infty \le 2\alpha, |||\Delta|||_{\mathrm{F}} \le D, \\ \sum_{j\in[d_2]} \Delta_{ij} = 0 \text{ for all } i \in [d_1], \mu |||\Delta|||_{\mathrm{nuc}} \le D^2 \Big\} \,. \tag{B.15}$$

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

Lemma B.1.3. *For* $(16 \min\{d_1, d_2\} \log d)/(3d_1) \leq k \leq d_1^2 \log d$,

$$\mathbb{P}\left\{ \sup_{\Delta \in \mathcal{B}(D)} \Big(\, \mathbb{E}[H(\Delta)] - H(\Delta) \, \Big) \geq \frac{e^{-4\alpha}}{4d_1 d_2} D^2 \right\} \leq \quad \exp\Big\{ - \frac{e^{-4\alpha} \, k \, D^4}{2^{19} \alpha^4 d_1 d_2^2} \Big\} \; . \tag{B.16}$$

Let $\eta = \exp\Big(- \frac{e^{-4\alpha} 4k(\beta - 1.002)\mu^4}{2^{19}\alpha^4 d_1 d_2^2} \Big)$. Applying the tail bound to (B.14), we get

$$\begin{aligned} \mathbb{P}\left\{ \exists \Delta \in \mathcal{A} \; , \; H(\Delta) < \frac{e^{-4\alpha}}{24 \, d_1 d_2} |||\Delta|||_{\rm F}^2 \right\} &\leq \sum_{\ell=1}^{\infty} \exp\Big\{ - \frac{e^{-4\alpha} k (\beta^\ell \mu)^4}{2^{19}\alpha^4 d_1 d_2^2} \Big\} \\ &\overset{(a)}{\leq} \sum_{\ell=1}^{\infty} \exp\Big\{ \frac{-e^{-4\alpha} 4k\ell(\beta - 1.002)\mu^4}{2^{19}\alpha^4 d_1 d_2^2} \Big\} \\ &\leq \frac{\eta}{1-\eta}, \end{aligned} \tag{B.17}$$

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,

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

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.

B.1.4 Proof of Lemma B.1.3

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 $Z - \mathbb{E}[Z] \leq$

$e^{-4\alpha}D^2/(64d_1d_2)$ with high probability, and $\mathbb{E}[Z] \le 9e^{-4\alpha}D^2/(40d_1d_2)$. This proves the desired claim in Lemma B.1.3.

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 [28, 29, 30], 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 $\chi_{i,\ell,\ell',\ell''} = \mathbb{I}\left(u_{i,\ell''} \ge \max\{u_{i,\ell},u_{i,\ell'}\}\right)$. 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 $x'_{i,\ell}$,

$$
\begin{aligned}
&\big|\, f\big(\, x_{1,1},\ldots,x_{i,\ell},\ldots,x_{d_1,k}\,\big) - f\big(\, x_{1,1},\ldots,x'_{i,\ell},\ldots,x_{d_1,k}\,\big)\,\big| \\
&= \big|\sup_{\Delta\in\mathcal{B}(D)} \left(\mathbb{E}\left[H(\Delta)\right] - H(\Delta,x_{1,1},\ldots,x_{i,\ell},\ldots,x_{d_1,k})\right) - \\
&\sup_{\Delta\in\mathcal{B}(D)} \big(\mathbb{E}\left[H(\Delta)\right] - H(\Delta,x_{1,1},\ldots,x'_{i,\ell},\ldots,x_{d_1,k})\big)\,\big| \\
&\le \sup_{\Delta\in\mathcal{B}(D)} \big|H(\Delta,x_{1,1},\ldots,x_{i,\ell},\ldots,x_{d_1,k}) - H(\Delta,x_{1,1},\ldots,x'_{i,\ell},\ldots,x_{d_1,k})\big| \\
&\overset{(a)}{\le} \frac{e^{-2\alpha}}{2\,k^3\,d_1} \sup_{\Delta\in\mathcal{B}(D)} \Big\{2\sum_{\ell'\in[k]} \langle\!\langle \Delta, e_{i,j_{i,\ell}} - e_{i,j_{i,\ell'}}\rangle\!\rangle^2 \sum_{\ell''=1}^{k} \chi_{i,\ell,\ell',\ell''} \\
&+ \sum_{\ell',\ell''\in[k]} \langle\!\langle \Delta, e_{i,j_{i,\ell'}} - e_{i,j_{i,\ell''}}\rangle\!\rangle^2 \chi_{i,\ell',\ell'',\ell}\Big\} \\
&\overset{(b)}{\le} \frac{8\alpha^2 e^{-2\alpha}}{k^3\,d_1}\Big\{2\sum_{\ell'\in[k]\setminus\{\ell\}}\sum_{\ell''=1}^{k}\chi_{i,\ell,\ell',\ell''} + \sum_{\ell',\ell''\in[k],\ell'\neq\ell'',}\chi_{i,\ell',\ell'',\ell}\Big\} \\
&\le \frac{16\alpha^2 e^{-2\alpha}}{k\,d_1}\,,
\end{aligned}
\tag{B.19}
$$

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

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq t\right\} \leq \exp\left(-\frac{k^2\, d_1^2\, t^2}{2^7\, \alpha^4 e^{-4\alpha} d_1 k}\right). \tag{B.20}$$

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

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

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

$$Y_{i,\ell,\ell',\ell''}(\Delta) \;\equiv\; (\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2 \chi_{i,\ell,\ell',\ell''}\,, \tag{B.21}$$

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 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: $T_1 = \{(1,2,3),(4,5,6)\}$, $T_2 = \{1,2,4),(3,5,6)\}$, $T_3 = \{(1,2,5),(3,4,6)\}$, $T_4 = \{(1,2,6),(3,4,5)\}$, $T_5 = \{(1,3,4),(2,5,6)\}$, $T_6 = \{(1,3,5),(2,4,6)\}$, $T_7 = \{(1,3,6),(2,4,5)\}$, $T_8 = \{(1,4,5),(2,3,6)\}$, $T_9 = \{(1,4,6),(2,3,5)\}$, and $T_{10} = \{(1,5,6),(2,3,4)\}$. 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 B.1.4. *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*

$$\{Y_{i,\ell,\ell',\ell''}\}_{i\in[d_1],(\ell,\ell',\ell'')\in T_a} \text{ are mutually independent.} \tag{B.22}$$

Now, we are ready to partition the summation.

$$\begin{aligned}
\mathbb{E}\big[Z\big] = {} & \frac{e^{-2\alpha}}{2\,k^3\,d_1}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)}\sum_{i\in[d_1]}\sum_{\ell,\ell',\ell''\in[k]}\big\{\mathbb{E}[Y_{i,\ell,\ell',\ell''}(\Delta)] - Y_{i,\ell,\ell',\ell''}(\Delta)\big\}\Big] \\
= {} & \frac{e^{-2\alpha}}{2\,k^3\,d_1}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)}\sum_{i\in[d_1]}\sum_{a\in[N]}\sum_{(\ell,\ell',\ell'')\in T_a}\big\{\mathbb{E}[Y_{i,\ell,\ell',\ell''}(\Delta)] - Y_{i,\ell,\ell',\ell''}(\Delta)\big\}\Big] \\
\leq {} & \frac{e^{-2\alpha}}{2\,k^3\,d_1}\sum_{a\in[N]}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)}\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\big\{\mathbb{E}[Y_{i,\ell,\ell',\ell''}(\Delta)] - Y_{i,\ell,\ell',\ell''}(\Delta)\big\}\Big] \\
\leq {} & \frac{e^{-2\alpha}}{k^3\,d_1}\sum_{a\in[N]}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)}\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\xi_{i,\ell,\ell',\ell''}Y_{i,\ell,\ell',\ell''}(\Delta)\Big] \\
= {} & \frac{e^{-2\alpha}}{k^3\,d_1}\sum_{a\in[N]}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)}\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\xi_{i,\ell,\ell',\ell''}(\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2\chi_{i,\ell,\ell',\ell''}\Big]\,,
\end{aligned} \tag{B.23}$$

where the first inequality follows from the fact that the sum of the supremum is no less than the supremum of the sum, and the second inequality follows from the 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 that

$$\begin{aligned}
& \mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)}\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\xi_{i,\ell,\ell',\ell''}(\Delta_{i,j_{i,\ell}} - \Delta_{i,j_{i,\ell'}})^2\chi_{i,\ell,\ell',\ell''}\Big] \\
& \leq\; 8\alpha\mathbb{E}\Big[\sup_{\Delta\in\mathcal{B}(D)}\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\xi_{i,\ell,\ell',\ell''}\,\chi_{i,\ell,\ell',\ell''}\,\langle\!\langle\Delta, e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\rangle\!\rangle\Big]\,.
\end{aligned} \tag{B.24}$$

Applying Hölder's inequality, we get that

$$\begin{aligned}
& \Big|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\xi_{i,\ell,\ell',\ell''}\,\chi_{i,\ell,\ell',\ell''}\,\langle\!\langle\Delta, e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\rangle\!\rangle\Big| \\
& \leq\; |\!|\!|\Delta|\!|\!|_{\rm nuc}\Big|\!\Big|\!\Big|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\xi_{i,\ell,\ell',\ell''}\,\chi_{i,\ell,\ell',\ell''}\,\big(e_i(e_{j_{i,\ell}} - e_{j_{i,\ell'}})^T\big)\Big|\!\Big|\!\Big|_2\,.
\end{aligned} \tag{B.25}$$

We are left to prove that the expected value of the right-hand side of the above inequality is bounded by $C|||\Delta|||_{\rm nuc}\sqrt{kd_1\log d/\min\{d_1,d_2\}}$ 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

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

almost surely, and

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

and

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

This gives

$$\begin{aligned}\sigma^2 =& \max\Bigg\{\Bigg|\Bigg|\Bigg|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\mathbb{E}[W_{i,\ell,\ell',\ell''}W^T_{i,\ell,\ell',\ell''}]\Bigg|\Bigg|\Bigg|_2\,,\\ &\Bigg|\Bigg|\Bigg|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}\mathbb{E}[W^T_{i,\ell,\ell',\ell''}W_{i,\ell,\ell',\ell''}]\Bigg|\Bigg|\Bigg|_2\Bigg\} &\text{(B.26)}\\ \le& \max\left\{2|T_a|\,,\,\frac{2d_1|T_a|}{d_2}\right\} = \frac{2d_1|T_a|}{\min\{d_1,d_2\}} \le \frac{2d_1k}{3\min\{d_1,d_2\}}\,, &\text{(B.27)}\end{aligned}$$

since we have designed T_a's such that $|T_a|\le k/3$. Applying matrix Bernstein inequality [26] yields the tail bound

$$\mathbb{P}\left\{\Bigg|\Bigg|\Bigg|\sum_{i\in[d_1]}\sum_{(\ell,\ell',\ell'')\in T_a}W_{i,\ell,\ell',\ell''}\Bigg|\Bigg|\Bigg|_2 \ge t\right\} \le \quad (d_1+d_2)\exp\left(\frac{-t^2/2}{\sigma^2+\sqrt{2}t/3}\right)\,. \tag{B.28}$$

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 $1-2d^{-3}$,

$$\left|\!\left|\!\left| \sum_{i\in[d_1]} \sum_{(\ell,\ell',\ell'')\in T_a} W_{i,\ell,\ell',\ell''} \right|\!\right|\!\right|_2 \leq \quad \max\left\{ \sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}}, \frac{16\sqrt{2}\log d}{3} \right\}. \tag{B.29}$$

It follows from the fact $\left|\!\left|\!\left| \sum_{i\in[d_1]} \sum_{(\ell,\ell',\ell'')\in T_a} W_{i,\ell,\ell',\ell''} \right|\!\right|\!\right|_2 \leq \sum_{i,(\ell,\ell',\ell'')} |\!|\!| W_{i,\ell,\ell',\ell''} |\!|\!|_2 \leq \sqrt{2}d_1k/3$ that

$$\begin{aligned}\mathbb{E}\left[\left|\!\left|\!\left| \sum_{i\in[d_1]} \sum_{(\ell,\ell',\ell'')\in T_a} W_{i,\ell,\ell',\ell''} \right|\!\right|\!\right|_2 \right] &\leq \max\left\{ \sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}}, \frac{16\sqrt{2}\log d}{3} \right\} + \frac{2\sqrt{2}d_1k}{3d^3} \\ &\leq 2\sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}}, \end{aligned} \tag{B.30}$$

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 (B.25), and then together with (B.24) and (B.23), this gives the following desired bound:

$$\begin{aligned}\mathbb{E}[Z] &\leq \sum_{a\in[N]} \sup_{\Delta\in\mathcal{B}(D)} \frac{16\alpha e^{-2\alpha}}{k^3\, d_1} \sqrt{\frac{32kd_1\log d}{3\min\{d_1,d_2\}}} |\!|\!|\Delta|\!|\!|_{\text{nuc}} \\ &\leq \sum_{a\in[N]} \frac{e^{-4\alpha}\sqrt{2}}{16\sqrt{3}k^2\, d_1\, d_2} \underbrace{\left(2^{10} e^{2\alpha} \alpha d_2 \sqrt{\frac{d_1\log d}{k\min\{d_1,d_2\}}} \right)}_{=\mu} |\!|\!|\Delta|\!|\!|_{\text{nuc}} \\ &\leq \frac{9e^{-4\alpha}D^2}{40d_1d_2}, \end{aligned} \tag{B.31}$$

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

B.1.5 Proof of Lemma B.1.4

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 (B.21). From the random utility model (RUM) interpretation of the MNL model presented in Section 1.1, 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 involves three teams. Let $T_{a,b}$ denote a set of triples playing at the same round, indexed by two integers $a \in \{3, \dots, 2k-3\}$ and $b \in \{5, \dots, 2k-1\}$. Hence, there are total $N = (2k-5)^2$ rounds.

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

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

We need to prove that there is no missing triple and no team plays twice in a single round. First, for any ordered triple (ℓ, ℓ', ℓ''), there exists $a \in \{3, \dots, 2k-3\}$ and $b \in \{5, \dots, 2k-1\}$ such that $\ell + \ell' = a$ and $\ell' + \ell'' = b$. Thus 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 is 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 want to cover the whole $[k] \times [k] \times [k]$, and not just the ordered triples. This issue can be resolved by simply taking all $T_{a,b}$'s from the above construction, and making six copies of each round, and permuting all the triples in each copy according to the same permutation over $\{1, 2, 3\}$. This operation 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.

B.2 Proof of Corollary 3.2.5: estimating approximate low-rank matrices

We follow closely the proof of a similar corollary in [13]. First fix a threshold $\tau > 0$, and set $r = \max\{j | \sigma_j(\Theta^*) > \tau\}$. With this choice of r, we have

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

Also, since $r\tau^q \le \sum_{j=1}^r \sigma_j(\Theta^*)^q \le \rho_q$, it follows that $\sqrt{r} \le \sqrt{\rho_q}\tau^{-q/2}$. Using these bounds, (3.5) is now

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

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

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

B.3 Proof of Theorem 4: information-theoretic lower bound for k-wise ranking

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

$$\sup_{\Theta^* \in \Omega_\alpha} \mathbb{P}\left\{\left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\mathrm{F}}^2 \ge \frac{\delta^2}{4}\right\} \;\ge\; \frac{1}{2} . \tag{B.32}$$

To prove the above claim, we follow the standard recipe of constructing a packing in Ω_α. Consider a family $\{\Theta^{(1)}, \dots, \Theta^{(M(\delta))}\}$ of $d_1 \times d_2$ dimensional matrices contained in Ω_α satisfying $|||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{F}} \ge \delta$ for all $\ell_1, \ell_2, \in [M(\delta)]$. We will use M to refer to $M(\delta)$ to 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 the 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

$$\sup_{\Theta^* \in \Omega_\alpha} \mathbb{P}\left\{\left|\!\left|\!\left|\widehat{\Theta} - \Theta^*\right|\!\right|\!\right|_{\mathrm{F}}^2 \ge \frac{\delta^2}{4}\right\} \;\ge\; \mathbb{P}\left\{\widehat{L} \ne L\right\} , \tag{B.33}$$

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

$$\begin{aligned}\mathbb{P}\left\{\widehat{L} \neq L | S(1), \ldots, S(d_1)\right\} &\geq 1 - \frac{I(\widehat{L}; L) + \log 2}{\log M} \\ &\geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1, \ell_2 \in [M]} D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) + \log 2}{\log M} \;,\end{aligned} \tag{B.34}$$

where $D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})$ denotes the Kullback-Leibler (KL) 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

$$\begin{aligned} I(\Sigma; L) &= \sum_{\ell \in [M], \Sigma} p(\Sigma|\ell) \frac{1}{M} \log \frac{p(\ell, \Sigma)}{p(\ell) p(\Sigma)} \\ &= \frac{1}{M} \sum_{\ell \in [M]} \sum_{\Sigma} p(\Sigma|\ell) \log \frac{p(\Sigma|\ell)}{\frac{1}{M} \sum_{\ell'} p(\Sigma|\ell')} \\ &\leq \frac{1}{M^2} \sum_{\ell, \ell' \in [M]} \sum_{\Sigma} p(\Sigma|\ell) \log \frac{p(\Sigma|\ell)}{p(\Sigma|\ell')} \\ &= \frac{1}{M^2} \sum_{\ell, \ell' \in [M]} D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) \;, \end{aligned} \tag{B.35}$$

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

$$
\begin{aligned}
D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)}) &\leq D_{\mathrm{KL}}(X^{(\ell_1)} \| X^{(\ell_2)}) \\
&= \sum_{i\in[d_1]} \sum_{j \in S_i} \left\{ e^{\Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)}} - (\Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)}) - 1 \right\} \\
&\leq \frac{e^{2\alpha}}{4\alpha^2} \sum_{i\in[d_1]} \sum_{j \in S_i} (\Theta_{ij}^{(\ell_1)} - \Theta_{ij}^{(\ell_2)})^2 \;, \qquad \text{(B.36)}
\end{aligned}
$$

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,

$$
\mathbb{E}_{S(1),\dots,S(d_1)}[D_{\mathrm{KL}}(\Theta^{(\ell_1)} \| \Theta^{(\ell_2)})] \;\leq\; \frac{e^{2\alpha}\, k}{4\,\alpha^2\, d_2} |||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{F}}^2 \,. \qquad \text{(B.37)}
$$

Combined with (B.34), we get that

$$
\begin{aligned}
\mathbb{P}\left\{ \widehat{L} \neq L \right\} &= \mathbb{E}_{S(1),\dots,S(d_1)}[\mathbb{P}\left\{ \widehat{L} \neq L | S(1), \dots, S(d_1) \right\}] \qquad \text{(B.38)} \\
&\geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1,\ell_2 \in [M]} (e^{2\alpha}k/(4\alpha^2 d_2)) |||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{F}}^2 + \log 2}{\log M} \,, \qquad \text{(B.39)}
\end{aligned}
$$

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

Lemma B.3.1. *Let $d_2 \geq d_1 \geq 607$ be positive integers. Then for each $r \in \{1, \dots, d_1\}$, and for any positive $\delta > 0$ there exists a family of $d_1 \times d_2$ dimensional matrices $\{\Theta^{(1)}, \dots, \Theta^{(M(\delta))}\}$ with cardinality $M(\delta) = \lfloor (1/4) \exp(rd_2/576) \rfloor$ such that each matrix is rank r and the following bounds hold:*

$$
\begin{aligned}
|||\Theta^{(\ell)}|||_{\mathrm{F}} &\leq \delta \;, \text{ for all } \ell \in [M] \qquad \text{(B.40)} \\
|||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{F}} &\geq \delta \;, \text{ for all } \ell_1, \ell_2 \in [M] \qquad \text{(B.41)} \\
\Theta^{(\ell)} &\in \Omega_{\tilde{\alpha}} \;, \text{ for all } \ell \in [M] \;, \qquad \text{(B.42)}
\end{aligned}
$$

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 implies that $|||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{F}}^2 \leq$

$4\delta^2$, which gives

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

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 (B.32) and (B.33), this proves that for all $\delta \leq \min\{\alpha d_2/(8\sqrt{2\log d}), \alpha e^{-\alpha}\sqrt{r\,d_2\,d/(2304\,k)}\}$,

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

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

B.3.1 Proof of Lemma B.3.1

Following the construction in [13], we use a 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:

$$\Theta^{(\ell)} \;\;=\;\; \frac{\delta}{\sqrt{rd_2}} U (V^{(\ell)})^T \Big(\mathbf{I}_{d_2\times d_2} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T\Big)\;, \qquad \text{(B.43)}$$

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

By construction, notice that $|\!|\!|\Theta^{(\ell)}|\!|\!|_{\mathrm{F}} = (\delta/\sqrt{rd_2})|\!|\!|(V^{(\ell)})^T(\mathbf{I} - (1/d_2)\mathbb{1}\mathbb{1}^T)|\!|\!|_{\mathrm{F}}$ $\leq \delta$, since $|\!|\!|V^{(\ell)}|\!|\!|_{\mathrm{F}} = \sqrt{rd_2}$ and $(\mathbf{I} - (1/d_2)\mathbb{1}\mathbb{1}^T)$ is a projection which can only decrease the norm.

Now, consider $|\!|\!|\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|\!|\!|^2_{\mathrm{F}} = (\delta^2/(rd_2))|\!|\!|(\mathbf{I} - (1/d_2)\mathbb{1}\mathbb{1}^T)(V^{(\ell_1)} - V^{(\ell_2)})|\!|\!|^2_{\mathrm{F}}$ $\equiv f(V^{(\ell_1)}, V^{(\ell_2)})$, 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

$$\begin{aligned}
&\big| f(V^{(\ell_1)}, V^{(\ell_2)}) - f(\widetilde{V}^{(\ell_1)}, \widetilde{V}^{(\ell_2)}) \big| \; = \\
&\quad \Bigg| \frac{\delta^2}{r\, d_2} \Big\vert\!\Big\vert\!\Big\vert (\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T)(V^{(\ell_1)} - V^{(\ell_2)}) \Big\vert\!\Big\vert\!\Big\vert_{\mathrm{F}}^2 \\
&\quad - \frac{\delta^2}{r\, d_2} \Big\vert\!\Big\vert\!\Big\vert (\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T)(V^{(\ell_1)} - V^{(\ell_2)} + 2e_{ij}) \Big\vert\!\Big\vert\!\Big\vert_{\mathrm{F}}^2 \Bigg| \\
&\quad = \Bigg| \frac{\delta^2}{r\, d_2} \Big\vert\!\Big\vert\!\Big\vert 2(\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T) e_{ij} \Big\vert\!\Big\vert\!\Big\vert_{\mathrm{F}}^2 + \frac{\delta^2}{r\, d_2} \langle\!\langle (\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T)(V^{(\ell_1)} - V^{(\ell_2)}), 2e_{ij} \rangle\!\rangle \Bigg| \\
&\quad \le \; \frac{4\,\delta^2}{r\, d_2} \; + \; \frac{\delta}{r\, d_2} \Big\vert\!\Big\vert\!\Big\vert (\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T)(V^{(\ell_1)} - V^{(\ell_2)}) \Big\vert\!\Big\vert\!\Big\vert_{\infty} \vert\!\vert\!\vert 2e_{ij} \vert\!\vert\!\vert_1 \\
&\quad \le \; \frac{12\,\delta^2}{r\, d_2} \; , \qquad\qquad \text{(B.44)}
\end{aligned}$$

where we used the fact that $(\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T)(V^{(\ell_1)} - V^{(\ell_2)})$ is entry-wise bounded by four. The expectation $\mathbb{E}[f(V^{(\ell_1)}, V^{(\ell_2)})]$ is

$$\begin{aligned}
\frac{\delta^2}{r\, d_2} \mathbb{E}\left[\Big\vert\!\Big\vert\!\Big\vert (\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T)(V^{(\ell_1)} - V^{(\ell_2)}) \Big\vert\!\Big\vert\!\Big\vert_{\mathrm{F}}^2 \right] &= \frac{2\delta^2}{r\, d_2} \mathbb{E}\left[\Big\vert\!\Big\vert\!\Big\vert (\mathbf{I} - \frac{1}{d_2}\mathbb{1}\mathbb{1}^T) V^{(\ell_1)} \Big\vert\!\Big\vert\!\Big\vert_{\mathrm{F}}^2 \right] \\
&= \frac{2\delta^2}{r\, d_2} \mathbb{E}\left[\vert\!\vert\!\vert V^{(\ell_1)} \vert\!\vert\!\vert_{\mathrm{F}}^2 \right] - \frac{2\delta^2}{r\, d_2^2} \mathbb{E}\left[\|\mathbb{1}^T V^{(\ell_1)}\|^2 \right] \\
&= \frac{2\,\delta^2\,(d_2 - 1)}{d_2} \; . \qquad\qquad \text{(B.45)}
\end{aligned}$$

Applying McDiarmid's inequality with bounded difference $12\delta^2/(r d_2)$, we get that

$$\mathbb{P}\left\{ \, f(V^{(\ell_1)}, V^{(\ell_2)}) \le 2\delta^2(1 - 1/d_2) - t \, \right\} \;\; \le \;\; \exp\left\{ - \frac{t^2\, r\, d_2}{144\, \delta^4} \right\} . \text{(B.46)}$$

Since there are fewer than $(M')^2$ pairs of (ℓ_1, ℓ_2), setting $t = (1 - 2/d_2)\delta^2$ and applying the union bound gives

$$\begin{aligned}
\mathbb{P}\left\{ \min_{\ell_1, \ell_2 \in [M']} \vert\!\vert\!\vert \Theta^{(\ell_1)} - \Theta^{(\ell_2)} \vert\!\vert\!\vert_{\mathrm{F}}^2 \ge \delta^2 \right\} &\ge 1 - \exp\left\{ - \frac{r\, d_2}{144} \Big(1 - \frac{2}{d_2} \Big)^2 + 2 \log M' \right\} \\
&\ge \frac{7}{8} \; , \qquad\qquad \text{(B.47)}
\end{aligned}$$

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 (3.4). Since we removed the mean such that $\Theta^{(\ell)}\mathbb{1} = 0$ 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 (B.49) for a fixed $\ell \in [M']$, and use this to show that there exists a large enough subset of matrices satisfying this bound. From (B.43), 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 [31], we have

$$\mathbb{P}\{|\langle\!\langle u_i, v_j \rangle\!\rangle| \geq t\} \;\leq\; 2\exp\Big\{-\frac{d_2\,t^2}{8\,r}\Big\} \,. \tag{B.48}$$

Notice that by the definition (B.43), $\max_{i,j}|\Theta^{(\ell)}_{ij}| \leq (2\delta/\sqrt{rd_2})\max_{i,j}|\langle\!\langle u_i, v_j\rangle\!\rangle|$. Setting $t = \sqrt{(32r/d_2)\log d_2}$ and taking the union bound over all $d_1 d_2$ indices, we get

$$\mathbb{P}\Big\{\max_{i,j}|\Theta^{(\ell)}_{ij}| \leq \frac{2\delta\sqrt{32\log d_2}}{d_2}\Big\} \geq \;\; 1 - 2d_1 d_2 \exp\Big\{-4\log d_2\Big\} \;\geq\; \frac{1}{2}\,, \tag{B.49}$$

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 (B.49) we get

$$\begin{aligned}&\mathbb{P}\Big\{\exists S \subset [M'] \text{ such that } |||\Theta^{(\ell)}|||_\infty \leq \frac{2\delta\sqrt{32\log d_2}}{d_2} \text{ for all } \ell \in S\Big\}\\ &\geq \sum_{m=M}^{M'}\binom{M'}{m}\Big(\frac{1}{2}\Big)^m\,,\end{aligned} \tag{B.50}$$

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

B.4 Proof of Theorem 5: pairwise rank breaking

Analogous to Section B.1, we define the gradient $\nabla\mathcal{L}(\Theta)$ as $\nabla_{ij}\mathcal{L} = \frac{\partial\mathcal{L}(\Theta)}{\partial\Theta_{ij}}$ and $\Delta \equiv \hat{\Theta} - \Theta^*$, and provide two main technical lemmas.

Lemma B.4.1. *If $\lambda \geq 2|||\nabla \mathcal{L}(\Theta^*)|||_2$, then we have,*

$$|||\Delta|||_{\text{nuc}} \leq 4\sqrt{2r}|||\Delta|||_{\text{F}} + 4 \sum_{j=\rho+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^*) \ , \tag{B.51}$$

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

Proof. This follows from the proof of Lemma 3.2.1, which only depends on the convexity of $\mathcal{L}(\Theta)$. □

Lemma B.4.2. *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 $1 - 2d^{-c}$,*

$$|||\nabla \mathcal{L}(\Theta^*)|||_2 \ \leq \ \sqrt{\frac{16(c+4)\log d}{k\, d_1^2}} \max\left\{\sqrt{\max\left\{\frac{1}{4}, \frac{d_1}{d_2}\right\}}, \frac{2}{3}\sqrt{\frac{2(c+4)\log d}{k}}\right\} \ . \tag{B.52}$$

The proof of this lemma is provided in Section B.4.1. 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 the first term,

$$\frac{2}{3}\sqrt{\frac{2(4+c)\log d}{k}} \leq \sqrt{\frac{1}{4}}\ . \tag{B.53}$$

(B.53) simplifies (B.52) as

$$\begin{aligned} |||\nabla \mathcal{L}(\Theta^*)|||_2 &\leq \sqrt{\frac{16(c+4)\log d}{k\, d_1^2} \max\left\{\frac{1}{4}, \frac{d_1}{d_2}\right\}} \\ &\leq \sqrt{\frac{32d\ (c+4)\log d}{k\, d_1^2\ d_2}} \\ &\overset{(a)}{\leq} \sqrt{32(c+4)}\lambda \ , \end{aligned} \tag{B.54}$$

where (a) is due to (3.22) .

For Lemma B.4.1 and further proof of Theorem 5 we want $\lambda \geq 2|||\nabla \mathcal{L}(\Theta)|||_2$; therefore, we assume that

$$\lambda \in [2\sqrt{32(c+4)}\lambda, c_p\lambda], \text{ for some } c_p \geq 2\sqrt{32(c+4)}\,. \tag{B.55}$$

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

$$\mu = 16\alpha \sqrt{\frac{48\, d_1 d_2^2 \log d}{k\; \min\{d_1, d_2\}}}\;. \tag{B.56}$$

Case 1: Assume $\mu|||\Delta|||_{\rm nuc} \leq |||\Delta|||_{\rm F}^2$.
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

$$\begin{aligned}\mathcal{L}(\hat{\Theta}) = \mathcal{L}(\Theta^*) - \langle\!\langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle\!\rangle\; + \qquad\qquad\qquad\qquad\qquad\qquad & \qquad (\text{B.57})\\ \frac{1}{2!\; d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{\Theta_{i,u_{i,m_1}}} e^{\Theta_{i,u_{i,m_2}}}}{\left(e^{\Theta_{i,u_{i,m_1}}} + e^{\Theta_{i,u_{i,m_2}}}\right)^2} \Big(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\Big)^2\,, &\end{aligned}$$

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 (B.57) as

$$\begin{aligned}&\frac{1}{2!\; d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{\Theta_{i,u_{i,m_1}}} e^{\Theta_{i,u_{i,m_2}}}}{\left(e^{\Theta_{i,u_{i,m_1}}} + e^{\Theta_{i,u_{i,m_2}}}\right)^2} \Big(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\Big)^2 \\ &\qquad \overset{(a)}{\geq} \frac{1}{2\; d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{-\alpha} e^{\alpha}}{(e^{-\alpha} + e^{\alpha})^2} \Big(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\Big)^2 \\ &\qquad \geq \frac{1}{2\; d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \frac{e^{-2\alpha}}{4} \Big(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\Big)^2\,,\end{aligned} \tag{B.58}$$

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 B.4.3. *For* $(4 \log d)/9 \leq k \leq \max\{d_1, d_2^2/d_1\} \log d$*, with probability* *at least* $1-2d^{-2^{13}}$,

$$\frac{1}{d_1\binom{k}{2}}\sum_{i=1}^{d_1}\sum_{(m_1,m_2)\in\mathcal{P}_0}\Big(\Delta_{i,u_{i,m_1}}-\Delta_{i,u_{i,m_2}}\Big)^2\geq\frac{1}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2\,, \tag{B.59}$$

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

$$\mathcal{A}=\Bigg\{\Delta\in\mathbb{R}^{d_1\times d_2}\;\Bigg|\;|||\Delta|||_\infty\leq 2\alpha,\ \sum_{j\in[d_2]}\Delta_{ij}=0\ \forall i\in[d_2],\ \textit{and,}$$
$$\mu|||\Delta|||_{\mathrm{nuc}}\leq|||\Delta|||_{\mathrm{F}}^2\Bigg\}\,. \tag{B.60}$$

The proof is given in Section B.4.2. Now using Lemma B.4.3 and (B.58) with high probability we get

$$\frac{1}{2!\;d_1\binom{k}{2}}\sum_{i=1}^{d_1}\sum_{(m_1,m_2)\in\mathcal{P}_0}\frac{e^{\Theta_{i,u_{i,m_1}}}e^{\Theta_{i,u_{i,m_2}}}}{\Big(e^{\Theta_{i,u_{i,m_1}}}+e^{\Theta_{i,u_{i,m_2}}}\Big)^2}\Big(\Delta_{i,u_{i,m_1}}-\Delta_{i,u_{i,m_1}}\Big)^2$$
$$\geq\frac{e^{-2\alpha}}{24\;d_1\;d_2}|||\Delta|||_{\mathrm{F}}^2\,. \tag{B.61}$$

Incorporating the above inequality in (B.57) we obtain

$$\frac{e^{-2\alpha}}{24\;d_1\;d_2}|||\Delta|||_{\mathrm{F}}^2\leq\mathcal{L}(\hat\Theta)-\mathcal{L}(\Theta^*)+\langle\!\langle\nabla\mathcal{L}(\Theta^*),\Delta\rangle\!\rangle\,. \tag{B.62}$$

From the definition of $\hat\Theta$, we have $\mathcal{L}(\hat\Theta)-\mathcal{L}(\Theta^*)\leq\lambda\left(|||\Theta^*|||_{\mathrm{nuc}}-\left|\left|\left|\hat\Theta\right|\right|\right|_{\mathrm{nuc}}\right)\leq\lambda|||\Delta|||_{\mathrm{nuc}}$, and we assume that $\lambda\geq 2\sqrt{32(c+1)}\;\lambda$, so that $\lambda\geq 2|||\nabla\mathcal{L}\left(\Theta^*\right)|||_2$ is true with a probability of at least $1-2d^{-c}$ from Lemma B.4.2. These inequalities give us the following with probability at least $1-2d^{-c}-2d^{-2^{13}}$.

$$\frac{e^{-2\alpha}}{24\;d_1\;d_2}|||\Delta|||_{\mathrm{F}}^2\leq\lambda|||\Delta|||_{\mathrm{nuc}}+|||\nabla\mathcal{L}(\Theta^*)|||_2|||\Delta|||_{\mathrm{nuc}}$$
$$\leq\frac{3\lambda}{2}|||\Delta|||_{\mathrm{nuc}}\,, \tag{B.63}$$

which gives us

$$\begin{aligned}|||\Delta|||_{\mathrm{F}}^2 &\leq 36e^{2\alpha}\ \lambda\ d_1 d_2\ |||\Delta|||_{\text{nuc}} \\ &\overset{(a)}{\leq} 36e^{2\alpha}\ c_p\ \lambda\ d_1 d_2\ |||\Delta|||_{\text{nuc}}\ , \end{aligned} \tag{B.64}$$

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

Case 2: Assume $|||\Delta|||_{\mathrm{F}}^2 \leq \mu |||\Delta|||_{\text{nuc}}$.
Here we prove that $\mu \leq 36\ e^{2\alpha}\ c_p\lambda\ d_1 d_2$.

$$\begin{aligned}\frac{\mu}{36\ e^{2\alpha}\ c_p\lambda\ d_1 d_2} &\overset{(a)}{\leq} \frac{\alpha}{e^{2\alpha}} \times \frac{16\sqrt{48}}{72\sqrt{32(c+4)}} \times \sqrt{\frac{d_1 d_2}{min\{d_1, d_2\} d}} \\ &\overset{(b)}{\leq} 1 \times \frac{16\sqrt{48}}{72\sqrt{32 \times 4}} \times \sqrt{\frac{\max\{d_1, d_2\}}{d}} \\ &\overset{(c)}{\leq} \sqrt{\frac{\max\{d_1, d_2\}}{2d}} \\ &\overset{(d)}{\leq} 1\ , \end{aligned} \tag{B.65}$$

where (a) is by substituting μ, λ, and c_p from (B.56), (3.22), and (B.55), respectively; (b) is because $x \leq e^x$; and (c) is because $d = (\max\{d_1, d_2\} + \min\{d_1, d_2\})/2$.

Now combining the above result with (B.51) we get with probability at least $1 - 2d^{-c} - 2d^{-2^{13}}$,

$$\frac{1}{d_1 d_2} |||\Delta|||_{\mathrm{F}}^2 \leq 144\sqrt{2} e^{2\alpha} c_p \lambda \sqrt{r} |||\Delta|||_{\mathrm{F}} + 144 e^{2\alpha} c_p \lambda \sum_{j=\rho+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*)\ . \tag{B.66}$$

B.4.1 Proof of Lemma B.4.2

From definition of $\mathcal{L}(\Theta)$ in (3.20), we get

$$\nabla \mathcal{L}_p(\Theta^*) = \frac{1}{d_1 \binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1, m_2) \in \mathcal{P}_0} \frac{e_i \left(e_{l_i(m_1,m_2)} - e_{h_i(m_1,m_2)}\right)^T}{1 + \exp\left(\Theta^*_{i, l_i(m_1,m_2)} - \Theta^*_{i, h_i(m_1,m_2)}\right)}\ , \tag{B.67}$$

where $\mathcal{P}_0 = \{(i, j) |\ 1 \leq i < j \leq k\}$. We use the matrix Bernstein inequality [26] for the sum of independent matrices. Similar to Lemma C.1.4, 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

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

such that

$$\nabla\mathcal{L}_p(\Theta^*) = \frac{1}{d_1\binom{k}{2}} \sum_{a=1}^{k-1} \tilde{Y}_a \ . \tag{B.68}$$

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 $\|\tilde{X}_{i,m_1,m_2}\|_2 \le \sqrt{2}$. We also have

$$\begin{aligned}
&\mathbb{E}\left[\tilde{X}_{i,m_1,m_2}\tilde{X}^T_{i,m_1,m_2}\right] \\
&\preceq 2\, e_i e_i^T\, \mathbb{E}\left[\mathbb{E}\left[\frac{\exp\left(\Theta^*_{i,l_i(m_1,m_2)}\right)^2}{\left(\exp\left(\Theta^*_{i,u_{i,m_1}}\right)+\exp\left(\Theta^*_{i,u_{i,m_2}}\right)\right)^2}\,\middle|\, u_{i,m_1}, u_{i,m_1}\right]\right] \\
&\overset{(a)}{=} 2\ e_i e_i^T\, \mathbb{E}\left[\frac{\exp\left(\Theta^*_{iu_{i,m_1}}\right)\exp\left(\Theta^*_{iu_{i,m_2}}\right)}{\left(\exp\left(\Theta^*_{i,u_{i,m_1}}\right)+\exp\left(\Theta^*_{i,u_{i,m_2}}\right)\right)^2}\right] \\
&\overset{(b)}{\preceq} \frac{1}{2} e_i e_i^T \ ,
\end{aligned} \tag{B.69}$$

where we get (a) from the MNL model for the random choice of $l_i(m_1,m_2)$ and (b) is due to the fact that $xy/(x+y)^2 \le 1/4$ for all $x, y > 0$. Define $p_{i,m_1,m_2} \equiv \left(\exp\left(\Theta^*_{i,u_{i,m_1}}\right) e_{u_{i,m_1}} + \exp\left(\Theta^*_{i,u_{i,m_2}}\right) e_{u_{i,m_2}}\right) / \left(\exp\left(\Theta^*_{i,u_{i,m_1}}\right) + \exp\left(\Theta^*_{i,u_{i,m_2}}\right)\right)$ to get

$$\begin{aligned}
\mathbb{E}\left[\tilde{X}^T_{i,m_1,m_2}\tilde{X}_{i,m_1,m_2}\right] &= \mathbb{E}\left[(e_{h_i(m_1,m_2)} - p_{i,m_1,m_2})(e_{h_i(m_1,m_2)} - p_{i,m_1,m_2})^T\right] \\
&= \mathbb{E}\left[e_{h_i(m_1,m_2)} e^T_{h_i(m_1,m_2)}\right] - \mathbb{E}\left[p_{i,m_1,m_2} p^T_{i,m_1,m_2}\right] \\
&\overset{(a)}{\preceq} \mathbb{E}\left[e_{u_{i,m_1}} e^T_{u_{i,m_1}} + e_{u_{i,m_2}} e^T_{u_{i,m_2}}\right] \\
&= \frac{2}{d_2} \mathbf{I}_{d_2\times d_2} \ ,
\end{aligned} \tag{B.70}$$

where (a) comes from the fact that $p_{i,m_1,m_2}p_{i,m_1,m_2}^T$ is a positive semi-definite matrix. Therefore using (B.69) and (B.70), we get

$$\begin{aligned}\sigma^2 \equiv & \left\{ \left\|\left\| \sum_{i\in[d_1],(m_1,m_2)\in\mathcal{P}_a} \mathbb{E}\left[\tilde{X}_{i,m_1,m_2}\tilde{X}_{i,m_1,m_2}^T\right] \right\|\right\|_2 , \right. \\ & \left. \left\|\left\| \sum_{i\in[d_1],(m_1,m_2)\in\mathcal{P}_a} \mathbb{E}\left[\tilde{X}_{i,m_1,m_2}^T\tilde{X}_{i,m_1,m_2}\right] \right\|\right\|_2 \right\} \\ & \leq k \max\left\{\frac{1}{4}, \frac{d_1}{d_2}\right\} . \end{aligned} \tag{B.71}$$

Define $\rho \equiv \max\{1/4, d_1/d_2\}$, then by the matrix Bernstein inequality [26], $\forall \ a \in [k-1]$,

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

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

$$t = \max\left\{\sqrt{4k\rho\left((1+c)\log d+\log(k-1)\right)}\,,\ \frac{4\sqrt{2}((1+c)\log d+\log(k-1))}{3}\right\} . \tag{B.72}$$

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

$$\begin{aligned}\mathbb{P}\left(\|\nabla\mathcal{L}_p(\Theta^*)\|_2 \geq \frac{k-1}{d_1\binom{k}{2}}t\right) &\leq \mathbb{P}\left(\sum_{a=1}^{k-1}\left\|\left|\tilde{Y}_a\right\|\right|_2 \geq (k-1)t\right) \\ &\overset{(a)}{\leq} \mathbb{P}\left(\max_{a\in[k-1]}\left\|\left|\tilde{Y}_a\right\|\right|_2 \geq t\right) \\ &\overset{(b)}{\leq} \sum_{a=1}^{k-1}\mathbb{P}\Big(\left\|\left|\tilde{Y}_a\right\|\right|_2 \geq t\Big) \\ &= 2\,d^{-c} , \end{aligned} \tag{B.73}$$

where we obtain (a) by the 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

$$(c+1)\log d + \log(k-1) \leq (c+4)\log d \;, \tag{B.74}$$

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.

B.4.2 Proof of Lemma B.4.3

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

$$\tilde{H}(\Delta) \;\equiv\; \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \Big(\Delta_{i,u_{i,m_1}} - \Delta_{i,u_{i,m_2}}\Big)^2 \tag{B.75}$$

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

$$\begin{aligned}\mathbb{E}\left[\tilde{H}(\Delta)\right] &= \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} \left[\frac{2}{d_2}\sum_{j\in[d_2]}\Delta_{ij}^2 - \frac{2}{d_2^2}\sum_{j\in[d_2]}\Delta_{ij}\sum_{j'\in[d_2]}\Delta_{ij'}\right] \\ &= \frac{2}{d_1 d_2}|||\Delta|||_{\mathrm{F}}^2 \;,\end{aligned} \tag{B.76}$$

where we used the fact that $\sum_j \Delta_{ij} = 0$. We want to upper bound the probability that $\tilde{H}(\Delta) \leq \frac{1}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2$ for some $\Delta \in \mathcal{A}$. As in the case of k-wise ranking we using the following peeling argument used in [13, Lemma 3], [27]. The strategy is to split this above event as the 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

$$\begin{aligned}\tilde{\mathcal{S}}_\ell = \Bigg\{\Delta \in \mathbb{R}^{d_1\times d_2} \;\Bigg|\; &|||\Delta|||_\infty \leq 2\alpha,\; \beta^{\ell-1}\mu \leq |||\Delta|||_{\mathrm{F}} \leq \beta^\ell \mu, \\ &\sum_{j\in[d_2]}\Delta_{ij} = 0 \text{ for all } i\in[d_2],\text{ and } |||\Delta|||_{\mathrm{nuc}} \leq \beta^{2\ell}\mu\Bigg\} \;,\end{aligned} \tag{B.77}$$

where $\beta = \sqrt{10/9}$ and $\ell \in \{1, 2, 3, \ldots\}$, which 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

$$\begin{aligned} \tilde{H}(\Delta) \leq \frac{1}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2 &\implies \frac{2}{d_1d_2}|||\Delta|||_{\mathrm{F}}^2 - \tilde{H}(\Delta) \geq \frac{5}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2 \\ &\implies \left(\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)\right) \geq \frac{5}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2 \,. \end{aligned} \tag{B.78}$$

Therefore using union bound we get

$$\begin{aligned} &\mathbb{P}\Big(\exists\ \Delta \in \mathcal{A}\ s.t.\ \tilde{H}(\Delta) \leq \frac{1}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2\Big) \\ &\quad \leq \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{\Delta\in\tilde{\mathcal{S}}_\ell} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{5}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2\right) \\ &\quad \overset{(a)}{\leq} \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{\Delta\in\tilde{\mathcal{S}}_\ell} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{3}{2d_1d_2}(\beta^\ell\mu)^2\right) \\ &\quad \overset{(b)}{\leq} \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{\Delta\in\tilde{\mathcal{B}}(\beta^\ell\mu)} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{3}{2d_1d_2}(\beta^\ell\mu)^2\right) \ , \end{aligned} \tag{B.79}$$

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

$$\begin{aligned} \tilde{\mathcal{B}}(D) = \Bigg\{\Delta \in \mathbb{R}^{d_1\times d_2}\ \Bigg|\ &|||\Delta|||_\infty \leq 2\alpha,\ |||\Delta|||_{\mathrm{F}} \leq D, \\ &\sum_{j\in[d_2]} \Delta_{ij} = 0 \text{ for all } i \in [d_2],\ \text{and } \mu|||\Delta|||_{\mathrm{nuc}} \leq D^2\Bigg\} \,, \end{aligned} \tag{B.80}$$

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

$$\frac{5}{3d_1d_2}|||\Delta|||_{\mathrm{F}}^2 \geq \frac{5}{3d_1d_2}(\beta^{\ell-1}\mu)^2 = \frac{3}{2d_1d_2}(\beta^\ell\mu)^2 \,, \tag{B.81}$$

and (b) is true because $\tilde{\mathcal{S}}_\ell \subset \tilde{\mathcal{B}}(\beta^\ell\mu)$.

Now we use following lemma to upper bound (B.79).

Lemma B.4.4. *For* $4(\log d)/3 \leq k \leq d^2 \log d$*,*

$$\mathbb{P}\left(\sup_{\Delta \in \tilde{\mathcal{B}}(D)} (\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)) \geq \frac{3}{2d_1 d_2} D^2 \right) \leq \exp\left(\frac{-kD^4}{2048\ \alpha^4\ d_1 d_2^2} \right) \tag{B.82}$$

Proof has been relegated to Section B.4.3. Now by (B.79) and Lemma B.4.4 we get

$$\begin{aligned}
\mathbb{P}\Big(\exists\ \Delta \in \mathcal{A}\ s.t.\ \tilde{H}(\Delta) \leq \frac{1}{3d_1 d_2} |||\Delta|||_{\mathrm{F}}^2\Big) &\leq \sum_{\ell=1}^{\infty} \exp\left(\frac{-k\left(\beta^\ell\ \mu\right)^4}{2048\ \alpha^4\ d_1 d_2^2} \right) \\
&\overset{(a)}{\leq} \sum_{\ell=1}^{\infty} \exp\left(\frac{-2^{13}\ 9\ \beta^{4\ell}\ d_1 d_2^2 \log^2 d}{k\ \min^2\{d_1, d_2\}} \right) \\
&\overset{(b)}{\leq} \sum_{\ell=1}^{\infty} \exp\left(\frac{-2^{13}\ 9\ 4\ell \times \frac{1}{36}\ d_1 d_2^2 \log^2 d}{k\ \min^2\{d_1, d_2\}} \right) \\
&\overset{(c)}{\leq} \sum_{\ell=1}^{\infty} \exp\left(-2^{13}\ \ell\ \log d\right) \\
&= \sum_{\ell=1}^{\infty} \left(\frac{1}{d^{2^{13}}} \right)^{\ell} \\
&\overset{(d)}{=} \frac{1/d^{2^{13}}}{1 - 1/d^{2^{13}}} \\
&\overset{(e)}{\leq} \frac{2}{d^{2^{13}}}\ ,
\end{aligned} \tag{B.83}$$

where we get (a) by substituting μ from (B.56); (b) by the fact that for $\beta = \sqrt{10/9}$ and $x \geq 1$, $\beta^x \geq x \log \beta \geq x(\beta - 1) \geq x/32$; (c) by assuming $k \leq \max\{d_1, d_2^2/d_1\} \log d$; (d) because we are summing an infinite geometric sequence with common ratio of $1/d^{2^{13}}$; and (e) because for $d \geq 2$, $1/d^{2^{13}}$ is less than $1/2$.

B.4.3 Proof of Lemma B.4.4

With a slight abuse of notations, let $\tilde{Z} \equiv \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \Big(\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta)\Big)$. 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 the 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:

$$\begin{aligned}
&|\tilde{Z}_1 - \tilde{Z}_2| \\
=&\Bigg| \max_{\Delta \in \tilde{\mathcal{B}}(D)} \left(\mathbb{E}\left[\tilde{H}(\Delta)\right] - \tilde{H}(\Delta, u_{1,1}, \ldots, u_{i',\ell'}, \ldots, u_{d_1,k}) \right) - \\
&\sup_{\Delta' \in \tilde{\mathcal{B}}(D)} \left(\mathbb{E}\left[\tilde{H}(\Delta')\right] - \tilde{H}(\Delta', u_{1,1}, \ldots, u'_{i',\ell'}, \ldots, u_{d_1,k}) \right) \Bigg| \\
\overset{(a)}{\leq}&\Bigg| \left(\mathbb{E}\left[\tilde{H}(\Delta^*)\right] - \tilde{H}(\Delta^*, u_{1,1}, \ldots, u_{i',\ell'}, \ldots, u_{d_1,k}) \right) - \\
&\left(\mathbb{E}\left[\tilde{H}(\Delta^*)\right] - \tilde{H}(\Delta^*, u_{1,1}, \ldots, u'_{i',\ell'}, \ldots, u_{d_1,k}) \right) \Bigg| \\
\leq& \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \left| \tilde{H}(\Delta, u_{1,1}, \ldots, u_{i',\ell'}, \ldots, u_{d_1,k}) - \tilde{H}(\Delta, u_{1,1}, \ldots, u'_{i',\ell'}, \ldots, u_{d_1,k}) \right| \\
\overset{(b)}{\leq}& \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \left| \frac{1}{d_1\binom{k}{2}} \sum_{\ell \neq \ell'} \left(\Delta_{i',u_{i',\ell}} - \Delta_{i',u_{i',\ell'}} \right)^2 - \left(\Delta_{i',u_{i',\ell}} - \Delta_{i',u'_{i',\ell'}} \right)^2 \right| \\
\overset{(c)}{\leq}& \frac{1}{d_1\binom{k}{2}} (k-1) \left(4\alpha\right)^2 = \frac{32\alpha^2}{d_1 k}. \qquad \text{(B.84)}
\end{aligned}$$

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 $\Delta_{iu_{i',\ell \neq \ell'}}$ and $\Delta_{i,\, u_{i',\ell'}}$, and (c) is because the maximum and the minimum value of difference square terms are $(4\alpha)^2$ and 0, respectively. Using McDiarmid's bounded differences inequality, we get

$$\mathbb{P}\{\tilde{Z} - \mathbb{E}\left[\tilde{Z}\right] \geq \epsilon\} \leq \exp\left(- \frac{2\epsilon^2}{d_1 k \left(\frac{32\alpha^2}{d_1 k}\right)^2} \right), \qquad \text{(B.85)}$$

because of (B.84) and the fact that there are $d_1 k$ random variables. We upper bound $\mathbb{E}\left[\tilde{Z}\right]$ as follows.

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{Z}\right] = \\
&\mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1\binom{k}{2}} \sum_{\substack{i\in[d_1] \\ (m_1,m_2)\in\mathcal{P}_0}} \mathbb{E}\left[\left(\Delta_{i,\, u_{i,m_1}} - \Delta_{i,\, u_{i,m_2}}\right)^2\right] - \left(\Delta_{i,\, u_{i,m_1}} - \Delta_{i,\, u_{i,m_2}}\right)^2 \\
&\overset{(a)}{\leq} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_0} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\, u_{i,m_1}} - \Delta_{i,\, u_{i,m_2}}\right)^2 \\
&\overset{(b)}{\leq} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{a=1}^{k-1} \sum_{(m_1,m_2)\in\mathcal{P}_a} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\, u_{i,m_1}} - \Delta_{i,\, u_{i,m_2}}\right)^2 \\
&\overset{(c)}{\leq} \sum_{a=1}^{k-1} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_a} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\, u_{i,m_1}} - \Delta_{i,\, u_{i,m_2}}\right)^2 ,
\end{aligned}
\tag{B.86}
$$

where (a) is by the 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 (B.68), and (c) is because of fact that the supremum of the 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 the Ledoux-Talagrand contraction inequality on (B.86) to get

$$
\begin{aligned}
E[\tilde{Z}] &\leq \sum_{a=1}^{k-1} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_a} 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\, u_{i,m_1}} - \Delta_{i,\, u_{i,m_2}}\right)^2 \\
&\leq \sum_{a=1}^{k-1} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \frac{1}{d_1\binom{k}{2}} \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_a} 4\alpha\ 2\tilde{\xi}_{i,m_1,m_2}\left(\Delta_{i,\, u_{i,m_1}} - \Delta_{i,\, u_{i,m_2}}\right) \\
&\overset{(a)}{\leq} \sum_{a=1}^{k-1} \frac{8\alpha}{d_1\binom{k}{2}} \mathbb{E} \sup_{\Delta \in \tilde{\mathcal{B}}(D)} \langle\!\langle \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_a} \tilde{W}_{i,m_1,m_2}, \Delta \rangle\!\rangle \\
&\overset{(b)}{\leq} \sum_{a=1}^{k-1} \frac{8\alpha}{d_1\binom{k}{2}} \mathbb{E}\left[\left|\left|\left| \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_a} \tilde{W}_{i,m_1,m_2} \right|\right|\right|_2\right] \sup_{\Delta \in \tilde{\mathcal{B}}(D)} |||\Delta|||_{\text{nuc}} ,
\end{aligned}
\tag{B.87}
$$

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 ($\langle\!\langle x, y \rangle\!\rangle \leq |\!|\!| x |\!|\!|_2 |\!|\!| y |\!|\!|_{\text{nuc}}$). Now we use Bernstein's inequality [26] to upperbound 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

$$\left|\!\left|\!\left| \tilde{W}_{i,m_1,m_2} \right|\!\right|\!\right|_2 \leq \sqrt{2} \,. \tag{B.88}$$

We also get

$$\begin{aligned}
\mathbb{E}\left[\tilde{W}_{i,m_1,m_2} \tilde{W}_{i,m_1,m_2}^T\right] &= 2 e_i e_i^T \mathbb{E}\left[1 - e_{u_{i,m_1}}^T e_{u_{i,m_2}}\right] \\
&\preceq e_i e_i^T \left(2 - \frac{2}{d_2}\right) \\
&\preceq 2 e_i e_i^T \,,
\end{aligned} \tag{B.89}$$

and

$$\begin{aligned}
\mathbb{E}\left[\tilde{W}_{i,m_1,m_2}^T \tilde{W}_{i,m_1,m_2}\right] &= \mathbb{E}\left[2 e_{u_{i,m_1}} e_{u_{i,m_1}}^T - 2 e_{u_{i,m_1}} e_{u_{i,m_2}}^T\right] \\
&\preceq \frac{2}{d_2} \mathbf{I}_{d_2 \times d_2} - \frac{2}{d_2^2} \mathbf{1}\mathbf{1}_{d_2 \times d_2} \\
&\preceq \frac{2}{d_2} \mathbf{I}_{d_2 \times d_2} \,.
\end{aligned} \tag{B.90}$$

Therefore, using (B.89) and (B.90), the standard deviation of $\sum_{(i,m_1,m_2)} Z_{i,m_2,m_2}$ is

$$\begin{aligned}
\sigma^2 &\leq \sum_{\substack{i \in [d_1] \\ (m_1,m_2) \in \mathcal{P}_a}} \max\left\{ \left|\!\left|\!\left| \mathbb{E}\left[\tilde{W}_{i,m_2,m_2} \tilde{W}_{i,m_2,m_2}^T\right] \right|\!\right|\!\right|_2, \left|\!\left|\!\left| \mathbb{E}\left[\tilde{W}_{i,m_2,m_2}^T \tilde{W}_{i,m_2,m_2}\right] \right|\!\right|\!\right|_2 \right\} \\
&\leq \frac{d_1 k}{2} \max\left\{ \frac{2}{d_1} |\!|\!| \mathbf{I} |\!|\!|_2, \frac{2}{d_2} |\!|\!| \mathbf{I} |\!|\!|_2 \right\} \\
&= \frac{k d_1}{\min\{d_1, d_2\}} \,.
\end{aligned} \tag{B.91}$$

By matrix Bernstein inequality [26], $\forall \;\; a \in [k-1]$,

$$
\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{i\in[d_1]} \sum_{(m_1,m_2)\in\mathcal{P}_a} \tilde{W}_{i,m_2,m_2} \right\|_2 > t \right) \\
&\quad \le (d_1+d_2)\exp\left(\frac{-t^2/2}{2kd_1/\min\{d_1,d_2\} + \sqrt{2}t/3} \right) , \qquad \text{(B.92)}
\end{aligned}
$$

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

$$
\begin{aligned}
t &= \max\left\{ \sqrt{\frac{8kd_1\left((1+c_1)\log d\right)}{\min\{d_1,d_2\}}} , \frac{4\sqrt{2}\left((1+c_1)\log d\right)}{3} \right\} \\
&= \sqrt{\frac{8kd_1\left((1+c_1)\log d\right)}{\min\{d_1,d_2\}}}, \text{when } k \ge 4(c_1+1)\log d/9 \; . \qquad \text{(B.93)}
\end{aligned}
$$

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

$$
\mathbb{E}\left[\left\| \sum_{i=1}^{d_1} \sum_{(m_1,m_2)\in\mathcal{P}_a} \tilde{W}_{i,m_2,m_2} \right\|_2 \right] \le \sqrt{\frac{8kd_1\left((1+c_1)\log d\right)}{\min\{d_1,d_2\}}} \;+\; \frac{2}{d^{c_1}}\frac{\sqrt{2}d_1k}{2} \, , \qquad \text{(B.94)}
$$

because from (B.88) we get $\left\| \sum_{\substack{i\in[d_1]\\(m_1,m_2)\in\mathcal{P}_a}} \tilde{W}_{i,m_2,m_2} \right\|_2 \le \sum_{\substack{i\in[d_1]\\(m_1,m_2)\in\mathcal{P}_a}} \left\| \tilde{W}_{i,m_2,m_2} \right\|_2$ $\le d_1k/2(\sqrt{2})$. From (B.87) and (B.94), putting $c_1 = 2$, we get

$$
\begin{aligned}
\mathbb{E}\left[\tilde{Z}\right] &\le \sum_{a=1}^{k-1} \frac{8\alpha}{d_1\binom{k}{2}} \left(\sqrt{\frac{24\ kd_1\log d}{\min\{d_1,d_2\}}} \;+\; \frac{\sqrt{2}d_1k}{d^2} \right) \sup_{\Delta\in\tilde{\mathcal{B}}(D)} |||\Delta|||_{\mathrm{nuc}} \\
&\overset{(a)}{\le} 8\alpha \left(2\sqrt{\frac{24\ \log d}{k\ d_1\min\{d_1,d_2\}}} \;+\; \frac{2\sqrt{2}}{d^2} \right) \frac{D^2}{\mu} \\
&\overset{(b)}{\le} 16\alpha\sqrt{\frac{48\log d}{k\ d_1\min\{d_1,d_2\}}} D^2 \frac{1}{16\alpha}\sqrt{\frac{k\ \min\{d_1,d_2\}}{48d_1d_2^2\log d}} \\
&= \frac{D^2}{d_1d_2} \, , \qquad \text{(B.95)}
\end{aligned}
$$

where (a) is obtained because of (B.80), which gives $\sup_{D\in\mathcal{B}(D)} |||\Delta|||_{\mathrm{nuc}} \le D^2/\mu$ and (b) can be obtained by assuming that $k \le d^2\log d$. Using the above bound in (B.85), we get

$$\mathbb{P}\{\tilde{Z} - D^2/(d_1 d_2) \geq \epsilon\} \leq \mathbb{P}\{\tilde{Z} - \mathbb{E}\left[\tilde{Z}\right] \geq \epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{d_1 k \left(\frac{32\alpha^2}{d_1 k}\right)^2}\right) , \tag{B.96}$$

and using $\epsilon = D^2/(2d_1 d_2)$ will get us the required bound.

APPENDIX C

PROOFS OF BUNDLED CHOICES

C.1 Proof of Theorem 6: performance guarantee for bundled choices

We use similar notations and techniques as the proof of Theorem 3 in Appendix B.1. From the definition of $\mathcal{L}(\Theta)$ in (4.3), we have for the true parameter Θ^*, the gradient evaluated at the true parameter is

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

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 $p_{j_1,j_2|S_i,T_i} \equiv \mathbb{P}\left\{(u_i, v_i) = (j_1, j_2)|S_i, T_i\right\} = e^{\Theta^*_{j_1,j_2}}/(\sum_{j_1'\in S_i, j_2'\in T_i} e^{\Theta^*_{j_1',j_2'}})$, 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

$$\begin{aligned}\frac{\partial^2\mathcal{L}(\Theta)}{\partial\Theta_{j_1,j_2}\,\partial\Theta_{j_1',j_2'}} \;&=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\big((j_1,j_2)\in S_i\times T_i\big)\frac{\partial p_{j_1,j_2|S_i,T_i}}{\partial\Theta_{j_1',j_2'}}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\big((j_1,j_2),(j_1',j_2')\in S_i\times T_i\big)\times\\ &\Big(p_{j_1,j_2|S_i,T_i}\mathbb{I}((j_1,j_2)=(j_1',j_2')) - p_{j_1,j_2|S_i,T_i}p_{j_1',j_2'|S_i,T_i}\Big)\;.\end{aligned} \tag{C.2}$$

We use $\nabla^2\mathcal{L}(\Theta) \in \mathbb{R}^{d_1d_2\times d_1d_2}$ to denote this Hessian. Let $\Delta = \Theta^* - \widehat{\Theta}$ where $\widehat{\Theta}$ is an optimal solution to the convex optimization in (4.2). 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 [26].

Lemma C.1.1. *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 $1 - 2d^{-c}$,*

$$\|\nabla \mathcal{L}(\Theta^*)\|_2 \;\leq\; \sqrt{\frac{4(1+c)e^{2\alpha}\max\{d_1,d_2\}\,\log d}{d_1\,d_2\,n}}\;. \tag{C.3}$$

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 a with relatively small nuclear norm, then we can prove restricted strong convexity, which gives the following bound.

Lemma C.1.2 (Restricted strong convexity for bundled choice). *Fix any $\Theta \in \Omega_\alpha$ and assume $(\min\{d_1,d_2\}/\min\{k_1,k_2\})\log d \leq n$ and $n \leq \min\{d^5 \log d, k_1 k_2 \max\{d_1^2, d_2^2\}\log d\}$. Under the random sampling model of choosing 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 1.1, we have, with probability larger than $1 - 2d^{-2^{25}}$,*

$$\mathrm{Vec}(\Delta)^T\, \nabla^2\mathcal{L}(\Theta)\,\mathrm{Vec}(\Delta) \geq \frac{e^{-2\alpha}}{8\,d_1\,d_2}\|\Delta\|_{\mathrm{F}}^2\,, \tag{C.4}$$

for all Δ in $\mathcal{A}$ where

$$\mathcal{A} = \left\{\Delta \in \mathbb{R}^{d_1\times d_2} \,\middle|\, \|\Delta\|_\infty \leq 2\alpha\,, \sum_{j_1\in[d_1],j_2\in[d_2]} \Delta_{j_1 j_2} = 0 \;\text{ and } \|\Delta\|_{\mathrm{F}}^2 \geq \mu\|\Delta\|_{\mathrm{nuc}}\right\}, \tag{C.5}$$

with

$$\mu \;\equiv\; 2^{10}\,\alpha\,d_1 d_2 \sqrt{\frac{\log d}{n\,\min\{d_1,d_2\}\,\min\{k_1,k_2\}}}\;. \tag{C.6}$$

Building on these lemmas, the proof of Theorem 6 is divided into the following two cases. In both cases, we will show that

$$\|\Delta\|_{\mathrm{F}}^2 \;\leq\; 12\,e^{2\alpha}c_1\lambda\,d_1 d_2\,\|\Delta\|_{\mathrm{nuc}}\,, \tag{C.7}$$

with high probability. Applying Lemma 3.2.1 proves the desired theorem. We are left to show (C.7) holds.

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

$$\mathcal{L}(\widehat{\Theta}) = \mathcal{L}(\Theta^*) - \langle\!\langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle\!\rangle + \frac{1}{2}\mathrm{Vec}(\Delta)\nabla^2\mathcal{L}(\Theta)\mathrm{Vec}^T(\Delta), \tag{C.8}$$

where $\Theta = a\widehat{\Theta} + (1-a)\Theta^*$ for some $a \in [0,1]$. It follows from Lemma C.1.2 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

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

By the assumption, we choose $\lambda \geq 8\lambda$. In view of Lemma C.1.1, 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 $\lambda \leq c_1\lambda$, this proves the desired bound in (C.7)
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 $\mu \leq 12\, e^{2\alpha} c_1 \lambda\, d_1 d_2$, and we get the same bound as in (C.7).

C.1.1 Proof of Lemma C.1.1

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 $e^{2\alpha}/(k_1k_2)$,

$$|||X_i|||_2 \;\leq\; 1 + \frac{e^{2\alpha}}{\sqrt{k_1k_2}} \,,$$

and

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}[X_i X_i^T] &= \sum_{i=1}^{n} (\mathbb{E}[e_{u_i} e_{u_i}^T] - p_i p_i^T) && \text{(C.9)}\\ &\preceq \sum_{i=1}^{n} \mathbb{E}[e_{u_i} e_{u_i}^T] && \text{(C.10)}\\ &\preceq \frac{e^{2\alpha}\, n}{d_1} \mathbf{I}_{d_1 \times d_1}\ , && \text{(C.11)}\end{aligned}$$

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,

$$\left|\!\left|\!\left| \sum_{i=1}^{n} \mathbb{E}[X_i X_i^T] \right|\!\right|\!\right|_2 \;\leq\; \frac{e^{2\alpha}\, n}{d_1}\ . \tag{C.12}$$

Similarly,

$$\left|\!\left|\!\left| \sum_{i=1}^{n} \mathbb{E}[X_i^T X_i] \right|\!\right|\!\right|_2 \;\leq\; \frac{e^{2\alpha}\, n}{d_2}\ . \tag{C.13}$$

Applying matrix Bernstein inequality [26], we get

$$\begin{aligned}&\mathbb{P}\left\{|\!|\!|\nabla\mathcal{L}(\Theta^*)|\!|\!|_2 > t\right\}\\ &\leq (d_1+d_2)\exp\left\{\frac{-n^2t^2/2}{(e^{2\alpha} n \max\{d_1,d_2\}/(d_1d_2)) \;+\; ((1+(e^{2\alpha}/\sqrt{k_1k_2}))nt/3)}\right\}\ ,\end{aligned} \tag{C.14}$$

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

$$\begin{aligned}t &= \max\left\{\sqrt{\frac{4(1+c)e^{2\alpha}\max\{d_1,d_2\}\log d}{d_1d_2n}}\,,\ \frac{4(1+c)(1+\frac{e^{2\alpha}}{\sqrt{k_1k_2}})\log d}{3n}\right\}\\ &= \sqrt{\frac{4(1+c)e^{2\alpha}\max\{d_1,d_2\}\log d}{d_1d_2n}}\ ,\end{aligned}$$

where the last equality follows from the assumption that
$n \geq (4(1+c)e^{2\alpha}d_1d_2\log d)/\max\{d_1,d_2\}$.

C.1.2 Proof of Lemma C.1.2

The quadratic form of the Hessian defined in (C.2) can be lower bounded by

$$\mathrm{Vec}(\Delta)^T\,\nabla^2\mathcal{L}(\Theta)\,\mathrm{Vec}(\Delta) \geq \underbrace{\frac{e^{-2\alpha}}{2\,k_1^2\,k_2^2\,n}\sum_{i=1}^{n}\sum_{j_1,j_1'\in S_i}\sum_{j_2,j_2'\in T_i}\left(\Delta_{j_1,j_2}-\Delta_{j_1',j_2'}\right)^2}_{\equiv H(\Delta)}\,, \tag{C.15}$$

which follows from Remark B.1.2. To lower bound $H(\Delta)$, we first compute the mean:

$$\begin{aligned}\mathbb{E}[H(\Delta)] &= \frac{e^{-2\alpha}}{2\,k_1^2\,k_2^2\,n}\sum_{i=1}^{n}\mathbb{E}\Big[\sum_{j_1,j_1'\in S_i}\sum_{j_2,j_2'\in T_i}\left(\Delta_{j_1,j_2}-\Delta_{j_1',j_2'}\right)^2\Big] &\text{(C.16)}\\ &= \frac{e^{-2\alpha}}{d_1\,d_2}|||\Delta|||_{\mathrm{F}}^2\,, &\text{(C.17)}\end{aligned}$$

where we used the fact that $\mathbb{E}[\sum_{j_1\in S_i,j_2\in T_i}\Delta_{j_1,j_2}] = \frac{k_1k_2}{d_1d_2}\sum_{\substack{j_1'\in[d_1]\\ j_2'\in[d_2]}}\Delta_{j_1',j_2'} = 0$ for $\Delta\in\Omega_{2\alpha}$ in (4.4).

We now prove that $H(\Delta)$ does not deviate from its mean too much. Suppose there exists a $\Delta\in\mathcal{A}$ defined in (C.5) such that (C.4) is violated, i.e. $H(\Delta) < (e^{-2\alpha}/(8k_1k_2d_1d_2))|||\Delta|||_{\mathrm{F}}^2$. In this case,

$$\mathbb{E}[H(\Delta)] - H(\Delta) \;\geq\; \frac{7\,e^{-2\alpha}}{8d_1d_2}|||\Delta|||_{\mathrm{F}}^2\,. \tag{C.18}$$

We will show that this happens with a small probability. We use the same peeling argument as in Appendix B.1 with

$$\mathcal{S}_\ell = \Big\{\Delta\in\mathbb{R}^{d_1\times d_2}\,|\,|||\Delta|||_\infty\leq 2\alpha, \beta^{\ell-1}\mu\leq|||\Delta|||_{\mathrm{F}}\leq\beta^\ell\mu,\\ \sum_{j_1\in[d_1],j_2\in[d_2]}\Delta_{j_1,j_2}=0,\text{ and }|||\Delta|||_{\mathrm{nuc}}\leq\beta^{2\ell}\mu\Big\}\,, \tag{C.19}$$

where $\beta=\sqrt{10/9}$ and for $\ell\in\{1,2,3,\ldots\}$, and μ is defined in (C.6). By the peeling argument, there exists an $\ell\in\mathbb{Z}_+$ such that $\Delta\in\mathcal{S}_\ell$ and

$$\mathbb{E}[H(\Delta)] - H(\Delta) \;\geq\; \frac{7\,e^{-2\alpha}}{8d_1d_2}\beta^{2\ell-2}(\mu)^2 \;\geq\; \frac{7\;e^{-2\alpha}}{9\,d_1d_2}\beta^{2\ell}(\mu)^2\,. \tag{C.20}$$

Applying the union bound over $\ell \in \mathbb{Z}_+$,

$$\begin{aligned}
\mathbb{P}&\left\{\exists \Delta \in \mathcal{A}\,,\; H(\Delta) < \frac{e^{-2\alpha}}{8\,d_1\,d_2} |||\Delta|||_{\rm F}^2 \right\} \\
&\leq \sum_{\ell=1}^{\infty} \mathbb{P}\left\{ \sup_{\Delta \in \mathcal{S}_\ell} \big(\, \mathbb{E}[H(\Delta)] - H(\Delta) \,\big) > \frac{7\,e^{-2\alpha}}{9 d_1 d_2} (\beta^\ell \mu)^2 \right\} \\
&\leq \;\sum_{\ell=1}^{\infty} \mathbb{P}\left\{ \sup_{\Delta \in \mathcal{B}(\beta^\ell \mu)} \big(\, \mathbb{E}[H(\Delta)] - H(\Delta) \,\big) > \frac{7 e^{-2\alpha}}{9 d_1 d_2} (\beta^\ell \mu)^2 \right\}\,,
\end{aligned} \tag{C.21}$$

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

$$\begin{aligned}
\mathcal{B}(D) = \Big\{\, \Delta \in \mathbb{R}^{d_1 \times d_2} \,\Big|\, &\|\Delta\|_\infty \leq 2\alpha, |||\Delta|||_{\rm F} \leq D, \\
&\sum_{j_1 \in [d_1], j_2 \in [d_2]} \Delta_{j_1 j_2} = 0, \mu |||\Delta|||_{\rm nuc} \leq D^2 \,\Big\}\,.
\end{aligned} \tag{C.22}$$

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

Lemma C.1.3. *For* $(\min\{d_1, d_2\}/\min\{k_1, k_2\}) \log d \leq n \leq d^5 \log d$,

$$\mathbb{P}\left\{ \sup_{\Delta \in \mathcal{B}(D)} \Big(\, \mathbb{E}[H(\Delta)] - H(\Delta) \,\Big) \geq \frac{e^{-2\alpha} D^2}{2 d_1 d_2} \right\} \leq \exp\left\{ -\frac{n\, \min\{k_1^2, k_2^2\}\, k_1 k_2\, D^4}{2^{10} \alpha^4 d_1^2 d_2^2} \right\}\,. \tag{C.23}$$

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 (C.21), we get

$$\begin{aligned}
\mathbb{P}&\left\{\exists \Delta \in \mathcal{A}\,,\; H(\Delta) < \frac{e^{-2\alpha}}{8\,d_1 d_2} |||\Delta|||_{\rm F}^2 \right\} \\
&\leq \sum_{\ell=1}^{\infty} \exp\left\{ -\frac{n\, k_1 k_2\, \min\{k_1^2, k_2^2\}\, (\beta^\ell \mu)^4}{2^{10} \alpha^4 d_1^2 d_2^2} \right\} \\
&\overset{(a)}{\leq} \sum_{\ell=1}^{\infty} \exp\left\{ -\frac{n k_1 k_2 \min\{k_1^2, k_2^2\} \ell (\beta - 1.002)(\mu)^4}{2^{10} \alpha^4 d_1^2 d_2^2} \right\} \leq \frac{\eta}{1-\eta},
\end{aligned} \tag{C.24}$$

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 $\beta - 1.002 \geq 2^{-5}$, and $n \leq k_1 k_2 \max\{d_1^2, d_2^2\} \log d$. 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.

C.1.3 Proof of Lemma C.1.3

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 $8\alpha^2 e^{-2\alpha} \frac{\max\{k_1,k_2\}}{k_1^2 k_2^2 n}$ when one of the $k_1 k_2 n$ independent random variables is changed, which gives

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq t\right\} \;\; \leq \;\; \exp\Big(- \frac{k_1^4 k_2^4 n^2 t^2}{64 \alpha^4 e^{-4\alpha} \max\{k_1^2, k_2^2\} k_1 k_2 n}\Big) \,. \quad \text{(C.25)}$$

With the choice of $t = D^2/(4e^{2\alpha}\, d_1 d_2)$, this gives

$$\mathbb{P}\left\{ Z - \mathbb{E}[Z] \geq \frac{e^{-2\alpha}}{4 d_1 d_2} D^2 \right\} \leq \exp\Big(- \frac{k_1^3 k_2^3 n D^4}{2^{10} \alpha^4 d_1^2 d_2^2 \max\{k_1^2, k_2^2\}}\Big) \,. \quad \text{(C.26)}$$

We first construct a partition of the space similar to Lemma B.1.4. Let

$$\tilde{k} \equiv \min\{k_1, k_2\} \,. \quad \text{(C.27)}$$

Lemma C.1.4. *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*

$$\{(\Delta_{j_{i,a}, j_{i,b}} - \Delta_{j_{i,a'}, j_{i,b'}})^2\}_{i \in [n], ((a,b),(a',b')) \in \mathcal{T}_\ell} \;\; \textit{are mutually independent} \quad \text{(C.28)}$$

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 $b \in [k_2]$.

$$
\begin{aligned}
&\mathbb{E}[Z] \\
&= \frac{e^{-2\alpha}}{2\,k_1^2\,k_2^2\,n}\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}(D)} \sum_{\substack{i\in[n]\\ a,a'\in[k_1]\\ b,b'\in[k_2]}} \mathbb{E}\big(\Delta_{j_{i,a},j_{i,b}} - \Delta_{j_{i,a'},j_{i,b'}}\big)^2 - \big(\Delta_{j_{i,a},j_{i,b}} - \Delta_{j_{i,a'},j_{i,b'}}\big)^2\right] \\
&\le \frac{e^{-2\alpha}}{2\,k_1^2\,k_2^2\,n}\sum_{\ell\in[N]}\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}(D)} \sum_{\substack{i\in[n]\\ (j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}} \mathbb{E}\big(\Delta_{j_1,j_2} - \Delta_{j_1',j_2'}\big)^2 - \big(\Delta_{j_1,j_2} - \Delta_{j_1',j_2'}\big)^2\right] \\
&\le \frac{e^{-2\alpha}}{k_1^2\,k_2^2\,n}\sum_{\ell\in[N]}\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i=1}^n \sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell} \xi_{i,j_1,j_2,j_1',j_2'}\big(\Delta_{j_1,j_2} - \Delta_{j_1',j_2'}\big)^2\right]\;,
\end{aligned}
\tag{C.29}
$$

where the first inequality follows from the fact that the supremum of the sum is smaller than the sum of supremum, and the second inequality follows from the standard symmetrization 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 that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i=1}^n \sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell} \xi_{i,j_1,j_2,j_1',j_2'}\big(\Delta_{j_1,j_2} - \Delta_{j_1',j_2'}\big)^2\right] \\
&\quad\le\; 8\alpha\,\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}(D)} \sum_{i=1}^n \sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell} \xi_{i,j_1,j_2,j_1',j_2'}\big(\Delta_{j_1,j_2} - \Delta_{j_1',j_2'}\big)\right] \\
&\quad\le\; 8\alpha\,\mathbb{E}\left[\sup_{\Delta\in\mathcal{B}(D)} |||\Delta|||_{\text{nuc}} \left|\left|\left|\sum_{i=1}^n \sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell} \xi_{i,j_1,j_2,j_1',j_2'}\big(e_{j_1,j_2} - e_{j_1',j_2'}\big)\right|\right|\right|_2\right] \\
&\quad\le\; \frac{8\alpha D^2}{\mu}\mathbb{E}\left[\left|\left|\left|\sum_{i=1}^n \sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell} \xi_{i,j_1,j_2,j_1',j_2'}\big(e_{j_1,j_2} - e_{j_1',j_2'}\big)\right|\right|\right|_2\right]\;,
\end{aligned}
\tag{C.30}
$$

where the second inequality follows for Hölder's inequality and the last inequality follows from $\mu|||\Delta|||_{\text{nuc}} \le D^2$ 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 $|||\xi_{i,j_1,j_2,j_1',j_2'}c|||_2 \le \sqrt{2}$ 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

$$\mathbb{P}\left\{\left|\!\left|\!\left|\sum_{i=1}^{n}\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\big(e_{j_1,j_2} - e_{j_1',j_2'}\big)\right|\!\right|\!\right|_2 > t\right\}$$

$$\leq (d_1+d_2)\exp\left\{\frac{-t^2/2}{\frac{2n\min\{k_1,k_2\}}{\min\{d_1,d_2\}} + \frac{\sqrt{2}t}{3}}\right\}\ , \qquad \text{(C.31)}$$

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 $\left|\!\left|\!\left|\sum_{i=1}^{n}\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\big(e_{j_1,j_2} - e_{j_1',j_2'}\big)\right|\!\right|\!\right|_2 \leq (n/\sqrt{2})\min\{k_1,k_2\}$, it follows that

$$\mathbb{E}\left[\left|\!\left|\!\left|\sum_{i=1}^{n}\sum_{(j_1,j_2,j_1',j_2')\in\mathcal{T}_\ell}\xi_{i,j_1,j_2,j_1',j_2'}\big(e_{j_1,j_2} - e_{j_1',j_2'}\big)\right|\!\right|\!\right|_2\right]$$

$$\leq\ \max\left\{\sqrt{\frac{64\,n\ \min\{k_1,k_2\}\log d}{\min\{d_1,d_2\}}}, (16\sqrt{2}/3)\log d\right\} + \frac{2n\min\{k_1,k_2\}}{\sqrt{2}d^7}$$

$$\leq\ \sqrt{\frac{66\,n\ \min\{k_1,k_2\}\log d}{\min\{d_1,d_2\}}} \qquad \text{(C.32)}$$

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 (C.29), and (C.30), we get that

$$\mathbb{E}[Z] \leq \frac{16e^{-2\alpha}\alpha D^2}{\mu}\sqrt{\frac{66\log d}{n\min\{k_1,k_2\}\min\{d_1,d_2\}}} \leq \frac{e^{-2\alpha}D^2}{4\,d_1 d_2}\ . \qquad \text{(C.33)}$$

C.2 Proof of Theorem 7: information-theoretic lower bound

This proof follows closely the proof of Theorem 4 in Appendix B.3. We apply the generalized Fano's inequality in the same way to get (B.34)

$$\mathbb{P}\left\{\widehat{L} \neq L\right\}\ \geq\ 1 - \frac{\binom{M}{2}^{-1}\sum_{\ell_1,\ell_2\in[M]} D_{\rm KL}(\Theta^{(\ell_1)}\|\Theta^{(\ell_2)}) + \log 2}{\log M}\ ,\text{(C.34)}$$

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_1k_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 sampling be $S \subset [d_1]$ and $T \subset [d_2]$ respectively. Then,

$$
\begin{aligned}
&D_{\mathrm{KL}}(\Theta^{(\ell_1)}\|\Theta^{(\ell_2)})\\
&\overset{(a)}{=} \frac{n}{\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \sum_{\substack{i\in S\\ j\in T}} \frac{e^{\Theta^{(\ell_1)}_{ij}}}{\sum_{\substack{i'\in S\\ j'\in T}} e^{\Theta^{(\ell_1)}_{i'j'}}} \log\left(\frac{e^{\Theta^{(\ell_1)}_{ij}} \sum_{\substack{i'\in S\\ j'\in T}} e^{\Theta^{(\ell_2)}_{i'j'}}}{e^{\Theta^{(\ell_2)}_{ij}} \sum_{\substack{i'\in S\\ j'\in T}} e^{\Theta^{(\ell_1)}_{i'j'}}}\right)\\
&\overset{(b)}{\le} \frac{n}{\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \left(\sum_{i,j} \frac{e^{2\Theta^{(\ell_1)}_{ij}} \sum_{i',j'} e^{\Theta^{(\ell_2)}_{i'j'}} - e^{\Theta^{(\ell_1)}_{ij}+\Theta^{(\ell_2)}_{ij}} \sum_{i',j'} e^{\Theta^{(\ell_1)}_{i'j'}}}{e^{\Theta^{(\ell_2)}_{ij}} \left(\sum_{i',j'} e^{\Theta^{(\ell_1)}_{i'j'}}\right)^2} \right)\\
&\overset{(c)}{\le} \frac{ne^{2\alpha}}{k_1^2k_2^2\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T}\sum_{i,j} \left(e^{2\Theta^{(\ell_1)}_{ij}-\Theta^{(\ell_2)}_{ij}} \sum_{i',j'} e^{\Theta^{(\ell_2)}_{i'j'}} - e^{\Theta^{(\ell_1)}_{ij}} \sum_{i',j'} e^{\Theta^{(\ell_1)}_{i'j'}} \right)\\
&= \frac{ne^{2\alpha}}{k_1^2k_2^2\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T} \left(\sum_{i',j'} e^{\Theta^{(\ell_2)}_{i'j'}} \sum_{i,j} \frac{\left(e^{\Theta^{(\ell_1)}_{ij}} - e^{\Theta^{(\ell_2)}_{ij}}\right)^2}{e^{\Theta^{(\ell_2)}_{ij}}} - \left(\sum_{i,j} e^{\Theta^{(\ell_1)}_{ij}} - e^{\Theta^{(\ell_2)}_{ij}}\right)^2 \right)\\
&\overset{(d)}{\le} \frac{ne^{4\alpha}}{k_1k_2\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T}\sum_{i,j} \left(e^{\Theta^{(\ell_1)}_{ij}} - e^{\Theta^{(\ell_2)}_{ij}}\right)^2\\
&\overset{(e)}{\le} \frac{ne^{5\alpha}}{k_1k_2\binom{d_1}{k_1}\binom{d_2}{k_2}} \sum_{S,T}\sum_{i,j} \left(\Theta^{(\ell_1)}_{ij} - \Theta^{(\ell_2)}_{ij}\right)^2\\
&\overset{(f)}{=} \frac{ne^{5\alpha}}{d_1d_2} \left|\!\left|\!\left| \Theta^{(\ell_1)}_{ij} - \Theta^{(\ell_2)}_{ij} \right|\!\right|\!\right|_{\mathrm{F}}^2
\end{aligned}
\tag{C.35}
$$

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 $\log(x) \le x-1$ with $x = (e^{\Theta^{(\ell_1)}_{ij}} \sum_{i'\in S, j'\in T} e^{\Theta^{(\ell_2)}_{i'j'}})/(e^{\Theta^{(\ell_2)}_{ij}} \sum_{i'\in S, j'\in T} e^{\Theta^{(\ell_1)}_{i'j'}})$. The constants at (c) are 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 $-\alpha \le x \le \alpha$ 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,

$$\mathbb{P}\Big\{\widehat{L} \neq L\Big\} \ge 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell_1,\ell_2 \in [M]} \frac{ne^{5\alpha}}{d_1 d_2} \Big|\!\Big|\!\Big| \Theta^{(\ell_2)}_{ij} - \Theta^{(\ell_2)}_{ij} \Big|\!\Big|\!\Big|_{\rm F}^2 + \log 2}{\log M}, \quad \text{(C.36)}$$

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

Lemma C.2.1. *Let $d_2 \ge d_1$ be sufficiently large positive integers. Then for each $r \in \{1, \dots, d_1\}$, and for any positive $\delta > 0$, there exists a family of $d_1 \times d_2$ dimensional matrices $\{\Theta^{(1)}, \dots, \Theta^{(M(\delta))}\}$ with cardinality $M(\delta) = \lfloor (1/4)\exp(rd_2/576) \rfloor$ such that each matrix is rank r and the following bounds hold:*

$$\begin{aligned} |\!|\!| \Theta^{(\ell)} |\!|\!|_{\rm F} &\le \delta\ , \ \textit{for all } \ell \in [M] && \text{(C.37)} \\ |\!|\!| \Theta^{(\ell_1)} - \Theta^{(\ell_2)} |\!|\!|_{\rm F} &\ge \frac{1}{2}\delta\ , \ \textit{for all } \ell_1, \ell_2 \in [M] && \text{(C.38)} \\ \Theta^{(\ell)} &\in \Omega_{\tilde\alpha}\ , \ \textit{for all } \ell \in [M]\ , && \text{(C.39)} \end{aligned}$$

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

Suppose $\delta \le \alpha d_2/(8\sqrt{2\log d})$ such that the matrices in the packing set are entry-wise bounded by α, then the above lemma C.2.1 implies that $|\!|\!| \Theta^{(\ell_1)} - \Theta^{(\ell_2)} |\!|\!|_{\rm F}^2 \le 4\delta^2$, which gives

$$\mathbb{P}\Big\{\widehat{L} \neq L\Big\} \ \ge\ 1 - \frac{\frac{e^{5\alpha} n 4\delta^2}{d_1 d_2} + \log 2}{\frac{rd_2}{576} - 2\log 2} \ \ge\ \frac{1}{2}\ , \quad \text{(C.40)}$$

where the last inequality holds for $\delta^2 \le (rd_1 d_2^2/(1152 e^{5\alpha} n))$ and assuming $rd_2 \ge 1600$. Together with (C.40) and (C.38), this inequality proves that for all $\delta \le \min\{\alpha d_2/(8\sqrt{2\log d}), rd_1 d_2^2/(1152 e^{5\alpha} n)\}$,

$$\inf_{\widehat\Theta} \sup_{\Theta^* \in \Omega_\alpha} \mathbb{E}\Big[\Big|\!\Big|\!\Big| \widehat\Theta - \Theta^* \Big|\!\Big|\!\Big|_{\rm F} \Big] \ \ge\ \delta/4\ .$$

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 a similar bound with d_1 and d_2 interchanged.

C.2.1 Proof of Lemma C.2.1

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:

$$\Theta^{(\ell)} \;=\; \frac{\delta}{\sqrt{rd_2}} U (V^{(\ell)})^T \Big(\mathbf{I}_{d_2 \times d_2} - \frac{\mathbb{1}^T U (V^{(\ell)})^T \mathbb{1}}{d_1 d_2} \mathbb{1}\mathbb{1}^T \Big) \,, \tag{C.41}$$

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 $V^{(\ell)}_{ij} \in \{-1,+1\}$ chosen independently and uniformly at random. By construction, notice that $|||\Theta^{(\ell)}|||_{\mathrm{F}} \leq (\delta/\sqrt{rd_2}) |||U(V^{(\ell)})^T|||_{\mathrm{F}} = \delta$.

Now, by triangular inequality, we have

$$\begin{aligned} & |||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{F}} \\ & \geq \frac{\delta}{\sqrt{rd_2}} |||U(V^{(\ell_1)} - V^{(\ell_2)})^T|||_{\mathrm{F}} - \frac{\delta\, |\mathbb{1}^T U (V^{(\ell_1)} - V^{(\ell_2)})^T \mathbb{1}|}{d_1 d_2 \sqrt{rd_2}} |||\mathbb{1}\mathbb{1}^T|||_{\mathrm{F}} \qquad \text{(C.42)} \\ & \geq \frac{\delta}{\sqrt{rd_2}} \underbrace{|||V^{(\ell_1)} - V^{(\ell_2)}|||_{\mathrm{F}}}_{A} - \frac{\delta}{\sqrt{r\, d_1\, d_2^2}} \big(\underbrace{|\mathbb{1}^T U (V^{(\ell_1)})^T \mathbb{1}|}_{B} + |\mathbb{1}^T U (V^{(\ell_2)})^T \mathbb{1}| \big) \,. \qquad \text{(C.43)} \end{aligned}$$

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. Thus with probability at least 3/4, there exists M matrices such that

$$|||\Theta^{(\ell_1)} - \Theta^{(\ell_2)}|||_{\mathrm{F}} \;\geq\; \delta\Big(1 - \sqrt{\frac{2^7 \log(32r)\, \log(32d)}{d_1 d_2}}\Big) \;\geq\; \frac{1}{2}\delta \,,$$

for all ℓ_1, $\ell_2 \in [M]$ and for sufficiently large d_1 and d_2.

Applying McDiarmid's inequality as in (B.47) in Appendix B.3, 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,

$$\mathbb{P}\Big\{|\mathbb{1}^T U (V^{(\ell)})^T \mathbb{1}| \leq 8\sqrt{2rd_2 \log(32r)\log(32d)}\Big\} \;\geq\; \frac{7}{8}\,. \tag{C.44}$$

Then using the similar technique as in (B.50), it follows that we can find $M = (1/4)M'$ matrices all satisfying this bound and also the bound on the max-entry in (C.45). We are left to prove (C.44). 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 [31], 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 (C.44).

We are left to prove that $\Theta^{(\ell)}$'s are in $\Omega_{(8\delta/d_2)\sqrt{2\log d_2}}$ as defined in (4.4). Similar to (B.49), applying Levy's concentration gives

$$\mathbb{P}\Big\{\max_{i,j}|\Theta_{ij}^{(\ell)}| \leq \frac{2\delta\sqrt{32\log d_2}}{d_2}\Big\} \geq\; 1 - 2\exp\Big\{-2\log d_2\Big\} \;\geq\; \frac{1}{2}\,, \tag{C.45}$$

for a fixed $\ell \in [M']$. Then using the similar technique as in (B.50), it follows that there exists $M = (1/4)M'$ matrices all satisfying this bound and also the bound on B in (C.44).

REFERENCES

[1] R. M. Bell and Y. Koren, "Lessons from the Netflix prize challenge," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 75–79, 2007.

[2] J. J. Bartholdi and J. B. Orlin, "Single transferable vote resists strategic voting," *Social Choice and Welfare*, vol. 8, no. 4, pp. 341–354, 1991.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[4] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright, "When is it better to compare than to score?" *arXiv.org*, 2014. [Online]. Available: https://arxiv.org/abs/1406.6618

[5] S. Negahban, S. Oh, K. K. Thekumparampil, and J. Xu, "Learning from comparisons and choices," *arXiv.org*, 2017. [Online]. Available: https://arxiv.org/abs/1704.07228

[6] S. Oh, K. K. Thekumparampil, and J. Xu, "Collaboratively learning preferences from ordinal data," in *Advances in Neural Information Processing Systems*, 2015, pp. 1900–1908.

[7] M. Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* l'Imprimerie Royale, 1785.

[8] D. McFadden, "Econometric models for probabilistic choice among products," *Journal of Business*, vol. 53, no. 3, pp. S13–S29, 1980.

[9] H. A. Soufiani, H. Diao, Z. Lai, and D. C. Parkes, "Generalized random utility models with multiple types," in *Advances in Neural Information Processing Systems*, 2013, pp. 73–81.

[10] K. T. Talluri and G. V. Ryzin, *The Theory and Practice of Revenue Management.* Springer, 2005.

[11] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[12] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[13] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[14] D. R. Hunter, "Mm algorithms for generalized bradley-terry models," *Annals of Statistics*, vol. 32, pp. 384–406, 2004.

[15] J. Guiver and E. Snelson, "Bayesian inference for plackett-luce ranking models," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 377–384.

[16] F. Caron and A. Doucet, "Efficient Bayesian inference for generalized Bradley–Terry models," *Journal of Computational and Graphical Statistics*, vol. 21, no. 1, pp. 174–196, 2012.

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

[18] R. Wu, J. Xu, R. Srikant, L. Massoulié, M. Lelarge, and B. Hajek, "Clustering and inference from pairwise comparisons," *arXiv.org*, 2015. [Online]. Available: https://arxiv.org/abs/1502.04631

[19] S. Oh and D. Shah, "Learning mixed multinomial logit model from ordinal data," in *Advances in Neural Information Processing Systems*, 2014, pp. 595–603.

[20] Y. Lu and S. N. Negahban, "Individualized rank aggregation using nuclear norm regularization," *arXiv.org*, 2014. [Online]. Available: https://arxiv.org/abs/1410.0860

[21] 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.

[22] A. Agarwal, S. Negahban, and M. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *Advances in Neural Information Processing Systems*, 2010, pp. 37–45.

[23] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA Journal of Numerical Analysis*, vol. 8, no. 1, pp. 141–148, 1988.

[24] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

[25] J. A. Tropp, "An introduction to matrix concentration inequalities," *arXiv.org*, 2015. [Online]. Available: https://arxiv.org/abs/1501.01571

[26] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[27] S. Van De Geer, *Empirical Processes in M-estimation.* Cambridge University Press, 2000.

[28] L. L. Thurstone, "A law of comparative judgment." *Psychological Review*, vol. 34, no. 4, p. 273, 1927.

[29] J. Marschak, "Binary-choice constraints and random utility indicators," in *Proceedings of a Symposium on Mathematical Methods in the Social Sciences*, vol. 7, 1960, pp. 19–38.

[30] D. R. Luce, *Individual Choice Behavior.* New York: Wiley, 1959.

[31] M. Ledoux, *The Concentration of Measure Phenomenon.* American Mathematical Soc., 2005.