

Individualized Rank Aggregation using Nuclear Norm Regularization

Yu Lu Sahand N. Negahban

Department of Statistics
Yale University

October 6, 2014

Technical Report
Department of Statistics, Yale University

Abstract

In recent years rank aggregation has received significant attention from the machine learning community. The goal of such a problem is to combine the (partially revealed) preferences over objects of a large population into a single, relatively consistent ordering of those objects. However, in many cases, we might not want a single ranking and instead opt for individual rankings. We study a version of the problem known as collaborative ranking. In this problem we assume that individual users provide us with pairwise preferences (for example purchasing one item over another). From those preferences we wish to obtain rankings on items that the users have not had an opportunity to explore. The results here have a very interesting connection to the standard matrix completion problem. We provide a theoretical justification for a nuclear norm regularized optimization procedure, and provide high-dimensional scaling results that show how the error in estimating user preferences behaves as the number of observations increase.

1 Introduction

We have seen a number of recent advancements to the theory of rank aggregation. This problem has a number of applications ranging from marketing and advertisements to competitions and election. The main question of rank aggregation is how to consistently combine various individual preferences. This type of data is frequently available to us: what webpage did a user select, who won the chess match, which movie did a user watch, etc.... All of these examples yield comparisons without explicitly revealing an underlying score. That is, only the preference is observed, not necessarily the strength of the preference (in the case of sports one might argue that the score indicates such a magnitude difference). Additionally, numeric scores have been shown to be inconsistent and subject to variations in calibration in various contexts. Given how natural the problem of rank aggregation is, there has been a wide body recent [9, 2] and classical work [3, 4, 7, 17] to understand how to consistently combine preferences. However, all of these methods have a major drawback: they aim to find *one* ranking. In many settings, various individuals will have separate preferences, and we wish to model those distinctions. For example, we might wish to provide personalized ads, search results, or movie recommendations on a per user basis. In standard contexts we assume that there is one consistent ranking that does well to approximate the behavior of all users, but these aggregation methods cannot model the discrepancies across users. Our goal is to understand how to analyze a method that has the flexibility to account for user differences and can be adaptive; that is, if there are no differences, then the method should have stronger performance guarantees. This task can be seen as rank aggregation analog to the standard collaborative filtering problem.

While there have been significant theoretical advances in the understanding of collaborative filtering, or more generally matrix completion [6, 13, 21], there has been far less work in understanding how to perform the proposed type of collaborative ranking. Recent work has demonstrated that taking rankings into consideration can significantly improve upon rating prediction accuracy [15, 28, 29, 30], thus it is a natural question to understand how such collaborative ranking methods might behave. One reason for this discrepancy is this theoretical understanding of single user rank aggregation is already

a very challenging problem as discussed above. Whereas, single rating aggregation is trivial: take an average. Another, possibly more interesting distinction is in the amount of apparent information made available. In the standard matrix completion setting we have direct (albeit noisy) access to the true underlying ratings. Therefore, if the noise is sufficiently small, we could order the information into a list. On the other hand, in the collaborative ranking problem we never have direct access to the true signal itself and only observe relative differences. In some sense, this is a harder problem [25] owing to the fact that the comparisons are in themselves functions of the underlying ratings. When we are given, for example, p ratings, then we can convert that to $\binom{p}{2}$ pairwise comparisons. This crude analysis seems to indicate that we would require far greater pairwise comparisons in order to recover the true underlying matrix. We will show that this increase in the number of examples is not required. In the sequel, we will show that under a natural choice model for collaborative ranking, the total number of comparisons needed to estimate the parameters is on the same order as the total number of explicit ratings observations required in the standard matrix completion literature. Thus, we demonstrate that collaborative ranking based pair-wise comparisons from a simple and natural model can yield very similar results as in the standard matrix completion setting.

Past Work As alluded to above there has been some work in understanding collaborative rankings and learning user preferences. The nuclear norm approach is fundamentally a regularized M -estimator [19]. The application of the nuclear norm approach to collaborative ranking was first proposed by Yi et al. [30]. Their work showed very good empirical evidence for using such a nuclear norm regularized based approach. However, that work left open the question of theoretical guarantees. Other results also assume that the underlying ratings are in fact available. However, rather than inferring unknown ratings their goal is to infer unknown ranked preferences *from* known ratings. That is, they wish to deduce if a user will prefer one item over another rather than guess what their ratings of that item might be [15, 29, 23]. The work by Weimer et. al. [29] also uses a nuclear norm regularization, but that work assumes access to the true underlying ratings, while we assume access only to pairwise preferences. Other algorithms aggregate users ratings by exploiting the similarity of users by nearest neighbor search [5, 28], low-rank matrix factorization [22, 23, 29], or probabilistic latent model [10, 16]. However, as noted, numeric ratings can be highly varied even when preferences are shared.

Pairwise preference based ranking methods can effectively address the limitations of rating based methods. Furthermore, numerical ratings can always be transformed into pairwise comparisons. Salimans et al. [24] use a bilinear model and do estimation in the Bayesian framework. Liu et al. [16] use the Bradley-Terry-Luce (BTL) Model. Rather than our low-rank setting, they characterize the similarity between different users by using a mixture model. Both methods are computationally inefficient. More important, all these methods fail to provide theoretical justifications of their algorithms.

There are some theoretical works for learning a single ranking list from pairwise comparisons. Work by Jamieson and Nowak [11] seeks to exploit comparisons to significantly reduce the number of samples required to obtain a good estimate of an individual’s utility function. Their method demonstrates that when the objects exist in a lower-dimensional space, then the number of queries required to learn the user’s utility significantly decreases. One drawback of their approach is that the authors must assume that descriptors or features for the underlying objects are provided; which is not necessarily the case in all contexts. Negahban et al. [18] propose the Rank Centrality algorithm and show rate optimal (up to log factors) error bounds of their algorithm under BTL model. They also provide theoretical analysis of penalized maximum likelihood estimator, which serves as an inspiration of our work.

Our contributions In this report, we present the first theoretical analysis of a collaborative ranking algorithm under a natural observation model. The algorithm itself is quite simple and falls into the framework of regularized M -estimator [19]. We provide finite sample guarantees that hold with high probability on recovering the underlying preference matrix. Furthermore, the techniques outlined in the proof section are general and can be applied to a variety of sampling operators for matrix completion. For example, a simple modification of our proof yields a different class of results for the “one-bit” matrix completion problem [8].

In the following we present an explicit description of our model in Section 2. In Section 3 we present the proposed estimation procedure that we wish to analyze. Finally, in Section 4 we provide a statement of the main theorem followed by experiments in Section 5. Finally, in Section 6 we present the proof.

Notation: For a positive integer n we will let $[n] = \{1, 2, \dots, n\}$ be the set of integers from 1 to n . For two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$ of commensurate dimensions, let $\langle\langle A, B \rangle\rangle = \text{trace}(A^T B)$ be the trace inner product. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ let $A_{i,j}$ denote the entry in the i^{th} row and j^{th} column of A . Take $\sigma_i(A)$ to be the i^{th} singular value of A where $\sigma_i(A) \geq \sigma_{i+1}(A)$. Let $\|A\|_2 = \sigma_1(A)$, $\|A\|_{\text{nuc}} = \sum_{j=1}^{\min(d_1, d_2)} \sigma_j(A)$ be the nuclear norm of A , i.e. the sum of the singular values of A , and $\|A\|_F = \sqrt{\langle\langle A, A \rangle\rangle} = \sqrt{\sum_{j=1}^{\min(d_1, d_2)} \sigma_j^2(A)}$ to be the Frobenius norm of A . Finally, we let $\|A\|_\infty = \max_{i,j} |A_{i,j}|$ to be the elementwise infinity norm of the matrix A .

2 Problem Statement and Model

In this section we provide a precise description of the underlying statistical model as well as our problem.

2.1 Data and Observation Model

Recall that each user provides a collection of pairwise preferences for various items. We assume that the data are the form $(X^{(i)}, y_i)$ where $X^{(i)} \in \mathbb{R}^{d_1 \times d_2}$. We assume that the i^{th} piece of data is a query to user $k(i)$ asking if she prefers item $l(i)$ to item $j(i)$. If she does, then $y_i = 1$, otherwise $y_i = 0$. In other words, $y_i = 1$ if user $k(i)$ prefers item $l(i)$ to item $j(i)$, otherwise $y_i = 0$. Let the underlying (unknown and unobservable) user preferences be encoded in the matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ such that $\Theta_{k,j}^*$ is the score that user k places on item j . We will also assume that $\|\Theta^*\|_F \leq 1$ to normalize the signal. For identifiability we assume that the sum of the rows of Θ^* is equal to zero. We must also assume that $\|\Theta^*\|_\infty \leq \frac{\alpha}{\sqrt{d_1 d_2}}$. Similar assumptions are made in the matrix completion literature and is known to control the “spikyness” of the matrix. Both of these assumptions are discussed in the sequel. For compactness in notation we let $X^{(i)} = \sqrt{d_1 d_2} e^{(k(i))} (e^{(l(i))} - e^{(j(i))})^T$ where $e^{(a)}$ is the standard basis vector that takes on the value 1 in the a^{th} entry and zeros everywhere else. Taking the trace inner product between Θ^* and $X^{(i)}$ yields

$$\langle\langle \Theta^*, X^{(i)} \rangle\rangle = \sqrt{d_1 d_2} (\Theta_{k(i), l(i)}^* - \Theta_{k(i), j(i)}^*)$$

and denotes the relative preference that user $k(i)$ has for item $l(i)$ versus $j(i)$. Our observation model takes the form

$$\mathbb{P}(y_i = 1 | l(i) = l, j(i) = j, k(i) = k) = \frac{\exp(\langle\langle \Theta^*, X^{(i)} \rangle\rangle)}{1 + \exp(\langle\langle \Theta^*, X^{(i)} \rangle\rangle)} \quad (1)$$

The above is the standard Bradley-Terry-Luce model for pairwise comparisons. In full generality, one can also consider the Thurstone models for pairwise preferences.

We shall take Θ^* to be low-rank or well approximate by a low-rank matrix. This is analogous to the matrix completion literature and models the fact that the underlying preferences are derived from latent low-dimensional factors. In this way, we can extract features on items and users without explicit domain knowledge.

Discussion of assumptions: In the above we assume that the ℓ_∞ norm of the matrix is bounded. This form of assumption is required for estimating the underlying parameters of the matrix and can be thought of as an incoherence requirement in order to ensure that the matrix itself is not orthogonal to the observation operator. For example, suppose that we have a matrix that is zeros everywhere except in one row where we have a single +1 and a single -1. In that case, we would never be able to recover those values from random samples without observing the entire matrix. Hence, the error bounds that

we derive will include some dependency on the infinity norm of the matrix. If generalization error bounds are the desired outcome, then such requirements can be relaxed at the expense of slower error convergence guarantees and no guarantees on individual parameter recovery. Also noted above is the requirement that the sum of each of the rows of Θ^* must be equal to 0. This assumption is natural owing to the fact that we can ever only observe the differences between the intrinsic item ratings. Hence, even if we could exactly observe all of those difference, the solution would not be unique up to linear offsets of each of the rows. We refer the reader to other work in matrix completion [6, 21] for a discussion of incoherence.

3 Estimation Procedure

We consider the following simple estimator for performing collaborating ranking. It is an example of a regularized M -estimator [19].

$$\hat{\Theta} = \operatorname{argmin}_{\Theta \in \Omega} \underbrace{\frac{1}{n} \sum_{i=1}^n \log(1 + \exp(\langle \Theta, X^{(i)} \rangle)) - y_i \langle \Theta, X^{(i)} \rangle + \lambda \|\Theta\|_{\text{nuc}}}_{\mathcal{L}_n(\Theta)}, \quad (2)$$

where $\mathcal{L}_n(\Theta)$ is the random loss function and

$$\Omega = \{A \in \mathbb{R}^{d_1 \times d_2} \mid \|A\|_{\infty} \leq \alpha, \text{ and } \forall j \in [d_1] \text{ we have } \sum_{k=1}^{d_2} A_{j,k} = 0\}$$

This method is a convex optimization procedure, and very much related to the matrix completion problems studied in the literature. A few things to note about the constraint set presented above. While in practice, we do not impose the ℓ_{∞} constraint, the theory requires us to impose the condition and an interesting line of work would be to remove such a constraint. A similar constraint appears in other matrix completion work [21]. As discussed above, the second condition is a fundamental one. It is required to guarantee identifiability in the problem even if infinite data were available.

The method itself has a very simple interpretation. The random loss function encourages the recovered parameters to match the observations. That is, if $y_i = 1$ then we expect that $\Theta_{k^{(i)}, l^{(i)}}^* > \Theta_{k^{(i)}, j^{(i)}}^*$. The second term is the nuclear norm and that encourages the underlying matrix Θ^* to be low-rank [6].

4 Main Results

In this section we present the main results of our paper, which demonstrates that we are able to recover the underlying parameters with very few total observations. The result is analogous to similar results presented for matrix completion [13, 12, 21],

Theorem 1. *Under the described sampling model, let $d = (d_1 + d_2)/2$, assume $n < d^2 \log d$, and take $\lambda \geq 32\sqrt{\frac{d \log d}{n}}$. Then, we have that the Frobenius norm of the error $\Delta = \hat{\Theta} - \Theta^*$ satisfies*

$$\|\Delta\|_F \leq c_1 \max\left(\alpha, \frac{1}{\psi(2\alpha)}\right) \max\left\{\sqrt{\frac{rd \log d}{n}}, \left(\sqrt{\frac{rd \log d}{n}} \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*)\right)^{1/2}\right\}$$

with probability at least $1 - \frac{2}{d^2}$ for some universal constant c_1 .

The above result demonstrates that we can obtain consistent estimates of the parameters Θ^* using the convex program outlined in the previous section. Furthermore, the error bound behaves as a parametric error rate, that is the error decays as $\frac{1}{n}$. The result also decomposes into two terms. The

first is the penalty for estimating a rank r matrix and the second is the price we pay for estimating an approximately low-rank matrix Θ^* with a rank r matrix. These results exactly match analogous results in the matrix completion literature barring one difference: there is also a dependency on the function ψ . However, this necessity is quite natural since if we are interested in parameter recovery, then it would be impossible to distinguish between extremely large parameters. Indeed, this observation is related to the problem of trying to measure the probability of a coin coming up heads when that probability is extremely close to one. Other results in matrix completion also discuss such a requirement as well as the influence of the spikyness parameter [13, 8]. The proof of this result, for which we provide an outline in Section 6, follows similar lines as other results for matrix completion.

5 Experiments

Here we present simulation results to demonstrate the accuracy of the error rate behavior predicted by Theorem 1. To make the results more clean, we consider the exact low rank case here, which means each individual user’s preference vector is the linear combination of r preference vectors. Then according to our main results, the empirical squared Frobenius norm error $\|\hat{\Theta} - \Theta^*\|_F^2$ under our estimation procedure (2) will be scaled as $\frac{rd \log d}{n}$. For all the experiments, we solved the convex program (2) by using proximal gradient descent with step-sizes from [1] for fast convergence via our own implementation in R.

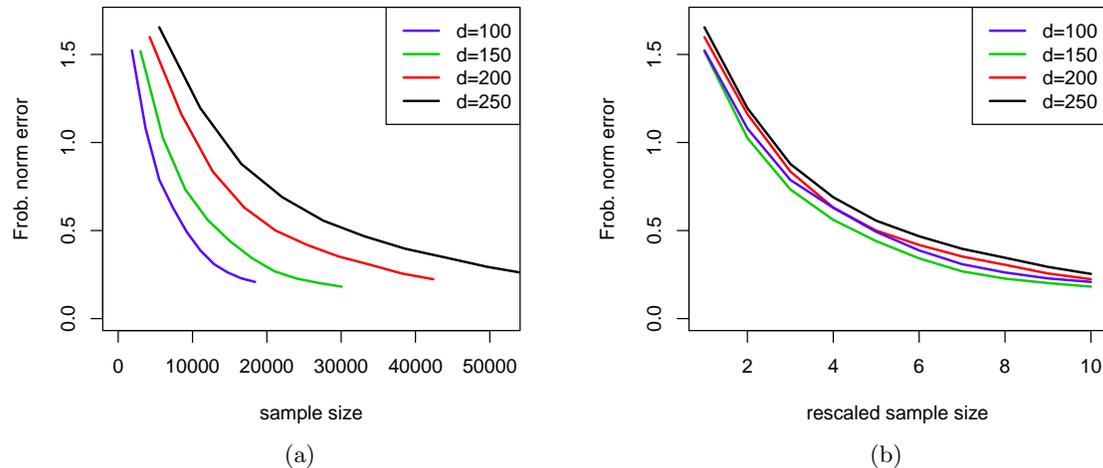


Figure 1: Plots of squared Frobenius norm error $\|\hat{\Theta} - \Theta^*\|_F^2$ when applying estimation procedure (2) on the exact low rank matrix. Each curve corresponds to a different problem size $d_1 = d_2 = d \in \{100, 150, 200, 250\}$ with a fixed rank $r = 4$. (a) Plots of Frobenius norm error against the raw sample size. As sample size increases, the error goes to zero. (b) Plots of the same Frobenius norm error against rescaled sample size $n/(rd \log d)$, all plots are aligned fairly well as expected by our theory.

In Figure 1 we report the results of four different problem sizes with equal user size d_1 and item size d_2 and the fixed rank r , where $d_1 = d_2 = d \in \{100, 150, 200, 250\}$, $r = 4$. For a given sample size d , we ran $T = 10$ trials and computed the squared Frobenius norm error $\|\hat{\Theta} - \Theta^*\|_F^2$ averaged over those trials. Panel (a) shows the plots of Frobenius norm error versus raw sample size. It shows the consistency of our estimation procedure because the Frobenius norm error goes to zero as sample size increases. And the curves shift to right as the problem dimension d increases, matching with the intuition that larger matrices require more samples. In panel (b), we plot the simulation results versus the rescaled sample size $N = n/(rd \log d)$. Consistent with the prediction of Theorem 1, the error plots are aligned fairly well and decay at the rate of $1/N$.

6 Proof of Main Result

We now present a proof of the main result. We will use the machinery developed by Negahban and Wainwright [21] and establish a Restricted Strong Convexity (RSC) for our loss. The proof follows standard techniques, with some care when handling the new observation operator.

6.1 Proof of Theorem 1

The key to establishing the RSC condition is to demonstrate that the error in the first order Taylor approximation of the loss is lower-bounded by some quadratic function. To that end we note that for $\Delta = \Theta - \Theta^*$ and by the Taylor expansion we have that

$$\mathcal{L}_n(\Theta) - \mathcal{L}_n(\Theta^*) - \langle \nabla \mathcal{L}_n(\Theta^*), \Delta \rangle = \frac{1}{2n} \sum_{i=1}^n \psi \left(\langle \Theta^*, X^{(i)} \rangle + s \langle \Delta, X^{(i)} \rangle \right) \left(\langle \Delta, X^{(i)} \rangle \right)^2, \quad (3)$$

where $s \in [0, 1]$ and

$$\psi(x) = \frac{\exp(x)}{(1 + \exp(x))^2}.$$

Now, we may apply the fact that both $\|\widehat{\Theta}\|_\infty, \|\Theta^*\|_\infty \leq \alpha/\sqrt{d_1 d_2}$ and that $\psi(x)$ is symmetric and decreases as x increases to obtain that equation (3) is lower-bounded by:

$$\frac{1}{2n} \sum_{i=1}^n \psi(2\alpha) \left(\langle \Delta, X^{(i)} \rangle \right)^2 \quad (4)$$

Therefore, it suffices to prove a lower-bound on $\frac{1}{2n} \left(\langle \Delta, X^{(i)} \rangle \right)^2$ for all possible vectors Δ . For that, we present the following lemma.

Lemma 1. *For $\|\Theta\|_\infty \leq r_3 := \frac{2\alpha}{\sqrt{d_1 d_2}}$, $d = (d_1 + d_2)/2$, and $n < d^2 \log d$. When $X^{(i)}$ are i.i.d observations we have with probability greater than $1 - 2d^{-2^{18}}$*

$$\frac{1}{n} \sum_{i=1}^n \left(\langle \Theta, X^{(i)} \rangle \right)^2 \geq \frac{1}{3} \|\Theta\|_F^2 \quad \text{for all } \Theta \text{ in } \mathcal{A}$$

where

$$\mathcal{A} = \left\{ \Theta \in \mathbb{R}^{d_1 \times d_2} \mid \|\Theta\|_\infty \leq r_3, \|\Theta\|_F^2 \geq 128\alpha \sqrt{\frac{d \log d}{n}} \|\Theta\|_{\text{nuc}} \text{ and } \forall j \in [d_1] \text{ we have } \sum_{k=1}^{d_2} \Theta_{j,k} = 0 \right\}$$

Another key element for establishing the error is the following upper-bound on the operator norm of a random matrix.

Lemma 2. *Consider the sampling model described above. Then for i.i.d. $(\xi_i, X^{(i)})$, where $|\xi_i| \leq \gamma$ and $\mathbb{E}[\xi_i | X^{(i)}] = 0$ we have that*

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_{i=1}^n \xi_i X^{(i)} \right\|_2 > 8\gamma \sqrt{\frac{d \log d}{n}} \right) \leq \frac{2}{d^2},$$

We these two ingredients in hand we may now prove the main result. The steps are a slight modification of the ones taken for standard matrix completion [20]. By the optimality of $\widehat{\Theta}$ we have

$$\mathcal{L}_n(\widehat{\Theta}) + \lambda \|\widehat{\Theta}\|_{\text{nuc}} \leq \mathcal{L}_n(\Theta^*) + \lambda \|\Theta^*\|_{\text{nuc}}$$

Let $\Delta = \widehat{\Theta} - \Theta^*$, then

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

By Taylor expansion, the left hand side is lower bounded by

$$\mathcal{L}_n(\widehat{\Theta}) - \mathcal{L}_n(\Theta^*) - \langle \nabla \mathcal{L}_n(\Theta^*), \Delta \rangle \geq \psi(2\alpha) \frac{1}{2n} \sum_{i=1}^n \left(\langle \Theta, X^{(i)} \rangle \right)^2$$

Hölder's inequality between the nuclear norm and operator norm yields

$$-\langle \nabla \mathcal{L}_n(\Theta^*), \Delta \rangle \leq \|\nabla \mathcal{L}_n(\Theta^*)\|_2 \|\Delta\|_{\text{nuc}}$$

By the triangle inequality $\|\Theta^*\|_{\text{nuc}} - \|\widehat{\Theta}\|_{\text{nuc}} \leq \|\Delta\|_{\text{nuc}}$. If we choose $\lambda > 2\|\nabla \mathcal{L}_n(\Theta^*)\|_2$, we have

$$\mathcal{L}_n(\widehat{\Theta}) - \mathcal{L}_n(\Theta^*) - \langle \nabla \mathcal{L}_n(\Theta^*), \Delta \rangle \leq 2\lambda \|\Delta\|_{\text{nuc}}$$

Now, the random matrix $\nabla \mathcal{L}_n(\Theta^*) = \frac{1}{n} \sum_{i=1}^n \left(\frac{\exp(\langle X^{(i)}, \Delta \rangle)}{1 + \exp(\langle X^{(i)}, \Delta \rangle)} - y_i \right) X^{(i)}$ and satisfies the conditions of Lemma 2 with $\gamma = 2$, so we can take $\lambda = 32\sqrt{\frac{d \log d}{n}}$

From Lemma 1 of Negahban and Wainwright [20], Δ can be decomposed into $\Delta' + \Delta''$, where Δ' has rank less than $2r$ and Δ'' satisfies

$$\|\Delta''\|_{\text{nuc}} \leq 3\|\Delta'\|_{\text{nuc}} + 4 \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*)$$

Then by the triangle inequality and $\|\Delta'\|_{\text{nuc}} \leq \sqrt{2r}\|\Delta'\|_F$

$$\|\Delta\|_{\text{nuc}} \leq 4\|\Delta'\|_{\text{nuc}} + 4 \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) \leq 4\sqrt{2r}\|\Delta\|_F + 4 \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) \quad (5)$$

Now depending on whether Δ belongs to set \mathcal{A} , we split into two cases.

Case 1: When $\Delta \notin \mathcal{A}$, $\|\Delta\|_F^2 \leq 128\alpha \|\Delta\|_{\text{nuc}} \sqrt{\frac{d \log d}{n}}$. From Equation (5), we get

$$\|\Delta\|_F \leq \alpha \max \left\{ 1024 \sqrt{\frac{rd \log d}{n}}, \left(512 \sqrt{\frac{rd \log d}{n}} \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) \right)^{1/2} \right\}$$

Case 2: Otherwise, from Lemma 1, with probability greater than $1 - 2d^{-2^{18}}$, $\mathcal{L}_n(\widehat{\Theta}) - \mathcal{L}_n(\Theta^*) - \langle \nabla \mathcal{L}_n(\Theta^*), \Delta \rangle \geq \frac{\psi(2\alpha)}{3} \|\Delta\|_F^2$. Therefore, the above equations yield

$$\|\Delta\|_F^2 \leq \frac{192}{\psi(2\alpha)} \sqrt{\frac{2rd \log d}{n}} \|\Delta\|_{\text{nuc}}.$$

Now, performing similar calculations as above we have

$$\|\Delta\|_F \leq \frac{1}{\psi(2\alpha)} \max \left\{ 1024 \sqrt{\frac{rd \log d}{n}}, \left(512 \sqrt{\frac{rd \log d}{n}} \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*) \right)^{1/2} \right\}.$$

Combining the two displays above yields the desired result.

6.2 Proof of Lemma 1

We use a peeling argument [27] as in Lemma 3 of [21] to prove Lemma 1. Before that, we first present the following lemma.

Lemma 3. Define the set

$$\mathcal{B}(D) = \left\{ \Theta \in \mathbb{R}^{d_1 \times d_2} \mid \|\Theta\|_\infty \leq r_3, \|\Theta\|_F \leq D, \|\Theta\|_{\text{nuc}} \leq \frac{D^2}{128\alpha} \sqrt{\frac{n}{d \log d}} \right\}$$

and

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

Then

$$\mathbb{P} \left\{ M(D) \geq \frac{3}{2} D^2 \right\} \leq \exp\left\{ -\frac{nD^4}{128\alpha^4} \right\}$$

Since for any $\Theta \in \mathcal{A}$,

$$\|\Theta\|_F^2 \geq 128\alpha \sqrt{\frac{d \log d}{n}} \|\Theta\|_{\text{nuc}} \geq 128\alpha \sqrt{\frac{d \log d}{n}} \|\Theta\|_F$$

then we have $\|\Theta\|_F \geq 128\alpha \sqrt{\frac{d \log d}{n}} := \mu$. Consider the sets

$$\mathcal{S}_\ell = \left\{ \Theta \in \mathbb{R}^{d_1 \times d_2} \mid \|\Theta\|_\infty \leq r_3, \beta^{\ell-1} \mu \leq \|\Theta\|_F \leq \beta^\ell \mu, \|\Theta\|_{\text{nuc}} \leq \frac{D^2}{128\alpha} \sqrt{\frac{n}{d \log d}} \right\}$$

where $\beta = \sqrt{\frac{10}{9}}$ and $\ell = 1, 2, 3, \dots$

Suppose there exists $\Theta \in \mathcal{A}$ such that $\frac{1}{n} \sum_{i=1}^n \left(\langle \Theta, X^{(i)} \rangle \right)^2 < \frac{1}{3} \|\Theta\|_F^2$. Since $\mathcal{A} \subseteq \bigcup_{\ell=1}^{\infty} \mathcal{S}_\ell \subseteq \bigcup_{\ell=1}^{\infty} \mathcal{B}(\beta^\ell \mu)$, there is some ℓ such that $\Theta \in \mathcal{B}(\beta^\ell \mu)$ and

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

Then by union bound, we have

$$\begin{aligned} & \mathbb{P} \left\{ \exists \Theta \in \mathcal{A}, \frac{1}{n} \sum_{i=1}^n \left(\langle \Theta, X^{(i)} \rangle \right)^2 < \frac{1}{3} \|\Theta\|_F^2 \right\} \\ & \leq \sum_{\ell=1}^{\infty} \mathbb{P} \left\{ M(\beta^\ell \mu) > \frac{3}{2} (\beta^\ell \mu)^2 \right\} \\ & \leq \sum_{\ell=1}^{\infty} \exp\left\{ -\frac{n(\beta^\ell \mu)^4}{128\alpha^4} \right\} \\ & \leq \sum_{\ell=1}^{\infty} \exp\left\{ -\frac{4\ell(\beta-1)n\mu^4}{128\alpha^4} \right\} \\ & \leq 2 \exp\left\{ -\frac{4(\beta-1)n\mu^4}{128\alpha^4} \right\} \\ & \leq 2 \exp\{-2^{18} \log d\} \end{aligned}$$

where the second inequality is Lemma 3, the third inequality is $\beta^\ell \geq \ell(\beta-1)$ and we use the fact that $n < d^2 \log d$ for the last inequality.

6.3 Proof of Lemma 3

Define

$$Z =: \frac{1}{d_1 d_2} M(D) = \sup_{\Theta \in \mathcal{B}(D)} \frac{1}{n} \sum_{i=1}^n \left[\mathbb{E} \left(\Theta_{k(i)l(i)} - \Theta_{k(i)j(i)} \right)^2 - \left(\Theta_{k(i)l(i)} - \Theta_{k(i)j(i)} \right)^2 \right]$$

Our goal will be to first show that Z concentrates around its mean and then upper bound the expectation. We prove the concentration results via the bounded differences inequality [14]; since Z is a symmetric function of its arguments, it suffices to establish the bounded differences property with respect to the first coordinate. Suppose we have two samples of $(k(i), l(i), j(i))_{i=1}^n$ that only differ at the first coordinate.

$$\begin{aligned} Z - Z' &\leq \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \left[\frac{1}{n} \sum_{i=1}^n (\Theta_{k'(i)l'(i)} - \Theta_{k'(i)j'(i)})^2 - \frac{1}{n} \sum_{i=1}^n (\Theta_{k(i)l(i)} - \Theta_{k(i)j(i)})^2 \right] \\ &= \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \frac{1}{n} \left((\Theta_{k'(1)l'(1)} - \Theta_{k'(1)j'(1)})^2 - (\Theta_{k(1)l(1)} - \Theta_{k(1)j(1)})^2 \right) \\ &\leq \frac{4r_3^2}{n} \end{aligned}$$

Then by the bounded differences inequality, we have

$$\mathbb{P}\{Z - \mathbb{E}Z \geq t\} \leq \exp\left\{-\frac{nt^2}{32r_3^4}\right\} \quad (6)$$

In order to upper bound $\mathbb{E}Z$, we use a standard symmetrization argument.

$$\begin{aligned} \mathbb{E}Z &= \mathbb{E} \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \frac{1}{n} \sum_{i=1}^n \left[\mathbb{E} \left(\Theta_{a(i)l(i)} - \Theta_{a(i)j(i)} \right)^2 - \left(\Theta_{a(i)l(i)} - \Theta_{a(i)j(i)} \right)^2 \right] \\ &\leq \mathbb{E} \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \frac{2}{n} \sum_{i=1}^n \varepsilon_i \left(\Theta_{a(i)l(i)} - \Theta_{a(i)j(i)} \right)^2 \\ &= \mathbb{E} \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \frac{2}{n} \sum_{i=1}^n \varepsilon_i \langle e_{k(i)}(e_{l(i)} - e_{j(i)})^T, \Theta \rangle^2 \end{aligned}$$

where ε_i are i.i.d. Rademacher random variables. Since $|\Theta_{a(i)l(i)} - \Theta_{a(i)j(i)}| \leq 2r_3$, we have by the Ledoux-Talagrand contraction inequality that

$$\mathbb{E} \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle e_{k(i)}(e_{l(i)} - e_{j(i)})^T, \Theta \rangle^2 \leq 4r_3 \mathbb{E} \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle e_{k(i)}(e_{l(i)} - e_{j(i)})^T, \Theta \rangle$$

By an application of Hölder's inequality we have that

$$\left| \sum_{i=1}^n \varepsilon_i \langle e_{k(i)}(e_{l(i)} - e_{j(i)})^T, \Theta \rangle \right| \leq \left\| \sum_{i=1}^n \varepsilon_i e_{k(i)}(e_{l(i)} - e_{j(i)})^T \right\|_2 \|\Theta\|_{\text{nuc}} \quad (7)$$

Let $W_i := \varepsilon_i e_{k(i)}(e_{l(i)} - e_{j(i)})^T$. W_i is a zero-mean random matrix, and since

$$\mathbb{E}[W_i W_i^T] = \mathbb{E}[e_{k(i)}(e_{l(i)} - e_{j(i)})^T (e_{l(i)} - e_{j(i)}) e_{k(i)}^T] = \left(2 - \frac{2}{d_2}\right) \frac{1}{d_1} \mathbf{I}_{d_1 \times d_1}$$

and

$$\mathbb{E}[W_i^T W_i] = \mathbb{E}[(e_{l(i)} - e_{j(i)}) e_{k(i)}^T e_{k(i)}(e_{l(i)} - e_{j(i)})^T] = \frac{2}{d_2} \mathbf{I}_{d_2 \times d_2} - \frac{2}{d_2^2} \mathbf{1}\mathbf{1}^T$$

we have

$$\sigma_i^2 = \max\{\|\mathbb{E}[W_i^T W_i]\|_2, \|\mathbb{E}[W_i W_i^T]\|_2\} \leq \max\left\{\frac{2}{d_2}, \left(2 - \frac{2}{d_2}\right) \frac{1}{d_1}\right\} \leq \frac{2}{\min\{d_1, d_2\}}$$

Notice $\|W_i\|_2 \leq 2$, thus, Lemma 4 yields the tail bound

$$\mathbb{P}\left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i e_{k(i)}(e_{l(i)} - e_{j(i)})^T \right\|_2 \geq t\right] \leq d_1 d_2 \max\left\{\exp\left(-\frac{nt^2 \min\{d_1, d_2\}}{8}\right), \exp\left(-\frac{nt}{4}\right)\right\} \quad (8)$$

Set $t = \sqrt{\frac{16 \log d_1 d_2}{n \min\{d_1, d_2\}}}$, we obtain with probability greater than $1 - \frac{1}{d_1 d_2}$,

$$\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i e_{k(i)} (e_{l(i)} - e_{j(i)})^T \right\|_2 \leq \sqrt{\frac{16 \log d_1 d_2}{n \min\{d_1, d_2\}}}$$

By the triangle inequality, $\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i e_{k(i)} (e_{l(i)} - e_{j(i)})^T \right\|_2 \leq \left\| \varepsilon_i e_{k(i)} (e_{l(i)} - e_{j(i)})^T \right\|_2 \leq 2$ and the fact $n \leq d^2 \log d$

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i e_{k(i)} (e_{l(i)} - e_{j(i)})^T \right\|_2 \leq \sqrt{\frac{16 \log d_1 d_2}{n \min\{d_1, d_2\}}} + \frac{2}{d_1 d_2} \leq 8 \sqrt{\frac{\log d_1 d_2}{n \min\{d_1, d_2\}}} \quad (9)$$

Putting those bounds together we have

$$\mathbb{E} \sup_{\Theta \in \mathcal{B}(\mathcal{D})} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(\Theta_{a(i)l(i)} - \Theta_{a(i)j(i)} \right)^2 \leq \sup_{\Theta \in \mathcal{B}(\mathcal{D})} 32r_3 \|\Theta\|_{\text{nuc}} \sqrt{\frac{\log d_1 d_2}{n \min\{d_1, d_2\}}} \leq \frac{D^2}{d_1 d_2}$$

Plug it into (6) and set $t = \frac{D^2}{2d_1 d_2}$, we get the result.

6.4 Ahlswede-Winter Matrix Bound

As in previous work [21] we also use a version of the Ahlswede-Winter concentration bound. We use a version due to Tropp [26].

Lemma 4 (Theorem 1.6 [26]). *Let W_i be independent $d_1 \times d_2$ zero-mean random matrices such that $\|W_i\|_2 \leq M$, and define*

$$\sigma_i^2 := \max\{\|\mathbb{E}[W_i^T W_i]\|_2, \|\mathbb{E}[W_i W_i^T]\|_2\}$$

as well as $\sigma^2 := \sum_{i=1}^n \sigma_i^2$. We have

$$\mathbb{P} \left[\left\| \sum_{i=1}^n W_i \right\|_2 \geq t \right] \leq (d_1 + d_2) \max\left\{ \exp\left(-\frac{t^2}{4\sigma^2}\right), \exp\left(-\frac{t}{2M}\right) \right\} \quad (10)$$

7 Discussion

In this paper we presented a theoretical justification for a ranking based collaborative filtering approach based on pairwise comparisons in contrast to other results that rely on knowing the underlying ratings. We provided the first convergence bounds for recovering the underlying user preferences of items and showed that those bounds are analogous to the ones originally developed for rating based matrix completion. The analysis here can also be extended to other observation models, for example to the ‘‘one-bit’’ matrix completion setting as well. However, that extension does not provide any additional insights beyond the analysis presented here. There remain a number of extensions for these methods including adaptive and active recommendations, skewed sampling distributions on the items, as well as different choice models. We leave such extensions for future work.

References

- [1] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.
- [2] A. Ammar and D. Shah. Ranking: Compare, don’t score. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 776–783. IEEE, 2011.
- [3] K. J. Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.

- [4] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, pages 324–345, 1952.
- [5] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.
- [6] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
- [7] M. D. Condorcet. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, 1785.
- [8] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2012.
- [9] J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 327–334, 2010.
- [10] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115, 2004.
- [11] K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2240–2248, 2011.
- [12] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.
- [13] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.
- [14] M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, Providence, RI, 2001.
- [15] N. N. Liu and Q. Yang. Eigenrank: a ranking-oriented approach to collaborative filtering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90. ACM, 2008.
- [16] N. N. Liu, M. Zhao, and Q. Yang. Probabilistic latent preference analysis for collaborative filtering. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 759–766. ACM, 2009.
- [17] R. D. Luce. *Individual choice behavior: A theoretical analysis*. Courier Dover Publications, 2005.
- [18] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2012.
- [19] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.
- [20] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.
- [21] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.
- [22] J. D. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

- [23] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008.
- [24] T. Salimans, U. Paquet, and T. Graepel. Collaborative learning of preference rankings. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 261–264. ACM, 2012.
- [25] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. When is it better to compare than to score? Technical report, UC Berkeley, July 2014. Preprint available at arXiv:1406.6618.
- [26] J. A. Tropp. User-friendly tail bounds for matrix martingales. Technical report, DTIC Document, 2011.
- [27] S. A. van de Geer. *Empirical Process Theory in M-estimation*. Cambridge University Press, 2000.
- [28] M. Volkovs and R. S. Zemel. Collaborative ranking with 17 parameters. In *Advances in Neural Information Processing Systems*, pages 2294–2302, 2012.
- [29] M. Weimer, A. Karatzoglou, Q. V. Le, and A. Smola. Maximum margin matrix factorization for collaborative ranking. *Advances in neural information processing systems*, 2007.
- [30] J. Yi, R. Jin, S. Jain, and A. Jain. Inferring users’ preferences from crowdsourced pairwise comparisons: A matrix completion approach. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.