The Annals of Statistics
2022, Vol. 50, No. 3, 1618–1652
https://doi.org/10.1214/21-AOS2166

PARTIAL RECOVERY FOR TOP-k RANKING: OPTIMALITY OF MLE AND SUBOPTIMALITY OF THE SPECTRAL METHOD

BY PINHAN CHEN[1,a], CHAO GAO[1,b] AND ANDERSON Y. ZHANG[2,c]

[1]*Department of Statistics, University of Chicago,* [a]*pinhanchen@uchicago.edu,* [b]*chaogao@uchicago.edu*
[2]*Department of Statistics, University of Pennsylvania,* [c]*ayz@wharton.upenn.edu*

Given partially observed pairwise comparison data generated by the Bradley–Terry–Luce (BTL) model, we study the problem of top-k ranking. That is, to optimally identify the set of top-k players. We derive the minimax rate with respect to a normalized Hamming loss. This provides the first result in the literature that characterizes the partial recovery error in terms of the proportion of mistakes for top-k ranking. We also derive the optimal signal to noise ratio condition for the exact recovery of the top-k set. The maximum likelihood estimator (MLE) is shown to achieve both optimal partial recovery and optimal exact recovery. On the other hand, we show another popular algorithm, the spectral method, is in general suboptimal. Our results complement the recent work (*Ann. Statist.* **47** (2019) 2204–2235) that shows both the MLE and the spectral method achieve the optimal sample complexity for exact recovery. It turns out the leading constants of the sample complexity are different for the two algorithms. Another contribution that may be of independent interest is the analysis of the MLE without any penalty or regularization for the BTL model. This closes an important gap between theory and practice in the literature of ranking.

1. Introduction. Given partially observed pairwise comparison data from n players, a central statistical question is how to optimally aggregate the comparison results and to find the leading top k players. This problem is known as top-k ranking, which has important applications in many areas such as web search [11, 12] and competitive sports [19, 22]. In this paper, our goal is to study the statistical limits of both *partial* and *exact* recovery of the top-k ranking problem.

We will focus on the popular Bradley–Terry–Luce (BTL) pairwise comparison model [3, 18]. That is, we observe L games played between i and j, and the outcome is modeled by

$$y_{ijl} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{w_i^*}{w_i^* + w_j^*}\right), \quad l = 1, \dots, L. \tag{1}$$

We only observe outcomes from a small subset of pairs. This subset $\mathcal{E}$ is modeled by edges generated by an Erdős–Rényi [13] random graph with connection probability p on the n players. More details of the model will be given in Section 2. With the observations $\{y_{ijl}\}_{(i,j)\in\mathcal{E}, l\in[L]}$, the goal is to reliably recover the set of top-k players with the largest skill parameters w_i^*.

Theoretical properties of the top-k ranking problem have been studied by [6–8, 14, 15, 21, 23] and references therein. The literature is mainly focused the problem of exact recovery. That is, to investigate the signal to noise ratio condition under which one can recovery the top-k set without any error in probability. For this purpose, the state-of-the-art result is obtained by the recent work [7]. It was shown by [7] that both the MLE and the spectral method can

Received July 2021; revised December 2021.
MSC2020 subject classifications. 62F07.
Key words and phrases. Top-k ranking, pairwise comparison, MLE, spectral method, partial recovery, exact recovery.

perfectly identify the top-k players under optimal sample complexity up to some constant factor. This discovery was also verified by a numerical experiment that shows almost identical performances of the two methods. The results of [7] lead to the following intriguing research questions. What is the leading constant factor of the optimal sample complexity? Are the MLE and the spectral method still optimal if we take the leading constant into consideration?

In this paper, we give complete answers to the above questions. Our results show that while the MLE achieves a leading constant that is information-theoretically optimal, the spectral method only achieves a suboptimal constant. In particular, the MLE achieves exact recovery when

$$npL\Delta^2 > 2.001V(\kappa)\big(\sqrt{\log k} + \sqrt{\log(n-k)}\big)^2, \tag{2}$$

and the spectral method requires

$$npL\Delta^2 > 2.001\overline{V}(\kappa)\big(\sqrt{\log k} + \sqrt{\log(n-k)}\big)^2.$$

In the above two formulas, Δ is the logarithmic gap of the skill parameters between the top-k group and the rest of the players. The parameter κ is the dynamic range of the skill vector that will be defined in Section 2. The performances of the two methods are precisely characterized by the two functions $V(\kappa)$ and $\overline{V}(\kappa)$, which are understood to be the effective variances of the two algorithms. The two functions satisfy the strict inequality that $\overline{V}(\kappa) > V(\kappa)$ for all $\kappa > 0$, and the equality $\overline{V}(\kappa) = V(\kappa)$ only holds when $\kappa = 0$. We also establish an information-theoretic lower bound that shows the MLE constant $V(\kappa)$ is optimal, and it characterizes the phase transition boundary of exact recovery for the top-k ranking problem.

We would like to emphasize that our results do not contradict the conclusions of [7]. On the contrary, the current paper complements and refines the results of [7]. The optimality claim made by [7] on both the MLE and the spectral method only refers to the order of the sample complexity. Our results show that the performances of the two algorithms can be drastically different when the dynamic range parameter κ is strictly positive. We are also able to explain why the numerical experiment conducted in [7] demonstrates nearly identical performances of the MLE and the spectral method. Note that the experiment in [7] was conducted with the skill parameters w_i^* only taking two possible values, e^{Δ} or 1, depending on whether i belongs to the top-k group or not. We show in Section 5 that this configuration of w^* is asymptotically equivalent to $\kappa = 0$, which is the only case that makes $\overline{V}(\kappa) = V(\kappa)$, and thus the nearly identical performances of the two algorithms are actually well expected by our theory. As long as w^* deviates from this simple two-piece structure, our extensive numerical experiments in this paper show that the MLE always performs better than the spectral method, and the advantage of the MLE is usually quite significant.

Another popular ranking algorithm is to select the k players with the most number of games won. This is often referred to as the Borda count [2] or the Copeland count [10]. This simple method does not require an explicit parametric model assumption, which makes it robust under potential model misspecification, an advantage over the MLE and the spectral method. The performance of the count method has been analyzed by [23] under a very general nonparametric setting. However, under the BTL model, it can be shown that the count method has an undesirable property that its error probability does not decrease to zero as L tends to infinity, which implies that it is not even rate-optimal under the BTL model. See Appendix F of the Supplementary Material [5] for an analysis of the count method.

In addition to the exact recovery results, we have also obtained a series of results for partial recovery. We observe that top-k ranking can be viewed as a clustering problem. That is, one wants to cluster the players into two groups of sizes k and $n-k$, respectively. Therefore, it is more natural to consider the problem of partial recovery by analyzing the proportion of

players that are clustered into a wrong group. Clearly, this problem is more relevant in practice, since one rarely expects any real application where top-k ranking can be done without any error. From a mathematical point of view, the partial recovery problem is more general and we will show in Section 3 that an optimal partial recovery error bound will lead to the optimal exact recovery condition (2). To the best of our knowledge, a systematic study of partial recovery for top-k ranking has never been done in the literature. Our paper is perhaps the first work that formulates the top-k ranking problem into a decision-theoretic framework and derives the minimax optimal partial recovery error rate. Similar to the results of exact recovery, we show that the MLE is also optimal for partial recovery. It has an exponential error bound with respect to a normalized Hamming loss. The error exponent is shown to depend on the variance function $V(\kappa)$. In comparison, the spectral method still achieves a suboptimal error rate for partial recovery, with the error exponent depending on $\overline{V}(\kappa)$.

Recently, a few papers provide sharp analysis of spectral methods on some high-dimensional estimation problems and show spectral methods can achieve optimal theoretical guarantees just as MLEs. For example, it was shown by [1] that spectral clustering achieves optimal community detection for a special class of stochastic block models (SBMs). The paper [16] proved spectral clustering is also optimal under Gaussian mixture models. We emphasize that the results of both papers imply that not only the order of the sample complexity of spectral clustering is optimal, but even the leading constant is optimal, at least in the setting of SBMs and Gaussian mixture models. The results of the current paper, however, show that the optimality of spectral methods may not hold under more complicated settings such as the BTL model.

Finally, we discuss another contribution of the paper that may be of independent interest. That is, we are able to give a sharp analysis of the MLE under the BTL model. Previous analyses of the MLE in the literature [7, 8, 21] all impose some additional regularization to address the challenge that the Hessian of the log-likelihood function is not well behaved. Whether the *vanilla* MLE works theoretically without any penalty or regularization remains an open problem. Our analysis solves this open problem by relating a regularized MLE to an ℓ_∞-constrained MLE. This allows us to show that the solution to the ℓ_∞-constrained MLE lies in the interior of the constraint. Thus, we can conclude that the ℓ_∞-constrained MLE is equivalent to the vanilla MLE in its original form. This equivalence then leads to the desired control of the spectrum of the Hessian matrix, which is the most critical step of our analysis.

The rest of the paper is organized as follows. We introduce the setting of the problem in Section 2. The results of the MLE and the spectral method will be given in Section 3 and Section 4, respectively. We then comprehensively compare the two methods in Section 5 by numerical experiments. Section 6 presents a minimax lower bound for partial recovery. In Section 7, we analyze the error rates of the MLE and the spectral method for each individual parameter. Due to the limit on pages, we include the proofs of the results of the MLE in Section 8 and the proofs of all the other results in the Supplementary Material.

We close this section by introducing some notation that will be used in the paper. For an integer d, we use $[d]$ to denote the set $\{1, 2, \ldots, d\}$. Given two numbers $a, b \in \mathbb{R}$, we use $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We also write $a_+ = \max(a, 0)$. For two positive sequences $\{a_n\}$, $\{b_n\}$, $a_n \lesssim b_n$ or $a_n = O(b_n)$ means $a_n \le C b_n$ for some constant $C > 0$ independent of n, $a_n = \Omega(b_n)$ means $b_n = O(a_n)$, and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We also write $a_n = o(b_n)$ when $\limsup_n \frac{a_n}{b_n} = 0$. For a set S, we use $\mathbb{I}\{S\}$ to denote its indicator function and $|S|$ to denote its cardinality. For a vector $v \in \mathbb{R}^d$, its norms are defined by $\|v\|_1 = \sum_{i=1}^d |v_i|$, $\|v\|^2 = \sum_{i=1}^d v_i^2$ and $\|v\|_\infty = \max_{1 \le i \le d} |v_i|$. The notation $\mathbb{1}_d$ means a d-dimensional column vector of all ones. For any $v \in \mathbb{R}^d$, we write $\mathsf{ave}(v) = d^{-1} \mathbb{1}_d^T v$. Given $p, q \in (0, 1)$, the Kullback–Leibler divergence is defined by $D(p\|q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$. For a natural number n, $\mathfrak{S}_n$ is the set of permutations on $[n]$. The notation $\mathbb{P}$ and

$\mathbb{E}$ are used for generic probability and expectation whose distribution is determined from the context.

2. Models and methods.

2.1. *The BTL model.* We start by introducing the setting of our problem. Consider n players, and each one is associated with a positive latent skill parameter w_i^* for $i \in [n]$. The comparison scheme of the n players is characterized by an Erdős–Rényi random graph $A \sim \mathcal{G}(n, p)$. That is, $A_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ for all $1 \le i < j \le n$. For a pair (i, j) that is connected by the random graph and $A_{ij} = 1$, we observe L games played between i and j. The outcome of the games is modeled by the Bradley–Terry–Luce (BTL) model (1). Our goal is to identify the top-k players whose skill parameters w_i^*'s have the largest values.

To formulate this problem from a decision-theoretic point of view, we reparametrize the BTL model (1) by a sorted vector θ^* and a rank vector r^*. A sorted vector θ^* satisfies $\theta_1^* \ge \theta_2^* \ge \cdots \ge \theta_n^*$, and a rank vector r^* is an element of permutation $r^* \in \mathfrak{S}_n$. Then the BTL model (1) can be equivalently written as

$$y_{ijl} \overset{\text{ind}}{\sim} \text{Bernoulli}\big(\psi\big(\theta_{r_i^*}^* - \theta_{r_j^*}^*\big)\big), \quad l = 1, \dots, L, \tag{3}$$

where $\psi(\cdot)$ is the sigmoid function $\psi(t) = \frac{1}{1+e^{-t}}$. In the original representation, we have $w_i^* = \exp(\theta_{r_i^*}^*)$ for all $i \in [n]$. With (3), the top-k ranking problem is to identify the subset $\{i \in [n] : r_i^* \le k\}$ from the random comparison data. This is a typical semiparametric problem because of the presence of the nuisance parameter θ^*.

2.2. *Loss function for top-k ranking.* Our goal is to study optimal top-k ranking in terms of both *partial* and *exact* recovery. We thus introduce a loss function to quantify the error of top-k ranking. Given any $\widehat{r}, r^* \in \mathfrak{S}_k$, define the normalized Hamming distance by

$$\mathsf{H}_k(\widehat{r}, r^*) = \frac{1}{2k}\left(\sum_{i=1}^n \mathbb{I}\{\widehat{r}_i > k, r_i^* \le k\} + \sum_{i=1}^n \mathbb{I}\{\widehat{r}_i \le k, r_i^* > k\}\right). \tag{4}$$

The definition (4) gives a natural loss function for top-k ranking, since $\mathsf{H}_k(\widehat{r}, r^*)$ can be equivalently written as the cardinality of the symmetric difference of the sets $\{i \in [n] : \widehat{r}_i \le k\}$ and $\{i \in [n] : r_i^* \le k\}$ normalized by $2k$. The value of $\mathsf{H}_k(\widehat{r}, r^*)$ is always within the unit interval $[0, 1]$. Moreover, $\mathsf{H}_k(\widehat{r}, r^*) = 0$ if and only if $\{i \in [n] : \widehat{r}_i \le k\} = \{i \in [n] : r_i^* \le k\}$.

The loss function (4) can be related to various quantities previously defined in the literature. One of the most popular distances to compare two rank vectors is the *Kendall tau distance* defined as

$$\mathsf{K}(\widehat{r}, r^*) = \frac{1}{n} \sum_{1 \le i < j \le n} \mathbb{I}\{\text{sign}(\widehat{r}_i - \widehat{r}_j)\,\text{sign}(r_i^* - r_j^*) < 0\}.$$

Since $\mathsf{K}(\widehat{r}, r^*)$ counts all pairwise differences in the ranking relation, it is a stronger distance than (4). While $\mathsf{K}(\widehat{r}, r^*) = 0$ requires $\widehat{r} = r^*$, $\mathsf{H}_k(\widehat{r}, r^*) = 0$ only requires the two top-k sets are identical regardless of the actual ranks of the members of the sets. In fact, the study of the BTL model under $\mathsf{K}(\widehat{r}, r^*)$, called full ranking, is also a very interesting problem, and will be considered in a different paper.

As we have discussed in Section 1, the top-k ranking problem can be thought of as a special variable selection problem. Variable selection under the normalized Hamming loss has recently been studied by [4, 20]. Consider either a Gaussian sequence model or a regression

model with coefficient vector $\beta^* \in \mathbb{R}^p$ that satisfies either $\beta_j^* = 0$ or $|\beta_j^*| > a$. The papers [4, 20] consider estimating β^* under the loss

$$\overline{\mathsf{H}}_s(\widehat{\beta}, \beta^*) = \frac{1}{2s}\left(\sum_{j=1}^{p} \mathbb{I}\{|\widehat{\beta}_j| > a, \beta_j^* = 0\} + \sum_{j=1}^{p} \mathbb{I}\{\widehat{\beta}_j = 0, |\beta_j^*| > a\}\right),$$

where s is the number of β_j^*'s that are not zero. One can clearly see the similarity between the two loss functions $\mathsf{H}_k(\widehat{r}, r^*)$ and $\overline{\mathsf{H}}_s(\widehat{\beta}, \beta^*)$. Similarly, the loss $\overline{\mathsf{H}}_s(\widehat{\beta}, \beta^*)$ only characterizes the estimation error of the set $\{j \in [p] : |\beta_j^*| > a\}$, and $\overline{\mathsf{H}}_s(\widehat{\beta}, \beta^*) = 0$ if and only if $\{j \in [p] : |\widehat{\beta}_j| > a\} = \{j \in [p] : |\beta_j^*| > a\}$.

2.3. *Parameter space.* For the nuisance parameter θ^* of the model (3), it is necessary that there exists a positive gap between θ_k^* and θ_{k+1}^* for the top-k set $\{i \in [n] : r_i^* \le k\}$ to be identifiable. We introduce a parameter space for this purpose. For any $0 \le \Delta \le \kappa$, define

$$\Theta(k, \Delta, \kappa) = \big\{\theta \in \mathbb{R}^n : \theta_1 \ge \cdots \ge \theta_n, \theta_k - \theta_{k+1} \ge \Delta, \theta_1 - \theta_n \le \kappa\big\}.$$

For any $\theta^* \in \Theta(k, \Delta, \kappa)$, a positive Δ guarantees that there is a separation between the group of top-k players and the rest. The number κ is called *dynamic range* of the problem.[1] This is a very important quantity, since it is closely related to the effective variance of the problem. Our results will give the exact dependence of the top-k ranking error on both Δ and κ.

2.4. *MLE and spectral method.* We study and compare the performances of two algorithms in the paper. The first algorithm is based on the maximum likelihood estimator (MLE). For each (i, j), we use the notation $\bar{y}_{ij} = \frac{1}{L}\sum_{l=1}^{L} y_{ijl}$. Throughout the paper, we adopt the convention of notation that $A_{ij} = A_{ji}$ and $\bar{y}_{ij} = 1 - \bar{y}_{ji}$. Then the negative log-likelihood function is given by

$$\ell_n(\theta) = \sum_{1 \le i < j \le n} A_{ij}\left[\bar{y}_{ij} \log \frac{1}{\psi(\theta_i - \theta_j)} + (1 - \bar{y}_{ij}) \log \frac{1}{1 - \psi(\theta_i - \theta_j)}\right]. \tag{5}$$

Define the MLE,

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta : \mathbb{1}_n^T \theta = 0} \ell_n(\theta). \tag{6}$$

It can be shown that $\widehat{\theta}$ is unique as long as the comparison graph is connected. Then set $\widehat{r}$ to be the rank of players based on $\widehat{\theta}$. In other words, find any $\widehat{r} \in \mathfrak{S}_n$ such that $\widehat{\theta}_{\widehat{\sigma}_1} \ge \cdots \ge \widehat{\theta}_{\widehat{\sigma}_n}$ is satisfied, where $\widehat{\sigma}$ is the inverse of $\widehat{r}$. We emphasize that the MLE (6) is written in its vanilla version, without any constraint or penalty. To the best of our knowledge, (6) has not been previously analyzed in the literature.

Another popular algorithm for ranking is the spectral method, also known as rank centrality proposed by [21]. Define a matrix $P \in \mathbb{R}^{n \times n}$ by

$$P_{ij} = \begin{cases} \dfrac{1}{d} A_{ij} \bar{y}_{ji}, & i \ne j, \\ 1 - \dfrac{1}{d} \displaystyle\sum_{l \in [n] \setminus \{i\}} A_{il} \bar{y}_{li}, & i = j, \end{cases} \tag{7}$$

where d needs to be at least the maximum degree of the random graph A. We just set $d = 2np$ throughout the paper. One can check that P is a transition matrix of a Markov chain. To see

[1] For readers who are familiar with [7], we note that our definitions of Δ and κ are slightly different from those in [7].

why P is useful, we can compute the conditional expectation of P given the random graph A,

$$P^*_{ij} = \begin{cases} \dfrac{1}{d} A_{ij} \psi(\theta^*_{r^*_j} - \theta^*_{r^*_i}), & i \neq j, \\ 1 - \dfrac{1}{d} \displaystyle\sum_{l \in [n] \setminus \{i\}} A_{il} \psi(\theta^*_{r^*_l} - \theta^*_{r^*_i}), & i = j. \end{cases}$$

The stationary distribution induced by the Markov chain P^* is

$$(\pi^*)^T = \left(\frac{\exp(\theta^*_{r^*_1})}{\sum_{i=1}^n \exp(\theta^*_{r^*_i})}, \ldots, \frac{\exp(\theta^*_{r^*_n})}{\sum_{i=1}^n \exp(\theta^*_{r^*_i})} \right).$$

One can easily check that $(\pi^*)^T P^* = (\pi^*)^T$. Since π^* preserves the order of $\{\theta^*_{r^*_i}\}$, the set with the k largest π^*_i's is the top-k group. With the sample version P, we can first compute its stationary distribution $\widehat{\pi}$, and then find any $\widehat{r} \in \mathfrak{S}_n$ such that $\widehat{\pi}_{\widehat{\sigma}_1} \geq \cdots \geq \widehat{\pi}_{\widehat{\sigma}_n}$, with $\widehat{\sigma}$ being the inverse of $\widehat{r}$.

3. Results for the MLE. We study the property of MLE in this section. Our first result gives theoretical guarantees for (6) under both ℓ_2 and ℓ_∞ loss functions.

THEOREM 3.1. *Assume $p \geq c_0 \frac{\log n}{n}$ for some sufficiently large constant $c_0 > 0$ and $\kappa \leq c_1$ for some constant $c_1 > 0$. Then, for the estimator $\widehat{\theta}$ defined by (6), we have*

$$\sum_{i=1}^n (\widehat{\theta}_i - \theta^*_{r^*_i})^2 \leq C \frac{1}{pL}, \tag{8}$$

$$\max_{i \in [n]} |\widehat{\theta}_i - \theta^*_{r^*_i}|^2 \leq C \frac{\log n}{npL}, \tag{9}$$

for some constant $C > 0$ only depending on c_1 with probability at least $1 - O(n^{-7})$ uniformly over all $r^* \in \mathfrak{S}_n$ and all $\theta^* \in \Theta(k, 0, \kappa)$ such that $\mathbb{1}_n^T \theta^* = 0$.

Let us give some comments on the assumptions and conclusions of Theorem 3.1. We have established that the MLE achieves the error rates $O(\frac{1}{pL})$ and $O(\frac{\log n}{npL})$ for the squared ℓ_2 loss and the squared ℓ_∞ loss, respectively. Both error rates are known to be optimal in the literature [7, 21]. Since the BTL model (3) is defined through pairwise differences of θ^*_i's, the model parameter is only identifiable up to a constant shift. We therefore require both $\mathbb{1}_n^T \widehat{\theta} = 0$ and $\mathbb{1}_n^T \theta^* = 0$ so that the two vectors are properly aligned. Note that the results for parameter estimation do not need a positive Δ, and we only assume $\theta^* \in \Theta(k, 0, \kappa)$. The condition $p \geq c_0 \frac{\log n}{n}$ is imposed for the random graph A to be well behaved in terms of both its degrees and the eigenvalues of the graph Laplacian. In fact, $p \gtrsim \frac{\log n}{n}$ is necessary to ensure the random graph is connected. Otherwise, ranking and parameter estimation would be impossible due to the identifiability issue caused by the lack of comparison between disconnected graph components. In the rest of the paper, some of the results will require a slightly stronger condition $\frac{np}{\log n} \to \infty$, but we will give very detailed remarks on when and why it will be needed. Last but not least, we require that the dynamic range κ to be bounded by a constant. One can certainly allow κ to tend to infinity, but the rates (8) and (9) would depend on κ exponentially [7, 21]. This is because the eigenvalues of the Hessian of the objective function of (6) will be exponentially small when κ diverges. In fact, when $\kappa \to \infty$, it is not clear whether MLE still leads to optimal error rates for parameter estimation. In this paper, we will

focus on the case $\kappa = O(1)$. We will see in later theorems that even with $\kappa = O(1)$, the exact value of κ still plays a fundamental role in top-k ranking.

To the best of our knowledge, Theorem 3.1 is the first result in the literature that gives optimal rates for parameter estimation by *vanilla* MLE under the BTL model. Previous results in the literature including [7, 8, 21] all work with regularized MLE,

$$\widehat{\theta}_\lambda = \underset{\theta:\mathbb{1}_n^T\theta=0}{\operatorname{argmin}}\left[\ell_n(\theta) + \frac{\lambda}{2}\|\theta\|^2\right]. \tag{10}$$

In particular, the recent paper [7] shows that $\widehat{\theta}_\lambda$ also achieves the optimal rates (8) and (9) for a λ that is chosen appropriately, though in practice it is known that the vanilla MLE performs very well. Theorem 3.1 shows that penalty is not needed for the MLE to be optimal, thus closing a gap between theory and practice.

The proof of Theorem 3.1 is built upon the elegant leave-one-out technique in [7]. We first show that with a sufficiently small λ, a (suboptimal) ℓ_∞ bound for $\widehat{\theta}_\lambda$ can be transferred to $\widehat{\theta}$. Then we apply a leave-one-out argument to derive the optimal rates (8) and (9). We also note that our leave-one-out argument is actually different from the form used in [7]. While the leave-one-out argument in [7] is applied together with a gradient descent analysis, we do not need to follow this gradient descent analysis because of the ℓ_∞ bound that has already been obtained. As a result, we are able to remove the additional technical assumption $\log L = O(\log n)$ that is imposed in [7]. A detailed analysis of the MLE will be given in Section 8.

Next, we study the theoretical property of $\widehat{r}$, the rank induced by the MLE $\widehat{\theta}$. Without loss of generality, let us assume $k \le \frac{n}{2}$ throughout the paper. The case $k > \frac{n}{2}$ can be dealt with by a symmetric bottom-k ranking problem. Before presenting the error bound for the loss function $\mathsf{H}_k(\widehat{r}, r^*)$, we need to introduce a few notation. We first define the effective variance of the MLE by

$$V(\kappa) = \max_{\substack{\kappa_1+\kappa_2\le\kappa \\ \kappa_1,\kappa_2\ge 0}} \frac{n}{k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2)}. \tag{11}$$

Recall that $\psi(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function so that $\psi'(t) = \psi(t)\psi(-t)$. Since $\kappa = O(1)$, we have $V(\kappa) \asymp 1$. Then the signal to noise ratio is defined by

$$\mathsf{SNR} = \frac{npL\Delta^2}{V(\kappa)}.$$

Note that SNR is a function of n, k, p, L, Δ, but we suppress the dependence for simplicity of notation. The following theorem shows that $\mathsf{H}_k(\widehat{r}, r^*)$ has an exponential rate with SNR appearing in the exponent.

THEOREM 3.2. *Assume $\frac{np}{\log n} \to \infty$ and $\kappa \le c_1$ for some constant $c_1 > 0$. Then, for the rank vector $\widehat{r}$ that is induced by the MLE (6), there exists some $\delta = o(1)$, such that*

$$\mathsf{H}_k(\widehat{r}, r^*) \le C\exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1-\delta)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1-\delta)\mathsf{SNR}}}\log\frac{n-k}{k}\right)_+^2\right), \tag{12}$$

for some constant $C > 0$ only depending on c_1 with probability $1 - o(1)$ uniformly over all $r^* \in \mathfrak{S}_n$ and all $\theta^* \in \Theta(k, \Delta, \kappa)$.

The error exponent of (12) is complicated. We present a special case of the bound when $k \asymp n$ to help understand the result.

COROLLARY 3.1. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$ *and* $k \asymp n$. *Then, as long as* $\mathsf{SNR} \to \infty$, *the rank vector* $\widehat{r}$ *induced by the MLE* (6) *satisfies*

$$\mathsf{H}_k(\widehat{r}, r^*) \le \exp\Big(-(1-o(1))\frac{\mathsf{SNR}}{8}\Big), \tag{13}$$

with probability $1-o(1)$ *uniformly over all* $r^* \in \mathfrak{S}_n$ *and all* $\theta^* \in \Theta(k, \Delta, \kappa)$.

Under the additional assumption $k \asymp n$, the top-k ranking problem can be viewed as a clustering or community detection problem, with the goal to divide the n players into two groups of sizes k and $n-k$, respectively. The exponential convergence rate (13) is in a typical form of optimal clustering error [17, 25]. It is intuitively clear that a larger SNR leads to a faster convergence rate. When the sizes of the two clusters are of different orders, one can obtain a more general convergence rate in the form of (12). The extra term $\log\frac{n-k}{k}$ characterizes the unbalancedness of the two clusters. We note that for variable selection under Hamming loss [4, 20], the optimal rate is very similar to the form of (12). This is because variable selection can also be thought of as clustering with two clusters of sizes s and $p-s$, whose orders can potentially be different.

Theorem 3.2 and Corollary 3.1 together reveal an interesting phenomenon for top-k ranking. The result shows that the top-k ranking problem can be very different for different orders of k. We note that in order to successfully identify the majority of the set $\{i \in [n] : r_i^* \le k\}$, we need to have $\mathsf{H}_k(\widehat{r}, r^*) \to 0$. When $k = n/4$, Corollary 3.1 shows that $\mathsf{H}_k(\widehat{r}, r^*) \to 0$ is achieved when $\mathsf{SNR} \to \infty$. In comparison, when $k = 5$, Theorem 3.2 shows that $\mathsf{H}_k(\widehat{r}, r^*) \to 0$ when $\mathsf{SNR} > (1+\epsilon)2\log n$ for some arbitrarily small constant $\epsilon > 0$. In other words, in terms of partial recovery consistency, top-quarter ranking is an easier problem than top-5 ranking. In general, a larger SNR is required for a smaller k according to the formula (12).

Compared with Theorem 3.1, we need a slightly stronger condition $\frac{np}{\log n} \to \infty$ for Theorem 3.2 and Corollary 3.1. If we only assume $p \ge c_0\frac{\log n}{n}$, the $1-\delta$ factor in the exponent of (12) can be replaced by $1-\epsilon$ with some ϵ of constant order. The constant ϵ can be made arbitrarily small as long as c_0 is sufficiently large.

The proof of Theorem 3.2 relies on a very interesting lemma that is stated below.

LEMMA 3.1. *Suppose* $\widehat{r}$ *is a rank vector induced by* $\widehat{\theta}$, *we then have*

$$\mathsf{H}_k(\widehat{r}, r^*) \le \frac{1}{k}\min_{t\in\mathbb{R}}\Big[\sum_{i:r_i^*\le k}\mathbb{I}\{\widehat{\theta}_i \le t\} + \sum_{i:r_i^*>k}\mathbb{I}\{\widehat{\theta}_i \ge t\}\Big].$$

The inequality holds for any $r^* \in \mathfrak{S}_n$.

We will prove Lemma 3.1 in Appendix E. This inequality shows that the error of ranking $\widehat{\theta}$ is bounded by the error of any thresholding rule. Using this result, we immediately obtain that

$$\mathbb{E}\mathsf{H}_k(\widehat{r}, r^*) \le \frac{1}{k}\min_{t\in\mathbb{R}}\Big[\sum_{i:r_i^*\le k}\mathbb{P}(\widehat{\theta}_i \le t) + \sum_{i:r_i^*>k}\mathbb{P}(\widehat{\theta}_i \ge t)\Big].$$

We then obtain the exponential error bound (12) by carefully analyzing the probability $\mathbb{P}(\widehat{\theta}_i \le t)$ (or $\mathbb{P}(\widehat{\theta}_i \ge t)$) for each $i \in [n]$. The analysis of $\mathbb{P}(\widehat{\theta}_i \le t)$ is quite involved. We need to first obtain a local linear expansion of the MLE at each coordinate, and then apply the leave-one-out technique introduced by [7] to decouple the dependence between the data and the coefficients of the local linear expansion. The details will be given in Section 8.

The result of Theorem 3.2 immediately implies a condition for exact recovery of the top-k set. By the definition of $\mathsf{H}_k(\widehat{r}, r^*)$, it is easy to see that

$$\mathsf{H}_k(\widehat{r}, r^*) \in \{0, (2k)^{-1}, 2(2k)^{-1}, 3(2k)^{-1}, \ldots, 1\}. \tag{14}$$

Then as long as $\mathsf{H}_k(\widehat{r}, r^*) < (2k)^{-1}$, we must have $\mathsf{H}_k(\widehat{r}, r^*) = 0$. Under the condition that the right-hand side of (12) is smaller than $(2k)^{-1}$, we obtain exact recovery of the top-k set. This result is stated as follows.

THEOREM 3.3. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, *and*

$$\frac{npL\Delta^2}{V(\kappa)} > (1+\epsilon)2\big(\sqrt{\log k} + \sqrt{\log(n-k)}\big)^2, \tag{15}$$

for some arbitrarily small constant $\epsilon > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the MLE* (6), *we have* $\mathsf{H}_k(\widehat{r}, r^*) = 0$ *with probability* $1 - o(1)$ *uniformly over all* $r^* \in \mathfrak{S}_n$ *and all* $\theta^* \in \Theta(k, \Delta, \kappa)$.

We remark that the condition $\frac{np}{\log n} \to \infty$ can be relaxed to $p \geq c_0 \frac{\log n}{n}$ for a sufficiently large constant c_0 without affecting the conclusion of Theorem 3.3. The result of Theorem 3.3 improves the exact recovery threshold obtained in the literature. The paper [7] proves that the MLE exactly recovers the top-k set when $npL\Delta^2 > C(\sqrt{\log k} + \sqrt{\log(n-k)})^2$ for some sufficiently large constant $C > 0$. We complement the result of [7] by showing that the leading constant should be $2V(\kappa)$, an increasing function of the dynamic range κ. Moreover, the symmetry of k and $n-k$ in (15) agrees with the understanding that top-k ranking and bottom-k ranking are mathematically equivalent.

The next theorem shows that the exact recovery threshold (15) is optimal, and cannot be further improved.

THEOREM 3.4. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, $(\log n)^8 = O(L)$ *and*

$$\frac{npL\Delta^2}{V(\kappa)} < (1-\epsilon)2\big(\sqrt{\log k} + \sqrt{\log(n-k)}\big)^2, \tag{16}$$

for some arbitrarily small constant $\epsilon > 0$. *Then we have*

$$\liminf_{n\to\infty} \inf_{\widehat{r}} \sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta(k,\Delta,\kappa)}} \mathbb{P}_{(\theta^*, r^*)}\big(\mathsf{H}_k(\widehat{r}, r^*) > 0\big) \geq 0.95,$$

where we use the notation $\mathbb{P}_{(\theta^*, r^*)}$ *for the data generating process* (3).

The proof of Theorem 3.4 relies on a precise lower bound characterization of the maximum of dependent binomial random variables. The extra assumption $L \gtrsim (\log n)^8$ allows us to apply a high-dimensional central limit theorem [9] for this purpose. Without this additional technical condition, we are not aware of any probabilistic tool to deal with maximum of dependent binomial random variables.

Theorem 3.3 and Theorem 3.4 together nail down the phase transition boundary of exact recovery, which is $\frac{npL\Delta^2}{V(\kappa)} = 2(\sqrt{\log k} + \sqrt{\log(n-k)})^2$. Thus, the MLE is an optimal procedure that achieves this boundary. The lower bound result of Theorem 3.4 also suggests that the partial recovery error rate obtained in Theorem 3.2 cannot be improved, since otherwise one would obtain a better SNR condition for exact recovery in Theorem 3.3. A rigorous minimax lower bound for partial recovery will be given in Section 6.

4. Results for the spectral method. In this section, we study the theoretical property of the spectral method, also known as rank centrality [21]. Let $\widehat{\pi}$ be the stationary distribution of the Markov chain with transition probability (7). The estimation error of $\widehat{\pi}$ has already been investigated by [7, 21]. For both ℓ_2 and ℓ_∞ loss functions, it has been shown by [7] that $\widehat{\pi}$ achieves the optimal rates (8) and (9) after an appropriate scaling. We therefore directly study the accuracy of the rank vector $\widehat{r}$ induced by $\widehat{\pi}$. This is where we can see the difference between the MLE and the spectral method.

We first define the effective variance of the spectral method,

$$\overline{V}(\kappa) = \max_{\substack{\kappa_1+\kappa_2\le\kappa \\ \kappa_1,\kappa_2\ge 0}} \frac{k\psi'(\kappa_1)(1+e^{\kappa_1})^2 + (n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2}{(k\psi(\kappa_1)+(n-k)\psi(-\kappa_2))^2/n}. \tag{17}$$

Note that $\overline{V}(\kappa) \asymp 1$ when $\kappa = O(1)$. The signal to noise ratio is defined by

$$\overline{\mathsf{SNR}} = \frac{npL\Delta^2}{\overline{V}(\kappa)}.$$

The error rate of the spectral method with respect to $\mathsf{H}_k(\widehat{r}, r^*)$ is stated as follows.

THEOREM 4.1. *Assume* $\frac{np}{\log n} \to \infty$ *and* $\kappa \le c_1$ *for some constant* $c_1 > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the stationary distribution of the Markov chain* (7), *there exists some* $\delta = o(1)$, *such that*

$$\mathsf{H}_k(\widehat{r}, r^*) \le C\exp\biggl(-\frac{1}{2}\biggl(\frac{\sqrt{(1-\delta)\overline{\mathsf{SNR}}}}{2} - \frac{1}{\sqrt{(1-\delta)\overline{\mathsf{SNR}}}}\log\frac{n-k}{k}\biggr)_+^2\biggr), \tag{18}$$

for some constant $C > 0$ *only depending on* c_1 *with probability* $1 - o(1)$ *uniformly over all* $r^* \in \mathfrak{S}_n$ *and all* $\theta^* \in \Theta(k, \Delta, \kappa)$.

The formula (18) characterizes the convergence rate of partial recovery of the top-k set by the spectral method. It can be compared with the MLE error bound (12). The only difference lies in the effective variance of the two methods. We will show in Lemma 5.1 that $\overline{V}(\kappa) \ge V(\kappa)$ and the equality only holds when $\kappa = 0$. Therefore, the spectral method is not optimal in general. Detailed comparisons of the two algorithms will be given in Section 5.

By the property (14), we immediately obtain an exact recovery result from Theorem 4.1.

THEOREM 4.2. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, *and*

$$\frac{npL\Delta^2}{\overline{V}(\kappa)} > (1+\epsilon)2\bigl(\sqrt{\log k} + \sqrt{\log(n-k)}\bigr)^2, \tag{19}$$

for some arbitrarily small constant $\epsilon > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the stationary distribution of the Markov chain* (7), *we have* $\mathsf{H}_k(\widehat{r}, r^*) = 0$ *with probability* $1 - o(1)$ *uniformly over all* $r^* \in \mathfrak{S}_n$ *and all* $\theta^* \in \Theta(k, \Delta, \kappa)$.

It has been shown in [7] that the spectral method exactly recovers the top-k set when $npL\Delta^2 > C(\sqrt{\log k} + \sqrt{\log(n-k)})^2$ for some sufficiently large constant $C > 0$. Without specifying the constant C, one cannot tell the difference between the MLE and the spectral method. In view of the lower bound result given by Theorem 3.4, the exact recovery threshold (19) of the spectral method does not achieve the phase transition boundary for a general κ. A careful reader may wonder whether this is resulted from a loose analysis in the proof. Our next result shows that the suboptimality of the spectral method is intrinsic.

THEOREM 4.3. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa \le c_1$ *for some constant* $c_1 > 0$, $k \to \infty$ *and*

$$\frac{npL\Delta^2}{\overline{V}(\kappa)} < (1-\epsilon)2\big(\sqrt{\log k} + \sqrt{\log(n-k)}\big)^2, \tag{20}$$

for some arbitrarily small constant $\epsilon > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the stationary distribution of the Markov chain* (7), *we have*

$$\liminf_{n\to\infty} \sup_{\substack{r^*\in\mathfrak{S}_n \\ \theta^*\in\Theta(k,\Delta,\kappa)}} \mathbb{P}_{(\theta^*,r^*)}\big(\mathsf{H}_k(\widehat{r}, r^*) > 0\big) \ge 0.95.$$

Moreover, there exists some $\delta = o(1)$, *such that*

$$\begin{aligned} &\sup_{\substack{r^*\in\mathfrak{S}_n \\ \theta^*\in\Theta(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r}, r^*) \\ &\ge C\exp\bigg(-\frac{1}{2}\bigg(\frac{\sqrt{(1+\delta)\overline{\mathsf{SNR}}}}{2} - \frac{1}{\sqrt{(1+\delta)\overline{\mathsf{SNR}}}}\log\frac{n-k}{k}\bigg)_+^2\bigg), \end{aligned} \tag{21}$$

for some constant $C > 0$ *only depending on* c_1 *and* ϵ.

Theorem 4.3 shows that the results of Theorem 4.1 and Theorem 4.2 on the performance of spectral method are sharp, under the additional condition that $k \to \infty$. The conclusion of Theorem 4.3 can also be extended to the case of $k = O(1)$ via a similar argument that is used in the proof of Theorem 3.4, as long as the technical condition $(\log n)^8 = O(np)$ is further imposed.

To close this section, we remark that all the theorems we have obtained for the spectral method can be stated under the weaker assumption $p \ge c_0\frac{\log n}{n}$ for some sufficiently large constant $c_0 > 0$, as long as the δ in (18) and (21) are replaced by some sufficiently small constant.

5. Comparison of the two methods. In this section, we compare the MLE and the spectral method based on the results obtained in Section 3 and Section 4. The statistical properties of the two methods in terms of partial and exact recovery are characterized by the two variance functions $V(\kappa)$ and $\overline{V}(\kappa)$, respectively. We first give a direct comparison of the two functions by plotting them together with different values of k/n. We observe in Figure 1 that $\overline{V}(\kappa) \ge V(\kappa)$ for all $\kappa \ge 0$. This inequality is rigorously established by the following lemma.

LEMMA 5.1. *For* $V(\kappa)$ *and* $\overline{V}(\kappa)$ *defined in* (11) *and* (17), *respectively, we have*

$$\overline{V}(\kappa) \ge V(\kappa),$$

for all $\kappa \ge 0$. *Moreover, the equality holds if and only if* $\kappa = 0$.

PROOF. By Jensen's inequality, we have

$$\frac{k\frac{e^{\kappa_1}}{(1+e^{\kappa_1})^2} + (n-k)\frac{e^{-\kappa_2}}{(1+e^{-\kappa_2})^2}}{ke^{\kappa_1} + (n-k)e^{-\kappa_2}} \ge \bigg(\frac{k\frac{e^{\kappa_1}}{1+e^{\kappa_1}} + (n-k)\frac{e^{-\kappa_2}}{1+e^{-\kappa_2}}}{ke^{\kappa_1} + (n-k)e^{-\kappa_2}}\bigg)^2. \tag{22}$$

Another way to see the above inequality is to construct a random variable X such that $\mathbb{P}(X = \frac{1}{1+e^{\kappa_1}}) = \frac{ke^{\kappa_1}}{ke^{\kappa_1}+(n-k)e^{-\kappa_2}}$ and $\mathbb{P}(X = \frac{1}{1+e^{-\kappa_2}}) = \frac{(n-k)e^{-\kappa_2}}{ke^{\kappa_1}+(n-k)e^{-\kappa_2}}$. Then (22) is equivalent to $\mathbb{E}X^2 \ge (\mathbb{E}X)^2$. The inequality (22) can be rearranged into

$$\frac{k\psi'(\kappa_1)(1+e^{\kappa_1})^2 + (n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2}{(k\psi(\kappa_1) + (n-k)\psi(-\kappa_2))^2/n} \ge \frac{n}{k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2)}. \tag{23}$$

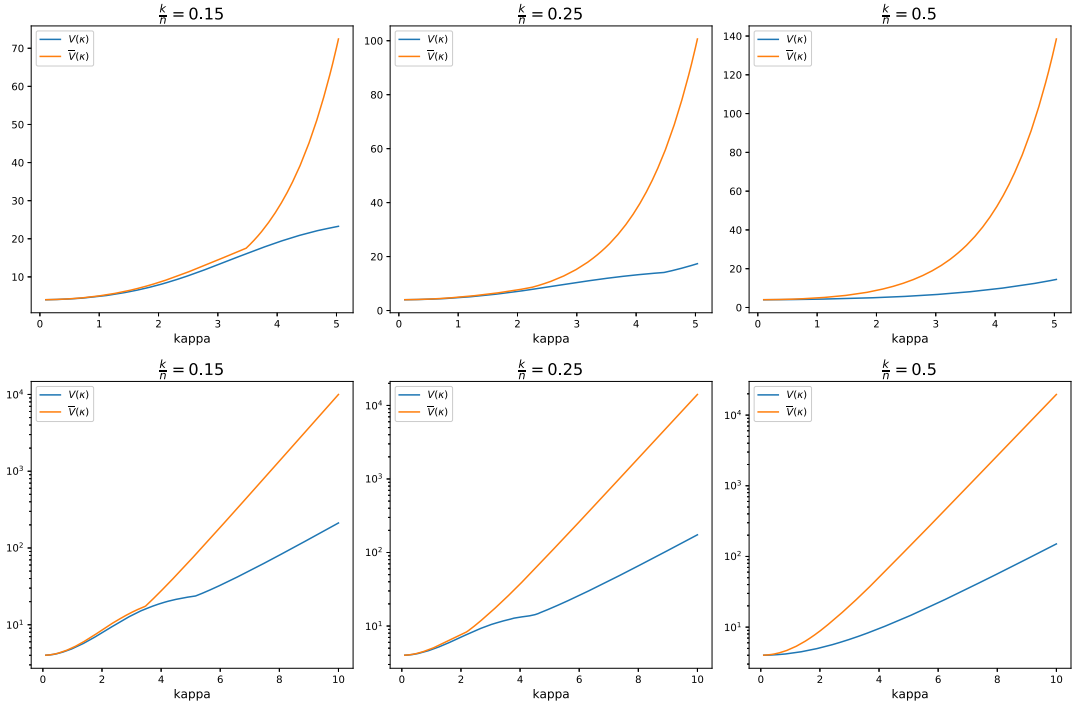

FIG. 1. *The functions $V(\kappa)$ and $\overline{V}(\kappa)$ with $k/n \in \{0.15, 0.25, 0.5\}$. In the first row, we plot the functions for $\kappa \in [0, 5]$. The second row plots the same functions for $\kappa \in [0, 10]$ in a logarithmic scale to better illustrate the global structure. It is very interesting that both $V(\kappa)$ and $\overline{V}(\kappa)$ have a point at which the derivative is not continuous. Before this critical point, the optimization of $V(\kappa)$ is achieved by $(\kappa_1^*, \kappa_2^*) = (0, \kappa)$. Right after the critical point, κ_1^* is immediately bounded away from 0 and κ_2^* is immediately bounded away from κ. The same property also holds for $\overline{V}(\kappa)$. Moreover, the critical point occurs earlier as k/n becomes larger (when $k/n \le 1/2$).*

Taking maximum over κ_1 and κ_2 on both sides, we obtain the inequality $\overline{V}(\kappa) \ge V(\kappa)$. When $\kappa = 0$, we obviously have $V(\kappa) = \overline{V}(\kappa)$. When $\kappa > 0$, we need to show $V(\kappa) \ne \overline{V}(\kappa)$. The optimization of $V(\kappa)$ must be achieved by some $(\kappa_1^*, \kappa_2^*) \ne (0, 0)$. For such (κ_1^*, κ_2^*), the constructed random variable X has a positive variance, and thus both inequalities (22) and (23) are strict. We then have

$$\begin{aligned}\overline{V}(\kappa) &\ge \frac{k\psi'(\kappa_1^*)(1+e^{\kappa_1^*})^2 + (n-k)\psi'(\kappa_2^*)(1+e^{-\kappa_2^*})^2}{(k\psi(\kappa_1^*) + (n-k)\psi(-\kappa_2^*))^2/n} \\ &> \frac{n}{k\psi'(\kappa_1^*) + (n-k)\psi'(\kappa_2^*)} \\ &= V(\kappa).\end{aligned}$$

The proof is complete. □

The comparison between $V(\kappa)$ and $\overline{V}(\kappa)$ shows that the spectral method is not optimal in general. It has a worse error exponent for partial recovery and requires a larger signal to noise ratio threshold for exact recovery. In fact, the difference $\overline{V}(\kappa) - V(\kappa)$ eventually grows exponentially fast as a function of κ. See Figure 1.

Note that both $V(\kappa)$ and $\overline{V}(\kappa)$ are the worst-case effective variances with respect to the parameter space $\Theta(k, \Delta, \kappa)$ for the two algorithms. In Section 7, we will further show that the MLE outperforms the spectral method for each $\theta^* \in \Theta(k, \Delta, \kappa)$. This conclusion is supported by extensive numerical experiments. We set $n = 200$, $p = 0.25$, $L = 20$ and $k = 50$ throughout the experiments.

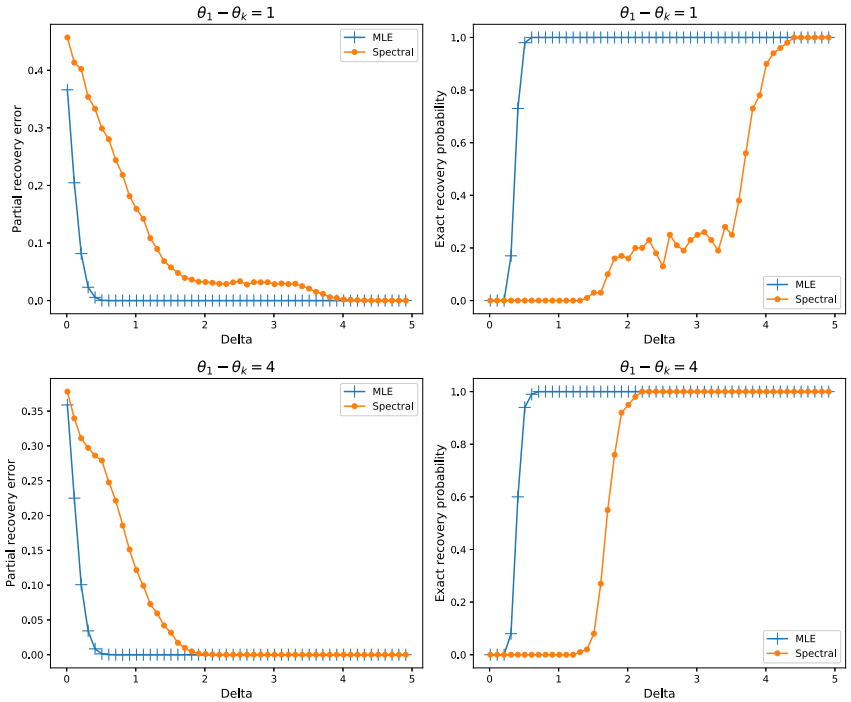

FIG. 2. *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is chosen to be a piecewise constant vector of four pieces of sizes* 25, 25, 75, 75. *The plots are obtained by averaging* 100 *independent experiments.*

In our first experiment, we consider $\theta^* \in \mathbb{R}^n$ that has four pieces, with the three change-points located at $\{25, 50, 200\}$. The values of the four pieces are set as 10, $10-\tau$, $10-\tau-\Delta$ and 0, respectively, where $\tau = \theta_1^* - \theta_k^* \in \{1, 4\}$ and Δ is varied from 0.01 to 5. We apply both the MLE and the spectral method to the data. Figure 2 shows the results for both partial and exact recovery. We observe that the MLE consistently outperforms the spectral method.

In the second experiment, we consider $\theta^* \in \mathbb{R}^n$ that has four pieces, with the three change-points located at $\{50(1-\rho), 50, 50+150\rho\}$. The values of the four pieces are set as 10, 6, $6-\Delta$ and 0, respectively. The parameter ρ is chosen in $\{0.1, 0.5, 0.9\}$ and Δ is varied from 0.01 to 3. The performance of the two methods for partial and exact recovery are plotted in Figure 3. Again, the MLE always outperforms the spectral method.

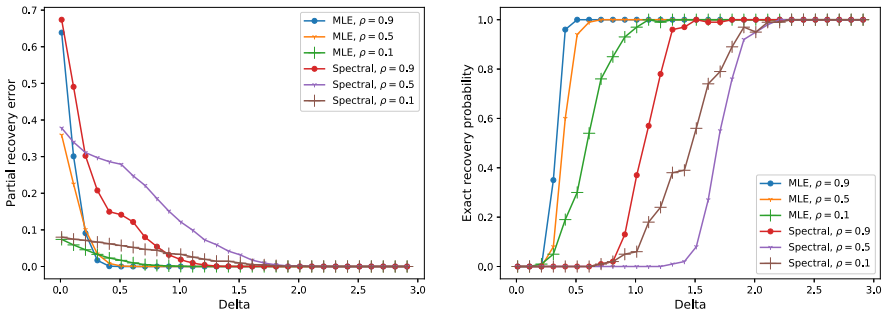

FIG. 3. *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is chosen to be a piecewise constant vector of four pieces of sizes* $50(1-\rho)$, 50ρ, $150(1-\rho)$, 150ρ. *The plots are obtained by averaging* 100 *independent experiments.*

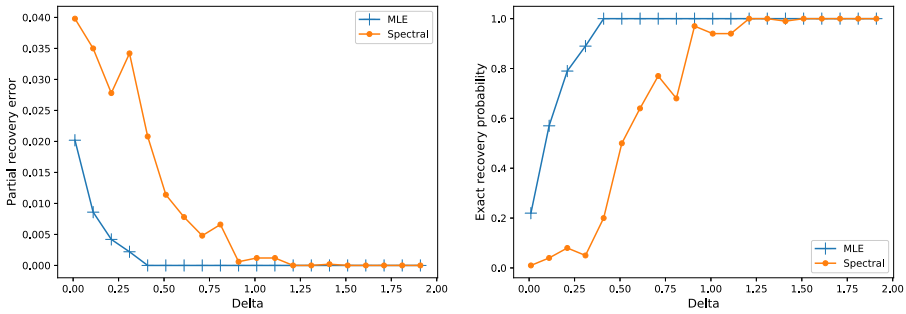

FIG. 4. *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is randomly generated from some distribution. The plots are obtained by averaging* 100 *independent experiments.*

Next, we consider a $\theta^* \in \mathbb{R}^n$ that has a more complicated structure. We fix $\theta_1^* = 10$, $\theta_{200}^* = 0$, generate $\theta_2^*, \ldots, \theta_{50}^*$ from Uniform[6, 10] and generate $\theta_{51}^*, \ldots, \theta_{199}^*$ from Uniform$[0, 6 - \Delta]$, and we vary Δ from 0.01 to 2. We find even for such randomly generated θ^*'s, the MLE always outperforms the spectral method. The results are summarized in Figure 4 for both partial and exact recovery.

In summary, we are able to confirm that the MLE is a much better algorithm than the spectral method under various scenarios. Our results complement the analysis in [7]. It is claimed in [7] that both the MLE and the spectral method are optimal in terms of the order of the exact recovery threshold. In addition, the paper conducts a very curious numerical experiment that shows the performances of the MLE and the spectral method are nearly identical. We note that the θ^* chosen in the numerical experiment of [7] is a piecewise constant vector with only two pieces. We will explain why this choice leads to nearly identical performances of the two algorithms. Let us first conduct a similar experiment to replicate this conclusion. We continue to use the setting $n = 200$, $p = 0.25$, $L = 20$ and $k = 50$. Then choose θ^* such that $\theta_1^* = \cdots = \theta_{50}^* = \Delta$ and $\theta_{51}^* = \cdots = \theta_{200}^* = 0$. Figure 5 plots the results of partial and exact recovery with Δ varied from 0.01 to 0.55. For both partial recovery and exact recovery, the results are indeed nearly identical for the two algorithms. This phenomenon can be easily explained by our theory. For $\theta^* \in \Theta(k, \Delta, \kappa)$ with only two pieces, we must have $\kappa = \Delta$. When $\Delta = o(1)$, we have $V(\kappa) = (1 + o(1))V(0)$ and $\overline{V}(\kappa) = (1 + o(1))\overline{V}(0)$. This leads to the relation $\overline{V}(\kappa) = (1 + o(1))V(\kappa)$, and thus the spectral method has the same asymptotic error exponent for partial recovery and achieves the optimal phase transition boundary for exact recovery. When Δ does not tend to zero but of a constant order, we have $\overline{\mathsf{SNR}} \gtrsim npL \gg \log n$,

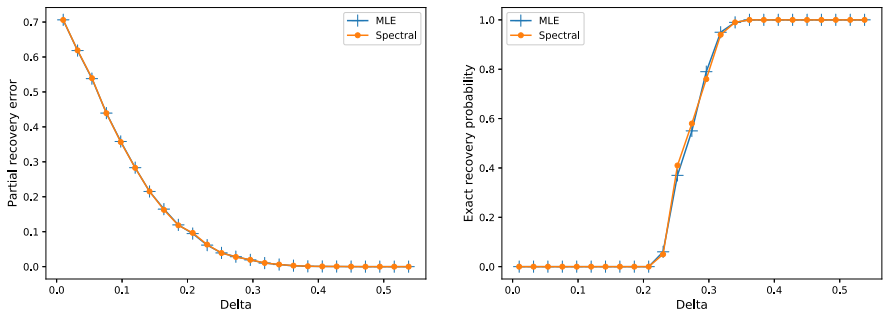

FIG. 5. *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is chosen to be a piecewise constant vector of two pieces of sizes* 50 *and* 150. *The plots are obtained by averaging* 100 *independent experiments.*

and the error bound (18) already leads to exact recovery because of the large value of $\overline{\mathsf{SNR}}$. In either case, the spectral method is optimal. Let us summarize the optimality of the spectral method under this special situation by the following corollary.

COROLLARY 5.1. *Assume* $\frac{np}{\log n} \to \infty$ *and* $\kappa = \Delta \le c_1$ *for some constant* $c_1 > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the stationary distribution of the Markov chain* (7), *there exists some* $\delta = o(1)$, *such that*

$$\mathsf{H}_k(\widehat{r}, r^*) \le C \exp\biggl(-\frac{1}{2}\biggl(\frac{\sqrt{(1-\delta)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1-\delta)\mathsf{SNR}}} \log \frac{n-k}{k}\biggr)_+^2\biggr),$$

for some constant $C > 0$ *only depending on* c_1 *with probability* $1 - o(1)$ *uniformly over all* $r^* \in \mathfrak{S}_n$ *and all* $\theta^* \in \Theta(k, \Delta, \Delta)$. *Moreover, as long as*

$$\frac{npL\Delta^2}{V(\kappa)} > (1+\epsilon)2\bigl(\sqrt{\log k} + \sqrt{\log(n-k)}\bigr)^2,$$

for some arbitrarily small constant $\epsilon > 0$. *Then* $\mathsf{H}_k(\widehat{r}, r^*) = 0$ *with probability* $1 - o(1)$ *uniformly over all* $r^* \in \mathfrak{S}_n$ *and all* $\theta^* \in \Theta(k, \Delta, \Delta)$.

To close this section, we remark that according to the equality condition of Lemma 5.1, the two-piece θ^*, or equivalently $\kappa = \Delta$, is essentially the only situation where the spectral method is optimal and performs as well as the MLE. Moreover, since both functions $V(\kappa)$ are $\overline{V}(\kappa)$ are increasing, the setting with $\kappa = \Delta$ leads to the smallest effective variance, and thus provides the two algorithms with the most favorable scenario.

6. Minimax lower bound of partial recovery. The purpose of this section is to show that the partial recovery error rate (12) achieved by the MLE cannot be improved from a minimax perspective. We are able to establish a matching lower bound for Theorem 3.2 using a slightly more general parameter space. Define

$$\Theta'(k, \Delta, \kappa) = \bigl\{\theta \in \mathbb{R}^n : \theta_1 \ge \cdots \ge \theta_n, \theta_k - \theta_{k+2} \ge \Delta, \theta_1 - \theta_n \le \kappa\bigr\}. \tag{24}$$

Compared with $\Theta(k, \Delta, \kappa)$, the new definition (24) imposes a gap between θ_k and θ_{k+2}. It is clear that $\Theta(k, \Delta, \kappa) \subset \Theta'(k, \Delta, \kappa)$, and the only difference of $\Theta'(k, \Delta, \kappa)$ is the ambiguity of θ_{k+1}. The player ranked at the $(k+1)$th position does not necessarily has a gap from either the top group or the bottom group. Though this additional uncertainty clearly better models scenarios in many real applications of top-k ranking, the main reason we adopt the slightly larger parameter space is to have a clean lower bound analysis. Directly establishing a lower bound for $\Theta(k, \Delta, \kappa)$ is still possible, but it requires some additional technical assumptions that make the problem unnecessarily involved.

Throughout this section, we assume that (16) holds. This is the regime of partial recovery, since exact recovery is impossible by Theorem 3.4. We first remark that with a slight modification of the proof of Theorem 3.2, the MLE can be shown to achieve the same error rate (12) over the parameter space $\Theta'(k, \Delta, \kappa)$ as well. Thus, the space $\Theta'(k, \Delta, \kappa)$ does not increase the statistical complexity of the problem.

Our lower bound analysis is based on the two least favorable vectors $\theta', \theta'' \in \Theta'(k, \Delta, \kappa)$. They are constructed as follows. Let $\rho = o(1)$ be a vanishing sequence that tends to zero with a sufficiently slow rate. We define κ_1^* and κ_2^* such that the optimization (11) is achieved at $(\kappa_1, \kappa_2) = (\kappa_1^*, \kappa_2^*)$. Then define $\theta_i' = \kappa_1^*$ for $1 \le i \le k - \rho k$, $\theta_i' = 0$ for $k - \rho k < i \le k$, $\theta_i' = -\Delta$ for $k < i \le k + \rho(n-k)$ and $\theta_i' = -\kappa_2^*$ for $k + \rho(n-k) < i \le n$. For θ'', we let

$\theta''_i = \theta'_i$ for all $i \in [n]\backslash\{k+1\}$ and set $\theta''_{k+1} = 0$. We will show that there exist $r', r'' \in \mathfrak{S}_n$, so that the hardness of top-k ranking is characterized by an optimal testing problem,

$$\inf_{0\le\phi\le 1}\left[\mathbb{E}_{(\theta'',r'')}\phi + \frac{n-k-1}{k}\mathbb{E}_{(\theta',r')}(1-\phi)\right]. \tag{25}$$

Moreover, there exists some $i \in [n]$, such that the two rank vectors r', r'' satisfy $\theta'_{r'_j} = \theta''_{r''_j}$ for all $j \in [n]\backslash\{i\}$. For the ith entry, we have $\theta'_{r'_i} = -\Delta$ and $\theta''_{r''_i} = 0$. The reduction of the top-k ranking problem to the testing problem (25) is the most important step in our lower bound analysis. A rigorous argument will be given in Appendix B.

The testing problem (25) can be roughly understood as to test whether the ith player belongs to the top-k set or not. The two hypotheses receive different weights 1 and $\frac{n-k-1}{k}$ because of the definition of the loss function $\mathsf{H}_k(\widehat{r}, r^*)$. The optimal procedure to (25) is given by the likelihood ratio test

$$\phi = \mathbb{I}\left\{\frac{d\mathbb{P}_{(\theta',r')}}{d\mathbb{P}_{(\theta'',r'')}} \ge \frac{k}{n-k-1}\right\},$$

according to Neyman–Pearson lemma. Since the vectors $\{\theta'_{r'_i}\}_{i\in[n]}$ and $\{\theta''_{r''_i}\}_{i\in[n]}$ only differ at the ith entry, the likelihood ratio statistic only depends on $\{\bar{y}_{ij}\}_{j\in[n]\backslash\{i\}}$ and $\{A_{ij}\}_{j\in[n]\backslash\{i\}}$. Therefore, the testing error (25) is relatively easy to quantify. A sharp lower bound can be obtained by a large deviation analysis.

THEOREM 6.1. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa \le c_1$ *for some constant* $c_1 > 0$, *and* (16) *holds for some arbitrarily small constant* $\epsilon > 0$. *Then there exists some* $\delta = o(1)$, *such that*

$$\inf_{\widehat{r}} \sup_{\substack{r^*\in\mathfrak{S}_n \\ \theta^*\in\Theta'(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r}, r^*)$$
$$\ge C\exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1+\delta)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta)\mathsf{SNR}}}\log\frac{n-k}{k}\right)_+^2\right),$$

for some constant $C > 0$ *only depending on* c_1 *and* ϵ.

7. Local error rates. So far, our study of the top-k ranking problem has been conducted under the minimax decision-theoretic framework laid out in Section 2. The upper and lower bounds for the MLE and the spectral method are established uniformly over the parameter space $\Theta(k, \Delta, \kappa)$. To complement the minimax results, in this section, we present local error rates for the MLE and the spectral method, which leads to a refined comparison between the two popular methods.

7.1. *Local error rate for the MLE.* To analyze the statistical property of the MLE for each individual θ, we first need to generalize the effective variance (11). For any $\theta \in \mathbb{R}^n$ and any $i \in [n]$, define

$$V_i(\theta) = \frac{n}{\sum_{j=1}^n \psi'(\theta_i - \theta_j)}. \tag{26}$$

With the help of $V_i(\theta)$, for any subset $S \subset [n]$, we also define

$$R_1(S, \theta, t, \delta) = \sum_{i\in S}\exp\left(-\frac{(1-\delta)(\theta_i - t)_+^2 npL}{2V_i(\theta)}\right), \tag{27}$$

$$R_2(S, \theta, t, \delta) = \sum_{i\in S}\exp\left(-\frac{(1-\delta)(t - \theta_i)_+^2 npL}{2V_i(\theta)}\right). \tag{28}$$

THEOREM 7.1. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa \le c_1$ *for some constant* $c_1 > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the MLE* (6), *any small constant* $0 < \delta < 0.1$, *any* $r^* \in \mathfrak{S}_n$ *and any* $\theta^* \in \Theta(k, 0, \kappa)$, *we have*

$$\mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \le C_1 \Big(\inf_t \frac{R_1([k], \theta^*, t, \delta) + R_2([n]\backslash[k], \theta^*, t, \delta)}{k} + n^{-3} \Big), \tag{29}$$

where $C_1 > 0$ *is a constant only depending on* c_1 *and* δ. *Moreover, we also have*

$$\mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \ge C_2 \Big(\inf_t \frac{R_1([k], \theta^*, t, -\delta) + R_2([n]\backslash[k], \theta^*, t, -\delta)}{k} \Big), \tag{30}$$

for some constant $C_2 > 0$ *only depending on* c_1 *and* δ, *if we additionally assume that* $\inf_t (R_1([k], \theta^*, t, -\delta) + R_2([n]\backslash[k], \theta^*, t, -\delta)) \to \infty$.

Theorem 7.1 gives matching upper and lower bounds for the error of the MLE for each individual $\theta^* \in \Theta(k, 0, \kappa)$ and $r^* \in \mathfrak{S}_n$, except for the additional n^{-3} term and an arbitrarily small δ. We remark that the n^{-3} term in the upper bound can be replaced by n^{-C} for an arbitrarily large constant C. The upper bound (29) can be viewed as an extension of Theorem 3.2, though the δ in Theorem 3.2 is allowed to vanish because of the less general setting. The lower bound (30) requires an extra condition $\inf_t (R_1([k], \theta^*, t, -\delta) + R_2([n]\backslash[k], \theta^*, t, -\delta)) \to \infty$, which implies the error rate is of higher order than $O(k^{-1})$. It plays the same role as the condition (20) in Theorem 4.3. This assumption covers most interesting partial recovery cases, since $O(k^{-1})$ is already the error rate of exact recovery.

Let t^* be a minimizer of the right-hand side of (29) or (30). Then we can interpret $\sum_{i=1}^k \exp(-\frac{(\theta_i^* - t^*)_+^2 npL}{2V_i(\theta^*)})$ as the order of the number of top k players that are ranked among the bottom group, and $\sum_{i=k+1}^n \exp(-\frac{(t^* - \theta_i^*)_+^2 npL}{2V_i(\theta^*)})$ as the order of the number of bottom $n-k$ players that are ranked in the top group.

A careful reader may notice that the error rate in Theorem 7.1 does not have a clear dependence on the signal gap $\theta_k^* - \theta_{k+1}^*$. This is because the current error rate depends on θ^* more explicitly rather than just the difference between θ_k^* and θ_{k+1}^*. Even when $\theta_k^* = \theta_{k+1}^*$, it is still possible that the right-hand side of (29) converges to zero as long as the majority of $\{\theta_i^*\}_{1 \le i \le k}$ are separated from most of $\{\theta_i^*\}_{k+1 \le i \le n}$.

7.2. *Local error rate for the spectral method.* To present a similar local error rate for the spectral method, we also need to generalize the effective variance (17). For any $\theta \in \mathbb{R}^n$ and any $i \in [n]$, define

$$\overline{V}_i(\theta) = \frac{n \sum_{j=1}^n \psi'(\theta_i - \theta_j)(1 + e^{\theta_j - \theta_i})^2}{(\sum_{j=1}^n \psi(\theta_j - \theta_i))^2}. \tag{31}$$

We also introduce two quantities similar to (27) and (28),

$$\overline{R}_1(S, \theta, t, \delta) = \sum_{i \in S} \exp\Big(-\frac{(1-\delta)(\theta_i - t)_+^2 npL}{2\overline{V}_i(\theta)} \Big),$$

$$\overline{R}_2(S, \theta, t, \delta) = \sum_{i \in S} \exp\Big(-\frac{(1-\delta)(t - \theta_i)_+^2 npL}{2\overline{V}_i(\theta)} \Big).$$

THEOREM 7.2. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa \le c_1$ *for some constant* $c_1 > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the stationary distribution of the Markov chain* (7), *any small*

constant $0 < \delta < 0.1$, *any* $r^* \in \mathfrak{S}_n$ *and any* $\theta^* \in \Theta(k, 0, \kappa)$, *we have*

$$\mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \le C_1 \biggl(\inf_t \frac{\overline{R}_1([k], \theta^*, t, \delta) + \overline{R}_2([n]\backslash[k], \theta^*, t, \delta)}{k} + n^{-3} \biggr), \tag{32}$$

where $C_1 > 0$ *is a constant only depending on* c_1 *and* δ. *Moreover, we also have*

$$\mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \ge C_2 \biggl(\inf_t \frac{\overline{R}_1([k], \theta^*, t, -\delta) + \overline{R}_2([n]\backslash[k], \theta^*, t, -\delta)}{k} \biggr), \tag{33}$$

for some constant $C_2 > 0$ *only depending on* c_1 *and* δ, *if we additionally assume that* $\inf_t (\overline{R}_1([k], \theta^*, t, -\delta) + \overline{R}_2([n]\backslash[k], \theta^*, t, -\delta)) \to \infty$.

Similar to Theorem 7.1, Theorem 7.2 also gives matching upper and lower bounds for the error of the spectral method for each individual $\theta^* \in \Theta(k, 0, \kappa)$ and $r^* \in \mathfrak{S}_n$.

Let us remark that the results of Theorem 7.1 and Theorem 7.2 can be further extended beyond the setting of the Erdős–Rényi graph and exactly L comparisons on each edge. To be specific, we can consider a random graph $A_{ij} \sim \text{Bernoulli}(p_{ij})$ independently for all $1 \le i < j \le n$. For each edge, we observe L_{ij} independent games. Then, as long as $\max_{ij} p_{ij} \le C \min_{ij} p_{ij}$ and $\max_{ij} L_{ij} \le C \min_{ij} L_{ij}$ hold for some constant $C > 0$, the results of Theorem 7.1 and Theorem 7.2 continue to hold with $\frac{V_i(\theta)}{npL}$ and $\frac{\overline{V}_i(\theta)}{npL}$ replaced by $\frac{\sum_{j\in[n]\backslash\{i\}} \frac{p_{ij}}{L_{ij}} \psi(\theta_i - \theta_j)\psi(\theta_j - \theta_i)}{(\sum_{j\in[n]\backslash\{i\}} p_{ij}\psi(\theta_i - \theta_j)\psi(\theta_j - \theta_i))^2}$ and $\frac{\sum_{j\in[n]\backslash\{i\}} \frac{p_{ij}}{L_{ij}} \psi(\theta_i - \theta_j)\psi(\theta_j - \theta_i)(1+e^{\theta_j - \theta_i})^2}{(\sum_{j\in[n]\backslash\{i\}} p_{ij}\psi(\theta_j - \theta_i))^2}$, respectively.

7.3. *Comparison of the two methods for each* θ^*. Theorem 7.1 and Theorem 7.2 allow us to give a refined comparison between the MLE and the spectral method. By ignoring the n^{-3} term in the upper bounds and the δ in each exponent, we can write the error rates of the MLE and the spectral method as

$$\inf_t \frac{1}{k} \biggl[\sum_{i=1}^k \exp\biggl(-\frac{(\theta_i^* - t)_+^2 npL}{2V_i(\theta^*)} \biggr) + \sum_{i=k+1}^n \exp\biggl(-\frac{(t - \theta_i^*)_+^2 npL}{2V_i(\theta^*)} \biggr) \biggr], \tag{34}$$

and

$$\inf_t \frac{1}{k} \biggl[\sum_{i=1}^k \exp\biggl(-\frac{(\theta_i^* - t)_+^2 npL}{2\overline{V}_i(\theta^*)} \biggr) + \sum_{i=k+1}^n \exp\biggl(-\frac{(t - \theta_i^*)_+^2 npL}{2\overline{V}_i(\theta^*)} \biggr) \biggr]. \tag{35}$$

It is clear that the only difference between (34) and (35) lies in the difference of the variance functions (26) and (31), whose comparison is given by the following lemma.

LEMMA 7.1. *For any* $\theta^* \in \mathbb{R}$ *and any* $i \in [n]$, *we have* $V_i(\theta^*) \le \overline{V}_i(\theta^*)$. *The equality holds if and only if* $\theta_1^* = \cdots = \theta_n^*$.

PROOF. Notice the following chain of equalities and inequality:

$$\begin{aligned} V_i(\theta^*) &= \frac{n}{\sum_{j\in[n]} \psi'(\theta_j^* - \theta_i^*)} \\ &= \frac{n(\sum_{j\in[n]} e^{\theta_j^* - \theta_i^*})}{(\sum_{j\in[n]} \psi'(\theta_j^* - \theta_i^*))(\sum_{j\in[n]} e^{\theta_j^* - \theta_i^*})} \\ &\le \frac{n(\sum_{j\in[n]} \psi'(\theta_j^* - \theta_i^*)(1 + e^{\theta_j^* - \theta_i^*})^2)}{(\sum_{j\in[n]} \psi(\theta_j^* - \theta_i^*))^2} \\ &= \overline{V}_i(\theta^*), \end{aligned} \tag{36}$$

where (36) is by the Cauchy–Schwarz inequality on the denominator. According to the equality condition of the Cauchy-Schwarz inequality, we know that $V_i(\theta^*) = \overline{V}_i(\theta^*)$ only when $\theta_1^* = \cdots = \theta_n^*$. $\square$

To close this section, we discuss two special cases of θ^*, under which the error rates recover the results of Theorem 3.2, Theorem 4.1 and Corollary 5.1.

EXAMPLE 7.1. According to the proof of Theorem 4.3 and the construction discussed in Section 6, the least favorable $\theta^* \in \Theta(k, \Delta, \kappa)$ takes the following form: $\theta_i^* = \kappa_1$ for all $1 \le i \le k - \rho k$, $\theta_i^* = 0$ for $k - \rho k < i \le k$, $\theta_i^* = -\Delta$ for $k < i \le k + \rho(n-k)$ and $\theta_i^* = -\kappa_2$ for $k + \rho(n-k) < i \le n$. Here, κ_1 and κ_2 are maximizers of either (11) for the MLE or (17) for the spectral method, and ρ is a sufficiently small constant. For this θ^*, the formulas (34) and (35) recover the minimax rates obtained in Theorem 3.2 and Theorem 4.1.

EXAMPLE 7.2. Another interesting θ^* is the two-piece model $\theta^* \in \Theta(k, \Delta, \Delta)$. By the translational invariance of the variance functions, we can consider $\theta_i^* = \Delta$ for all $1 \le i \le k$ and $\theta_i^* = 0$ for all $k < i \le n$. We discuss the consequence of this choice of θ^* under two situations. First, consider $\Delta = o(1)$, and one can check that $V_i(\theta^*) = (1 + o(1))4$ and $\overline{V}_i(\theta^*) = (1 + o(1))4$ for all $i \in [n]$, which implies the equivalence of error rates of the MLE and the spectral method. Second, consider Δ lower bounded by some constant. In this case, both the formulas (34) and (35) are $o(k^{-1})$, which implies both the MLE and the spectral method achieve exact recovery with high probability. As shown in Corollary 5.1, the spectral method is actually optimal for $\theta^* \in \Theta(k, \Delta, \Delta)$. We are therefore able to give a theoretical justification of the numerical experiment of [7].

8. Analysis of the MLE. In this section, we analyze the MLE (6), and prove Theorem 3.1, Theorem 3.2 and Theorem 3.3. Since the BTL model (3) is invariant to a shift of the model parameter, we can assume $\mathbb{1}_n^T \theta^* = 0$ without loss of generality. For simplicity of notation, we also assume $r_i^* = i$ for each $i \in [n]$, and thus we have $\theta_{r_i^*}^* = \theta_i^*$. Recall the convention of notation that $A_{ij} = A_{ji}$ and $\bar{y}_{ij} = 1 - \bar{y}_{ji}$ for any $i < j$. We also set $A_{ii} = 0$ for all $i \in [n]$. Throughout the analysis, we will repeatedly use the properties that both $\psi(t)$ and $\psi'(t)$ are bounded continuous functions with bounded Lipschitz constants.

The section is organized as follows. We will first give a brief overview of the techniques and the main steps of the analysis in Section 8.1. We then present a few technical lemmas in Section 8.2. In Section 8.3, we establish an important result on the ℓ_∞ bound of the MLE. Theorem 3.1 will be proved in Section 8.4. Finally, we prove Theorem 3.2 and Theorem 3.3 in Section 8.5.

8.1. *Overview of the techniques.* A major difficulty of analyzing the MLE is to control the spectrum of the Hessian matrix of the negative log-likelihood function. Recall the definition of $\ell_n(\theta)$ in (5). Its Hessian $\nabla^2 \ell_n(\theta) = H(\theta) \in \mathbb{R}^{n \times n}$ is given by the formula

$$H_{ij}(\theta) = \begin{cases} \displaystyle\sum_{l \in [n] \setminus \{i\}} A_{il} \psi'(\theta_i - \theta_l), & i = j, \\ -A_{ij} \psi'(\theta_i - \theta_j), & i \neq j. \end{cases}$$

It can be viewed as the Laplacian of the weighted random graph $\{\psi'(\theta_i - \theta_j) A_{ij}\}$. For θ that satisfies $\max_{i<j} |\theta_i - \theta_j| = O(1)$, the spectrum of $H(\theta)$ can be well controlled via some standard random matrix tool [24]. The property $\max_{i<j} |\theta_i - \theta_j| = O(1)$ certainly holds for $\theta^* \in \Theta(k, \Delta, \kappa)$. However, when analyzing the Taylor expansion of $\ell_n(\theta)$, we actually need to understand $H(\theta)$ for θ that is a convex combination between $\widehat{\theta}$ and θ^*. Since the MLE

is defined without any constraint or regularization, there is no such control for $\widehat{\theta}$. Our first step is to establish the following proposition that shows $\|\widehat{\theta} - \theta^*\|_\infty$ is bounded with high probability even though the MLE has no constraint or regularization.

PROPOSITION 8.1. *Under the setting of Theorem 3.1, we have*

$$\|\widehat{\theta} - \theta^*\|_\infty \le 5, \tag{37}$$

with probability at least $1 - O(n^{-7})$.

The proof of Proposition 8.1 borrows strength from the property of a regularized MLE. Recall the definition of $\widehat{\theta}_\lambda$ in (10). This is the version of MLE that has been analyzed by [7]. We will choose $\lambda = n^{-1}$ in order that $\widehat{\theta}_\lambda$ is close to $\widehat{\theta}$. Following the techniques in [7], we can first show $\|\widehat{\theta}_\lambda - \theta^*\|_\infty \le 4$ with high probability. The presence of the penalty in (10) is crucial for the result $\|\widehat{\theta}_\lambda - \theta^*\|_\infty \le 4$ to be established. Next, we have an argument to show that the two estimators $\widehat{\theta}_\lambda$ and $\widehat{\theta}$ are sufficiently close. This leads to the bound (37). A detailed proof of Proposition 8.1 will be given in Section 8.3.

The result of Proposition 8.1 is arguably the most important step in the analysis of the MLE. It directly leads to the control of the spectrum of $H(\theta)$. Then the first bound (8) of Theorem 3.1 can be obtained by a Taylor expansion of the objective function $\ell_n(\theta)$. The second bound (9) of Theorem 3.1 and Theorem 3.2 requires an entrywise analysis of $\widehat{\theta}$, and is therefore more complicated. We need to take advantage of the powerful leave-one-out argument in [7]. The intuition of the leave-one-out technique has been thoroughly discussed in [7], and we do not repeat it here. We would like to emphasize that our version of the leave-one-out argument is in fact different from the form introduced in [7]. We do not need to combine the leave-one-out argument with a gradient descent analysis as in [7]. This helps us to avoid the extra technical condition $\log L = O(\log n)$ in [7] when proving the theorems.

8.2. *Some technical lemmas.* Let us present a few technical lemmas that facilitate our analysis of the MLE. The first two lemmas are concentration properties of the random graph $A \sim \mathcal{G}(n, p)$. We define $\mathcal{L}_A = D - A$ to be the graph Laplacian of A, where D is a diagonal matrix whose entries are given by $D_{ii} = \sum_{j \in [n] \setminus \{i\}} A_{ij}$.

LEMMA 8.1. *Assume* $p \ge \frac{c_0 \log n}{n}$ *for some sufficiently large* $c_0 > 0$. *We then have*

$$\frac{1}{2}np \le \min_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} A_{ij} \le \max_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} A_{ij} \le 2np,$$

and

$$\lambda_{\min,\perp}(\mathcal{L}_A) = \min_{u \ne 0 : \mathbb{1}_n^T u = 0} \frac{u^T \mathcal{L}_A u}{\|u\|^2} \ge \frac{np}{2},$$

$$\lambda_{\max}(\mathcal{L}_A) = \max_{u \ne 0} \frac{u^T \mathcal{L}_A u}{\|u\|^2} \le 2np$$

with probability at least $1 - O(n^{-10})$.

LEMMA 8.2. *Assume* $p \ge \frac{c_0 \log n}{n}$ *for some sufficiently large* $c_0 > 0$. *For any fixed* $\{w_{ij}\}$, *we have*

$$\max_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} w_{ij}^2 (A_{ij} - p)^2 \le Cnp \max_{i,j \in [n]} |w_{ij}|^2,$$

and

$$\max_{i\in[n]}\Big(\sum_{j\in[n]\setminus\{i\}} w_{ij}(A_{ij}-p)\Big)^2 \le C(\log n)^2 \max_{i,j\in[n]}|w_{ij}|^2 + Cp\log n \max_{i\in[n]}\sum_{j\in[n]} w_{ij}^2,$$

for some constant $C>0$ *with probability at least* $1-O(n^{-10})$.

With $\lambda_{\min,\perp}(\mathcal{L}_A)$ shown to be well behaved, the next lemma establishes a similar control for $\lambda_{\min,\perp}(H(\theta))$.

LEMMA 8.3. *Assume* $p \ge \frac{c_0\log n}{n}$ *for some sufficiently large* $c_0>0$. *For any* $\theta\in\mathbb{R}^n$ *that satisfies* $\max_{i\in[n]}\theta_i - \min_{i\in[n]}\theta_i \le M$, *we have*

$$\lambda_{\min,\perp}\big(H(\theta)\big) \ge \frac{1}{8}e^{-M}np,$$

with probability at least $1-O(n^{-10})$.

Finally, we need a few concentration inequalities.

LEMMA 8.4. *Assume* $\kappa = O(1)$ *and* $p \ge \frac{c_0\log n}{n}$ *for some sufficiently large* $c_0>0$. *Then we have*

$$\sum_{i=1}^n\Big(\sum_{j\in[n]\setminus\{i\}} A_{ij}\big(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)\big)\Big)^2 \le C\frac{n^2p}{L},$$

$$\max_{i\in[n]}\Big(\sum_{j\in[n]\setminus\{i\}} A_{ij}\big(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)\big)\Big)^2 \le C\frac{np\log n}{L},$$

$$\max_{i\in[n]}\sum_{j\in[n]\setminus\{i\}} A_{ij}\big(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)\big)^2 \le C\frac{np}{L},$$

for some constant $C>0$ *with probability at least* $1-O(n^{-10})$ *uniformly over all* $\theta^*\in\Theta(k,0,\kappa)$.

The proofs of the four lemmas above will be given in Appendix E.

8.3. *Proof of Proposition* 8.1. As we have outlined in Section 8.1, the main argument to bound $\|\widehat{\theta}-\theta^*\|_\infty$ is to first derive a bound for $\|\widehat{\theta}_\lambda-\theta^*\|_\infty$, where $\widehat{\theta}_\lambda$ is the penalized MLE defined in (10). Then we only need to show $\widehat{\theta}_\lambda$ and $\widehat{\theta}$ are close with λ as small as $\lambda=n^{-1}$. We first state a lemma that bounds $\|\widehat{\theta}_\lambda-\theta^*\|_\infty$.

LEMMA 8.5. *Under the setting of Theorem* 3.1, *for the estimator* $\widehat{\theta}_\lambda$ *with* $\lambda=n^{-1}$, *we have*

$$\|\widehat{\theta}_\lambda-\theta^*\|_\infty \le 4,$$

with probability at least $1-O(n^{-7})$.

We first prove Proposition 8.1 with the help of Lemma 8.5. The proof of Lemma 8.5 largely follows the arguments in [7] that analyze the regularized MLE. Since we only need to show $\|\widehat{\theta}_\lambda-\theta^*\|_\infty\le 4$ rather than the optimal rate, the condition on L imposed by [7] is not needed anymore. This requires a few minor changes in the proof of [7]. We include the detailed proof of Lemma 8.5 in Appendix D in the Supplementary Material to be self-contained.

PROOF OF PROPOSITION 8.1. Define a constraint MLE as

$$\widehat{\theta}^{\mathsf{con}} = \operatorname*{argmin}_{\mathbb{1}_n^T\theta=0:\|\theta-\theta^*\|_\infty\le 5} \ell_n(\theta). \tag{38}$$

By Lemma 8.5, $\widehat{\theta}_\lambda$ is feasible for the constraint of (38). We then have

$$\ell_n(\widehat{\theta}_\lambda) \ge \ell_n\big(\widehat{\theta}^{\mathsf{con}}\big). \tag{39}$$

We apply Taylor expansion, and obtain

$$\ell_n\big(\widehat{\theta}^{\mathsf{con}}\big) = \ell_n(\widehat{\theta}_\lambda) + \big(\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda\big)^T \nabla\ell_n(\widehat{\theta}_\lambda) + \frac{1}{2}\big(\widehat{\theta}_\lambda - \widehat{\theta}^{\mathsf{con}}\big)^T H(\xi)\big(\widehat{\theta}_\lambda - \widehat{\theta}^{\mathsf{con}}\big),$$

where ξ is a convex combination of $\widehat{\theta}^{\mathsf{con}}$ and $\widehat{\theta}_\lambda$. By Lemma 8.5, we know that $\|\widehat{\theta}_\lambda - \theta^*\|_\infty \le 4$. We also have $\|\widehat{\theta}^{\mathsf{con}} - \theta^*\|_\infty \le 5$ by the definition of $\widehat{\theta}^{\mathsf{con}}$. Thus, $\|\xi - \theta^*\|_\infty \le 5$. By Lemma 8.3, we get the lower bound

$$\ell_n\big(\widehat{\theta}^{\mathsf{con}}\big) \ge \ell_n(\widehat{\theta}_\lambda) + \big(\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda\big)^T \nabla\ell_n(\widehat{\theta}_\lambda) + c_1 np\big\|\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda\big\|^2, \tag{40}$$

for some constant $c_1 > 0$. By (39) and (40), we have

$$\big\|\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda\big\|^2 \le \frac{|(\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda)^T \nabla\ell_n(\widehat{\theta}_\lambda)|}{c_1 np}.$$

By the Cauchy–Schwarz inequality and the fact that $\nabla\ell_n(\widehat{\theta}_\lambda) + \lambda\widehat{\theta}_\lambda = 0$, we have

$$\big\|\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda\big\|^2 \le \frac{\|\nabla\ell_n(\widehat{\theta}_\lambda)\|^2}{(c_1 np)^2} = \frac{\lambda^2\|\widehat{\theta}_\lambda\|^2}{(c_1 np)^2} \lesssim \frac{n\lambda^2}{(c_1 np)^2} \lesssim n^{-1}.$$

Finally, since

$$\big\|\widehat{\theta}^{\mathsf{con}} - \theta^*\big\|_\infty \le \big\|\widehat{\theta}_\lambda - \theta^*\big\|_\infty + \big\|\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda\big\| \le 4 + \frac{c_2}{\sqrt{n}} \le \frac{9}{2},$$

the minimizer of (38) is in the interior of the constraint. By the convexity of (38), we have $\widehat{\theta}^{\mathsf{con}} = \widehat{\theta}$, and thus the desired conclusion $\|\widehat{\theta} - \theta^*\|_\infty \le 5$ is obtained. □

8.4. *Proof of Theorem* 3.1. We give separate proofs for the conclusions (8) and (9) in this section.

PROOF OF (8) OF THEOREM 3.1. By the definition of $\widehat{\theta}$, we have $\ell_n(\theta^*) \ge \ell_n(\widehat{\theta})$. We then apply Taylor expansion and obtain

$$\ell_n(\widehat{\theta}) = \ell_n(\theta^*) + \big(\widehat{\theta} - \theta^*\big)^T \nabla\ell_n(\theta^*) + \frac{1}{2}\big(\widehat{\theta} - \theta^*\big)^T H(\xi)\big(\widehat{\theta} - \theta^*\big),$$

where ξ is a convex combination of $\widehat{\theta}$ and θ^*. By Proposition 8.1, we have $\|\widehat{\theta} - \theta^*\|_\infty \le 5$, which implies $\|\xi - \theta^*\|_\infty \le 5$. Thus, we can apply Lemma 8.3 and get $\frac{1}{2}(\widehat{\theta} - \theta^*)^T H(\xi)(\widehat{\theta} - \theta^*) \ge c_1 np\|\widehat{\theta} - \theta^*\|^2$ for some constant $c_1 > 0$. Together with $\ell_n(\theta^*) \ge \ell_n(\widehat{\theta})$ and a Cauchy–Schwarz inequality, we have $\|\widehat{\theta} - \theta^*\|^2 \le \frac{\|\nabla\ell_n(\theta^*)\|^2}{(c_1 np)^2}$. Use (S128) and Lemma 8.4, we obtain the desired conclusion that $\|\widehat{\theta} - \theta^*\|^2 \lesssim \frac{1}{Lp}$. □

The proof of (9) is more involved. It is based on a leave-one-out argument that is very different from the one used in [7]. Let us decompose the objective function $\ell_n(\theta)$ as

$$\ell_n(\theta) = \ell_n^{(-m)}(\theta_{-m}) + \ell_n^{(m)}(\theta_m|\theta_{-m}), \tag{41}$$

where we use $\theta_m \in \mathbb{R}$ for the mth entry of θ and $\theta_{-m} \in \mathbb{R}^{n-1}$ for the remaining entries. The two functions in (41) are defined as

$$\ell_n^{(-m)}(\theta_{-m}) = \sum_{1\le i<j\le n: i,j\neq m} A_{ij}\Big[\bar{y}_{ij}\log\frac{1}{\psi(\theta_i-\theta_j)} + (1-\bar{y}_{ij})\log\frac{1}{1-\psi(\theta_i-\theta_j)}\Big],$$

$$\ell_n^{(m)}(\theta_m|\theta_{-m}) = \sum_{j\in[n]\setminus\{m\}} A_{mj}\Big[\bar{y}_{mj}\log\frac{1}{\psi(\theta_m-\theta_j)} + (1-\bar{y}_{mj})\log\frac{1}{1-\psi(\theta_m-\theta_j)}\Big].$$

Define

$$\theta_{-m}^{(m)} = \operatorname*{argmin}_{\theta_{-m}:\|\theta_{-m}-\theta_{-m}^*\|_\infty\le 5} \ell_n^{(-m)}(\theta_{-m}). \tag{42}$$

We first present an ℓ_2 norm bound for $\theta_{-m}^{(m)}$. We also use $H^{(-m)}(\theta_{-m})$ for the Hessian matrix $\nabla^2\ell_n^{(-m)}(\theta_{-m})$.

LEMMA 8.6. *Under the setting of Theorem 3.1, there exists some constant $C>0$ such that*

$$\max_{m\in[n]}\big\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big\|^2 \le C\frac{1}{pL},$$

with probability at least $1-O(n^{-9})$, where $a_m = \mathsf{ave}(\theta_{-m}^{(m)} - \theta_m^*)$.

PROOF. The proof is very similar to that of (8), since $\theta_{-m}^{(m)}$ can be thought of as a constrained MLE on a subset of the data. By the definition of $\theta_{-m}^{(m)}$, we have

$$\begin{aligned}\ell_n^{(-m)}(\theta_{-m}^*) &\ge \ell_n^{(-m)}(\theta_{-m}^{(m)})\\ &= \ell_n^{(-m)}(\theta_{-m}^*) + \big(\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big)^T\nabla\ell_n^{(-m)}(\theta_{-m}^*)\\ &\quad + \frac{1}{2}\big(\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big)^T H^{(-m)}(\xi)\big(\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big),\end{aligned}$$

where ξ is a convex combination of $\theta_{-m}^{(m)}$ and θ_{-m}^*. In the above Taylor expansion, we have also used the property that $\ell_n^{(-m)}(\theta_{-m}) = \ell_n^{(-m)}(\theta_{-m}+c\mathbb{1}_{n-1})$, $\nabla\ell_n^{(-m)}(\theta_{-m}) = \nabla\ell_n^{(-m)}(\theta_{-m}+c\mathbb{1}_{n-1})$ and $H^{(-m)}(\theta_{-m}) = H^{(-m)}(\theta_{-m}+c\mathbb{1}_{n-1})$ for any $c\in\mathbb{R}$. Since $\|\xi-\theta_{-m}^*\|_\infty \le \|\theta_{-m}^{(m)}-\theta_{-m}^*\|_\infty \le 5$, we can apply Lemma 8.3 to the subset of the data, and obtain

$$\begin{aligned}&\frac{1}{2}\big(\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big)^T H^{(-m)}(\xi)\big(\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big)\\ &\quad\ge c_1np\big\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big\|^2,\end{aligned}$$

with probability at least $1-O(n^{-10})$ for some constant $c_1>0$. By the Cauchy–Schwarz inequality, we have

$$\big\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\big\|^2 \le \frac{\|\nabla\ell_n^{(-m)}(\theta_{-m}^*)\|^2}{(c_1np)^2}.$$

Apply (S128) and Lemma 8.4 to the subset of the data, and we obtain that $\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m\mathbb{1}_{n-1}\|^2 \le C\frac{1}{pL}$ with probability at least $1-O(n^{-10})$. Finally, a union bound argument leads to the desired result. □

With the help of Lemma 8.6, we are ready to prove (9).

PROOF OF (9) OF THEOREM 3.1. By Proposition 8.1, we have $\|\widehat{\theta}_{-m} - \theta^*_{-m}\|_\infty \le \|\widehat{\theta} - \theta^*\|_\infty \le 5$, and thus $\widehat{\theta}_{-m}$ is feasible for the constraint of (42). By the definition of $\theta^{(m)}_{-m}$, we have

$$\begin{aligned}\ell_n^{(-m)}(\widehat{\theta}_{-m}) &\ge \ell_n^{(-m)}\big(\theta^{(m)}_{-m}\big)\\ &= \ell_n^{(-m)}(\widehat{\theta}_{-m}) + \big(\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big)^T \nabla \ell_n^{(-m)}(\widehat{\theta}_{-m})\\ &\quad + \frac{1}{2}\big(\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big)^T H^{(-m)}(\xi)\big(\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big),\end{aligned}$$

where $\bar{a}_m = \mathsf{ave}(\theta^{(m)}_{-m} - \widehat{\theta}_{-m})$ and ξ is a convex combination of $\theta^{(m)}_{-m}$ and $\widehat{\theta}_{-m}$. Since both $\theta^{(m)}_{-m}$ and $\widehat{\theta}_{-m}$ satisfy the constraint of (42), we must have $\|\xi - \theta^*_{-m}\|_\infty \le 5$. Then we can apply Lemma 8.3 to the subset of the data, and obtain

$$\begin{aligned}&\frac{1}{2}\big(\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big)^T H^{(-m)}(\xi)\big(\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big)\\ &\quad \ge c_1 np \big\|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big\|^2,\end{aligned}$$

for some constant $c_1 > 0$. By the Cauchy–Schwarz inequality, we have

$$\big\|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big\|^2 \le \frac{\|\nabla \ell_n^{(-m)}(\widehat{\theta}_{-m})\|^2}{(c_1 np)^2}.$$

For each $i \in [n]\backslash\{m\}$, by the decomposition (41), we have

$$\frac{\partial}{\partial \theta_i}\ell_n^{(-m)}(\theta_{-m}) = \frac{\partial}{\partial \theta_i}\ell_n(\theta) - \frac{\partial}{\partial \theta_i}\ell_n^{(m)}(\theta_m|\theta_{-m}).$$

Since $\nabla \ell_n(\widehat{\theta}) = 0$, we have

$$\frac{\partial}{\partial \theta_i}\ell_n^{(-m)}(\theta_{-m})_{|\theta=\widehat{\theta}} = -\frac{\partial}{\partial \theta_i}\ell_n^{(m)}(\theta_m|\theta_{-m})_{|\theta=\widehat{\theta}} = -A_{mi}\big(\bar{y}_{mi} - \psi(\widehat{\theta}_m - \widehat{\theta}_i)\big).$$

We therefore have the bound,

$$\begin{aligned}\big\|\nabla \ell_n^{(-m)}(\widehat{\theta}_{-m})\big\|^2 &= \sum_{i\in[n]\backslash\{m\}} A_{mi}\big(\bar{y}_{mi} - \psi(\widehat{\theta}_m - \widehat{\theta}_i)\big)^2\\ &\le 2\sum_{i\in[n]\backslash\{m\}} A_{mi}\big(\bar{y}_{mi} - \psi\big(\theta^*_m - \theta^*_i\big)\big)^2\\ &\quad + 2\sum_{i\in[n]\backslash\{m\}} A_{mi}\big(\psi\big(\theta^*_m - \theta^*_i\big) - \psi(\widehat{\theta}_m - \widehat{\theta}_i)\big)^2\\ &\le 2\sum_{i\in[n]\backslash\{m\}} A_{mi}\big(\bar{y}_{mi} - \psi\big(\theta^*_m - \theta^*_i\big)\big)^2 + 2\|\widehat{\theta} - \theta^*\|_\infty^2 \sum_{i\in[n]\backslash\{m\}} A_{mi}\\ &\le 2\sum_{i\in[n]\backslash\{m\}} A_{mi}\big(\bar{y}_{mi} - \psi\big(\theta^*_m - \theta^*_i\big)\big)^2 + 4np\|\widehat{\theta} - \theta^*\|_\infty^2,\end{aligned}$$

where the last inequality is by Lemma 8.1. This implies

$$\begin{aligned}\max_{m\in[n]}\big\|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\big\|^2 &\le \frac{\max_{m\in[n]}\sum_{i\in[n]\backslash\{m\}} A_{mi}\big(\bar{y}_{mi} - \psi\big(\theta^*_m - \theta^*_i\big)\big)^2}{(c_1 np)^2/2}\\ &\quad + \frac{\|\widehat{\theta} - \theta^*\|_\infty^2}{c_1^2 np/4}.\end{aligned}$$

Since we need a bound for $\max_{m\in[n]}\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\|^2$, we need to quantify the difference between a_m and $\bar{a}_m$. Recall that $a_m=\mathsf{ave}(\theta^{(m)}_{-m}-\theta^*_m)$. Since $\mathbb{1}_n^T\widehat{\theta}=\mathbb{1}_n^T\theta^*=0$, we have

$$\|a_m\mathbb{1}_{n-1}-\bar{a}_m\mathbb{1}_{n-1}\|^2=(n-1)\big(\mathsf{ave}(\widehat{\theta}_{-m}-\theta^*_{-m})\big)^2=\frac{(\widehat{\theta}_m-\theta^*_m)^2}{n-1}\le\frac{\|\widehat{\theta}-\theta^*\|^2_\infty}{n-1}.$$

We then have

$$\begin{aligned}&\max_{m\in[n]}\big\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\big\|^2\\&\le C_1\frac{\max_{m\in[n]}\sum_{i\in[n]\setminus\{m\}}A_{mi}(\bar{y}_{mi}-\psi(\theta^*_m-\theta^*_i))^2}{n^2p^2}\\&\quad+C_1\frac{\|\widehat{\theta}-\theta^*\|^2_\infty}{np},\end{aligned}\tag{43}$$

for some constant $C_1>0$.

Next, let us derive a bound for $\|\widehat{\theta}-\theta^*\|^2_\infty$ in terms of $\max_{m\in[n]}\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\|^2$. We introduce the notation:

$$f^{(m)}(\theta_m|\theta_{-m})=\frac{\partial}{\partial\theta_m}\ell^{(m)}_n(\theta_m|\theta_{-m})=-\sum_{i\in[n]\setminus\{m\}}A_{mi}\big(\bar{y}_{mi}-\psi(\theta_m-\theta_i)\big),$$

$$g^{(m)}(\theta_m|\theta_{-m})=\frac{\partial^2}{\partial\theta_m^2}\ell^{(m)}_n(\theta_m|\theta_{-m})=\sum_{i\in[n]\setminus\{m\}}A_{mi}\psi(\theta_m-\theta_i)\psi(\theta_i-\theta_m).$$

By the definition of $\widehat{\theta}$, we know that $\ell_n(\widehat{\theta})=\min_{\theta:\mathbb{1}_n^T\theta=0}\ell_n(\theta)$. Since $\ell_n(\theta)=\ell_n(\theta+c\mathbb{1}_n)$ for any $c\in\mathbb{R}$, we also have $\ell_n(\widehat{\theta})=\min_\theta\ell_n(\theta)$. This allows us to compare the value of the objective $\ell_n(\theta)$ at $\widehat{\theta}$ with any vector that is not necessarily centered. We then have

$$\ell^{(m)}_n(\theta^*_m|\widehat{\theta}_{-m})+\ell^{(-m)}_n(\widehat{\theta}_{-m})\ge\ell_n(\widehat{\theta}),$$

which implies

$$\begin{aligned}\ell^{(m)}_n\big(\theta^*_m|\widehat{\theta}_{-m}\big)&\ge\ell^{(m)}_n(\widehat{\theta}_m|\widehat{\theta}_{-m})\\&=\ell^{(m)}_n\big(\theta^*_m|\widehat{\theta}_{-m}\big)+\big(\widehat{\theta}_m-\theta^*_m\big)f^{(m)}\big(\theta^*_m|\widehat{\theta}_{-m}\big)+\frac{1}{2}\big(\widehat{\theta}_m-\theta^*_m\big)^2g^{(m)}(\xi|\widehat{\theta}_{-m}),\end{aligned}$$

where ξ is a scalar between θ^*_m and $\widehat{\theta}_m$. By Proposition 8.1, $|\xi-\theta^*_m|\le|\widehat{\theta}_m-\theta^*_m|\le\|\widehat{\theta}-\theta^*\|_\infty\le5$. Therefore, for any $i\in[n]\setminus\{m\}$, $|\xi-\widehat{\theta}_i|\le|\xi-\theta^*_m|+|\theta^*_m-\theta^*_i|+|\widehat{\theta}_i-\theta^*_i|\le10+\kappa$. This implies $\frac{1}{2}g^{(m)}(\xi|\widehat{\theta}_{-m})\ge c_2np$ for some constant $c_2>0$ with the help of Lemma 8.1. We then have the bound

$$\big(\widehat{\theta}_m-\theta^*_m\big)^2\le\frac{|f^{(m)}(\theta^*_m|\widehat{\theta}_{-m})|^2}{(c_2np)^2}.\tag{44}$$

We bound $|f^{(m)}(\theta^*_m|\widehat{\theta}_{-m})|$ by

$$\begin{aligned}\big|f^{(m)}\big(\theta^*_m|\widehat{\theta}_{-m}\big)\big|&=\Big|\sum_{i\in[n]\setminus\{m\}}A_{mi}\big(\bar{y}_{mi}-\psi\big(\theta^*_m-\widehat{\theta}_i\big)\big)\Big|\\&\le\Big|\sum_{i\in[n]\setminus\{m\}}A_{mi}\big(\bar{y}_{mi}-\psi\big(\theta^*_m-\theta^*_i\big)\big)\Big|\end{aligned}\tag{45}$$

$$+\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\psi\big(\theta^*_m-\theta^*_i\big)-\psi\big(\theta^*_m-\theta^{(m)}_i+a_m\big)\big)\right| \tag{46}$$

$$+\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\psi\big(\theta^*_m-\theta^{(m)}_i+a_m\big)-\psi\big(\theta^*_m-\widehat{\theta}_i\big)\big)\right|. \tag{47}$$

We use Lemma 8.2 to bound (46). We have

$$\begin{aligned}
&\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\psi\big(\theta^*_m-\theta^*_i\big)-\psi\big(\theta^*_m-\theta^{(m)}_i+a_m\big)\big)\right| \\
&\le p\left|\sum_{i\in[n]\setminus\{m\}} \big(\psi\big(\theta^*_m-\theta^*_i\big)-\psi\big(\theta^*_m-\theta^{(m)}_i+a_m\big)\big)\right| \qquad (48)\\
&\quad+\left|\sum_{i\in[n]\setminus\{m\}} (A_{mi}-p)\big(\psi\big(\theta^*_m-\theta^*_i\big)-\psi\big(\theta^*_m-\theta^{(m)}_i+a_m\big)\big)\right| \qquad (49)\\
&\le p\sqrt{n}\big\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\big\|+C_2\log n\big\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\big\|_\infty \\
&\quad+C_2\sqrt{p\log n}\big\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\big\| \\
&\le \big(p\sqrt{n}+C_2\sqrt{p\log n}\big)\big\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\big\|+C_2\log n\big\|\widehat{\theta}-\theta^*\big\|_\infty \\
&\quad+C_2\log n\big\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\big\|.
\end{aligned}$$

With the help of 8.1, we can also bound (47), and we get

$$\begin{aligned}
&\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\psi\big(\theta^*_m-\theta^{(m)}_i+a_m\big)-\psi\big(\theta^*_m-\widehat{\theta}_i\big)\big)\right| \\
&\le\sqrt{\sum_{i\in[n]\setminus\{m\}} A_{mi}}\big\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\big\| \qquad (50)\\
&\le C_3\sqrt{np}\big\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\big\|.
\end{aligned}$$

Plug the bounds into (44), and we have

$$\begin{aligned}
\big\|\widehat{\theta}-\theta^*\big\|_\infty \le{}& \frac{\max_{m\in[n]}\big|\sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\bar{y}_{mi}-\psi\big(\theta^*_m-\theta^*_i\big)\big)\big|}{c_2np} \\
&+\frac{\big(p\sqrt{n}+C_2\sqrt{p\log n}\big)\max_{m\in[n]}\big\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\big\|}{c_2np} \\
&+\frac{\big(C_2\log n+C_3\sqrt{np}\big)\big\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\big\|}{c_2np}+\frac{C_2\log n\big\|\widehat{\theta}-\theta^*\big\|_\infty}{c_2np}.
\end{aligned}$$

Since $np\ge c_0\log n$ for some sufficiently large c_0, we obtain the bound

$$\begin{aligned}
\big\|\widehat{\theta}-\theta^*\big\|_\infty \le{}& C_4\frac{\max_{m\in[n]}\big|\sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\bar{y}_{mi}-\psi\big(\theta^*_m-\theta^*_i\big)\big)\big|}{np} \\
&+C_4\frac{p\sqrt{n}\max_{m\in[n]}\big\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\big\|}{np} \qquad (51)\\
&+C_4\frac{\big(\log n+\sqrt{np}\big)\big\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\big\|}{np}.
\end{aligned}$$

Let us plug the above bound into (43). Then, after some rearrangement, we obtain

$$\begin{aligned}\max_{m\in[n]}\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\| \le C_5\frac{\max_{m\in[n]}\sqrt{\sum_{i\in[n]\setminus\{m\}}A_{mi}(\bar{y}_{mi}-\psi(\theta^*_m-\theta^*_i))^2}}{np}\\ + C_5\frac{\max_{m\in[n]}|\sum_{i\in[n]\setminus\{m\}}A_{mi}(\bar{y}_{mi}-\psi(\theta^*_m-\theta^*_i))|}{np\sqrt{np}}\\ + C_5\frac{\max_{m\in[n]}\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\|}{n\sqrt{p}}.\end{aligned}$$

By Lemma 8.4 and Lemma 8.6, we have

$$\max_{m\in[n]}\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\| \le C_7\sqrt{\frac{1}{npL}}. \tag{52}$$

Now we can plug the bound (52) back into (51), and together with Lemma 8.4 and Lemma 8.6, we have

$$\|\widehat{\theta}-\theta^*\|_\infty \le C_8\sqrt{\frac{\log n}{npL}}, \tag{53}$$

which is the desired conclusion. Tracking all the probabilistic events that we have used in the proof, we can conclude that both (52) and (53) hold with probability at least $1-O(n^{-7})$. □

8.5. *Proofs of Theorem 3.2 and Theorem 3.3.* In the proof of (9), we have established the byproduct (52). This bound turns out to be extremely important for us to establish the result of Theorem 3.2. We therefore list it, together with its consequence, as a lemma.

LEMMA 8.7. *Under the setting of Theorem 3.1, there exists some constant $C>0$ such that*

$$\max_{m\in[n]}\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\|^2 \le C\frac{1}{npL},$$

$$\max_{m\in[n]}\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\|^2_\infty \le C\frac{\log n}{npL},$$

with probability at least $1-O(n^{-7})$, where $a_m=\mathsf{ave}(\theta^{(m)}_{-m}-\theta^*_m)$ and $\theta^{(m)}_{-m}$ is defined by (42).

PROOF. The first conclusion has been established in (52). The second conclusion is a consequence of the inequality

$$\max_{m\in[n]}\|\theta^{(m)}_{-m}-\theta^*_{-m}-a_m\mathbb{1}_{n-1}\|^2_\infty \le 2\max_{m\in[n]}\|\theta^{(m)}_{-m}-\widehat{\theta}_{-m}-a_m\mathbb{1}_{n-1}\|^2+2\|\widehat{\theta}-\theta^*\|^2_\infty,$$

and (53). □

Now we are ready to prove Theorem 3.2.

PROOF OF THEOREM 3.2. When the error exponent is of constant order, the bound is also a constant, and the result already holds since $\mathsf{H}_k(\widehat{r},r^*)\le 1$. Therefore, we only need to consider the case when the error exponent tends to infinity. We first introduce some notation. Define

$$\eta=\frac{1}{2}-\frac{V(\kappa)}{(1-\bar{\delta})\Delta^2 npL}\log\frac{n-k}{k}, \tag{54}$$

where $\bar{\delta} = o(1)$ is chosen such that $\eta > 0$. The specific choice of $\bar{\delta}$ will be specified in the proof. Then let

$$\bar{\Delta}_i = \begin{cases} \min\Big(\eta(\theta_k^* - \theta_{k+1}^*) + \theta_i^* - \theta_k^*, \Big(\dfrac{\log n}{np}\Big)^{1/4}\Big), & 1 \le i \le k, \\ \min\Big((1-\eta)(\theta_k^* - \theta_{k+1}^*) + \theta_{k+1}^* - \theta_i^*, \Big(\dfrac{\log n}{np}\Big)^{1/4}\Big), & k+1 \le i \le n. \end{cases} \tag{55}$$

Since the diverging error exponent implies $\mathsf{SNR} \to \infty$, we have $\min_{i\in[n]} \bar{\Delta}_i^2 Lnp \to \infty$ and $\max_{i\in[n]} \bar{\Delta}_i \to 0$.

The proof involves several steps. In the first step, we need to derive a sharp probabilistic bound for $|f^{(m)}(\theta_m^* | \widehat{\theta}_{-m})|$. In the proof of (9) of Theorem 3.1, we have shown that $|f^{(m)}(\theta_m^* | \widehat{\theta}_{-m})|$ can be bounded by the sum of (45), (47), (48) and (49). For (45), we can use Hoeffding's inequality and Lemma 8.1 and obtain the bound

$$\Big| \sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\bar{y}_{mi} - \psi(\theta_m^* - \theta_i^*)\big)\Big| \le C_1 \sqrt{\frac{x \sum_{i\in[n]\setminus\{m\}} A_{mi}}{L}} \le C_2 \sqrt{\frac{xnp}{L}}, \tag{56}$$

with probability at least $1 - O(n^{-10}) - e^{-x}$. Take $x = \bar{\Delta}_m^{3/2} Lnp$, and we have

$$\Big| \sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\bar{y}_{mi} - \psi(\theta_m^* - \theta_i^*)\big)\Big| \le C_2 \sqrt{\bar{\Delta}_m^{3/2} (np)^2}, \tag{57}$$

with probability at least $1 - O(n^{-10}) - e^{-\bar{\Delta}_m^{3/2} Lnp}$. Since we have already shown (47) can be bounded by (50) with probability at least $1 - O(n^{-10})$, an application of Lemma 8.7 implies that

$$\Big| \sum_{i\in[n]\setminus\{m\}} A_{mi}\big(\psi(\theta_m^* - \theta_i^{(m)} + a_m) - \psi(\theta_m^* - \widehat{\theta}_i)\big)\Big| \le C_3 \sqrt{\frac{1}{L}}, \tag{58}$$

with probability at least $1 - O(n^{-7})$. By the Cauchy–Schwarz inequality, we can bound (48) by $p\sqrt{n}\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1}\|$. With the help of Lemma 8.6, we have

$$p\Big| \sum_{i\in[n]\setminus\{m\}} \big(\psi(\theta_m^* - \theta_i^*) - \psi(\theta_m^* - \theta_i^{(m)} + a_m)\big)\Big| \le C_4 \sqrt{\frac{np}{L}}, \tag{59}$$

with probability at least $1 - O(n^{-9})$. For (49), we use Bernstein's inequality, and we have

$$\begin{aligned} &\Big| \sum_{i\in[n]\setminus\{m\}} (A_{mi} - p)\big(\psi(\theta_m^* - \theta_i^*) - \psi(\theta_m^* - \theta_i^{(m)} + a_m)\big)\Big| \\ &\quad \le C_5 \sqrt{px}\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1}\| + C_5 x \|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1}\|_\infty, \end{aligned} \tag{60}$$

with probability at least $1 - e^{-x}$. We choose $x = \min(\bar{\Delta}_m^2 Lnp \frac{np}{\log n}, 7\log n)$. Then, with the help of Lemma 8.6 and Lemma 8.7, we have

$$\begin{aligned} &\Big| \sum_{i\in[n]\setminus\{m\}} (A_{mi} - p)\big(\psi(\theta_m^* - \theta_i^*) - \psi(\theta_m^* - \theta_i^{(m)} + a_m)\big)\Big| \\ &\quad \le C_6 \frac{1}{\sqrt{L}} \sqrt{\min\Big(\bar{\Delta}_m^2 Lnp \frac{np}{\log n}, 7\log n\Big)} \\ &\qquad + C_6 \sqrt{\frac{\log n}{npL}} \min\Big(\bar{\Delta}_m^2 Lnp \frac{np}{\log n}, 7\log n\Big), \end{aligned} \tag{61}$$

with probability at least $1 - O(n^{-7}) - \exp(-\bar{\Delta}_m^2 npL\frac{np}{\log n})$. Combining the bounds (57)–(61), we obtain a bound for $|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|$. This also implies a bound for $|\widehat{\theta}_m - \theta_m^*|$ because of the inequality (44).

In the second step, we define

$$\bar{\theta}_m = \theta_m^* - \frac{f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})}{g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})}. \tag{62}$$

We need to show $\bar{\theta}_m$ and $\widehat{\theta}_m$ are close. By Proposition 8.1, $\|\widehat{\theta} - \theta^*\|_\infty \le 5$, and thus $g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) \ge c_1 np$ for some constant $c_1 > 0$, so that we have the bound $|\bar{\theta}_m - \theta_m^*| \le \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}$. In fact, given the inequality (44), we can choose c_1 to be sufficiently small so that $|\widehat{\theta}_m - \theta_m^*| \le \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}$ is also true. Therefore, we can express $\bar{\theta}_m$ and $\widehat{\theta}_m$ as

$$\bar{\theta}_m = \operatorname*{argmin}_{|\theta_m - \theta_m^*| \le \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}} \bar{\ell}_n^{(m)}(\theta_m|\widehat{\theta}_{-m}),$$

$$\widehat{\theta}_m = \operatorname*{argmin}_{|\theta_m - \theta_m^*| \le \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}} \ell_n^{(m)}(\theta_m|\widehat{\theta}_{-m}),$$

where

$$\bar{\ell}_n^{(m)}(\theta_m|\widehat{\theta}_{-m}) = \ell_n^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) + (\theta_m - \theta_m^*) f^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) + \frac{1}{2}(\theta_m - \theta_m^*)^2 g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}).$$

Recall the definition of $\ell_n^{(m)}(\theta_m|\theta_{-m})$ in (41) and the display afterwards. We will show $\bar{\theta}_m$ and $\widehat{\theta}_m$ are close by bounding the difference between the two objective functions. By Taylor expansion, we have

$$|\ell_n^{(m)}(\theta_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\theta_m|\widehat{\theta}_{-m})| = \frac{1}{2}(\theta_m - \theta_m^*)^2 |g^{(m)}(\xi|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|,$$

where ξ is a scalar between θ_m and θ_m^*. We then have

$$\begin{aligned}
&|g^{(m)}(\xi|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})| \\
&= \Big| \sum_{i\in[n]\setminus\{m\}} A_{mi}\psi(\xi - \widehat{\theta}_i)\psi(\widehat{\theta}_i - \xi) - \sum_{i\in[n]\setminus\{m\}} A_{mi}\psi(\theta_m^* - \widehat{\theta}_i)\psi(\widehat{\theta}_i - \theta_m^*) \Big| \\
&\le |\xi - \theta_m^*| \sum_{i\in[n]\setminus\{m\}} A_{mi} \\
&\le C_7 |\theta_m - \theta_m^*| np,
\end{aligned}$$

where the last inequality uses Lemma 8.1. Therefore, for any θ_m that satisfies $|\theta_m - \theta_m^*| \le \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}$, the difference between the two objective functions can be bounded by

$$|\ell_n^{(m)}(\theta_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\theta_m|\widehat{\theta}_{-m})| \le \frac{C_7 np}{2}|\theta_m - \theta_m^*|^3 \le \frac{C_7 np}{2}\Big(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}\Big)^3.$$

By Pythagorean identity, $\bar{\ell}_n^{(m)}(\widehat{\theta}_m|\widehat{\theta}_{-m}) = \bar{\ell}_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) + \frac{1}{2}g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})(\widehat{\theta}_m - \bar{\theta}_m)^2$. Then

$$\begin{aligned}
&\frac{1}{2}g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})(\widehat{\theta}_m - \bar{\theta}_m)^2 \\
&\quad = \bar{\ell}_n^{(m)}(\widehat{\theta}_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m})
\end{aligned}$$

$$\le \ell_n^{(m)}(\widehat{\theta}_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) + \frac{C_7 np}{2}\left(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}\right)^3$$

$$\le \ell_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) + \frac{C_7 np}{2}\left(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}\right)^3$$

$$\le 2\frac{C_7 np}{2}\left(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}\right)^3.$$

Since $g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) \ge c_1 np$, we obtain the bound

$$(\widehat{\theta}_m - \bar{\theta}_m)^2 \le \frac{2C_7}{c_1^4}\left(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{np}\right)^3.$$

Since $|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|$ has been shown to be bounded by the sum of (57)–(61), we have

$$|\widehat{\theta}_m - \bar{\theta}_m| \le \delta\bar{\Delta}_m, \tag{63}$$

for some $\delta = o(1)$ with probability at least $1 - O(n^{-7}) - \exp(-\bar{\Delta}_m^{3/2} Lnp) - \exp(-\bar{\Delta}_m^2 np \times L\frac{np}{\log n})$ under the condition that $\bar{\Delta}_m = o(1)$ and $\frac{np}{\log n} \to \infty$.

In the third step, we need to show that $\frac{f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})}{g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})}$ in the definition of $\bar{\theta}_m$ can be replaced by $\frac{f^{(m)}(\theta_m^*|\theta_{-m}^*)}{g^{(m)}(\theta_m^*|\theta_{-m}^*)}$ with a negligible error. By triangle inequality, we can bound $|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - f^{(m)}(\theta_m^*|\theta_{-m}^*)|$ by the sum of (58), (59) and (61). Given that $g^{(m)}(\theta_m^*|\theta_{-m}^*) \gtrsim np$, we have

$$\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - f^{(m)}(\theta_m^*|\theta_{-m}^*)|}{g^{(m)}(\theta_m^*|\theta_{-m}^*)} \le \delta\bar{\Delta}_m, \tag{64}$$

for some $\delta = o(1)$ with probability at least $1 - O(n^{-7}) - \exp(-\bar{\Delta}_m^2 npL\frac{np}{\log n})$ under the assumption that $npL\bar{\Delta}_m^2 \to \infty$ and $\frac{np}{\log n} \to \infty$. Note that we can choose the same δ to accommodate the two bounds (63) and (64). We also need to give a sharp approximation to $g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})$. We have

$$\begin{aligned}
&|g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\theta_{-m}^*)| \\
&\le |g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\theta_{-m}^{(m)} - a_m\mathbb{1}_{n-1})| \\
&\quad + |g^{(m)}(\theta_m^*|\theta_{-m}^{(m)} - a_m\mathbb{1}_{n-1}) - g^{(m)}(\theta_m^*|\theta_{-m}^*)| \\
&\le \sqrt{\sum_{i\in[n]\setminus\{m\}} A_{mi}}\,\|\theta_{-m}^{(m)} - a_m\mathbb{1}_{n-1} - \widehat{\theta}_{-m}\| + p\sqrt{n}\|\theta_{-m}^{(m)} - a_m\mathbb{1}_{n-1} - \theta^*\| \\
&\quad + \sum_{i\in[n]\setminus\{m\}} (A_{mi} - p)|\theta_i^{(m)} - a_m - \theta_i^*|.
\end{aligned}$$

By Lemma 8.1, Lemma 8.6 and Lemma 8.7, the first two terms can be bounded by $C_8\sqrt{\frac{np}{L}}$ with probability at least $1 - O(n^{-7})$. To bound the third term, we can use Lemma 8.2, and then $\sum_{i\in[n]\setminus\{m\}}(A_{mi} - p)|\theta_i^{(m)} - a_m - \theta_i^*|$ can be bounded by

$$C_8\sqrt{p\log n}\|\theta_{-m}^{(m)} - a_m\mathbb{1}_{n-1} - \theta^*\| + C_8\log n\|\theta_{-m}^{(m)} - a_m\mathbb{1}_{n-1} - \theta^*\|_\infty,$$

with probability at least $1 - O(n^{-10})$. By Lemma 8.6 and Lemma 8.7, the above display is at most $C_9\sqrt{\frac{\log n}{L}} + C_9\frac{(\log n)^{3/2}}{\sqrt{npL}}$ with probability at least $1 - O(n^{-7})$. Combining our bounds,

we obtain

$$|g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\theta_{-m}^*)| \lesssim \sqrt{\frac{np}{L}} + \frac{(\log n)^{3/2}}{\sqrt{npL}}. \tag{65}$$

Since $g^{(m)}(\theta_m^*|\theta_{-m}^*) \gtrsim np$, we have

$$\frac{|g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\theta_{-m}^*)|}{g^{(m)}(\theta_m^*|\theta_{-m}^*)} \le \delta, \tag{66}$$

for some $\delta = o(1)$ with probability at least $1 - O(n^{-7})$. Note that we can choose the same δ to accommodate the three bounds (63), (64) and (66).

In the last step, we will apply Lemma 3.1 with $t = (1-\eta)\theta_k^* + \eta\theta_{k+1}^*$ to complete the proof. Recall the definition of η in (54). For any $i \le k$, we have

$$\begin{aligned}
&\mathbb{P}\big(\widehat{\theta}_i \le (1-\eta)\theta_k^* + \eta\theta_{k+1}^*\big)\big) \\
&\quad\le \mathbb{P}\big(\widehat{\theta}_i - \theta_i^* \le -\eta\big(\theta_k^* - \theta_{k+1}^*\big) - \big(\theta_i^* - \theta_k^*\big)\big) \\
&\quad\le \mathbb{P}\big(\bar{\theta}_i - \theta_i^* \le -(1-\delta)\bar{\Delta}_i\big) + \mathbb{P}\big(|\bar{\theta}_i - \widehat{\theta}_i| > \delta\bar{\Delta}_i\big) \\
&\quad\le \mathbb{P}\bigg(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -\big(1+\delta^2-3\delta\big)\bar{\Delta}_i\bigg) \\
&\qquad+ \mathbb{P}\big(|\bar{\theta}_i - \widehat{\theta}_i| > \delta\bar{\Delta}_i\big) \\
&\qquad+ \mathbb{P}\bigg(\frac{|g^{(i)}(\theta_i^*|\widehat{\theta}_{-i}) - g^{(i)}(\theta_i^*|\theta_{-i}^*)|}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} > \delta\bigg) \\
&\qquad+ \mathbb{P}\bigg(\frac{|f^{(i)}(\theta_i^*|\widehat{\theta}_{-i}) - f^{(i)}(\theta_i^*|\theta_{-i}^*)|}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} > \delta\bar{\Delta}_i\bigg) \\
&\quad\le \mathbb{P}\bigg(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -(1-3\delta)\bar{\Delta}_i\bigg) + O\big(n^{-7}\big) \\
&\qquad+ \exp\big(-\bar{\Delta}_i^{3/2}Lnp\big) + \exp\bigg(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\bigg),
\end{aligned} \tag{67}$$

where the last inequality is due to (63), (64) and (66). Define the event

$$\mathcal{A}_i = \bigg\{A : \bigg|\frac{\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi(\theta_i^* - \theta_j^*)\psi(\theta_j^* - \theta_i^*)}{p\sum_{j\in[n]\setminus\{i\}}\psi(\theta_i^* - \theta_j^*)\psi(\theta_j^* - \theta_i^*)} - 1\bigg| \le \delta\bigg\}.$$

By Bernstein's inequality, we have $\mathbb{P}(A \in \mathcal{A}_i^c) \le O(n^{-7})$ for some $\delta = o(1)$. Again, we shall adjust the value of δ so that (63), (64) and (66) are still true. We then have

$$\begin{aligned}
&\mathbb{P}\bigg(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -(1-3\delta)\bar{\Delta}_i\bigg) \\
&\quad\le \sup_{A\in\mathcal{A}_i}\mathbb{P}\bigg(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))}{\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi(\theta_i^* - \theta_j^*)\psi(\theta_j^* - \theta_i^*)} \le -(1-3\delta)\bar{\Delta}_i | A\bigg) \\
&\qquad+ \mathbb{P}\big(A \in \mathcal{A}_i^c\big) \\
&\quad\le \sup_{A\in\mathcal{A}_i}\exp\bigg(-\frac{\frac{1}{2}(1-3\delta)^2\bar{\Delta}_i^2(L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^* - \theta_j^*))^2}{L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^* - \theta_j^*) + \frac{1-3\delta}{3}\bar{\Delta}_i L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^* - \theta_j^*)}\bigg)
\end{aligned} \tag{68}$$

$$+ O(n^{-7})$$

$$= \exp\Big(-\frac{1+o(1)}{2}\bar{\Delta}_i^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*)\Big) + O(n^{-7}) \tag{69}$$

$$\le \exp\Big(-\frac{1+o(1)}{2}\big(\eta(\theta_k^* - \theta_{k+1}^*) + (\theta_i^* - \theta_k^*)\big)^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*)\Big) \tag{70}$$

$$+ O(n^{-7})$$

$$\le \exp\Big(-\frac{1+o(1)}{2}\big(\bar{\Delta} + (\theta_i^* - \theta_k^*)\big)^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*)\Big) + O(n^{-7}). \tag{71}$$

The bound (68) is by Bernstein's inequality. We then use the definition of $\mathcal{A}_i$ to obtain the expression (69). To see why (70) is true, note that when $\bar{\Delta}_i^2 = \sqrt{\frac{\log n}{np}}$, the first term of (69) can be absorbed into $O(n^{-7})$. Finally, in (71), we have used the notation $\bar{\Delta} = \min(\eta(\theta_k^* - \theta_{k+1}^*), (\frac{\log n}{np})^{1/4})$. For each $j \in [n]$, define

$$h_j(t) = (\bar{\Delta} + t)^2 \psi'(t + \theta_k^* - \theta_j^*), \quad \text{for all } t \ge 0.$$

The derivative of this function is

$$h_j'(t) = (\bar{\Delta} + t)\psi'(t + \theta_k^* - \theta_j^*)\big[2 + (\bar{\Delta} + t)\big(1 - 2\psi\big(t + \theta_k^* - \theta_j^*\big)\big)\big].$$

Since $\max_{j,k} |\theta_k^* - \theta_j^*| = O(1)$, we can find a sufficiently small constant $c_2 > 0$, such that $h_j(t)$ is increasing on $[0, c_2]$. Moreover, there exists another small constant $c_3 > 0$ such that $\min_{t\in(c_2,\kappa]} h_j(t) \ge c_3$. With this fact, we can bound the exponent of (71) as

$$\begin{aligned}
&\big(\bar{\Delta} + (\theta_i^* - \theta_k^*)\big)^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*) \\
&\quad \ge Lp \sum_{j\in[n]\setminus\{i\}} \min\big(\bar{\Delta}^2 \psi'(\theta_k^* - \theta_j^*), c_3\big) \\
&\quad \ge Lp(k-1)\min\big(\bar{\Delta}^2\psi'(\theta_1^* - \theta_k^*), c_3\big) + Lp(n-k)\min\big(\bar{\Delta}^2\psi'(\theta_k^* - \theta_n^*), c_3\big) \qquad (72) \\
&\quad = Lp\bar{\Delta}^2\big((k-1)\psi'(\theta_1^* - \theta_k^*) + (n-k)\psi'(\theta_k^* - \theta_n^*)\big) \\
&\quad \ge (1+o(1))Lp\min\Big(\eta^2\Delta^2, \sqrt{\frac{\log n}{np}}\Big)\frac{n}{V(\kappa)}
\end{aligned}$$

where the equality (72) uses the fact that $\bar{\Delta} \to 0$. Therefore, we can further bound (71) as

$$\begin{aligned}
&\exp\Big(-\frac{1+o(1)}{2}Lp\min\Big(\eta^2\Delta^2, \sqrt{\frac{\log n}{np}}\Big)\frac{n}{V(\kappa)}\Big) + O(n^{-7}) \\
&\quad \le \exp\Big(-\frac{(1+o(1))\eta^2\Delta^2 npL}{2V(\kappa)}\Big) + O(n^{-7}).
\end{aligned}$$

The last inequality holds because when $\min(\eta^2\Delta^2, \sqrt{\frac{\log n}{np}}) = \sqrt{\frac{\log n}{np}}$, the first term becomes $\exp(-\frac{(1+o(1))L\sqrt{np\log n}}{2V(\kappa)})$, which can be absorbed by $O(n^{-7})$. Since $\exp(-\bar{\Delta}_i^{3/2}Lnp) + \exp(-\bar{\Delta}_i^2 npL\frac{np}{\log n}) \le \exp(-\frac{(1+o(1))\eta^2\Delta^2 npL}{2V(\kappa)}) + O(n^{-7})$, we have

$$\mathbb{P}\big(\widehat{\theta}_i \le (1-\eta)\theta_k^* + \eta\theta_{k+1}^*\big)\big) \le \exp\Big(-\frac{(1-\delta')\eta^2\Delta^2 npL}{2V(\kappa)}\Big) + O(n^{-7}), \tag{73}$$

with some $\delta' = o(1)$ for all $i \le k$. With a similar argument, we also have

$$\mathbb{P}\big(\widehat{\theta}_i \ge (1-\eta)\theta_k^* + \eta\theta_{k+1}^*\big)\big) \le \exp\bigg(-\frac{(1-\delta')(1-\eta)^2\Delta^2 npL}{2V(\kappa)}\bigg) + O\big(n^{-7}\big), \tag{74}$$

for all all $i \ge k+1$. It can be checked that the δ' above is independent of the $\bar{\delta}$ in the definition of η. Now we can choose η as in (54) with $\bar{\delta} = \delta'$. By Lemma 3.1, we have

$$\begin{aligned}\mathbb{E}\mathsf{H}_k(\widehat{r}, r^*) &\le \exp\bigg(-\frac{(1-\bar{\delta})\eta^2\Delta^2 npL}{2V(\kappa)}\bigg) \\ &\quad + \frac{n-k}{k}\exp\bigg(-\frac{(1-\bar{\delta})(1-\eta)^2\Delta^2 npL}{2V(\kappa)}\bigg) + O\big(n^{-7}\big) \\ &\le 2\exp\bigg(-\frac{1}{2}\bigg(\frac{\sqrt{(1-\bar{\delta})\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1-\bar{\delta})\mathsf{SNR}}}\log\frac{n-k}{k}\bigg)^2\bigg) + O\big(n^{-7}\big).\end{aligned}$$

By Markov's inequality, the above bound implies

$$\mathsf{H}_k(\widehat{r}, r^*) \le \exp\bigg(-\frac{1}{2}\bigg(\frac{\sqrt{(1-\delta_1)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1-\delta_1)\mathsf{SNR}}}\log\frac{n-k}{k}\bigg)^2\bigg) + O\big(n^{-6}\big),$$

for some $\delta_1 = o(1)$ with high probability. One can take, for example,

$$\delta_1 = \bar{\delta} + \frac{1}{\frac{\sqrt{(1-\bar{\delta})\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1-\bar{\delta})\mathsf{SNR}}}\log\frac{n-k}{k}}.$$

When $O(n^{-6})$ dominates the bound, we have $\mathsf{H}_k(\widehat{r}, r^*) = O(n^{-6})$, which implies $\mathsf{H}_k(\widehat{r}, r^*) = 0$ since $\mathsf{H}_k(\widehat{r}, r^*) \in \{0, (2k)^{-1}, 2(2k)^{-1}, 3(2k)^{-1}, \dots, 1\}$. Therefore, we always have

$$\mathsf{H}_k(\widehat{r}, r^*) \le 2\exp\bigg(-\frac{1}{2}\bigg(\frac{\sqrt{(1-\delta_1)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1-\delta_1)\mathsf{SNR}}}\log\frac{n-k}{k}\bigg)^2\bigg),$$

with high probability for some $\delta_1 = o(1)$. The proof is complete. □

PROOF OF THEOREM 3.3. With some rearrangements, the condition is equivalent to

$$\frac{npL\Delta^2}{2(1+\epsilon)V(\kappa)}\bigg(\frac{1}{2} - \frac{(1+\epsilon)V(\kappa)}{npL\Delta^2}\log\frac{n-k}{k}\bigg)^2 > \log k.$$

Since ϵ is a constant, it implies

$$\frac{npL\Delta^2}{2V(\kappa)}\bigg(\frac{1}{2} - \frac{V(\kappa)}{(1-\delta)npL\Delta^2}\log\frac{n-k}{k}\bigg)^2 > (1+\epsilon)\log k,$$

for any $\delta = o(1)$. Therefore, $\mathsf{H}_k(\widehat{r}, r^*) = o(k^{-1})$ when $k \to \infty$. Given the fact that $\mathsf{H}_k(\widehat{r}, r^*) \in \{0, (2k)^{-1}, 2(2k)^{-1}, 3(2k)^{-1}, \dots, 1\}$, we must have $\mathsf{H}_k(\widehat{r}, r^*) = 0$. When $k = O(1)$, the condition implies $\frac{npL\Delta^2}{2V(\kappa)} > (1+\epsilon')\log n$ for some constant $\epsilon' > 0$. This leads to the fact that $(\frac{1}{2} - \frac{V(\kappa)}{(1-\delta)npL\Delta^2}\log\frac{n-k}{k})^2 > c_1$ for some constant $c_1 > 0$. Therefore, $\mathsf{H}_k(\widehat{r}, r^*) = o(1) = o(k^{-1})$, which implies $\mathsf{H}_k(\widehat{r}, r^*) = 0$. □

Funding. The first and second authors were supported in part by NSF CAREER award DMS-1847590 and NSF Grant CCF-1934931.

The third author was supported in part by NSF Grant DMS-2112988.

SUPPLEMENTARY MATERIAL

Supplement to "Partial recovery for top-k ranking: Optimality of MLE and suboptimality of the spectral method" (DOI: 10.1214/21-AOS2166SUPP; .pdf). The supplement [5] includes all the technical proofs. In Appendix A, we first give proofs for all the results established in Section 4: Theorem 4.1, Theorem 4.2 and Theorem 4.3. After that, we prove Theorem 3.4 and Theorem 6.1 in Appendix B. We then include the proofs of Theorem 7.1 and Theorem 7.2 in Appendix C, the proof of Lemma 8.5 in Appendix D and the proofs of all the other technical lemmas in Appendix E. The count method is discussed in Appendix F.

REFERENCES

[1] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

[2] BORDA, J. D. (1784). Mémoire sur les élections au scrutin. Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784).

[3] 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 https://doi.org/10.2307/2334029

[4] BUTUCEA, C., NDAOUD, M., STEPANOVA, N. A. and TSYBAKOV, A. B. (2018). Variable selection with Hamming loss. *Ann. Statist.* **46** 1837–1875. MR3845003 https://doi.org/10.1214/17-AOS1572

[5] CHEN, P., GAO, C. and ZHANG, A. Y (2022). Supplement to "Partial Recovery for Top-k Ranking: Optimality of MLE and SubOptimality of the Spectral Method." https://doi.org/10.1214/21-AOS2166SUPP

[6] 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 https://doi.org/10.1137/1.9781611974782.81

[7] CHEN, Y., FAN, J., MA, C. and WANG, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.* **47** 2204–2235. MR3953449 https://doi.org/10.1214/18-AOS1745

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

[9] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

[10] COPELAND, A. (1951). A 'reasonable'social welfare function, Seminar on mathematics in social sciences, University of Michigan. Cited Indirectly from Its Mention by Luce and Raiffa (1957) 358.

[11] COSSOCK, D. and ZHANG, T. (2006). Subset ranking using regression. In *Learning Theory*. *Lecture Notes in Computer Science* **4005** 605–619. Springer, Berlin. MR2280634 https://doi.org/10.1007/11776420_44

[12] DWORK, C., KUMAR, R., NAOR, M. and SIVAKUMAR, D. (2001). Rank aggregation methods for the web. In *Proceedings of the* 10*th International Conference on World Wide Web* 613–622.

[13] ERDŐS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 17–61. MR0125031

[14] 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.

[15] JANG, M., KIM, S., SUH, C. and OH, S. (2017). Optimal sample complexity of m-wise data for top-k ranking. In *Advances in Neural Information Processing Systems* 1686–1696.

[16] LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *Ann. Statist.* **49** 2506–2530. MR4338373 https://doi.org/10.1214/20-aos2044

[17] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. ArXiv preprint. Available at arXiv:1612.02099.

[18] LUCE, R. D. (2012). *Individual Choice Behavior*: *A Theoretical Analysis*. Courier Corporation.

[19] MOTEGI, S. and MASUDA, N. (2012). A network-based dynamical ranking system for competitive sports. *Sci. Rep.* **2** 904. https://doi.org/10.1038/srep00904

[20] NDAOUD, M. and TSYBAKOV, A. B. (2020). Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inf. Theory* **66** 2517–2532. MR4087700 https://doi.org/10.1109/TIT.2020.2965738

[21] NEGAHBAN, S., OH, S. and SHAH, D. (2017). Rank centrality: Ranking from pairwise comparisons. *Oper. Res.* **65** 266–287. MR3613103 https://doi.org/10.1287/opre.2016.1534

[22] SHA, L., LUCEY, P., YUE, Y., CARR, P., ROHLF, C. and MATTHEWS, I. (2016). Chalkboarding: A new spatiotemporal query paradigm for sports play retrieval. In *Proceedings of the* 21*st International Conference on Intelligent User Interfaces* 336–347.

[23] SHAH, N. B. and WAINWRIGHT, M. J. (2017). Simple, robust and optimal ranking from pairwise comparisons. *J. Mach. Learn. Res.* **18** 199. MR3827087

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

[25] ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428