The Annals of Statistics
2019, Vol. 47, No. 4, 2204–2235
https://doi.org/10.1214/18-AOS1745

SPECTRAL METHOD AND REGULARIZED MLE ARE BOTH OPTIMAL FOR TOP-K RANKING[1]

BY YUXIN CHEN[*,2], JIANQING FAN[†,*,3], CONG MA[*] AND KAIZHENG WANG[*]

Princeton University[*] and Fudan University[†]

This paper is concerned with the problem of top-K ranking from pairwise comparisons. Given a collection of n items and a few pairwise comparisons across them, one wishes to identify the set of K items that receive the highest ranks. To tackle this problem, we adopt the logistic parametric model—the Bradley–Terry–Luce model, where each item is assigned a latent preference score, and where the outcome of each pairwise comparison depends solely on the relative scores of the two items involved. Recent works have made significant progress toward characterizing the performance (e.g., the mean square error for estimating the scores) of several classical methods, including the spectral method and the maximum likelihood estimator (MLE). However, where they stand regarding top-K ranking remains unsettled.

We demonstrate that under a natural random sampling model, the spectral method alone, or the regularized MLE alone, is minimax optimal in terms of the sample complexity—the number of paired comparisons needed to ensure exact top-K identification, for the fixed dynamic range regime. This is accomplished via optimal control of the entrywise error of the score estimates. We complement our theoretical studies by numerical experiments, confirming that both methods yield low entrywise errors for estimating the underlying scores. Our theory is established via a novel leave-one-out trick, which proves effective for analyzing both iterative and noniterative procedures. Along the way, we derive an elementary eigenvector perturbation bound for probability transition matrices, which parallels the Davis–Kahan $\sin\Theta$ theorem for symmetric matrices. This also allows us to close the gap between the ℓ_2 error upper bound for the spectral method and the minimax lower limit.

1. Introduction. Imagine we have a large collection of n items, and we are given partially revealed comparisons between pairs of items. These paired comparisons are collected in a nonadaptive fashion, and could be highly noisy and incomplete. The aim is to aggregate these partial preferences so as to identify the K items that receive the highest ranks. This problem, which is called *top-K*

Received August 2017; revised July 2018.

[1]Author names are sorted alphabetically.

[2]Supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ARO Grant W911NF-18-1-0303, by the ONR Grant N00014-19-1-2120, and by the Princeton SEAS innovation award.

[3]Supported in part by NSF Grants DMS-1662139 and DMS-1712591, ONR Grant N00014-19-1-2120, and NIH Grant 2R01-GM072611-13.

MSC2010 subject classifications. Primary 62F07; secondary 62B10.

Key words and phrases. Top-K ranking, pairwise comparisons, spectral method, regularized MLE, entrywise perturbation, leave-one-out analysis, reversible Markov chains.

rank aggregation, finds applications in numerous contexts, including web search [Dwork et al. (2001)], recommendation systems [Baltrunas, Makcinskas and Ricci (2010)], sports competition [Masse (1997)], to name just a few. The challenge is both statistical and computational: how can one achieve reliable top-K ranking from a minimal number of pairwise comparisons, while retaining computational efficiency?

1.1. *Popular approaches.* To address the aforementioned challenge, many prior approaches have been put forward based on certain statistical models. Arguably one of the most widely used parametric models is the Bradley–Terry–Luce (BTL) model [Bradley and Terry (1952), Luce (1959)], which assigns a latent preference score $\{w_i^*\}_{1\le i\le n}$ to each of the n items. The BTL model posits that: the chance of each item winning a paired comparison is determined by the relative scores of the two items involved, or more precisely,

$$\mathbb{P}\{\text{item } j \text{ is preferred over item } i\} = \frac{w_j^*}{w_i^* + w_j^*} \tag{1.1}$$

in each comparison of item i against item j. The items are repeatedly compared in pairs according to this parametric model. The task then boils down to identifying the K items with the highest preference scores, given these pairwise comparisons.

Among the ranking algorithms tailored to the BTL model, the following two procedures have received particular attention, both of which rank the items based on appropriate estimates of the latent preference scores.

(1) *Spectral method.* By connecting the winning probability in (1.1) with the transition probability of a reversible Markov chain, the spectral method attempts recovery of $\{w_i^*\}$ via the leading left eigenvector of a sample transition matrix. This procedure, also known as Rank Centrality [Negahban, Oh and Shah (2017)], bears similarity to the PageRank algorithm.

(2) *Maximum likelihood estimator (MLE).* This approach proceeds by finding the score assignment that maximizes the likelihood function [Ford (1957)]. When parameterized appropriately, solving the MLE becomes a convex program, and hence is computationally feasible. There are also important variants of the MLE that enforce additional regularization.

Details are postponed to Section 2.2. In addition to their remarkable practical applicability, these two ranking paradigms are appealing in theory as well. For instance, both of them provably achieve intriguing ℓ_2 *accuracy* when estimating the latent preference scores [Negahban, Oh and Shah (2017)].

Nevertheless, the ℓ_2 error for estimating the latent scores merely serves as a "meta-metric" for the ranking task, which does not necessarily reveal the accuracy of top-K identification. In fact, given that the ℓ_2 loss only reflects the estimation error in some average sense, it is certainly possible that an algorithm obtains minimal

ℓ_2 estimation loss but incurs (relatively) large errors when estimating the scores of the highest ranked items. Interestingly, a recent work Chen and Suh (2015) demonstrates that: a careful combination of the spectral method and the coordinate-wise MLE is optimal for top-K ranking. This leaves open the following natural questions: *where does the spectral alone, or the MLE alone, stand in top-K ranking? Are they capable of attaining exact top-K recovery from minimal samples?* These questions form the primary objectives of our study.

As we will elaborate later, the spectral method part of the preceding questions was recently explored by Jang, Kim, Suh and Oh (2016), for a regime where a relatively large fraction of item pairs have been compared. However, it remains unclear how well the spectral method can perform in a much broader —and often much more challenging—regime, where the fraction of item pairs being compared may be vanishingly small. Additionally, the ranking accuracy of the MLE (and its variants) remains unknown.

1.2. *Main contributions*. The central focal point of the current paper is to assess the accuracy of both the spectral method and the regularized MLE in top-K identification. Assuming that the pairs of items being compared are randomly selected and that the preference scores fall within a *fixed* dynamic range, our paper delivers a somewhat surprising message:

> *Both the spectral method and the regularized MLE achieve perfect identification of top-K ranked items under optimal sample complexity (up to some constant factor)!*

It is worth emphasizing that these two algorithms succeed even under the sparsest possible regime, a scenario where only an exceedingly small fraction of pairs of items have been compared. This calls for precise control of the entrywise error—as opposed to the ℓ_2 loss—for estimating the scores. To this end, our theory is established upon a novel *leave-one-out argument*, which might shed light on how to analyze the entrywise error for more general optimization problems.

As a byproduct of the analysis, we derive an elementary eigenvector perturbation bound for (asymmetric) probability transition matrices, which parallels Davis–Kahan's $\sin\Theta$ theorem for symmetric matrices. This simple perturbation bound immediately leads to an improved ℓ_2 error bound for the spectral method, which allows to close the gap between the theoretical performance of the spectral method and the minimax lower limit.

1.3. *Notation*. Before proceeding, we introduce a some notation that will be useful throughout. To begin with, for any strictly positive probability vector $\boldsymbol{\pi} \in \mathbb{R}^n$, we define the inner product space indexed by $\boldsymbol{\pi}$ as a vector space in $\mathbb{R}^n$ endowed with the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{\pi}} = \sum_{i=1}^n \pi_i x_i y_i$. The corresponding vector norm and the induced matrix norm are defined respectively as $\|\boldsymbol{x}\|_{\boldsymbol{\pi}} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{\pi}}}$ and $\|\boldsymbol{A}\|_{\boldsymbol{\pi}} = \sup_{\|\boldsymbol{x}\|_{\boldsymbol{\pi}}=1} \|\boldsymbol{x}^\top \boldsymbol{A}\|_{\boldsymbol{\pi}}$.

Additionally, the notation $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ means there is a constant $c > 0$ such that $|f(n)| \le c|g(n)|$, $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ means there is a constant $c > 0$ such that $|f(n)| \ge c|g(n)|$, $f(n) = \Theta(g(n))$ or $f(n) \asymp g(n)$ means that there exist constants $c_1, c_2 > 0$ such that $c_1|g(n)| \le |f(n)| \le c_2|g(n)|$, and $f(n) = o(g(n))$ means $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

Given a graph $\mathcal{G}$ with vertex set $\{1, 2, \ldots, n\}$ and edge set $\mathcal{E}$, we denote by $\boldsymbol{L}_{\mathcal{G}} = \sum_{(i,j)\in\mathcal{E}, i>j} (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top$ the (unnormalized) Laplacian matrix [Chung (1997)] associated with it, where $\{\boldsymbol{e}_i\}_{1\le i\le n}$ are the standard basis vectors in $\mathbb{R}^n$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with n real eigenvalues, we let $\lambda_1(\boldsymbol{A}) \ge \lambda_2(\boldsymbol{A}) \ge \cdots \ge \lambda_n(\boldsymbol{A})$ be the eigenvalues sorted in descending order.

2. Statistical models and main results.

2.1. *Problem setup*. We begin with a formal introduction of the Bradley–Terry–Luce parametric model for binary comparisons.

Preference scores. As introduced earlier, we assume the existence of a positive latent score vector

$$\boldsymbol{w}^* = \big[w_1^*, \ldots, w_n^*\big]^\top \tag{2.1}$$

that comprises the underlying preference scores $\{w_i^* > 0\}_{1\le i\le n}$ assigned to each of the n items. Alternatively, it is sometimes more convenient to reparameterize the score vector by

$$\boldsymbol{\theta}^* = \big[\theta_1^*, \ldots, \theta_n^*\big]^\top \qquad \text{where } \theta_i^* = \log w_i^*. \tag{2.2}$$

These scores are assumed to fall within a *dynamic range* given by

$$w_i^* \in [w_{\min}, w_{\max}], \quad \text{or} \quad \theta_i^* \in [\theta_{\min}, \theta_{\max}] \tag{2.3}$$

for all $1 \le i \le n$ and for some $w_{\min} > 0$, $w_{\max} > 0$, $\theta_{\min} = \log w_{\min}$ and $\theta_{\max} = \log w_{\max}$. We also introduce the *condition number* as

$$\kappa := w_{\max}/w_{\min}. \tag{2.4}$$

Notably, the current paper primarily focuses on the case with a *fixed* dynamic range (i.e., κ is a fixed constant independent of n), although we will also discuss extensions to the large dynamic range regime in Section 3. Without loss of generality, it is assumed that

$$w_{\max} \ge w_1^* \ge w_2^* \ge \cdots \ge w_n^* \ge w_{\min}, \tag{2.5}$$

meaning that items 1 through K are the desired top-K ranked items.

Comparison graph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ stand for a comparison graph, where the vertex set $\mathcal{V} = \{1, 2, \ldots, n\}$ represents the n items of interest. The items i and j are compared if and only if (i, j) falls within the edge set $\mathcal{E}$. Unless otherwise noted, we assume that $\mathcal{G}$ is drawn from the Erdős–Rényi random graph $\mathcal{G}_{n,p}$, such that an edge between any pair of vertices is present independently with some probability p. In words, p captures the fraction of item pairs being compared.

Pairwise comparisons. For each $(i, j) \in \mathcal{E}$, we obtain L independent paired comparisons between items i and j. Let $y_{i,j}^{(l)}$ be the outcome of the ℓth comparison, which is independently drawn as

$$y_{i,j}^{(l)} \overset{\text{ind.}}{=} \begin{cases} 1 & \text{with probability } \dfrac{w_j^*}{w_i^* + w_j^*} = \dfrac{e^{\theta_j^*}}{e^{\theta_i^*} + e^{\theta_j^*}}, \\ 0 & \text{else.} \end{cases} \tag{2.6}$$

By convention, we set $y_{i,j}^{(l)} = 1 - y_{j,i}^{(l)}$ for all $(i, j) \in \mathcal{E}$ throughout the paper. This is also known as the *logistic* pairwise comparison model, due to its strong resemblance to logistic regression. It is self-evident that the sufficient statistics under this model are given by

$$\boldsymbol{y} := \{y_{i,j} \mid (i, j) \in \mathcal{E}\} \qquad \text{where } y_{i,j} := \frac{1}{L} \sum_{l=1}^{L} y_{i,j}^{(l)}. \tag{2.7}$$

To simplify the notation, we shall also take $y_{i,j}^* := \frac{w_j^*}{w_i^* + w_j^*} = \frac{e^{\theta_j^*}}{e^{\theta_i^*} + e^{\theta_j^*}}$.

Goal. The goal is to identify the *set* of top-K ranked items—that is, the *set* of K items that enjoy the largest preference scores—from the pairwise comparison data $\boldsymbol{y}$.

2.2. *Algorithms.*

2.2.1. *The spectral method*: *Rank centrality.*

The spectral ranking algorithm, or Rank Centrality [Negahban, Oh and Shah (2017)], is motivated by the connection between the pairwise comparisons and a random walk over a directed graph. The algorithm starts by converting the pairwise comparison data $\boldsymbol{y}$ into a transition matrix $\boldsymbol{P} = [P_{i,j}]_{1 \le i,j \le n}$ in such a way that

$$P_{i,j} = \begin{cases} \dfrac{1}{d} y_{i,j} & \text{if } (i, j) \in \mathcal{E}, \\ 1 - \dfrac{1}{d} \displaystyle\sum_{k:(i,k) \in \mathcal{E}} y_{i,k} & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{2.8}$$

Algorithm 1 Spectral method (Rank Centrality)

Input the comparison graph $\mathcal{G}$, sufficient statistics $\boldsymbol{y}$, and the normalization factor d.

Define the probability transition matrix $\boldsymbol{P} = [P_{i,j}]_{1\le i,j\le n}$ as in (2.8).

Compute the leading left eigenvector $\boldsymbol{\pi}$ of $\boldsymbol{P}$.

Output the K items that correspond to the K largest entries of $\boldsymbol{\pi}$.

for some given normalization factor $d > 0$, and then proceeds by computing the stationary distribution $\boldsymbol{\pi} \in \mathbb{R}^n$ of the Markov chain induced by $\boldsymbol{P}$. As we shall see later, the parameter d is taken to be on the same order of the maximum vertex degree of $\mathcal{G}$ while ensuring the nonnegativity of $\boldsymbol{P}$. As asserted by Negahban, Oh and Shah (2017), $\boldsymbol{\pi}$ is a faithful estimate of $\boldsymbol{w}^*$ up to some global scaling. The algorithm is summarized in Algorithm 1.

To develop some intuition regarding why this spectral algorithm gives a reasonable estimate of $\boldsymbol{w}^*$, it is perhaps more convenient to look at the population transition matrix $\boldsymbol{P}^* = [P^*_{i,j}]_{1\le i,j\le n}$:

$$P^*_{i,j} = \begin{cases} \dfrac{1}{d}\dfrac{w^*_j}{w^*_i + w^*_j} & \text{if } (i,j) \in \mathcal{E}, \\ 1 - \dfrac{1}{d}\displaystyle\sum_{k:(i,k)\in\mathcal{E}} \dfrac{w^*_k}{w^*_i + w^*_k} & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

which coincides with $\boldsymbol{P}$ by taking $L \to \infty$. It can be seen that the normalized score vector

$$\boldsymbol{\pi}^* := \frac{1}{\sum_{i=1}^n w^*_i}\big[w^*_1, w^*_2, \ldots, w^*_n\big]^\top \tag{2.9}$$

is the stationary distribution of the Markov chain induced by the transition matrix $\boldsymbol{P}^*$, since $\boldsymbol{P}^*$ and $\boldsymbol{\pi}^*$ are in detailed balance, namely,

$$\pi^*_i P^*_{i,j} = \pi^*_j P^*_{j,i}, \qquad \forall (i,j). \tag{2.10}$$

As a result, one expects the stationary distribution of the sample version $\boldsymbol{P}$ to form a good estimate of $\boldsymbol{w}^*$, provided the sample size is sufficiently large.

2.2.2. *The regularized MLE.* Under the BTL model, the negative log-likelihood function conditioned on $\mathcal{G}$ is given by (up to some global scaling)

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) :&= -\sum_{(i,j)\in\mathcal{E}, i>j} \left\{ y_{j,i} \log \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} + (1 - y_{j,i}) \log \frac{e^{\theta_j}}{e^{\theta_i} + e^{\theta_j}} \right\} \\ &= \sum_{(i,j)\in\mathcal{E}, i>j} \big\{ -y_{j,i}(\theta_i - \theta_j) + \log\big(1 + e^{\theta_i - \theta_j}\big) \big\}. \end{aligned} \tag{2.11}$$

The regularized MLE then amounts to solving the following convex program:

$$\text{minimize}_{\boldsymbol{\theta}\in\mathbb{R}^n} \mathcal{L}_\lambda(\boldsymbol{\theta}; \boldsymbol{y}) := \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) + \frac{1}{2}\lambda\|\boldsymbol{\theta}\|_2^2, \tag{2.12}$$

for a regularization parameter $\lambda > 0$. As will be discussed later, we shall adopt the choice $\lambda \asymp \sqrt{\frac{np\log n}{L}}$ throughout this paper. For the sake of brevity, we let $\boldsymbol{\theta}$ represent the resulting penalized maximum likelihood estimate whenever it is clear from the context. Similar to the spectral method, one reports the K items associated with the K largest entries of $\boldsymbol{\theta}$.

2.3. *Main results.* The most challenging part of top-K ranking is to distinguish the Kth and the $(K+1)$th items. In fact, the score difference of these two items captures the distance between the item sets $\{1,\dots,K\}$ and $\{K+1,\dots,n\}$. Unless their latent scores are sufficiently separated, the finite-sample nature of the model would make it infeasible to distinguish these two critical items. With this consideration in mind, we define the following separation measure:

$$\Delta_K := \frac{w_K^* - w_{K+1}^*}{w_{\max}}. \tag{2.13}$$

This metric turns out to play a crucial role in determining the minimal sample complexity for perfect top-K identification.

The main finding of this paper concerns the optimality of both the spectral method and the regularized MLE in the presence of a fixed dynamic range [i.e., $\kappa = O(1)$]. Recall that under the BTL model, the total number N of samples we collect concentrates sharply around its mean, namely,

$$N = \big(1+o(1)\big)\mathbb{E}[N] = \big(1+o(1)\big)n^2pL/2 \tag{2.14}$$

occurs with high probability. Our main result is stated in terms of the sample complexity required for exact top-K identification.

THEOREM 2.1. *Consider the pairwise comparison model specified in Section 2.1 with $\kappa = O(1)$. Suppose that $p > \frac{c_0\log n}{n}$ and that*

$$\frac{n^2pL}{2} \geq \frac{c_1 n\log n}{\Delta_K^2} \tag{2.15}$$

for some sufficiently large positive constants c_0 and c_1. Further assume $L \leq c_2\cdot n^{c_3}$ for any absolute constants $c_2, c_3 > 0$. With probability exceeding $1 - O(n^{-5})$, the set of top-K ranked items can be recovered exactly by the spectral method given in Algorithm 1, and by the regularized MLE given in (2.12). Here, we take $d = c_d np$ in the spectral method and $\lambda = c_\lambda\sqrt{\frac{np\log n}{L}}$ in the regularized MLE, where $c_d \geq 2$ and $c_\lambda > 0$ are some absolute constants.

REMARK 2.2. We emphasize that $p \geq \frac{c_0 \log n}{n}$ for $c_0 \geq 1$ is a fundamental requirement for the ranking task. In fact, if $p < (1-\varepsilon)\frac{\log n}{n}$ for any constant $\varepsilon > 0$, then the comparison graph $\mathcal{G} \sim \mathcal{G}_{n,p}$ is disconnected with high probability. This means that there exists at least one isolated item (which has not been compared with any other item) and cannot be ranked.

REMARK 2.3. In fact, the assumption that $L \leq c_2 \cdot n^{c_3}$ for any absolute constants $c_2, c_3 > 0$ is not needed for the spectral method.

REMARK 2.4. Here, we assume the same number of comparisons L to simplify the presentation as well as the proof. The result still holds true if we have distinct $L_{i,j}$'s for each $i \neq j$, as long as $n^2 p \min_{i \neq j} L_{i,j} \gtrsim \frac{n \log n}{\Delta_K^2}$.

Theorem 2.1 asserts that both the spectral method and the regularized MLE achieve a sample complexity on the order of $\frac{n \log n}{\Delta_K^2}$. Encouragingly, this sample complexity coincides with the minimax limit identified in Chen and Suh (2015), Theorem 2, in the fixed dynamic range, that is, $\kappa = O(1)$.

THEOREM 2.5 [Chen and Suh (2015)]. *Fix $\varepsilon \in (0, \frac{1}{2})$, and suppose that*

$$n^2 pL \leq 2c_2 \frac{(1-\varepsilon) n \log n - 2}{\Delta_K^2}, \tag{2.16}$$

where $c_2 = w_{\min}^4/(4w_{\max}^4)$. Then for any ranking procedure ψ, one can find a score vector $\boldsymbol{w}^*$ with separation Δ_K such that ψ fails to retrieve the top-K items with probability at least ε.

We are now positioned to compare our results with Jang et al. (2016), which also investigates the accuracy of the spectral method for top-K ranking. Specifically, Theorem 3 in Jang et al. (2016) establishes the optimality of the spectral method for the relatively dense regime where $p \gtrsim \sqrt{\frac{\log n}{n}}$. In this regime, however, the total sample size necessarily exceeds

$$n^2 pL/2 \geq n^2 p/2 \gtrsim \sqrt{n^3 \log n}, \tag{2.17}$$

which rules out the possibility of achieving minimal sample complexity if Δ_K is sufficiently large. For instance, consider the case where $\Delta_K \asymp 1$, then the optimal sample size—as revealed by Theorem 2.1 or Chen and Suh (2015), Theorem 1—is on the order of

$$(n \log n)/\Delta_K^2 \asymp n \log n,$$

which is a factor of $\sqrt{\frac{n}{\log n}}$ lower than the bound in (2.17). By contrast, our results hold all the way down to the sparsest possible regime where $p \asymp \frac{\log n}{n}$, confirming the optimality of the spectral method even for the most challenging scenario. Furthermore, we establish that the regularized MLE shares the same optimality guarantee as the spectral method, which was previously out of reach.

2.4. *Optimal control of entrywise estimation errors*. In order to establish the ranking accuracy as asserted by Theorem 2.1, the key is to obtain precise control of the ℓ_∞ loss of the score estimates. Our results are as follows.

THEOREM 2.6 (Entrywise error of the spectral method). *Consider the pairwise comparison model in Section 2.1 with $\kappa = O(1)$. Suppose $p > \frac{c_0 \log n}{n}$ for some sufficiently large constant $c_0 > 0$. Choose $d = c_d np$ for some constant $c_d \geq 2$ in Algorithm 1. Then the spectral estimate $\boldsymbol{\pi}$ satisfies*

$$\frac{\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty}{\|\boldsymbol{\pi}^*\|_\infty} \lesssim \sqrt{\frac{\log n}{npL}} \tag{2.18}$$

with probability $1 - O(n^{-5})$, where $\boldsymbol{\pi}^*$ is the normalized score vector [*cf.* (2.9)].

THEOREM 2.7 (Entrywise error of the regularized MLE). *Consider the pairwise comparison model specified in Section 2.1 with $\kappa = O(1)$. Suppose that $p \geq \frac{c_0 \log n}{n}$ for some sufficiently large constant $c_0 > 0$ and that $L \leq c_2 \cdot n^{c_3}$ for any absolute constants $c_2, c_3 > 0$. Set the regularization parameter to be $\lambda = c_\lambda \sqrt{\frac{np \log n}{L}}$ for some absolute constant $c_\lambda > 0$. Then the regularized MLE $\boldsymbol{\theta}$ satisfies*

$$\frac{\|e^{\boldsymbol{\theta}} - e^{\boldsymbol{\theta}^* - \overline{\theta}^* \mathbf{1}}\|_\infty}{\|e^{\boldsymbol{\theta}^* - \overline{\theta}^* \mathbf{1}}\|_\infty} \lesssim \sqrt{\frac{\log n}{npL}}$$

with probability exceeding $1 - O(n^{-5})$, where $\overline{\theta}^* := \frac{1}{n} \mathbf{1}^\top \boldsymbol{\theta}^*$ and $e^{\boldsymbol{\theta}} := [e^{\theta_1}, \ldots, e^{\theta_n}]^\top$.

Theorems 2.6–2.7 indicate that if the number of comparisons associated with each item—which concentrates around npL—exceeds the order of $\log n$, then both methods are able to achieve a small ℓ_∞ error when estimating the scores.

Recall that the ℓ_2 estimation error of the spectral method has been characterized by Negahban, Oh and Shah (2017) (or Theorem 5.2 of this paper that improves it by removing the logarithmic factor), which obeys

$$\frac{\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_2}{\|\boldsymbol{\pi}^*\|_2} \lesssim \sqrt{\frac{\log n}{npL}} \tag{2.19}$$

with high probability. Similar theoretical guarantees have been derived for another variant of the MLE (the constrained version) under a uniform sampling model as well [Negahban, Oh and Shah (2017)]. In comparison, our results indicate that the estimation errors for both algorithms are almost evenly spread out across all coordinates rather than being localized or clustered. Notably, the pointwise errors revealed by Theorems 2.6–2.7 immediately lead to exact top-K identification as claimed by Theorem 2.1.

PROOF OF THEOREM 2.1. In what follows, we prove the theorem for the spectral method part. The regularized MLE part follows from an almost identical argument, and hence is omitted.

Since the spectral algorithm ranks the items in accordance with the score estimate $\boldsymbol{\pi}$, it suffices to demonstrate that

$$\pi_i - \pi_j > 0, \qquad \forall 1 \le i \le K, K+1 \le j \le n.$$

To this end, we first apply the triangle inequality to get

$$\begin{aligned} \frac{\pi_i - \pi_j}{\|\boldsymbol{\pi}^*\|_\infty} &\ge \frac{\pi_i^* - \pi_j^*}{\|\boldsymbol{\pi}^*\|_\infty} - \frac{|\pi_i - \pi_i^*|}{\|\boldsymbol{\pi}^*\|_\infty} - \frac{|\pi_j - \pi_j^*|}{\|\boldsymbol{\pi}^*\|_\infty} \\ &\ge \Delta_K - \frac{2\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty}{\|\boldsymbol{\pi}^*\|_\infty}. \end{aligned} \tag{2.20}$$

In addition, it follows from Theorem 2.6 as well as our sample complexity assumption that

$$\frac{\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty}{\|\boldsymbol{\pi}^*\|_\infty} \lesssim \sqrt{\frac{\log n}{npL}} \quad \text{and} \quad n^2 pL \gtrsim \frac{n \log n}{\Delta_K^2}.$$

These conditions taken collectively imply that $\frac{\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty}{\|\boldsymbol{\pi}^*\|_\infty} < \frac{1}{2}\Delta_K$ as long as $\frac{npL\Delta_K^2}{\log n}$ exceeds some sufficiently large constant. Substitution into (2.20) reveals that $\pi_i - \pi_j > 0$, as claimed. □

2.5. *Heuristic arguments.* We pause to develop some heuristic explanation as to why the estimation errors are expected to be spread out across all entries. For simplicity, we focus on the case where $p = 1$ and L is sufficiently large, so that $\boldsymbol{y}$ and $\boldsymbol{P}$ sharply concentrate around $\boldsymbol{y}^*$ and $\boldsymbol{P}^*$, respectively.

We begin with the spectral algorithm. Since $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$ are respectively the invariant distributions of the Markov chains induced by $\boldsymbol{P}$ and $\boldsymbol{P}^*$, we can decompose

$$(\boldsymbol{\pi} - \boldsymbol{\pi}^*)^\top = \boldsymbol{\pi}^\top \boldsymbol{P} - \boldsymbol{\pi}^{*\top} \boldsymbol{P}^* = (\boldsymbol{\pi} - \boldsymbol{\pi}^*)^\top \boldsymbol{P} + \underbrace{\boldsymbol{\pi}^{*\top}(\boldsymbol{P} - \boldsymbol{P}^*)}_{:=\boldsymbol{\xi}}. \tag{2.21}$$

When $p = 1$ and $\frac{w_{\max}}{w_{\min}} \asymp 1$, the entries of $\boldsymbol{\pi}^*$ (resp., the off-diagonal entries of $\boldsymbol{P}^*$ and $\boldsymbol{P} - \boldsymbol{P}^*$) are all of the same order and, as a result, the energy of the uncertainty

term $\boldsymbol{\xi}$ is spread out (using standard concentration inequalities). In fact, we will demonstrate in Section 5.2 that

$$\frac{\|\boldsymbol{\xi}\|_\infty}{\|\boldsymbol{\pi}^*\|_\infty} \lesssim \sqrt{\frac{\log n}{npL}} \asymp \frac{\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_2\sqrt{\log n}}{\|\boldsymbol{\pi}^*\|_2}, \tag{2.22}$$

which coincides with the optimal rate. Further, if we look at each entry of (2.21), then for all $1 \le m \le n$,

$$\begin{aligned}\pi_m - \pi_m^* &= \big[(\boldsymbol{\pi}-\boldsymbol{\pi}^*)^\top \boldsymbol{P}\big]_m + \xi_m \\ &= \underbrace{(\pi_m - \pi_m^*)P_{m,m}}_{\text{contraction}} \\ &\quad + \underbrace{[P_{1,m},\ldots,P_{m-1,m},0,P_{m+1,m},\ldots,P_{n,m}]\begin{bmatrix}\pi_1-\pi_1^*\\ \vdots \\ \pi_n-\pi_n^*\end{bmatrix}}_{\text{error averaging}} + \xi_m.\end{aligned} \tag{2.23}$$

By construction of the transition matrix, one can easily verify that $P_{m,m}$ is bounded away from 1 and $P_{j,m} \asymp \frac{1}{n}$ for all $j \ne m$. As a consequence, the identity $\boldsymbol{\pi}^\top \boldsymbol{P} = \boldsymbol{\pi}^\top$ allows one to treat each $\pi_m - \pi_m^*$ as a mixture of three effects: (i) the first term of (2.23) behaves as an entrywise contraction of the error; (ii) the second term of (2.23) is a (nearly uniformly weighted) average of the errors over all coordinates, which can essentially be treated as a *smoothing* operator applied to the error components and (iii) the uncertainty term ξ_m. Rearranging terms in (2.23), we are left with

$$(1-P_{m,m})|\pi_m - \pi_m^*| \lesssim \frac{1}{n}\sum_{i=1}^n |\pi_i - \pi_i^*| + \xi_m, \qquad \forall m \tag{2.24}$$

which further gives

$$\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_\infty \lesssim \frac{1}{n}\sum_{i=1}^n |\pi_i - \pi_i^*| + \|\boldsymbol{\xi}\|_\infty. \tag{2.25}$$

There are two possibilities compatible with this bound (2.25): (1) $\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_\infty \lesssim \frac{1}{n}\sum_{i=1}^n |\pi_i - \pi_i^*|$, and (2) $\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_\infty \lesssim \|\boldsymbol{\xi}\|_\infty \lesssim \frac{\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_2}{\|\boldsymbol{\pi}^*\|_2}\|\boldsymbol{\pi}^*\|_\infty$ by (2.22). In either case, the errors are fairly delocalized, revealing that

$$\frac{\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_\infty}{\|\boldsymbol{\pi}^*\|_\infty} \lesssim \left\{\frac{1}{n}\frac{\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_1}{\|\boldsymbol{\pi}^*\|_\infty}, \frac{\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_2\sqrt{\log n}}{\|\boldsymbol{\pi}^*\|_2}\right\}.$$

We now move on to the regularized MLE, following a very similar argument. By the optimality condition that $\nabla \mathcal{L}_\lambda(\boldsymbol{\theta}) = \mathbf{0}$, one can derive (for some η to be

specified later)

$$\begin{aligned}\boldsymbol{\theta}-\boldsymbol{\theta}^* &= \boldsymbol{\theta}-\eta\nabla\mathcal{L}_\lambda(\boldsymbol{\theta})-\boldsymbol{\theta}^* \\ &= \boldsymbol{\theta}-\eta\nabla\mathcal{L}_\lambda(\boldsymbol{\theta})-\big(\boldsymbol{\theta}^*-\eta\nabla\mathcal{L}_\lambda(\boldsymbol{\theta}^*)\big)-\underbrace{\eta\nabla\mathcal{L}_\lambda(\boldsymbol{\theta}^*)}_{:=\boldsymbol{\zeta}} \\ &\approx \big(\boldsymbol{I}-\eta\nabla^2\mathcal{L}_\lambda(\boldsymbol{\theta}^*)\big)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)-\boldsymbol{\zeta}.\end{aligned}$$

Write $\nabla^2\mathcal{L}_\lambda(\boldsymbol{\theta}^*)=\boldsymbol{D}-\boldsymbol{A}$, where $\boldsymbol{D}$ and $\boldsymbol{A}$ denote respectively the diagonal and off-diagonal parts of $\nabla^2\mathcal{L}_\lambda(\boldsymbol{\theta}^*)$. Under our assumptions, one can check that $D_{m,m}\asymp n$ for all $1\le m\le n$ and $A_{j,m}\asymp 1$ for any $j\neq m$. With this notation in place, one can write the entrywise error as follows:

$$\theta_m-\theta_m^*=(1-\eta D_{m,m})(\theta_m-\theta_m^*)+\sum_{j:j\neq m}\eta A_{j,m}(\theta_j-\theta_j^*)-\zeta_m.$$

By choosing $\eta=c_2/n$ for some sufficiently small constant $c_2>0$, we get $1-\eta D_{m,m}<1$ and $\eta A_{j,m}\asymp 1/n$. Therefore, the right-hand side of the above relation also comprises a contraction term as well as an error smoothing term, similar to (2.23). Carrying out the same argument as for the spectral method, we see that the estimation errors of the regularized MLE are expected to be spread out.

2.6. *Numerical experiments.* It is worth noting that extensive numerical experiments on both synthetic and real data have already been conducted in Negahban, Oh and Shah (2017) to confirm the practicability of both the spectral method and the regularized MLE. See also Chen and Suh (2015) for the experiments on the Spectral-MLE algorithm. This section provides some additional simulations to complement their experimental results as well as our theory. Throughout the experiments, we set the number of items n to be 200, while the number of repeated comparisons L and the edge probability p can vary with the experiments. Regarding the tuning parameters, we choose $d=2d_{\max}$ in the spectral method where $d_{\max}$ is the maximum degree of the graph and $\lambda=2\sqrt{\frac{np\log n}{L}}$ in the regularized MLE, which are consistent with the configurations considered in the main theorems. Additionally, we also display the experimental results for the unregularized MLE, that is, $\lambda=0$. All of the results are averaged over 100 Monte Carlo simulations.

We first investigate the ℓ_∞ error of the spectral method and the (regularized) MLE when estimating the preference scores. To this end, we generate the latent scores w_i^* $(1\le i\le n)$ independently and uniformly at random over the interval $[0.5, 1]$. Figure 1(a) [resp., Figure 1(b)] displays the entrywise error in the spectral score estimation as the number of repeated comparisons L (resp., the edge probability p) varies. As is seen from the plots, the ℓ_∞ error of all methods gets smaller as p and L increase, confirming our results in Theorems 2.6–2.7. Next, we show in Figure 1(c) the relative ℓ_∞ error while fixing the total number of samples (i.e., n^2pL). It can be seen that the performance almost does not change if the sample

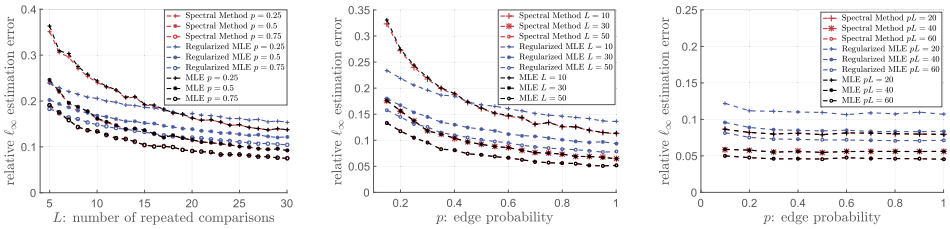

FIG. 1. *Empirical performance of the spectral method and the* (*regularized*) *MLE*: (a) ℓ_∞ *error versus* L, (b) ℓ_∞ *error versus* p *and* (c) ℓ_∞ *error versus* n^2pL.

complexity n^2pL remains the same. It is also interesting to see that the ℓ_∞ error of the spectral method and the MLE are very similar. In addition, Figure 2 illustrates the relative ℓ_∞ error and the relative ℓ_2 error in score estimation for all three methods. As we can see, the relative ℓ_∞ errors are not much larger than the relative ℓ_2 errors (recall that $n = 200$), thus offering empirical evidence that the errors in the score estimates are spread out across all entries.

Further, we examine the top-K ranking accuracy of all three methods. Here, we fix $p = 0.25$ and $L = 20$, set $K = 10$ and let $w_i^* = 1$ for all $1 \le i \le K$ and $w_j^* = 1 - \Delta$ for all $K + 1 \le j \le n$. By construction, the score separation satisfies $\Delta_K = \Delta$. Figure 3 illustrates the accuracy in identifying the top-K ranked items. The performance of them improves when the score separation becomes larger, which matches our theory in Theorem 2.1.

2.7. *Other related works.* The problem of ranking based on partial preferences has received much attention during the past decade. Two types of observation models have been considered: (1) *the cardinal-based model*, where users provide explicit numerical ratings of the items, (2) *the ordinal-based model*, where users are asked to make comparative measurements. See Ammar and Shah (2011) for detailed comparisons between them.

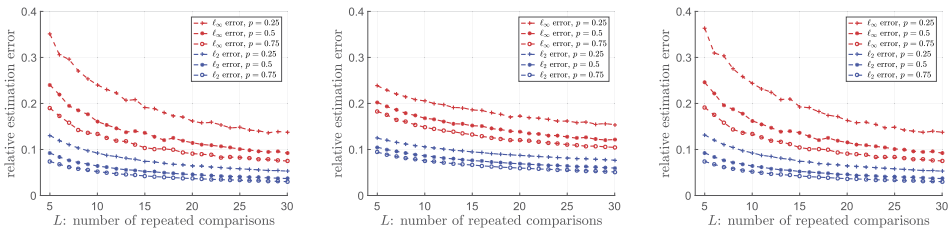

FIG. 2. *Comparisons between the relative* ℓ_∞ *error and the relative* ℓ_2 *error for* (a) *the spectral method*, (b) *the regularized MLE and* (c) *the MLE.*

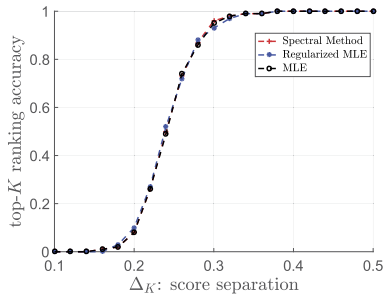

FIG. 3. *The top-K ranking accuracy of both the spectral method and the regularized MLE.*

In terms of the ordinal-based model—and in particular, ranking from pairwise comparisons—both parametric and nonparametric models have been extensively studied. For example, Hunter (2004) examined variants of the parametric BTL model, and established the convergence properties of the minorization-maximization algorithm for computing the MLE. Moreover, the BTL model falls under the category of *low-rank parametric models*, since the preference matrix is generated by passing a rank-2 matrix through the logistic link function [Rajkumar and Agarwal (2016)]. Additionally, the work Jiang et al. (2011) proposed a least-squares type method to estimate the full ranking, which generalizes the simple Borda count algorithm [Ammar and Shah (2011)]. For many of these algorithms, the sample complexities needed for perfect total ranking were determined by Rajkumar and Agarwal (2014), although the top-K ranking accuracy was not considered there.

Going beyond the parametric models, a recent line of works Chen et al. (2017), Pananjady et al. (2017), Shah and Wainwright (2015), Shah et al. (2017) considered the nonparametric *stochastically transitive model*, where the only model assumption is that the comparison probability matrix follows certain transitivity rules. This type of model subsumes the BTL model as a special case. For instance, Shah and Wainwright (2015) suggested a simple counting-based algorithm which can reliably recover the top-K ranked items for various models. However, the sampling paradigm considered therein is quite different from ours in the sparse regime; for instance, their model does not come close to the setting where p is small but L is large, which is the most challenging regime of the model adopted in our paper and Chen and Suh (2015), Negahban, Oh and Shah (2017).

All of the aforementioned papers concentrate on the case where there is a single ground-truth ordering. It would also be interesting to investigate the scenarios where different users might have different preference scores. To this end, Lu and Negahban (2014), Negahban et al. (2017) imposed the low-rank structure on the underlying preference matrix and adopted the nuclear-norm relaxation approach to recover the users' preferences. Additionally, several papers explored the ranking problem for the more general Plackett–Luce model [Hajek, Oh and Xu (2014),

Soufiani et al. (2013)], in the presence of adaptive sampling [Agarwal et al. (2017), Busa-Fekete et al. (2013), Heckel et al. (2016), Jamieson and Nowak (2011)], for the crowdsourcing scenario [Chen et al. (2013)], and in the adversarial setting [Suh, Tan and Zhao (2017)]. These are beyond the scope of the present paper.

Speaking of the error metric, the ℓ_∞ norm is appropriate for top-K ranking problem and other learning problems as well. In particular, ℓ_∞ perturbation bounds for eigenvectors of symmetric matrices [Abbe et al. (2017), Eldridge, Belkin and Wang (2017), Fan, Wang and Zhong (2018), Koltchinskii and Lounici (2016)] and singular vectors of general matrices [Koltchinskii and Xia (2016)] have been studied. In stark contrast, we study the ℓ_∞ norm errors of the leading eigenvector of a class of asymmetric matrices (probability transition matrix) and the regularized MLE. Furthermore, most existing results require the expectations of data matrices to have low rank, at least approximately. We do not impose such assumptions.

When it comes to the technical tools, it is worth noting that the leave-one-out idea has been invoked to analyze random designs for other high-dimensional problems, for example, robust M-estimators [El Karoui (2018)], confidence intervals for Lasso [Javanmard and Montanari (2018)], likelihood ratio test [Sur, Chen and Candès (2017)] and nonconvex statistical learning [Chen et al. (2018), Ma et al. (2017)]. In particular, Zhong and Boumal (2017) and Abbe et al. (2017) use it to precisely characterize entrywise behavior of eigenvectors of a large class of symmetric random matrices, which improves upon prior ℓ_∞ eigenvector analysis. Consequently, they are able to show the sharpness of spectral methods in many popular models. Our introduction of leave-one-out auxiliary quantities is similar in spirit to these papers.

Finally, the family of spectral methods has been successfully applied in numerous applications, for example, matrix completion [Keshavan, Montanari and Oh (2010)], phase retrieval [Chen and Candès (2017)], graph clustering [Abbe et al. (2017), Rohe, Chatterjee and Yu (2011)] and joint alignment [Chen and Candes (2016)]. All of them are designed based on the eigenvectors of some symmetric matrix, or the singular vectors if the matrix of interest is asymmetric. Our paper contributes to this growing literature by establishing a sharp *eigenvector* perturbation analysis framework for an important class of *asymmetric* matrices—the probability transition matrices.

3. Extension: General dynamic range. All of the preceding results concern the regime with a fixed dynamic range [i.e., $\kappa = O(1)$]. This section moves on to discussing the case with large κ.

To start with, by going through the same proof technique, we can readily obtain—in the general κ setting—the following performance guarantees for both the spectral estimate $\boldsymbol{\pi}$ and the regularized MLE $\boldsymbol{\theta}$.

THEOREM 3.1. *Consider the pairwise comparison model in Section 2.1. Suppose that* $p > \frac{c_0\kappa^5 \log n}{n}$ *for some sufficiently large constant* $c_0 > 0$, *and choose*

$d = c_d np$ for some constant $c_d \geq 2$ in Algorithm 1. Then with probability exceeding $1 - O(n^{-5})$:

1. *the spectral estimate $\boldsymbol{\pi}$ satisfies*

$$\frac{\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty}{\|\boldsymbol{\pi}^*\|_\infty} \lesssim \kappa \sqrt{\frac{\log n}{npL}},$$

where $\boldsymbol{\pi}^*$ is the normalized score vector as defined in (2.9).

2. *the set of top-K ranked items can be recovered exactly by the spectral method given in Algorithm 1, as long as*

$$\frac{n^2 pL}{2} \geq c_1 \frac{\kappa^2 n \log n}{\Delta_K^2}$$

for some sufficiently large constant $c_1 > 0$.

THEOREM 3.2. *Consider the pairwise comparison model in Section 2.1. Suppose that $p \geq \frac{c_0 \kappa^4 \log n}{n}$ for some sufficiently large constant $c_0 > 0$ and that $L \leq c_2 \cdot n^{c_3}$ for any absolute constants $c_2, c_3 > 0$. Set the regularization parameter to be $\lambda = c_\lambda \frac{1}{\log \kappa} \sqrt{\frac{np \log n}{L}}$ for some absolute constant $c_\lambda > 0$. Then with probability exceeding $1 - O(n^{-5})$:*

1. *the regularized MLE $\boldsymbol{\theta}$ satisfies*

$$\frac{\|e^{\boldsymbol{\theta}} - e^{\boldsymbol{\theta}^* - \overline{\theta}^* \mathbf{1}}\|_\infty}{\|e^{\boldsymbol{\theta}^* - \overline{\theta}^* \mathbf{1}}\|_\infty} \lesssim \kappa^2 \sqrt{\frac{\log n}{npL}}$$

where $\overline{\theta}^* := \frac{1}{n} \mathbf{1}^\top \boldsymbol{\theta}^*$ and $e^{\boldsymbol{\theta}} := [e^{\theta_1}, \ldots, e^{\theta_n}]^\top$.

2. *the set of top-K ranked items can be recovered exactly by the regularized MLE given in (2.12), as long as*

$$\frac{n^2 pL}{2} \geq c_1 \frac{\kappa^4 n \log n}{\Delta_K^2}$$

for some sufficiently large constant $c_1 > 0$.

REMARK 3.3. The guarantees on exact top-K recovery for both the spectral method and the regularized MLE are immediate consequences of their ℓ_∞ error bound, as we have argued in Section 2.4. Hence we will focus on proving the ℓ_∞ error bound in Sections 5–6.

Notably, the achievability bounds for top-K ranking in Theorems 3.1–3.2 do not match the lower bound asserted in Theorem 2.5 in terms of κ. This is partly because the separation measure Δ_K fails to capture the information bottleneck for

the general κ setting. In light of this, we introduce the following new measure that seems to be a more suitable metric to reflect the hardness of the top-K ranking problem:

$$\Delta_K^* := \frac{w_K^* - w_{K+1}^*}{w_{K+1}^*} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{w_{K+1}^* w_i^*}{(w_K^* + w_i^*)^2}}, \tag{3.1}$$

which will be termed the *generalized separation measure*. Informally, $(\Delta_K^*)^2$ is a reasonably tight upper bound on certain normalized KL divergence metric [see the proof of Theorem 3.4 for details in the Supplemental Material [Chen, Fan, Ma and Wang (2019)]]. With this metric in place, we derive another lower bound as follows.

THEOREM 3.4. *Fix $\varepsilon \in (0, \frac{1}{2})$, and let $\mathcal{G} \sim \mathcal{G}_{n,p}$. Consider any preference score vector $\boldsymbol{w}^*$, and let Δ_K^* denote its generalized separation. If*

$$n^2 pL \le \frac{\varepsilon^2}{2} \frac{n}{(\Delta_K^*)^2},$$

then there exists another preference score vector $\tilde{\boldsymbol{w}}$ with the same generalized separation Δ_K^* and different top-K items such that $P_{\mathrm{e}}(\psi) \ge \frac{1-\varepsilon}{2}$ for any ranking scheme ψ. Here, $P_{\mathrm{e}}(\psi)$ represents the probability of error in distinguishing these two vectors given $\boldsymbol{y}$.

The preceding sample complexity lower bound scales inversely proportionally to $(\Delta_K^*)^2$. To see why this generalized measure may be more suitable compared to the original separation metric, we single out three examples in the Supplementary Material. Unfortunately, our current analyses do not yield a matching upper bound with respect to Δ_K^* unless κ is a constant. For instance, the analysis of the spectral method relies on the eigenvector perturbation bound (Theorem 5.1), where the spectral gap and matrix perturbation play a crucial rule. However, the current results for controlling these quantities have explicit dependency on κ [Negahban, Oh and Shah (2017)]. It is not clear whether we could incorporate the new measure to eliminate such dependency on κ. This calls for more refined analysis techniques, which we leave for future investigation.

Moreover, it is not obvious whether the spectral method alone or the regularized MLE alone can achieve the minimal sample complexity in the general κ regime. It is possible that one needs to first screen out those items with extremely high or low scores using methods like Borda count [Ammar and Shah (2012)], as advocated by [Chen and Suh (2015), Negahban, Oh and Shah (2017), Jang et al. (2016)]. All in all, finding tight upper bounds for general κ remains an open question.

4. Discussion. This paper justifies the optimality of both the spectral method and the regularized MLE for top-K rank aggregation for the fixed dynamic range case. Our theoretical studies are by no means exhaustive, and there are numerous directions that would be of interest for future investigations. We point out a few possibilities as follows.

General condition number κ. As mentioned before, our current theory is optimal in the presence of a fixed dynamic range with $\kappa = O(1)$. We have also made a first attempt in considering the large κ regime. It is desirable to characterize the statistical and computational limits for more general κ.

Goodness-of-fit. Throughout this paper, we have assumed the BTL model captures the randomness underlying the data we collect. A practical question is whether the real data actually follows the BTL model. It would be interesting to investigate how to test the goodness-of-fit of this model.

Unregularized MLE. We have studied the optimality of the regularized MLE with the regularization parameter $\lambda \asymp \sqrt{\frac{np\log n}{L}}$. Our analysis relies on the regularization term to obtain convergence of the gradient descent algorithm (see Lemma 6.7). It is natural to ask whether such a regularization term is necessary or not. This question remains open.

More general comparison graphs. So far, we have focused on a tractable but somewhat restrictive comparison graph, namely, the Erdős–Rényi random graph. It would certainly be important to understand the performance of both methods under a broader family of comparison graphs, and to see which algorithms would enable optimal sample complexities under general sampling patterns.

Entrywise perturbation analysis for convex optimization. This paper provides the ℓ_∞ perturbation analysis for the regularized MLE using the leave-one-out trick as well as an inductive argument along the algorithmic updates. We expect this analysis framework to carry over to a much broader family of convex optimization problems, which may in turn offer a powerful tool for showing the stability of optimization procedures in an entrywise fashion.

5. Analysis for the spectral method. This section is devoted to proving Theorem 3.1, and hence Theorem 2.6, which characterizes the pointwise error of the spectral estimate.

5.1. *Preliminaries.* Here, we gather some preliminary facts about reversible Markov chains as well as the Erdős–Rényi random graph.

The first important result concerns the eigenvector perturbation for probability transition matrices, which can be treated as the analogue of the celebrated Davis-Kahan $\sin\Theta$ theorem [Davis and Kahan (1970)]. Due to its potential importance for other problems, we promote it to a theorem as follows.

THEOREM 5.1 (Eigenvector perturbation). *Suppose that $\boldsymbol{P}$, $\hat{\boldsymbol{P}}$ and $\boldsymbol{P}^*$ are probability transition matrices with stationary distributions $\boldsymbol{\pi}$, $\hat{\boldsymbol{\pi}}$, $\boldsymbol{\pi}^*$, respectively.*

Also, assume that $\boldsymbol{P}^*$ *represents a reversible Markov chain. When* $\|\boldsymbol{P}-\hat{\boldsymbol{P}}\|_{\boldsymbol{\pi}^*} < 1-\max\{\lambda_2(\boldsymbol{P}^*), -\lambda_n(\boldsymbol{P}^*)\}$, *it holds that*

$$\|\boldsymbol{\pi}-\hat{\boldsymbol{\pi}}\|_{\boldsymbol{\pi}^*} \le \frac{\|\boldsymbol{\pi}^\top(\boldsymbol{P}-\hat{\boldsymbol{P}})\|_{\boldsymbol{\pi}^*}}{1-\max\{\lambda_2(\boldsymbol{P}^*), -\lambda_n(\boldsymbol{P}^*)\}-\|\boldsymbol{P}-\hat{\boldsymbol{P}}\|_{\boldsymbol{\pi}^*}}.$$

Several remarks regarding Theorem 5.1 are in order. First, in contrast to standard perturbation results like Davis–Kahan's $\sin\Theta$ theorem, our theorem involves three matrices in total, where $\boldsymbol{P}$, $\hat{\boldsymbol{P}}$ and $\boldsymbol{P}^*$ can all be arbitrary. For example, one may choose $\boldsymbol{P}^*$ to be the population transition matrix, and $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ as two finite-sample versions associated with $\boldsymbol{P}^*$. Second, we only impose reversibility on $\boldsymbol{P}^*$, whereas $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ need not induce reversible Markov Chains. Third, Theorem 5.1 allows one to derive the ℓ_2 estimation error in Negahban, Oh and Shah (2017) directly without resorting to the power method; in fact, our ℓ_2 estimation error bound improves upon Negahban, Oh and Shah (2017) by some logarithmic factor.

THEOREM 5.2. *Consider the pairwise comparison model specified in Section 2.1 with* $\kappa = O(1)$. *Suppose* $p \ge c_0\frac{\log n}{n}$ *for some sufficiently large constant* $c_0 > 0$ *and* $d \ge c_d np$ *for* $c_d \ge 2$ *in Algorithm 1. With probability exceeding* $1-O(n^{-5})$, *one has*

$$\frac{\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_2}{\|\boldsymbol{\pi}^*\|_2} \lesssim \frac{1}{\sqrt{npL}}.$$

Notably, Theorem 5.2 matches the minimax lower bound derived in Negahban, Oh and Shah (2017), Theorem 3. As far as we know, this is the first result that demonstrates the orderwise optimality of the spectral method when measured by the ℓ_2 loss.

The next result is concerned with the concentration of the vertex degrees in an Erdős–Rényi random graph.

LEMMA 5.3 (Degree concentration). *Suppose that* $\mathcal{G} \sim \mathcal{G}_{n,p}$. *Let* d_i *be the degree of node* i, $d_{\min} = \min_{1\le i\le n} d_i$ *and* $d_{\max} = \max_{1\le i\le n} d_i$. *If* $p \ge \frac{c_0\log n}{n}$ *for some sufficiently large constant* $c_0 > 0$, *then the following event*

$$\mathcal{A}_0 = \left\{\frac{np}{2} \le d_{\min} \le d_{\max} \le \frac{3np}{2}\right\} \tag{5.1}$$

obeys

$$\mathbb{P}(\mathcal{A}_0) \ge 1 - O\big(n^{-10}\big).$$

PROOF. The proof follows from the standard Chernoff bound and is hence omitted. □

Since d is chosen to be $c_d np$ for some constant $c_d \ge 2$, we have, by Lemma 5.3, that the maximum vertex degree obeys $d_{\max} < d$ with high probability.

5.2. *Proof outline of Theorem 3.1.* In this subsection, we outline the proof of Theorem 3.1.

Recall that $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_n]^\top$ and $\boldsymbol{\pi}^* = [\pi_1^*, \ldots, \pi_n^*]^\top$ are the stationary distributions associated with $\boldsymbol{P}$ and $\boldsymbol{P}^*$, respectively. This gives

$$\boldsymbol{\pi}^\top \boldsymbol{P} = \boldsymbol{\pi}^\top \quad \text{and} \quad \boldsymbol{\pi}^{*\top} \boldsymbol{P}^* = \boldsymbol{\pi}^{*\top}.$$

For each $1 \le m \le n$, one can decompose

$$\begin{aligned}
\pi_m - \pi_m^* &= \boldsymbol{\pi}^\top \boldsymbol{P}_{\cdot m} - \boldsymbol{\pi}^{*\top} \boldsymbol{P}^*_{\cdot m} \\
&= \boldsymbol{\pi}^{*\top} (\boldsymbol{P}_{\cdot m} - \boldsymbol{P}^*_{\cdot m}) + (\boldsymbol{\pi} - \boldsymbol{\pi}^*)^\top \boldsymbol{P}_{\cdot m} \\
&= \underbrace{\sum_j \pi_j^* (P_{j,m} - P^*_{j,m})}_{:= I_1^m} + \underbrace{(\pi_m - \pi_m^*) P_{m,m}}_{:= I_2^m} \\
&\quad + \sum_{j: j \ne m} (\pi_j - \pi_j^*) P_{j,m},
\end{aligned}$$

where $\boldsymbol{P}_{\cdot m}$ (resp., $\boldsymbol{P}^*_{\cdot m}$) denotes the mth column of $\boldsymbol{P}$ (resp., $\boldsymbol{P}^*$). Then it boils down to controlling I_1^m, I_2^m and $\sum_{j: j \ne m} (\pi_j - \pi_j^*) P_{j,m}$.

1. Since $\boldsymbol{\pi}^*$ is deterministic while $\boldsymbol{P}$ is random, we can easily control I_1^m using Hoeffding's inequality. The bound is the following.

LEMMA 5.4. *With probability exceeding* $1 - O(n^{-5})$, *one has*

$$\max_m |I_1^m| \lesssim \sqrt{\frac{\log n}{Ld}} \|\boldsymbol{\pi}^*\|_\infty.$$

2. Next, we show the term I_2^m behaves as a contraction of $|\pi_m - \pi_m^*|$.

LEMMA 5.5. *With probability exceeding* $1 - O(n^{-5})$, *there exists some constant* $c > 0$ *such that for all* $1 \le m \le n$,

$$|I_2^m| \le \left(1 - \frac{np}{2(1+\kappa)d} + c\sqrt{\frac{\log n}{Ld}}\right) |\pi_m - \pi_m^*|.$$

3. The statistical dependency between $\boldsymbol{\pi}$ and $\boldsymbol{P}$ introduces difficulty in obtaining a sharp estimate of the third term $\sum_{j: j \ne m} (\pi_j - \pi_j^*) P_{j,m}$. Nevertheless, the leave-one-out technique helps us decouple the dependency and obtain effective control of this term. The key component of the analysis is the Introduction of a new probability transition matrix $\boldsymbol{P}^{(m)}$, which is a leave-one-out version of the original matrix $\boldsymbol{P}$. More precisely, $\boldsymbol{P}^{(m)}$ replaces all of the transition probabilities

involving the mth item with their expected values (unconditional on $\mathcal{G}$); that is, for any $i \neq j$,

$$P_{i,j}^{(m)} := \begin{cases} P_{i,j}, & i \neq m,\ j \neq m, \\ \dfrac{p}{d} y_{i,j}^*, & i = m \text{ or } j = m \end{cases}$$

with $y_{i,j}^* := \frac{w_j^*}{w_i^* + w_j^*}$. For any $1 \le i \le n$, set

$$P_{i,i}^{(m)} := 1 - \sum_{j:j\neq i} P_{i,j}^{(m)}$$

in order to ensure that $\boldsymbol{P}^{(m)}$ is a probability transition matrix. In addition, we let $\boldsymbol{\pi}^{(m)}$ be the stationary distribution of the Markov chain induced by $\boldsymbol{P}^{(m)}$. As will be demonstrated later, the main advantages of introducing $\boldsymbol{\pi}^{(m)}$ are two-fold: (1) the original spectral estimate $\boldsymbol{\pi}$ is very well approximated by $\boldsymbol{\pi}^{(m)}$, and (2) $\boldsymbol{\pi}^{(m)}$ is statistically independent of the connectivity of the mth node and the comparisons with regards to the mth item. Now we further decompose $\sum_{j:j\neq m}(\pi_j - \pi_j^*)P_{j,m}$:

$$\sum_{j:j\neq m} (\pi_j - \pi_j^*) P_{j,m} = \underbrace{\sum_{j:j\neq m} (\pi_j - \pi_j^{(m)}) P_{j,m}}_{:=I_3^m} + \underbrace{\sum_{j:j\neq m} (\pi_j^{(m)} - \pi_j^*) P_{j,m}}_{:=I_4^m}.$$

4. For I_3^m, we apply the Cauchy–Schwarz inequality to obtain that with probability at least $1 - O(n^{-10})$,

$$|I_3^m| \le \|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}\|_2 \Big(\sum_{j:j\neq m} P_{j,m}^2 \Big)^{\frac{1}{2}} \overset{\text{(i)}}{\le} \frac{1}{\sqrt{d}} \|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}\|_2,$$

where (i) follows from the fact that $P_{j,m} \le \frac{1}{d}$ for all $j \neq m$ and $d_{\max} \le d$ on the event $\mathcal{A}_0$ (defined in Lemma 5.3). Consequently, it suffices to control the ℓ_2 difference between the original spectral estimate $\boldsymbol{\pi}$ and its leave-one-out version $\boldsymbol{\pi}^{(m)}$. This is accomplished in the following lemma.

LEMMA 5.6. *Suppose that $np\gamma^2 > c\kappa \log n$ for some sufficiently large constant $c > 0$. With probability at least $1 - O(n^{-5})$,*

$$\|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}\|_2 \le \frac{16\sqrt{\kappa}}{\gamma} \sqrt{\frac{\log n}{Ld}} \|\boldsymbol{\pi}^*\|_\infty + \|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty, \tag{5.2}$$

where $\kappa = w_{\max}/w_{\min}$ and $\gamma = 1 - \max\{\lambda_2(\boldsymbol{P}^*), -\lambda_n(\boldsymbol{P}^*)\} - \|\boldsymbol{P} - \boldsymbol{P}^*\|_{\boldsymbol{\pi}^*}$.

Using Negahban, Oh and Shah (2017), Lemmas 3 and 4, and our Lemma 5.3, we can bound γ from below:

LEMMA 5.7 [Spectral gap, Negahban, Oh and Shah (2017)]. *Under the model specified in Section 2.1, if* $p \geq c_0 \frac{\log n}{n} \max\{1, \frac{\kappa^5}{L}\}$ *for some sufficiently large constant* $c_0 > 0$, *then with probability at least* $1 - O(n^{-5})$,

$$\gamma := 1 - \max\{\lambda_2(\boldsymbol{P}^*), -\lambda_n(\boldsymbol{P}^*)\} - \|\boldsymbol{P} - \boldsymbol{P}^*\|_{\boldsymbol{\pi}^*} \geq \frac{1}{2\kappa^2}\frac{d_{\min}}{d_{\max}} \geq \frac{1}{6\kappa^2}.$$

5. In order to control I_4^m, we exploit the statistical independence between $\boldsymbol{\pi}^{(m)}$ and $\boldsymbol{P}_{\cdot m}$. Specifically, we demonstrate the following.

LEMMA 5.8. *Suppose that* $p > \frac{c_0 \log n}{n}$ *for some sufficiently large constant* $c_0 > 0$. *With probability at least* $1 - O(n^{-10})$,

$$|I_4^m| \lesssim \frac{1}{\sqrt{n}}\|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}\|_2 + \sqrt{\frac{\log n}{Ld}}\|\boldsymbol{\pi}^*\|_\infty + \frac{\sqrt{np\log n} + \log n}{d}\|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}^*\|_\infty.$$

The above bound depends on both $\|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}\|_2$ and $\|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}^*\|_\infty$. We can invoke Lemma 5.6 and the inequality $\|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}^*\|_\infty \leq \|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}\|_2 + \|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty$ to reach

$$\begin{aligned}
|I_4^m| &\lesssim \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{np\log n} + \log n}{d}\right)\|\boldsymbol{\pi}^{(m)} - \boldsymbol{\pi}\|_2 + \sqrt{\frac{\log n}{Ld}}\|\boldsymbol{\pi}^*\|_\infty \\
&\quad + \frac{\sqrt{np\log n} + \log n}{d}\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty \\
&\lesssim \left\{\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{np\log n} + \log n}{d}\right)\frac{\sqrt{\kappa}}{\gamma} + 1\right\}\sqrt{\frac{\log n}{Ld}}\|\boldsymbol{\pi}^*\|_\infty \\
&\quad + \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{np\log n} + \log n}{d}\right)\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty.
\end{aligned}$$

6. Finally, we put the preceding bounds together. When $\frac{np}{\kappa^5 \log n}$ is large enough, with high probability, for some absolute constants $c_1, c_2, c_3 > 0$ one has

$$\begin{aligned}
&\left(\frac{np}{2(1+\kappa)d} - c_1\sqrt{\frac{\log n}{Ld}}\right)|\pi_m - \pi_m^*| \\
&\quad \leq \left\{c_2 + \left(\frac{1}{\sqrt{d}} + \frac{c_3}{\sqrt{n}} + c_3\frac{\sqrt{np\log n} + \log n}{d}\right)\frac{16\sqrt{\kappa}}{\gamma}\right\}\sqrt{\frac{\log n}{Ld}}\|\boldsymbol{\pi}^*\|_\infty \\
&\qquad + \left(\frac{1}{\sqrt{d}} + \frac{c_3}{\sqrt{n}} + 2c_3\frac{\sqrt{np\log n} + \log n}{d}\right)\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_\infty.
\end{aligned}$$

simultaneously for all $1 \le m \le n$. By taking the maximum over m on the left-hand side and combining terms, we get

$$\underbrace{\left(\frac{np}{2(1+\kappa)d} - c_1\sqrt{\frac{\log n}{Ld}} - \frac{1}{\sqrt{d}} - \frac{c_3}{\sqrt{n}} - 2c_3\frac{\sqrt{np\log n}+\log n}{d}\right)}_{:=\alpha_1}\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_\infty$$
$$\le \underbrace{\left\{c_2 + \left(\frac{1}{\sqrt{d}} + \frac{c_3}{\sqrt{n}} + c_3\frac{\sqrt{np\log n}+\log n}{d}\right)\frac{16\sqrt{\kappa}}{\gamma}\right\}}_{:=\alpha_2}\sqrt{\frac{\log n}{Ld}}\|\boldsymbol{\pi}^*\|_\infty.$$

Hence, as long as $\frac{np}{\kappa^5 \log n}$ is sufficiently large, one has

$$c_1\sqrt{\frac{\log n}{Ld}} + \frac{1}{\sqrt{d}} + \frac{c_3}{\sqrt{n}} + 2c_3\frac{\sqrt{np\log n}+\log n}{d} \lesssim \sqrt{\frac{\log n}{np}} \lesssim \frac{1}{\sqrt{\kappa^5}},$$

which further leads to $\alpha_1 \gtrsim 1/\kappa$, $\alpha_2 \lesssim 1$, and

$$\|\boldsymbol{\pi}-\boldsymbol{\pi}^*\|_\infty \lesssim \kappa\sqrt{\frac{\log n}{Ld}}\|\boldsymbol{\pi}^*\|_\infty \asymp \kappa\sqrt{\frac{\log n}{npL}}\|\boldsymbol{\pi}^*\|_\infty.$$

This completes the proof of Theorem 3.1 and Theorem 2.6.

6. Analysis for the regularized MLE. This section establishes the ℓ_∞ error of the regularized MLE as claimed in Theorem 3.2 (and also Theorem 2.7). Recall that in Theorem 3.2, we compare the regularized MLE $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^* - \overline{\theta}^*\mathbf{1}$. Therefore, without loss of generality we can assume that

$$\mathbf{1}^\top\boldsymbol{\theta}^* = 0. \tag{6.1}$$

This combined with the fact that $\theta_{\max} - \theta_{\min} = \log\kappa$ reveals that

$$\|\boldsymbol{\theta}^*\|_\infty \le \log\kappa \quad \text{and} \quad \|\boldsymbol{\theta}^*\|_2 \le \sqrt{n}\log\kappa.$$

In addition, we assume that $L = O(n^5)$ in this section. It is straightforward to extend the proof to cover $L \le c_2 \cdot n^{c_3}$ for any constants $c_2, c_3 > 0$.

6.1. *Preliminaries and notation.* Before proceeding to the proof, we gather some basic facts. To begin with, the gradient and the Hessian of $\mathcal{L}(\cdot; \boldsymbol{y})$ in (2.11) can be computed as

$$\nabla\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{(i,j)\in\mathcal{E}, i>j}\left\{-y_{j,i} + \frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j}}\right\}(\boldsymbol{e}_i - \boldsymbol{e}_j); \tag{6.2}$$

$$\nabla^2\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{(i,j)\in\mathcal{E}, i>j}\frac{e^{\theta_i}e^{\theta_j}}{(e^{\theta_i}+e^{\theta_j})^2}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top. \tag{6.3}$$

Here, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ stand for the canonical basis vectors in $\mathbb{R}^n$. When evaluated at the truth $\boldsymbol{\theta}^*$, the size of the gradient can be controlled as follows.

LEMMA 6.1. *Let λ be as specified in Theorem 3.2. The following event*

$$\mathcal{A}_2 := \left\{ \|\nabla \mathcal{L}_\lambda(\boldsymbol{\theta}^*; \boldsymbol{y})\|_2 \lesssim \sqrt{\frac{n^2 p \log n}{L}} \right\} \tag{6.4}$$

occurs with probability exceeding $1 - O(n^{-10})$.

The following lemmas characterize the smoothness and the strong convexity of the function $\mathcal{L}_\lambda(\cdot; \boldsymbol{y})$. In the sequel, we denote by $\boldsymbol{L}_\mathcal{G} = \sum_{(i,j)\in\mathcal{E}, i>j} (\boldsymbol{e}_i - \boldsymbol{e}_j) \times (\boldsymbol{e}_i - \boldsymbol{e}_j)^\top$ the (unnormalized) Laplacian matrix [Chung (1997)] associated with $\mathcal{G}$. For any matrix $\boldsymbol{A}$, we let

$$\lambda_{\min,\perp}(\boldsymbol{A}) := \min\{\mu \mid \boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} \geq \mu \|\boldsymbol{z}\|_2^2 \text{ for all } \boldsymbol{z} \text{ with } \mathbf{1}^\top \boldsymbol{z} = 0\}, \tag{6.5}$$

namely, the smallest eigenvalue when restricted to vectors orthogonal to $\mathbf{1}$.

LEMMA 6.2. *Suppose that $p > \frac{c_0 \log n}{n}$ for some sufficiently large constant $c_0 > 0$. Then on the event $\mathcal{A}_0$ as defined in (5.1), one has*

$$\lambda_{\max}\big(\nabla^2 \mathcal{L}_\lambda(\boldsymbol{\theta}; \boldsymbol{y})\big) \leq \lambda + np, \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^n.$$

PROOF. Note that $\frac{e^{\theta_i} e^{\theta_j}}{(e^{\theta_i} + e^{\theta_j})^2} \leq \frac{1}{4}$. It follows immediately from the Hessian in (6.3) that

$$\lambda_{\max}\big(\nabla^2 \mathcal{L}_\lambda(\boldsymbol{\theta}; \boldsymbol{y})\big) \leq \lambda + \frac{1}{4}\|\boldsymbol{L}_\mathcal{G}\| \leq \lambda + \frac{1}{2} d_{\max},$$

where $d_{\max}$ is the maximum vertex degree in the graph $\mathcal{G}$. In addition, on the event $\mathcal{A}_0$ we have $d_{\max} \leq 2np$, which completes the proof. □

LEMMA 6.3. *For all $\boldsymbol{\theta} \in \mathbb{R}^n$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty \leq C$ for some $C \geq 0$, we have*

$$\lambda_{\min,\perp}\big(\nabla^2 \mathcal{L}_\lambda(\boldsymbol{\theta}; \boldsymbol{y})\big) \geq \lambda + \frac{1}{4\kappa e^{2C}} \lambda_{\min,\perp}(\boldsymbol{L}_\mathcal{G}).$$

LEMMA 6.4. *Let $\mathcal{G} \sim \mathcal{G}_{n,p}$, and suppose that $p > \frac{c_0 \log n}{n}$ for some sufficiently large constant $c_0 > 0$. Then one has*

$$\mathbb{P}\big(\lambda_{\min,\perp}(\boldsymbol{L}_\mathcal{G}) \geq np/2\big) \geq 1 - O\big(n^{-10}\big).$$

PROOF. Note that $\lambda_{\min,\perp}(\boldsymbol{L}_\mathcal{G})$ is exactly the spectral gap of the Laplacian matrix. See Tropp (2015), Section 5.3.3, for the derivation of this lemma. □

By combining Lemma 6.3 with Lemma 6.4, we reach the following result.

Algorithm 2 Gradient descent for computing the regularized MLE

Initialize $\boldsymbol{\theta}^0 = \boldsymbol{\theta}^*$.

for $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_t \nabla \mathcal{L}_\lambda\big(\boldsymbol{\theta}^t; \boldsymbol{y}\big); \tag{6.6}$$

COROLLARY 6.5. *Under the assumptions of Lemma 6.4, with probability exceeding* $1 - O(n^{-10})$, *one has*

$$\lambda_{\min,\perp}\big(\nabla^2 \mathcal{L}_\lambda(\boldsymbol{\theta}; \boldsymbol{y})\big) \geq \lambda + \frac{1}{8\kappa e^{2C}} np$$

simultaneously for all $\boldsymbol{\theta}$ *obeying* $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty \leq C$ *for some* $C \geq 0$.

6.2. *Proof outline of Theorem 3.2.* This subsection outlines the main steps for establishing Theorem 3.2.

Rather than directly resorting to the optimality condition, we adopt an algorithmic perspective to analyze the regularized MLE $\boldsymbol{\theta}$. Specifically, we consider the standard gradient descent algorithm that is expected to converge to the minimizer $\boldsymbol{\theta}$, and analyze the trajectory of this iterative algorithm instead. The algorithm is stated in Algorithm 2.

Notably, this gradient descent algorithm is not practical since the initial point is set to be $\boldsymbol{\theta}^*$. Nevertheless, it is helpful for analyzing the statistical accuracy of the regularized MLE $\boldsymbol{\theta}$. In what follows, we shall adopt a time-invariant step size rule:

$$\eta_t \equiv \eta = 1/(\lambda + np), \qquad t = 0, 1, 2, \ldots \tag{6.7}$$

Our proof can be divided into three steps:

I. establish—via standard optimization theory—that the output $\boldsymbol{\theta}^T$ of Algorithm 2 is sufficiently close to the regularized MLE $\boldsymbol{\theta}$, namely,

$$\big\|\boldsymbol{\theta}^T - \boldsymbol{\theta}\big\|_\infty \leq \big\|\boldsymbol{\theta}^T - \boldsymbol{\theta}\big\|_2 \leq C_0 \kappa^2 \sqrt{\frac{\log n}{npL}} \tag{6.8}$$

for $T = n^5$, where $C_0 > 0$ is some absolute constant;

II. use the leave-one-out argument to demonstrate that: the output $\boldsymbol{\theta}^T$ is close to the truth $\boldsymbol{\theta}^*$ in an entrywise fashion, that is,

$$\big\|\boldsymbol{\theta}^T - \boldsymbol{\theta}^*\big\|_\infty \leq C_4 \kappa^2 \sqrt{\frac{\log n}{npL}}$$

for some universal constant $C_4 > 0$. Combining this with (6.8) yields

$$\big\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\big\|_\infty \lesssim \kappa^2 \sqrt{\frac{\log n}{npL}};$$

III. the final step is to translate the perturbation bound on $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty$ to $\|e^{\boldsymbol{\theta}} - e^{\boldsymbol{\theta}^*}\|_\infty$ as claimed in the theorem.

Before continuing, we single out an important fact that will be used throughout the proof.

FACT 6.6. Suppose $\mathbf{1}^\top \boldsymbol{\theta}^* = 0$. Then we have $\mathbf{1}^\top \boldsymbol{\theta}^t = 0$ for all $t \geq 0$.

6.3. *Step I.* The first step relies heavily on optimization theory, namely the theory of gradient descent on strongly convex and smooth functions.

1. It is seen that the sequence $\{\boldsymbol{\theta}^t\}_{t=1}^\infty$ converges geometrically fast to the regularized MLE $\boldsymbol{\theta}$, a property that is standard in convex optimization literature. This claim is summarized in the following lemma.

LEMMA 6.7. *On the event $\mathcal{A}_0$ as defined in (5.1), one has*

$$\|\boldsymbol{\theta}^t - \boldsymbol{\theta}\|_2 \leq \rho^t \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}\|_2,$$

where $\rho = 1 - \frac{\lambda}{\lambda + np}$.

PROOF. This result directly follows from the smoothness property (see Lemma 6.2), the trivial strong convexity of $\mathcal{L}_\lambda(\boldsymbol{\theta}; \boldsymbol{y})$ [$\nabla^2 \mathcal{L}_\lambda(\boldsymbol{\theta}; \boldsymbol{y}) \succeq \lambda \boldsymbol{I}_n$, $\forall \boldsymbol{\theta}$], as well as the convergence property of the gradient descent algorithm [e.g., Bubeck (2015), Theorem 3.10]. □

A direct consequence of this convergence result and Fact 6.6 is that $\mathbf{1}^\top \boldsymbol{\theta} = 0$ for the regularized MLE $\boldsymbol{\theta}$.

2. We then control $\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}\|_2$. Recall that $\boldsymbol{\theta}^0 = \boldsymbol{\theta}^*$, and we have the following.

LEMMA 6.8. *On the event $\mathcal{A}_2$ as defined in (6.4), there exists some constant $c_2 > 0$ such that*

$$\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}\|_2 = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq c_2 \sqrt{n} \log \kappa.$$

3. The previous two claims taken together lead us to conclude that

$$\begin{aligned}
&\|\boldsymbol{\theta}^T - \boldsymbol{\theta}\|_2 \\
&\quad \leq \rho^T \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}\|_2 \leq \rho^T c_2 \sqrt{n} \log \kappa = c_2 \left(1 - \frac{\lambda}{\lambda + np}\right)^T \sqrt{n} \log \kappa \\
&\quad \leq c_2 \exp\left(-\frac{T\lambda}{\lambda + np}\right) \sqrt{n} \kappa^2 \\
&\quad \leq c_2 \exp\left(-\frac{T}{c_3 \log \kappa} \sqrt{\frac{\log n}{npL}}\right) \sqrt{n} \kappa^2 \qquad \left(\text{by } \lambda \asymp \frac{1}{\log \kappa} \sqrt{\frac{np \log n}{L}} < np\right)
\end{aligned}$$

$$\le c_2 \exp\Big(-\frac{T}{c_4 \log n}\sqrt{\frac{\log n}{npL}}\Big)\sqrt{n}\kappa^2 \qquad \Big(\text{by } \kappa^4 \lesssim \frac{np}{\log n} \le \frac{n}{\log n}\Big)$$

$$\le C_0 \kappa^2 \sqrt{\frac{\log n}{npL}},$$

for some constants $c_3, c_4, C_0 > 0$, $L \lesssim n^5$ and sufficiently large T (recall that $T = n^5$). The above bounds are somewhat loose, but they suffice for our purpose. We then naturally obtain

$$\big\|\boldsymbol{\theta}^T - \boldsymbol{\theta}\big\|_\infty \le \big\|\boldsymbol{\theta}^T - \boldsymbol{\theta}\big\|_2 \le C_0 \kappa^2 \sqrt{\frac{\log n}{npL}}$$

as claimed. This completes the first step of the proof.

6.4. *Step II.* The purpose of this step is to show that all iterates $\{\boldsymbol{\theta}^t\}_{0\le t\le T}$ are sufficiently close to $\boldsymbol{\theta}^*$ in terms of the ℓ_∞-norm distance. To facilitate analysis, for each $1 \le m \le n$, we introduce a leave-one-out sequence $\{\boldsymbol{\theta}^{t,(m)}\}$ constructed via the following update rule:

$$\boldsymbol{\theta}^{t+1,(m)} = \boldsymbol{\theta}^{t,(m)} - \eta \nabla \mathcal{L}_\lambda^{(m)}\big(\boldsymbol{\theta}^{t,(m)}\big), \tag{6.9}$$

where $\boldsymbol{\theta}^{0,(m)} = \boldsymbol{\theta}^0 = \boldsymbol{\theta}^*$ and

$$\begin{aligned}\mathcal{L}_\lambda^{(m)}(\boldsymbol{\theta}; \boldsymbol{y}) :=& \sum_{(i,j)\in\mathcal{E}, i>j, i\ne m, j\ne m} \big\{-y_{j,i}(\theta_i - \theta_j) + \log\big(1 + e^{\theta_i - \theta_j}\big)\big\} \\ &+ \sum_{i: i\ne m} p\bigg\{-\frac{e^{\theta_i^*}}{e^{\theta_i^*} + e^{\theta_m^*}}(\theta_i - \theta_m) + \log\big(1 + e^{\theta_i - \theta_m}\big)\bigg\} \\ &+ \frac{1}{2}\lambda\|\boldsymbol{\theta}\|_2^2.\end{aligned} \tag{6.10}$$

Here, the leave-one-out loss function $\mathcal{L}_\lambda^{(m)}(\boldsymbol{\theta}; \boldsymbol{y})$ replaces all log-likelihood components involving the mth item with their expected values (unconditional on $\mathcal{G}$). For any $1 \le m \le n$, the auxiliary sequence $\{\boldsymbol{\theta}^{t,(m)}\}$ serves as a reasonably good proxy for $\{\boldsymbol{\theta}^t\}$, while remaining statistically independent of $\{y_{i,m} \mid (i,m) \in \mathcal{E}\}$.

Our proof in this step is inductive in nature. For the sake of clarity, we first list all induction hypotheses needed in our analysis:

$$\big\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\big\|_2 \le C_1 \kappa \sqrt{\frac{\log n}{pL}}, \tag{6.11a}$$

$$\max_{1\le m\le n} \big|\theta_m^{t,(m)} - \theta_m^*\big| \le C_2 \kappa^2 \sqrt{\frac{\log n}{npL}}, \tag{6.11b}$$

$$\max_{1\le m\le n}\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t,(m)}\right\|_2\le C_3\kappa\sqrt{\frac{\log n}{npL}}, \tag{6.11c}$$

$$\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^*\right\|_\infty\le C_4\kappa^2\sqrt{\frac{\log n}{npL}}, \tag{6.11d}$$

where $C_1,\dots,C_4>0$ are some absolute constants. We aim to show that if the iterates at the tth iteration—that is, $\boldsymbol{\theta}^t$ and $\{\boldsymbol{\theta}^{t,(m)}\}_{1\le m\le n}$ —satisfy the induction hypotheses (6.11), then the $(t+1)$th iterates continue to satisfy these hypotheses. Clearly, it suffices to justify (6.11) for all $0\le t\le T=n^5$.

Before we dive into the inductive arguments, there are a few direct consequences of (6.11) that are worth listing. We gather them in the next lemma.

LEMMA 6.9. *Suppose the induction hypotheses* (6.11) *hold true for the* t*th iteration, then there exist some universal constants* $C_5,C_6>0$ *such that the following two bounds hold*:

$$\max_{1\le m\le n}\left\|\boldsymbol{\theta}^{t,(m)}-\boldsymbol{\theta}^*\right\|_\infty\le C_5\kappa^2\sqrt{\frac{\log n}{npL}}, \tag{6.12a}$$

$$\max_{1\le m\le n}\left\|\boldsymbol{\theta}^{t,(m)}-\boldsymbol{\theta}^*\right\|_2\le C_6\kappa\sqrt{\frac{\log n}{pL}}. \tag{6.12b}$$

Note that the base case (i.e., the case for $t=0$) is trivially true due to the same initial points, namely, $\boldsymbol{\theta}^{0,(m)}=\boldsymbol{\theta}^0=\boldsymbol{\theta}^*$ for all $1\le m\le n$. We start with the first induction hypothesis (6.11a), which is supplied below.

LEMMA 6.10. *Suppose the induction hypotheses* (6.11) *hold true for the* t*th iteration, then with probability at least* $1-O(n^{-10})$, *one has*

$$\left\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^*\right\|_2\le C_1\kappa\sqrt{\frac{\log n}{pL}},$$

as long as the step size obeys $0<\eta\le\frac{1}{\lambda+np}$ *and* $C_1>0$ *is sufficiently large.*

The remaining induction steps are provided in the following lemmas.

LEMMA 6.11. *Suppose the induction hypotheses* (6.11) *hold true for the* t*th iteration, then with probability at least* $1-O(n^{-10})$, *one has*

$$\max_{1\le m\le n}\left|\theta_m^{t+1,(m)}-\theta_m^*\right|\le C_2\kappa^2\sqrt{\frac{\log n}{npL}},$$

with the proviso that $0<\eta\le\frac{1}{\lambda+np}$ *and* $C_2\gtrsim C_6+c_\lambda$.

LEMMA 6.12. *Suppose the induction hypotheses* (6.11) *hold true for the* t*th iteration, then with probability at least* $1 - O(n^{-10})$, *one has*

$$\max_{1\le m\le n} \left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^{t+1,(m)}\right\|_2 \le C_3\kappa\sqrt{\frac{\log n}{npL}},$$

as long as the step size obeys $0 < \eta \le \frac{1}{\lambda+np}$ *and* $C_3 > 0$ *is sufficiently large.*

LEMMA 6.13. *Suppose the induction hypotheses* (6.11) *hold true for the* t*th iteration, then with probability at least* $1 - O(n^{-10})$, *one has*

$$\left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\right\|_\infty \le C_4\kappa^2\sqrt{\frac{\log n}{npL}}$$

for any $C_4 \ge C_3 + C_2$.

Taking the union bound over $T = n^5$ iterations yields that with probability at least $1 - O(n^{-5})$,

$$\left\|\boldsymbol{\theta}^{T} - \boldsymbol{\theta}^*\right\|_\infty \le C_4\kappa^2\sqrt{\frac{\log n}{npL}},$$

which together with the conclusion in Step I results in

$$\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\right\|_\infty \le \left\|\boldsymbol{\theta}^{T} - \boldsymbol{\theta}^*\right\|_\infty + \left\|\boldsymbol{\theta}^{T} - \boldsymbol{\theta}\right\|_\infty \le (C_0 + C_4)\kappa^2\sqrt{\frac{\log n}{npL}}. \tag{6.13}$$

6.5. *Step III.* It remains to show that

$$\frac{\|e^{\boldsymbol{\theta}} - e^{\boldsymbol{\theta}^*}\|_\infty}{\|e^{\boldsymbol{\theta}^*}\|_\infty} \lesssim \kappa^2\sqrt{\frac{\log n}{npL}}.$$

Toward this end, we observe that for each $1 \le m \le n$,

$$\frac{|e^{\theta_m} - e^{\theta_m^*}|}{e^{\theta_{\max}}} = \frac{|e^{\tilde{\theta}_m}(\theta_m - \theta_m^*)|}{e^{\theta_{\max}}} \le \frac{e^{\theta_{\max}+\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_\infty}\cdot|\theta_m - \theta_m^*|}{e^{\theta_{\max}}},$$

where $\tilde{\theta}_m$ is between θ_m and θ_m^*, and $\theta_{\max}$ is the largest entry of θ^*. Continuing the derivation and using (6.13), we arrive at

$$\max_{1\le m\le n}\frac{|e^{\theta_m} - e^{\theta_m^*}|}{e^{\theta_{\max}}} \le \frac{e^{\theta_{\max}+\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_\infty}}{e^{\theta_{\max}}}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\right\|_\infty \lesssim \kappa^2\sqrt{\frac{\log n}{npL}}$$

as long as $\kappa^2\sqrt{\frac{\log n}{npL}}$ is small enough. This completes the proof of Theorem 3.2.

SUPPLEMENTARY MATERIAL

Additional Proofs (DOI: 10.1214/18-AOS1745SUPP; .pdf). Additional proofs of the results in the paper can be found in the Supplementary Material.

REFERENCES

ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. ArXiv preprint. Available at ArXiv:1709.09565.

AGARWAL, A., AGARWAL, S., ASSADI, S. and KHANNA, S. (2017). Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory* 39–75.

AMMAR, A. and SHAH, D. (2011). Ranking: Compare, don't score. In 2011 49*th Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 776–783. DOI:10.1109/Allerton.2011.6120246.

AMMAR, A. and SHAH, D. (2012). Efficient rank aggregation using partial data. In *SIGMETRICS* **40** 355–366. ACM, New York.

BALTRUNAS, L., MAKCINSKAS, T. and RICCI, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems*. *RecSys* '10 119–126. ACM, New York.

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925

BUBECK, S. (2015). Convex optimization: Algorithms and complexity. *Found*. *Trends Mach*. *Learn*. **8** 231–357.

BUSA-FEKETE, R., SZÖRÉNYI, B., WENG, P., CHENG, W. and HÜLLERMEIER, E. (2013). Top-k selection based on adaptive sampling of noisy preferences. In *International Conference on Machine Learning*.

CHEN, Y. and CANDES, E. (2016). The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Comm*. *Pure Appl*. *Math*. To appear.

CHEN, Y. and CANDÈS, E. J. (2017). Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm*. *Pure Appl*. *Math*. **70** 822–883. MR3628877

CHEN, Y. and SUH, C. (2015). Spectral MLE: Top-K rank aggregation from pairwise comparisons. In *International Conference on Machine Learning* 371–380.

CHEN, X., BENNETT, P. N., COLLINS-THOMPSON, K. and HORVITZ, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *ACM International Conference on Web Search and Data Mining* 193–202. ACM, New York.

CHEN, X., GOPI, S., MAO, J. and SCHNEIDER, J. (2017). Competitive analysis of the top-K ranking problem. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* 1245–1264. SIAM, Philadelphia, PA. MR3627810

CHEN, Y., CHI, Y., FAN, J. and MA, C. (2018). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. Available at ArXiv:1803.07726.

CHEN, Y., FAN, J., MA, C. and WANG, K. (2019). Supplement to "Spectral method and regularized MLE are both optimal for top-K ranking." DOI:10.1214/18-AOS1745SUPP.

CHUNG, F. R. K. (1997). *Spectral Graph Theory*. *CBMS Regional Conference Series in Mathematics* **92**. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the Amer. Math. Soc., Providence, RI. MR1421568

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J*. *Numer*. *Anal*. **7** 1–46. MR0264450

DWORK, C., KUMAR, R., NAOR, M. and SIVAKUMAR, D. (2001). Rank aggregation methods for the Web. In *International Conference on World Wide Web* 613–622.

EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab*. *Theory Related Fields* **170** 95–175. MR3748322

ELDRIDGE, J., BELKIN, M. and WANG, Y. (2017). Unperturbed: Spectral analysis beyond Davis–Kahan. ArXiv preprint. Available at ArXiv:1706.06516.

FAN, J., WANG, W. and ZHONG, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application. *J. Mach. Learn. Res.* **18** 1–42.

FORD, L. R. JR. (1957). Solution of a ranking problem from binary comparisons. *Amer. Math. Monthly* **64** 28–33. MR0097876

HAJEK, B., OH, S. and XU, J. (2014). Minimax-optimal inference from partial rankings. In *Neural Information Processing Systems* 1475–1483.

HECKEL, R., SHAH, N. B., RAMCHANDRAN, K. and WAINWRIGHT, M. J. (2016). Active ranking from pairwise comparisons and when parametric assumptions don't help. ArXiv preprint. Available at ArXiv:1606.08842.

HUNTER, D. R. (2004). MM algorithms for generalized Bradley–Terry models. *Ann. Statist.* **32** 384–406. MR2051012

JAMIESON, K. G. and NOWAK, R. D. (2011). Active ranking using pairwise comparisons. In *Neural Information Processing Systems* 2240–2248.

JANG, M., KIM, S., SUH, C. and OH, S. (2016). Top-K ranking from pairwise comparisons: When spectral ranking is optimal. ArXiv preprint. Available at arXiv:1603.04153.

JAVANMARD, A. and MONTANARI, A. (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622. MR3851749

JIANG, X., LIM, L.-H., YAO, Y. and YE, Y. (2011). Statistical ranking and combinatorial Hodge theory. *Math. Program.* **127** 203–244. MR2776715

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022

KOLTCHINSKII, V. and LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.* **52** 1976–2013. MR3573302

KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII. Progress in Probability* **71** 397–423. Springer, Cham. MR3565274

LU, Y. and NEGAHBAN, S. N. (2014). Individualized rank aggregation using nuclear norm regularization. ArXiv preprint. Available at ArXiv:1410.0860.

LUCE, R. D. (1959). *Individual Choice Behavior*: *A Theoretical Analysis*. Wiley, New York; Chapman & Hall, London. MR0108411

MA, C., WANG, K., CHI, Y. and CHEN, Y. (2017). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. ArXiv preprint. Available at ArXiv:1711.10467.

MASSE, K. (1997). Statistical models applied to the rating of sports teams. Technical Report, Bluefield College, Bluefield, VA.

NEGAHBAN, S., OH, S. and SHAH, D. (2017). Rank centrality: Ranking from pairwise comparisons. *Oper. Res.* **65** 266–287. MR3613103

NEGAHBAN, S., OH, S., THEKUMPARAMPIL, K. K. and XU, J. (2017). Learning from comparisons and choices. ArXiv preprint. Available at ArXiv:1704.07228.

PANANJADY, A., MAO, C., MUTHUKUMAR, V., WAINWRIGHT, M. J. and COURTADE, T. A. (2017). Worst-case vs average-case design for estimation from fixed pairwise comparisons. ArXiv preprint. Available at ArXiv:1707.06217.

RAJKUMAR, A. and AGARWAL, S. (2014). A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning* I-118–I-126.

RAJKUMAR, A. and AGARWAL, S. (2016). When can we rank well from comparisons of $O(n \log n)$ non-actively chosen pairs? In *Conference on Learning Theory* 1376–1401.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

SHAH, N. B. and WAINWRIGHT, M. J. (2015). Simple, robust and optimal ranking from pairwise comparisons. ArXiv preprint. Available at arXiv:1512.08949.

SHAH, N. B., BALAKRISHNAN, S., GUNTUBOYINA, A. and WAINWRIGHT, M. J. (2017). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Trans. Inform. Theory* **63** 934–959. MR3604649

SOUFIANI, H. A., CHEN, W. Z., PARKES, D. C. and XIA, L. (2013). Generalized method-of-moments for rank aggregation. In *Proceedings of the* 26*th International Conference on Neural Information Processing Systems. NIPS'*13 2706–2714.

SUH, C., TAN, V. Y. F. and ZHAO, R. (2017). Adversarial top-K ranking. *IEEE Trans. Inform. Theory* **63** 2201–2225. MR3626865

SUR, P., CHEN, Y. and CANDÈS, E. J. (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled Chi-square. ArXiv preprint. Available at ArXiv:1706.01191.

TROPP, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* **8** 1–230.

ZHONG, Y. and BOUMAL, N. (2017). Near-optimal bounds for phase synchronization. Available at arXiv:1703.06605.

Y. CHEN
DEPARTMENT OF ELECTRICAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: yuxin.chen@princeton.edu

J. FAN
SCHOOL OF DATA SCIENCE
FUDAN UNIVERSITY
SHANGHAI
CHINA
AND
DEPARTMENT OF OPERATIONS RESEARCH AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu

C. MA
K. WANG
DEPARTMENT OF OPERATIONS RESEARCH AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: congm@princeton.edu
kaizheng@princeton.edu