THE UNIVERSITY OF CHICAGO

RANK, RANKING AND PHASE TRANSITION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

PINHAN CHEN

CHICAGO, ILLINOIS
MARCH 2022

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my advisor, Prof. Chao Gao, for his consistent support, encouragement, and professionalism throughout my academic life. His acute insight and enthusiasm for statistics research not only lighted up my road, but also taught me what it takes to be a mature researcher. While giving me a great deal of freedom to explore my passion and interest, his hands-on guidance and detailed discussions have left an indelible trace on my future career. I would also like to thank my co-author, Prof. Anderson Ye Zhang, without whom many parts of this dissertation could not be accomplished, for many of his brilliant insights during our collaboration. I would like to thank Prof. Wei Biao Wu for his guidance in my very first research during my first year of Ph.D. Also, I'm greatly indebted to have him and Prof. Rina Foygel Barber, a world-renowned statistician I admire a lot, being on my dissertation committee and spending time reading this dissertation and providing valuable feedback. Besides, I would like to thank all faculty and staff in the Department of Statistics where I spent an unforgettable five years. I have been very fortunate to be in many friendships during my time in the University of Chicago. Thank you for your support along the way and bringing me plenty of happy memories. Finally, I would like to thank my wonderful and loving parents. Without their selfless love and endless support, I would not be here today.

ABSTRACT

The thesis consists of three topics. The first two topics are both about ranking under the Bradley-Terry-Luce (BTL) model. We first study the problem of top-k ranking, which is to optimally identify the set of top-k players from pairwise comparisons. We derive the minimax rate with respect to a normalized Hamming loss. 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 sub-optimal. Then we come to the problem of full ranking, which needs to provide the full rank of all players instead of just the top k. The minimax rate of this ranking problem is derived with respect to the Kendall's tau distance. The minimax rate of full ranking under this loss exhibits a phase transition between an exponential rate and a polynomial rate depending on the magnitude of the signal-to-noise ratio of the problem. To achieve the minimax rate, we propose a divide-and-conquer ranking algorithm that first divides the n players into groups of similar skills and then computes local MLE within each group. The third topic, instead, is about rank, where a Bayesian method for low-rank matrix estimation is proposed. We also explore the possibility of rank adaptation by proposing a rank-adaptive prior and have some preliminary results for the special case when the underlying signal is of rank 1.

CHAPTER 1

INTRODUCTION

Given partially observed pairwise comparison data from n players, we are interested in ranking the players according to their skills by aggregating the comparison results. This high-dimensional statistical estimation problem has important applications in many areas such as recommendation systems [7, 18], sports and gaming [9, 33, 58, 83, 84, 95], web search [32, 42], social choices [74, 79, 81, 82, 92], psychology [28, 76, 102], information retrieval [19, 72], etc. Two natural questions people can ask in the ranking problem is how to identify the top-k players, which we refer to as top-k ranking and how to give a full rank of all players, which we call full ranking. In this thesis, we will investigate both of the two problems thoroughly.

To formulate the ranking problem mathematically, we focus on arguably one of the most widely used parametric models, the Bradley-Terry-Luce (BTL) model [11, 77]. That is, suppose each player is associated with a skill parameter $w_i^* > 0$, we observe L games played between players i and j, and the outcome is modeled by

$$y_{ijl} \overset{ind}{\sim} \text{Bernoulli}\left(\frac{w_i^*}{w_i^* + w_j^*}\right), \quad l \in [L]. \tag{1.1}$$

where $[L] = \{1, 2, ..., L\}$ and $i < j$. We only observe outcomes from a small subset of pairs. This subset E is modeled by edges generated by an Erdős-Rényi random graph [44] with connection probability p on the n players. More details of the model will be given in later chapters. With the observations $\{y_{ijl}\}_{(i,j)\in E, l\in[L]}$, our goal is to optimally recover either the set of players with the largest skill parameters w_i^*'s or the full rank of w_i^*'s.

Besides the above mentioned two ranking problems, we have also investigated the problem of low-rank matrix estimation. Low-rank matrices find its applications in many areas such as machine learning [5, 41], signal processing [2, 36], finance [54, 78], and quantum states

tomography [55], etc. The central statistical problem is to estimate low-rank matrices from noisy or even incomplete and indirect observations. At its most general form, the statistical model of low-rank matrix estimation can be written as

$$X = \mathcal{A}(M_0) + Z \tag{1.2}$$

where M_0 is some unknown $p \times q$ matrix with a potentially low-rank structure, $\mathcal{A}$ is a given linear measurement operator (can be fixed or random) that maps $\mathbb{R}^{p\times q}$ to $\mathbb{R}^{s\times t}$, $Z \in \mathbb{R}^{s\times t}$ is the noise and $X \in \mathbb{R}^{s\times t}$ is our observation. In this thesis, we study the simplest possible specialization of (1.2) when $\mathcal{A}$ is the identity operator which corresponds to direct observation of a noisy version of each entry of M_0. While the statistical properties of this problem have been extensively studied in the frequentists' world, little is known about its Bayesian counterpart. We aim to analyze a Bayesian procedure in low-rank matrix estimation leveraging the posterior contraction framework [52]. A more detailed introduction of this topic will be outlined in Chapter 4.

We briefly introduce the content of the remaining chapters.

In Chapter 2, we study the problem of top-k ranking. 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 sub-optimal. Our results complement the recent work by [25] 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. This chapter is adapted from the author's paper [21].

In Chapter 3, we come to the problem of full ranking. For the first time in the literature, the minimax rate of this ranking problem is derived with respect to the Kendall's tau distance that measures the difference between two rank vectors by counting the number of inversions. The minimax rate of ranking exhibits a transition between an exponential rate and a polynomial rate depending on the magnitude of the signal-to-noise ratio of the problem. To the best of our knowledge, this phenomenon is unique to full ranking and has not been seen in any other statistical estimation problem. To achieve the minimax rate, we propose a divide-and-conquer ranking algorithm that first divides the n players into groups of similar skills and then computes local MLE within each group. The optimality of the proposed algorithm is established by a careful approximate independence argument between the two steps. This chapter is adapted from the author's paper [22].

In Chapter 4, we discuss a Bayesian procedure to estimate low-rank matrices. Given the data matrix generated from a low-rank matrix perturbed by Gaussian noise, we propose a Bayesian estimator with optimal posterior contraction rate. We also explore the possibility of rank adaptation by proposing a rank-adaptive prior and have some preliminary results for the special case when the underlying signal is of rank 1.

CHAPTER 2

PARTIAL RECOVERY FOR TOP-K RANKING: OPTIMALITY OF MLE AND SUB-OPTIMALITY OF SPECTRAL METHOD

2.1 Introduction

In this chapter, our goal is to study the statistical limits of both *partial* and *exact* recovery of the top-k ranking problem under the BTL model.

Theoretical properties of the top-k ranking problem have been studied by [24, 62, 97, 23, 63, 86, 25] 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 exactly with high probability. For this purpose, the state-of-the-art result is obtained by the recent work [25]. It was shown by [25] that both the MLE and the spectral method can 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 [25] 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 sub-optimal constant. In particular, the MLE achieves exact recovery when

$$npL\Delta^2 > 2.001V(\kappa)\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2, \tag{2.1}$$

and the spectral method requires

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

In the above two formulas, the constant 2.001 should be understood as any constant larger than 2. 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 2.2. 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 [25]. On the contrary, the current paper complements and refines the results of [25]. The optimality claim made by [25] 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 [25] demonstrates nearly identical performances of the MLE and the spectral method. Note that the experiment in [25] 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 2.5 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 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 dominates the spectral method, and the advantage of the MLE is usually quite significant.

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 2.3 that an optimal partial recovery error bound will lead to the optimal exact recovery condition (2.1). 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 sub-optimal 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 [73] 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 [24, 86, 25] 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 2.2. The results of the MLE and the spectral method will be given in Section 2.3 and Section 2.4, respectively. We then comprehensively compare the two methods in Section 2.5 by numerical experiments. Section 2.6 presents a minimax lower bound for partial recovery. In Section 2.7, we analyze the error rates of the MLE and the spectral method for each individual parameter. The proofs of our main results are given in Sections 2.8-2.11, with Section 2.8 for the analysis of the MLE, Section 2.9 for the analysis of the spectral method, Section 2.10 for the proofs of the lower bounds, and Section 2.11 for the proof of local error rates. Finally, a few technical lemmas will be given and proved in Section 2.12.

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 $\{1, 2, ..., d\}$. 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 $a_n \leq Cb_n$ 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.

2.2 Models and Methods

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, $A_{ij} \overset{iid}{\sim} \text{Bernoulli}(p)$ 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 (1.1). 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 (1.1) 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 (1.1) can be equivalently written as

$$y_{ijl} \overset{ind}{\sim} \text{Bernoulli}(\psi(\theta^*_{r_i^*} - \theta^*_{r_j^*})), \quad l = 1, \cdots, L. \tag{2.2}$$

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 (3.2), 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^*.

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

$$\mathsf{H}_k(\widehat{r}, r^*) = \frac{1}{2k}\left(\sum_{i=1}^{n} \mathbb{I}\left\{\widehat{r}_i > k, r_i^* \le k\right\} + \sum_{i=1}^{n} \mathbb{I}\left\{\widehat{r}_i \le k, r_i^* > k\right\}\right). \tag{2.3}$$

The definition (3.6) 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 (3.6) 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's tau distance*, defined as

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

Since $\mathsf{K}(\widehat{r}, r^*)$ counts all pairwise differences in the ranking relation, it is a stronger distance than (3.6). 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 Chapter 3.

As we have discussed in Section 2.1, 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 [15, 85]. 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

[15, 85] consider estimating β^* under the loss

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

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 $\{j \in [p] : |\widehat{\beta}_j| > a\} = \{j \in [p] : |\beta_j^*| > a\}$.

Parameter Space. For the nuisance parameter θ^* of the model (3.2), 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

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

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.

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 any $i < j$, we use the notation $\bar{y}_{ij} = \frac{1}{L}\sum_{l=1}^{L} y_{ijl}$. Then, the negative log-likelihood function

1. For readers who are familiar with [25], we note that our definitions of Δ and κ are slightly different from those in [25].

is given by

$$\ell_n(\theta) = \sum_{1\le i<j\le n} A_{ij}\left[\bar{y}_{ij}\log\frac{1}{\psi(\theta_i-\theta_j)} + (1-\bar{y}_{ij})\log\frac{1}{1-\psi(\theta_i-\theta_j)}\right]. \tag{2.4}$$

Define the MLE,

$$\widehat{\theta} \in \underset{\theta:\mathbb{1}_n^T\theta=0}{\operatorname{argmin}}\ \ell_n(\theta). \tag{2.5}$$

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 (2.5) is written in its vanilla version, without any constraint or penalty. To the best of our knowledge, (2.5) has not been previously analyzed in the literature.

Another popular algorithm for ranking is the spectral method, also known as Rank Centrality proposed by [86]. Define a matrix $P \in \mathbb{R}^{n\times n}$ by

$$P_{ij} = \begin{cases} \frac{1}{d}A_{ij}\bar{y}_{ji}, & i \ne j, \\ 1 - \frac{1}{d}\sum_{l\in[n]\setminus\{i\}} A_{il}\bar{y}_{li}, & i = j, \end{cases} \tag{2.6}$$

where d needs to be at least the maximum degree of the random graph A. Throughout the paper, we just set $d = 2np$, same as the convention in [25], which is because $2np > \max_{i\in[n]}\sum_{j\in[n]\setminus\{i\}} A_{ij}$ with high probability due to concentration. We also adopt the convention of notation that $A_{ij} = A_{ji}$ and $\bar{y}_{ij} = 1 - \bar{y}_{ji}$ for each (i,j) pair. One can check that P is a transition matrix of a Markov chain. To see why P is useful, we can compute the conditional expectation of P given the random graph A,

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

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

$$(\pi^*)^T = \left(\frac{\exp(\theta^*_{r_1^*})}{\sum_{i=1}^n \exp(\theta^*_{r_i^*})}, \cdots, \frac{\exp(\theta^*_{r_n^*})}{\sum_{i=1}^n \exp(\theta^*_{r_i^*})} \right).$$

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 $\widehat{r}$.

2.3 Results for the MLE

We study the property of MLE in this section. Our first result gives theoretical guarantees for (2.5) under both ℓ_2 and ℓ_∞ loss functions.

Theorem 2.3.1. *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 (2.5), we have*

$$\sum_{i=1}^n (\widehat{\theta}_i - \theta^*_{r_i^*})^2 \leq C \frac{1}{pL}, \tag{2.7}$$

$$\max_{i \in [n]} |\widehat{\theta}_i - \theta^*_{r_i^*}|^2 \leq C \frac{\log n}{npL}, \tag{2.8}$$

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 2.3.1. We have established that the MLE achieves the error rates $O\left(\frac{1}{pL}\right)$ and $O\left(\frac{\log n}{npL}\right)$ for the squared ℓ_2 loss and the squared ℓ_∞ loss, respectively. Both error rates are known to be optimal in the literature [86, 25]. Since the BTL model (3.2) 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 (2.7) and (2.8) would depend on κ exponentially [86, 25]. This is because the eigenvalues of the Hessian of the objective function of (2.5) 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 2.3.1 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 [24, 86, 25] all work with regularized MLE

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta : \mathbb{1}_n^T \theta = 0} \left[\ell_n(\theta) + \frac{\lambda}{2} \|\theta\|^2 \right]. \tag{2.9}$$

In particular, the recent paper [25] shows that $\widehat{\theta}_\lambda$ also achieves the optimal rates (2.7) and (2.8) for a λ that is chosen appropriately, though in practice it is known that the vanilla MLE performs very well. Theorem 2.3.1 shows that penalty is not needed for the MLE to be optimal, thus closing a gap between theory and practice.

The proof of Theorem 2.3.1 is built upon the elegant leave-one-out technique in [25]. We

first show that with a sufficiently small λ, a (sub-optimal) 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 (2.7) and (2.8). We also note that our leave-one-out argument is actually different from the form used in [25]. While the leave-one-out argument in [25] 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 [25]. A detailed analysis of the MLE will be given in Section 2.8.

Next, we study the theoretical property of $\widehat{r}$, the rank induced by the MLE $\widehat{\theta}$. Without loss of generality, let us assume $k \leq \frac{n}{2}$ 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

$$V(\kappa) = \max_{\substack{\kappa_1+\kappa_2 \leq \kappa \\ \kappa_1, \kappa_2 \geq 0}} \frac{n}{k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2)}. \tag{2.10}$$

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 2.3.2. *Assume $\frac{np}{\log n} \to \infty$ and $\kappa \leq c_1$ for some constant $c_1 > 0$. Then, for the*

rank vector $\widehat{r}$ that is induced by the MLE (2.5), there exists some $\delta = o(1)$, such that

$$\mathsf{H}_k(\widehat{r}, r^*) \le C \exp\left(-\frac{1}{2} \left(\frac{\sqrt{(1-\delta)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1-\delta)\mathsf{SNR}}} \log \frac{n-k}{k} \right)_+^2 \right), \tag{2.11}$$

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

Corollary 2.3.1. *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 (2.5) satisfies*

$$\mathsf{H}_k(\widehat{r}, r^*) \le \exp\left(-(1 - o(1)) \frac{\mathsf{SNR}}{8} \right), \tag{2.12}$$

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 (2.12) is in a typical form of optimal clustering error [108, 75]. 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 (2.11). The extra term $\log \frac{n-k}{k}$ characterizes the unbalancedness of the two clusters. We note that for variable selection under Hamming loss [15, 85], the optimal rate is very similar to the form of (2.11). 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 2.3.2 and Corollary 2.3.1 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 $\{i \in [n] : r_i^* \leq k\}$, we need to have $\mathsf{H}_k(\widehat{r}, r^*) \to 0$. When $k = n/4$, Corollary 2.3.1 shows that $\mathsf{H}_k(\widehat{r}, r^*) \to 0$ is achieved when $\mathsf{SNR} \to \infty$. In comparison, when $k = 5$, Theorem 2.3.2 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 (2.11).

Compared with Theorem 2.3.1, we need a slightly stronger condition $\frac{np}{\log n} \to \infty$ for Theorem 2.3.2 and Corollary 2.3.1. If we only assume $p \geq c_0 \frac{\log n}{n}$, the $1-\delta$ factor in the exponent of (2.11) 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 2.3.2 relies on a very interesting lemma that is stated below.

Lemma 2.3.1. *Suppose $\widehat{r}$ is a rank vector induced by $\widehat{\theta}$, we then have*

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

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

We will prove Lemma 2.3.1 in Section 2.12. 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

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

We then obtain the exponential error bound (2.11) by carefully analyzing the probability $\mathbb{P}(\widehat{\theta}_i \leq t)$ (or $\mathbb{P}(\widehat{\theta}_i \geq t)$) for each $i \in [n]$. The analysis of $\mathbb{P}(\widehat{\theta}_i \leq t)$ 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 [25] to decouple the dependence between the data

and the coefficients of the local linear expansion. The details will be given in Section 2.8.

The result of Theorem 2.3.2 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

$$\mathsf{H}_k(\widehat{r}, r^*) \in \{0, (2k)^{-1}, 2(2k)^{-1}, 3(2k)^{-1}, \cdots, 1\}. \tag{2.13}$$

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 (2.11) is smaller than $(2k)^{-1}$, we obtain exact recovery of the top-k set. This result is stated as follows.

Theorem 2.3.3. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, *and*

$$\frac{npL\Delta^2}{V(\kappa)} > (1+\epsilon)2\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2, \tag{2.14}$$

for some arbitrarily small constant $\epsilon > 0$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the MLE (2.5), 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 universal constant c_0 without affecting the conclusion of Theorem 2.3.3. The result of Theorem 2.3.3 improves the exact recovery threshold obtained in the literature. The paper [25] proves that the MLE exactly recovers the top-k set when $npL\Delta^2 > C\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2$ for some sufficiently large constant $C > 0$. We complement the result of [25] 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 (2.14) agrees with the understanding that top-k ranking and bottom-k ranking are mathematically equivalent.

The next theorem shows that the exact recovery threshold (2.14) is optimal, and cannot be further improved.

Theorem 2.3.4. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, $(\log n)^8 = O(L)$, *and*

$$\frac{npL\Delta^2}{V(\kappa)} < (1-\epsilon)2\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2, \tag{2.15}$$

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

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

where we use the notation $\mathbb{P}_{(\theta^*, r^*)}$ *for the data generating process (3.2).*

The proof of Theorem 2.3.4 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 [27] for this purpose. Without this additional technical condition, we are not aware of any probabilistic tool to deal with the maximum of dependent binomial random variables.

Theorem 2.3.3 and Theorem 2.3.4 together nail down the phase transition boundary of exact recovery, which is $\frac{npL\Delta^2}{V(\kappa)} = 2\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2$. Thus, the MLE is an optimal procedure that achieves this boundary. The lower bound result of Theorem 2.3.4 also suggests that the partial recovery error rate obtained in Theorem 2.3.2 cannot be improved, since otherwise one would obtain a better SNR condition for exact recovery in Theorem 2.3.3. A rigorous minimax lower bound for partial recovery will be given in Section 2.6.

2.4 Results for the Spectral Method

In this section, we study the theoretical property of the spectral method, also known as rank centrality [86]. Let $\widehat{\pi}$ be the stationary distribution of the Markov chain with transition probability (3.5). The estimation error of $\widehat{\pi}$ has already been investigated by [86, 25]. For both ℓ_2 and ℓ_∞ loss functions, it has been shown by [25] that $\widehat{\pi}$ achieves the optimal rates

(2.7) and (2.8) 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,

$$\overline{V}(\kappa) = \max_{\substack{\kappa_1+\kappa_2\le\kappa \\ \kappa_1,\kappa_2\ge 0}} \frac{k\psi'(\kappa_1)(1+e^{\kappa_1})^2 + (n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2}{(k\psi(\kappa_1)+(n-k)\psi(-\kappa_2))^2/n}. \tag{2.16}$$

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 2.4.1. *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 (3.5), there exists some $\delta = o(1)$, such that*

$$H_k(\widehat{r}, r^*) \le C\exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1-\delta)\overline{SNR}}}{2} - \frac{1}{\sqrt{(1-\delta)\overline{SNR}}}\log\frac{n-k}{k}\right)_+^2\right), \tag{2.17}$$

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 (2.17) 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 (2.11). The only difference lies in the effective variance of the two methods. We will show in Lemma 2.5.1 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 2.5.

By the property (2.13), we immediately obtain an exact recovery result from Theorem 2.4.1.

Theorem 2.4.2. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, *and*

$$\frac{npL\Delta^2}{\overline{V}(\kappa)} > (1+\epsilon)2\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2, \tag{2.18}$$

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 (3.5), 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 [25] that the spectral method exactly recovers the top-k set when $npL\Delta^2 > C\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2$ 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 2.3.4, the exact recovery threshold (2.18) 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 sub-optimality of the spectral method is intrinsic.

Theorem 2.4.3. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa \le c_1$ *for some constant* $c_1 > 0$, $k \to \infty$ *and*

$$\frac{npL\Delta^2}{\overline{V}(\kappa)} < (1-\epsilon)2\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)^2, \tag{2.19}$$

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 (3.5), we have*

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

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

$$\sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*, r^*)} H_k(\widehat{r}, r^*) \geq C \exp\left(-\frac{1}{2} \left(\frac{\sqrt{(1+\delta)\overline{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta)\overline{SNR}}} \log \frac{n-k}{k} \right)_+^2 \right), \tag{2.20}$$

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

Theorem 2.4.3 shows that the results of Theorem 2.4.1 and Theorem 2.4.2 on the performance of spectral method are sharp, under the additional condition that $k \to \infty$. The conclusion of Theorem 2.4.3 can also be extended to the case of $k = O(1)$ via a similar argument that is used in the proof of Theorem 2.3.4, 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 $p \geq c_0 \frac{\log n}{n}$ for some sufficiently large universal constant $c_0 > 0$, as long as the δ in (2.17) and (2.20) are replaced by some sufficiently small constant.

2.5 Comparison of the Two Methods

In this section, we compare the MLE and the spectral method based on the results obtained in Section 2.3 and Section 2.4. 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 2.1 that $\overline{V}(\kappa) \geq V(\kappa)$ for all $\kappa \geq 0$. This inequality is rigorously established by the following lemma.

Lemma 2.5.1. *For $V(\kappa)$ and $\overline{V}(\kappa)$ defined in (2.10) and (2.16), respectively, we have*

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

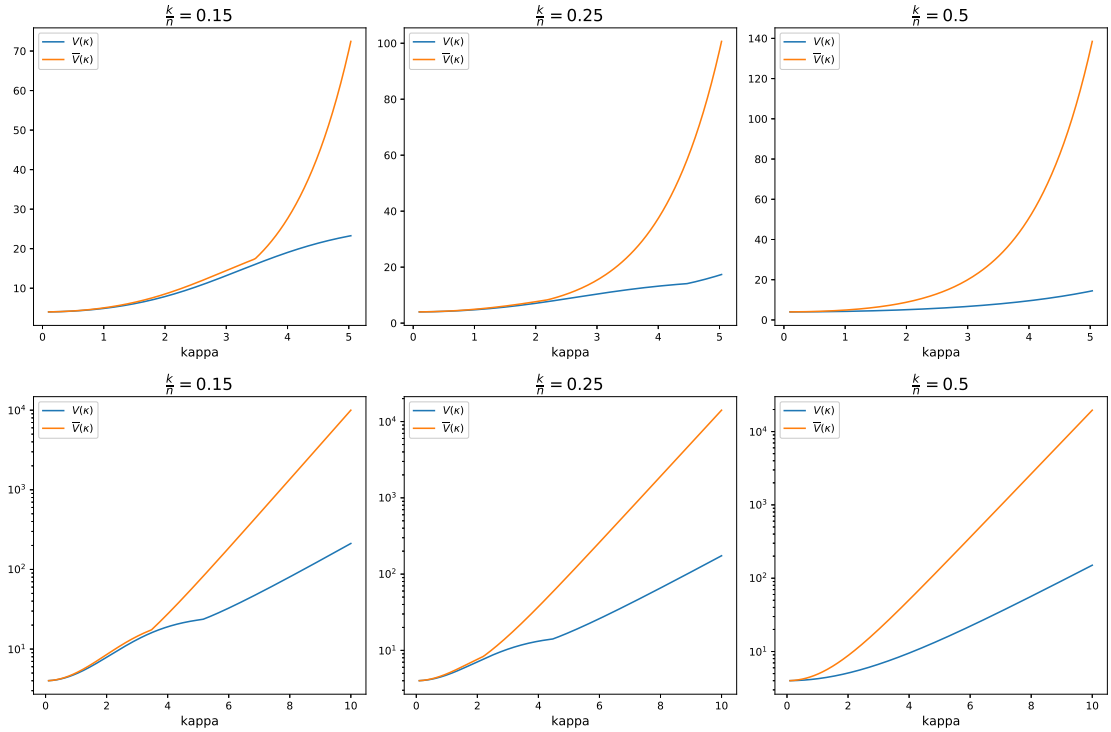

Figure 2.1: *The functions $V(\kappa)$ and $\overline{V}(\kappa)$ with $k/n \in \{0.15, 0.25, 0.5\}$. In the first row, we plot the functions for $\kappa \in [0, 5]$. The second row plots the same functions for $\kappa \in [0, 10]$ in a logarithmic scale to better illustrate the global structure. Recall the definition of $V(\kappa)$ in (2.10) and $\overline{V}(\kappa)$ in (2.16). 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 $k/n \le 1/2$).*

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

Proof. Recall the definition of $V(\kappa)$ in (2.10) and $\overline{V}(\kappa)$ in (2.16) and the roles of κ_1, κ_2 in the maximization. By Jensen's inequality, we have

$$\frac{k\frac{e^{\kappa_1}}{(1+e^{\kappa_1})^2} + (n-k)\frac{e^{-\kappa_2}}{(1+e^{-\kappa_2})^2}}{ke^{\kappa_1} + (n-k)e^{-\kappa_2}} \ge \left(\frac{k\frac{e^{\kappa_1}}{1+e^{\kappa_1}} + (n-k)\frac{e^{-\kappa_2}}{1+e^{-\kappa_2}}}{ke^{\kappa_1} + (n-k)e^{-\kappa_2}} \right)^2. \tag{2.21}$$

Another way to see the above inequality is to construct a random variable X such that $\mathbb{P}\left(X = \frac{1}{1+e^{\kappa_1}}\right) = \frac{ke^{\kappa_1}}{ke^{\kappa_1}+(n-k)e^{-\kappa_2}}$ and $\mathbb{P}\left(X = \frac{1}{1+e^{-\kappa_2}}\right) = \frac{(n-k)e^{-\kappa_2}}{ke^{\kappa_1}+(n-k)e^{-\kappa_2}}$. Then, (2.21) is

equivalent to $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$. The inequality (2.21) can be rearranged into

$$\frac{k\psi'(\kappa_1)(1+e^{\kappa_1})^2 + (n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2}{(k\psi(\kappa_1)+(n-k)\psi(-\kappa_2))^2/n} \geq \frac{n}{k\psi'(\kappa_1)+(n-k)\psi'(\kappa_2)}. \tag{2.22}$$

Taking maximum over κ_1 and κ_2 on both sides, we obtain the inequality $\overline{V}(\kappa) \geq V(\kappa)$. When $\kappa = 0$, we obviously have $V(\kappa) = \overline{V}(\kappa)$. When $\kappa > 0$, we need to show $V(\kappa) \neq \overline{V}(\kappa)$. The optimization of $V(\kappa)$ must be achieved by some $(\kappa_1^*, \kappa_2^*) \neq (0,0)$. For such (κ_1^*, κ_2^*), the constructed random variable X has a positive variance, and thus both inequalities (2.21) and (2.22) are strict. We then have

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

The proof is complete. □

The comparison between $V(\kappa)$ and $\overline{V}(\kappa)$ shows that the spectral method is not optimal in general. The rate in (2.17) and (2.20) has a worse error exponent than that of (2.11) 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 2.1.

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 2.7, 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.

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 2.2 shows the results for both

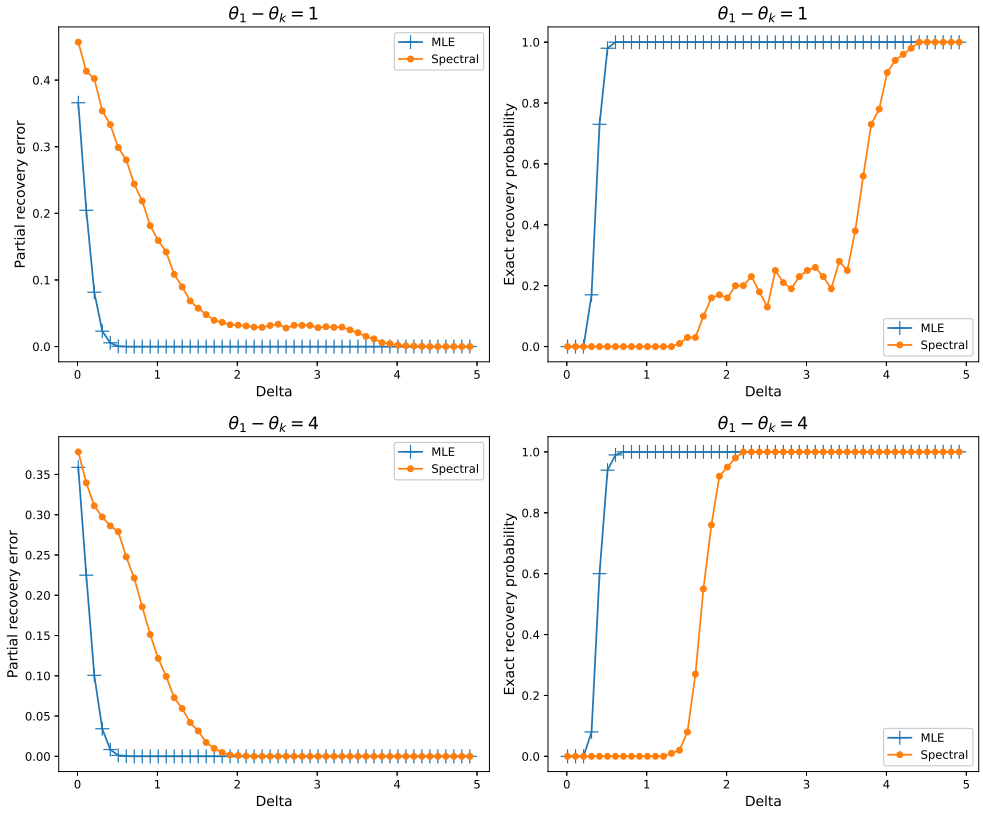

Figure 2.2: *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is chosen to be a piecewise constant vector of four pieces of sizes $25, 25, 75, 75$. The plots are obtained by averaging 100 independent experiments.*

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 2.3. Again, the MLE always outperforms the spectral method.

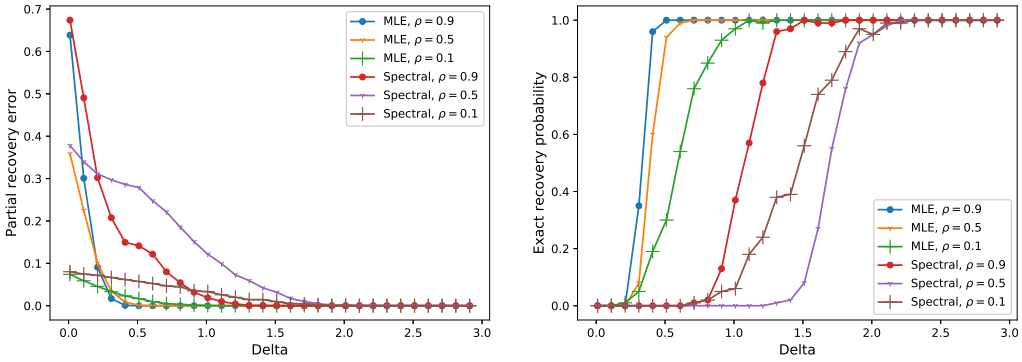

Figure 2.3: *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is chosen to be a piecewise constant vector of four pieces of sizes $50(1-\rho), 50\rho, 150(1-\rho), 150\rho$. The plots are obtained by averaging* 100 *independent 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 $\theta^*_2, \cdots, \theta^*_{50}$ from Uniform$[6, 10]$ and generate $\theta^*_{51}, \cdots, \theta^*_{199}$ 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

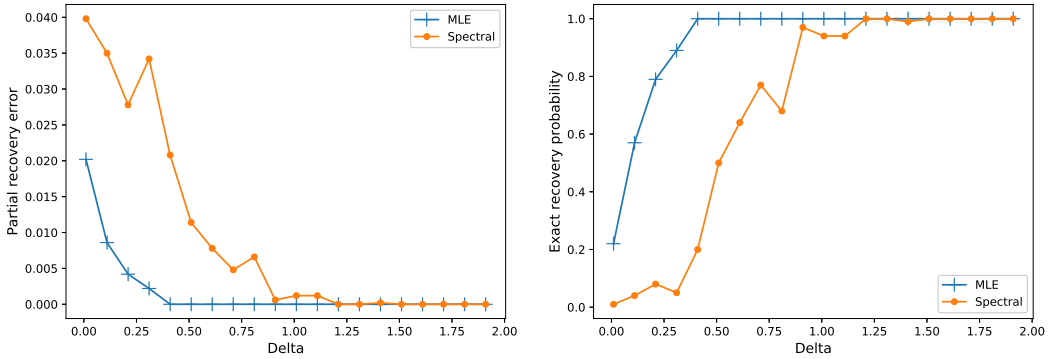

Figure 2.4: *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is randomly generated from some distribution. The plots are obtained by averaging* 100 *independent experiments.*

always outperforms the spectral method. The results are summarized in Figure 2.4 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 [25]. It is claimed in [25] 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 [25] 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 2.5 plots the results of partial and exact recovery with Δ varied from 0.01 to 0.55. For both partial

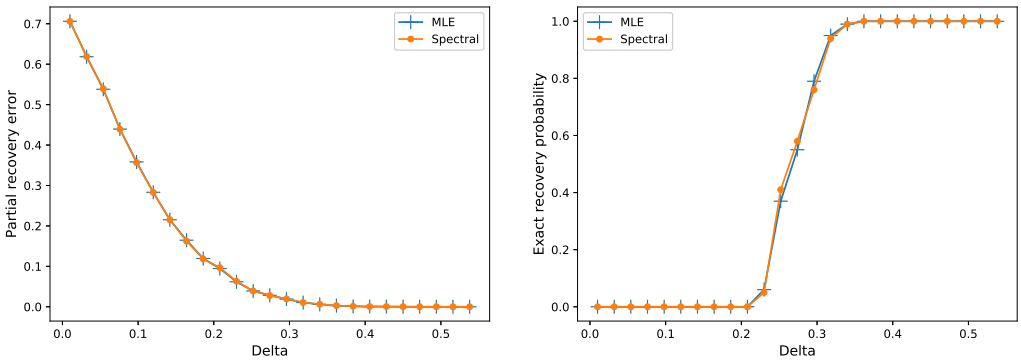

Figure 2.5: *The partial recovery error (left) and the exact recovery probability (right) for the MLE and the spectral method. The parameter θ^* is chosen to be a piecewise constant vector of two pieces of sizes* 50 *and* 150. *The plots are obtained by averaging* 100 *independent experiments.*

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 $\overline{\mathsf{SNR}} \gtrsim npL \gg \log n$, and the error bound (2.17) 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 2.5.1. *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 (3.5), there exists some $\delta = o(1)$, such that*

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

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

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

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 2.5.1, 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.

2.6 Minimax Lower Bound of Partial Recovery

The purpose of this section is to show that the partial recovery error rate (2.11) achieved by the MLE cannot be improved from a minimax perspective. We are able to establish

a matching lower bound for Theorem 2.3.2 using a slightly more general parameter space. Define

$$\Theta'(k,\Delta,\kappa) = \{\theta \in \mathbb{R}^n : \theta_1 \geq \cdots \geq \theta_n, \theta_k - \theta_{k+2} \geq \Delta, \theta_1 - \theta_n \leq \kappa\}. \tag{2.23}$$

Compared with $\Theta(k,\Delta,\kappa)$, the new definition (2.23) 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 (2.15) holds. This is the regime of partial recovery, since exact recovery is impossible by Theorem 2.3.4. We first remark that with a slight modification of the proof of Theorem 2.3.2, the MLE can be shown to achieve the same error rate (2.11) 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 (2.10) is achieved at $(\kappa_1,\kappa_2) = (\kappa_1^*,\kappa_2^*)$. Then, define $\theta_i' = \kappa_1^*$ for $1 \leq i \leq k - \rho k$, $\theta_i' = 0$ for $k - \rho k < i \leq k$, $\theta_i' = -\Delta$ for $k < i \leq k + \rho(n-k)$ and $\theta_i' = -\kappa_2^*$ for $k + \rho(n-k) < i \leq n$. 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,

$$\inf_{0\leq\phi\leq 1}\left[\mathbb{E}_{(\theta'',r'')}\phi + \frac{n-k-1}{k}\mathbb{E}_{(\theta',r')}(1-\phi)\right]. \tag{2.24}$$

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 (2.24) is the most important step in our lower bound analysis. A rigorous argument will be given in Section 2.10.

The testing problem (2.24) 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 (2.24) is given by the likelihood ratio test

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

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 (2.24) is relatively easy to quantify. A sharp lower bound can be obtained by a large deviation analysis.

Theorem 2.6.1. *Assume $\frac{np}{\log n} \to \infty$, $\kappa \leq c_1$ for some constant $c_1 > 0$, and (2.15) holds for some arbitrarily small constant $\epsilon > 0$. Then, there exists some $\delta = o(1)$, such that*

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

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

2.7 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 2.2. 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.

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 (2.10). For any $\theta \in \mathbb{R}^n$ and any $i \in [n]$, define

$$V_i(\theta) = \frac{n}{\sum_{j=1}^n \psi'(\theta_i - \theta_j)}. \tag{2.25}$$

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

$$R_1(S, \theta, t, \delta) = \sum_{i \in S} \exp\left(-\frac{(1-\delta)(\theta_i - t)_+^2 npL}{2V_i(\theta)}\right), \tag{2.26}$$

$$R_2(S, \theta, t, \delta) = \sum_{i \in S} \exp\left(-\frac{(1-\delta)(t - \theta_i)_+^2 npL}{2V_i(\theta)}\right). \tag{2.27}$$

Theorem 2.7.1. *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 (2.5), any small constant $0 < \delta < 0.1$, any $r^* \in \mathfrak{S}_n$ and any $\theta^* \in \Theta(k, 0, \kappa)$, we have*

$$\mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \le C_1 \left(\inf_t \frac{R_1([k], \theta^*, t, \delta) + R_2([n]\backslash[k], \theta^*, t, \delta)}{k} + n^{-3} \right), \tag{2.28}$$

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

$$\mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \ge C_2 \left(\inf_t \frac{R_1([k], \theta^*, t, -\delta) + R_2([n]\backslash[k], \theta^*, t, -\delta)}{k} \right), \tag{2.29}$$

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 2.7.1 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 (2.28) can be viewed as an extension of Theorem 2.3.2, though the δ in Theorem 2.3.2 is allowed to vanish because of the less general setting. The lower bound (2.29) 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 (2.19) in Theorem 2.4.3. 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 (2.28) or (2.29). Then, we can interpret $\sum_{i=1}^{k} \exp\left(-\frac{(\theta_i^* - t^*)_+^2 npL}{2V_i(\theta^*)}\right)$ as the order of the number of top k players that are ranked among the bottom group, and $\sum_{i=k+1}^{n} \exp\left(-\frac{(t^* - \theta_i^*)_+^2 npL}{2V_i(\theta^*)}\right)$ 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 2.7.1 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 (2.28) converges to zero as long as the majority of $\{\theta_i^*\}_{1 \le i \le k}$ are separated from most of $\{\theta_i^*\}_{k+1 \le i \le n}$.

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 (2.16). For any $\theta \in \mathbb{R}^n$ and any $i \in [n]$, define

$$\overline{V}_i(\theta) = \frac{n \sum_{j=1}^{n} \psi'(\theta_i - \theta_j)(1 + e^{\theta_j - \theta_i})^2}{\left(\sum_{j=1}^{n} \psi(\theta_j - \theta_i)\right)^2}. \tag{2.30}$$

We also introduce two quantities similar to (2.26) and (2.27),

$$\begin{aligned}\overline{R}_1(S,\theta,t,\delta) &= \sum_{i\in S}\exp\left(-\frac{(1-\delta)(\theta_i-t)_+^2 npL}{2\overline{V}_i(\theta)}\right),\\ \overline{R}_2(S,\theta,t,\delta) &= \sum_{i\in S}\exp\left(-\frac{(1-\delta)(t-\theta_i)_+^2 npL}{2\overline{V}_i(\theta)}\right).\end{aligned}$$

Theorem 2.7.2. *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 (3.5), any small constant $0<\delta<0.1$, any $r^*\in\mathfrak{S}_n$ and any $\theta^*\in\Theta(k,0,\kappa)$, we have*

$$\mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r},r^*)\le C_1\left(\inf_t\frac{\overline{R}_1([k],\theta^*,t,\delta)+\overline{R}_2([n]\backslash[k],\theta^*,t,\delta)}{k}+n^{-3}\right), \tag{2.31}$$

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

$$\mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r},r^*)\ge C_2\left(\inf_t\frac{\overline{R}_1([k],\theta^*,t,-\delta)+\overline{R}_2([n]\backslash[k],\theta^*,t,-\delta)}{k}\right), \tag{2.32}$$

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 2.7.1, Theorem 2.7.2 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 2.7.1 and Theorem 2.7.2 can be further extended beyond the setting of 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 2.7.1 and Theorem 2.7.2 continue to hold with $\frac{V_i(\theta)}{npL}$ and $\frac{\overline{V}_i(\theta)}{npL}$ replaced

by $\frac{\sum_{j\in[n]\setminus\{i\}} \frac{p_{ij}}{L_{ij}}\psi(\theta_i-\theta_j)\psi(\theta_j-\theta_i)}{\left(\sum_{j\in[n]\setminus\{i\}} p_{ij}\psi(\theta_i-\theta_j)\psi(\theta_j-\theta_i)\right)^2}$ and $\frac{\sum_{j\in[n]\setminus\{i\}} \frac{p_{ij}}{L_{ij}}\psi(\theta_i-\theta_j)\psi(\theta_j-\theta_i)\left(1+e^{\theta_j-\theta_i}\right)^2}{\left(\sum_{j\in[n]\setminus\{i\}} p_{ij}\psi(\theta_j-\theta_i)\right)^2}$, respectively.

Comparison of the Two Methods for each θ^*. Theorem 2.7.1 and Theorem 2.7.2 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

$$\inf_t \frac{1}{k}\left[\sum_{i=1}^{k}\exp\left(-\frac{(\theta_i^*-t)_+^2 npL}{2V_i(\theta^*)}\right)+\sum_{i=k+1}^{n}\exp\left(-\frac{(t-\theta_i^*)_+^2 npL}{2V_i(\theta^*)}\right)\right], \tag{2.33}$$

and

$$\inf_t \frac{1}{k}\left[\sum_{i=1}^{k}\exp\left(-\frac{(\theta_i^*-t)_+^2 npL}{2\overline{V}_i(\theta^*)}\right)+\sum_{i=k+1}^{n}\exp\left(-\frac{(t-\theta_i^*)_+^2 npL}{2\overline{V}_i(\theta^*)}\right)\right]. \tag{2.34}$$

It is clear that the only difference between (2.33) and (2.34) lies in the difference of the variance functions (2.25) and (2.30), whose comparison is given by the following lemma.

Lemma 2.7.1. *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 $\theta_1^* = ... = \theta_n^*$.*

Proof. Notice the following chain of equalities and inequality,

$$\begin{aligned} V_i(\theta^*) &= \frac{n}{\sum_{j\in[n]}\psi'(\theta_j^*-\theta_i^*)} \\ &= \frac{n\left(\sum_{j\in[n]} e^{\theta_j^*-\theta_i^*}\right)}{\left(\sum_{j\in[n]}\psi'(\theta_j^*-\theta_i^*)\right)\left(\sum_{j\in[n]} e^{\theta_j^*-\theta_i^*}\right)} \\ &\le \frac{n\left(\sum_{j\in[n]}\psi'(\theta_j^*-\theta_i^*)(1+e^{\theta_j^*-\theta_i^*})^2\right)}{\left(\sum_{j\in[n]}\psi(\theta_j^*-\theta_i^*)\right)^2} \qquad (2.35) \\ &= \overline{V}_i(\theta^*), \end{aligned}$$

where (2.35) is by 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 $\theta_1^* = ... = \theta_n^*$. □

To close this section, we discuss two special cases of θ^*, under which the error rates recover the results of Theorem 2.3.2, Theorem 2.4.1, and Corollary 2.5.1.

Example 2.7.1. *According to the proof of Theorem 2.4.3 and the construction discussed in Section 2.6, 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 (2.10) for the MLE or (2.16) for the spectral method, and* ρ *is a sufficiently small constant. For this* θ^*, *the formulas (2.33) and (2.34) recover the minimax rates obtained in Theorem 2.3.2 and Theorem 2.4.1.*

Example 2.7.2. *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 (2.33) and (2.34) are* $o(k^{-1})$, *which implies both the MLE and the spectral method achieve exact recovery with high probability. As shown in Corollary 2.5.1, 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 [25].*

2.8 Analysis of the MLE

In this section, we analyze the MLE (2.5), and prove Theorem 2.3.1, Theorem 2.3.2, and Theorem 2.3.3. Since the BTL model (3.2) 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 2.8.1. We then present a few technical lemmas in Section 2.8.2. In Section 2.8.3, we establish an important result on the ℓ_∞ bound of the MLE. Theorem 2.3.1 will be proved in Section 2.8.4. Finally, we prove Theorem 2.3.2 and Theorem 2.3.3 in Section 2.8.5.

2.8.1 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 (2.4). Its Hessian $\nabla^2 \ell_n(\theta) = H(\theta) \in \mathbb{R}^{n \times n}$ is given by the formula

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

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 [103]. 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 2.8.1. *Under the setting of Theorem 2.3.1, we have*

$$\|\widehat{\theta} - \theta^*\|_\infty \leq 5, \tag{2.36}$$

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

The proof of Proposition 2.8.1 borrows strength from the property of a regularized MLE. Recall the definition of $\widehat{\theta}_\lambda$ in (2.9). This is the version of MLE that has been analyzed by [25]. We will choose $\lambda = n^{-1}$ in order that $\widehat{\theta}_\lambda$ is close to $\widehat{\theta}$. Following the techniques in [25], we can first show $\|\widehat{\theta}_\lambda - \theta^*\|_\infty \leq 4$ with high probability. The presence of the penalty in (2.9) is crucial for the result $\|\widehat{\theta}_\lambda - \theta^*\|_\infty \leq 4$ 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 (2.36). A detailed proof of Proposition 2.8.1 will be given in Section 2.8.3.

The result of Proposition 2.8.1 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 (2.7) of Theorem 2.3.1 can be obtained by a Taylor expansion of the objective function $\ell_n(\theta)$. The second bound (2.8) of Theorem 2.3.1 and Theorem 2.3.2 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 [25]. The intuition of the leave-one-out technique has been thoroughly discussed in [25], 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 [25]. We do not need to combine the leave-one-out argument with a gradient descent analysis as in [25]. This helps us to avoid the extra technical condition $\log L = O(\log n)$ in [25] when proving the theorems.

2.8.2 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 2.8.1. *Assume $p \geq \frac{c_0 \log n}{n}$ for some sufficiently large $c_0 > 0$. We then have*

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

and

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

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

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

Lemma 2.8.2. *Assume $p \geq \frac{c_0 \log n}{n}$ for some sufficiently large $c_0 > 0$. For any fixed $\{w_{ij}\}$, we have*

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

and

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

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

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

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

Finally, we need a few concentration inequalities.

Lemma 2.8.4. *Assume $\kappa = O(1)$ and $p \geq \frac{c_0 \log n}{n}$ for some sufficiently large $c_0 > 0$. Then, we have*

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

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

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

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 Section 2.12.

2.8.3 Proof of Proposition 2.8.1

As we have outlined in Section 2.8.1, 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 (2.9). 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 2.8.5. *Under the setting of Theorem 2.3.1, for the estimator $\widehat{\theta}_\lambda$ with $\lambda = n^{-1}$, we have*

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

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

We first prove Proposition 2.8.1 with the help of Lemma 2.8.5. We then prove Lemma 2.8.5 at the end of this section.

Proof of Proposition 2.8.1. Define a constraint MLE as

$$\widehat{\theta}^{\mathsf{con}} = \underset{\mathbb{1}_n^T\theta=0:\|\theta-\theta^*\|_\infty\leq 5}{\operatorname{argmin}} \ell_n(\theta). \tag{2.37}$$

By Lemma 2.8.5, $\widehat{\theta}_\lambda$ is feasible for the constraint of (2.37). We then have

$$\ell_n(\widehat{\theta}_\lambda) \geq \ell_n(\widehat{\theta}^{\mathsf{con}}). \tag{2.38}$$

We apply Taylor expansion, and obtain

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

where ξ is a convex combination of $\widehat{\theta}^{\mathsf{con}}$ and $\widehat{\theta}_\lambda$. By Lemma 2.8.5, we know that $\|\widehat{\theta}_\lambda - \theta^*\|_\infty \leq 4$. We also have $\|\widehat{\theta}^{\mathsf{con}} - \theta^*\|_\infty \leq 5$ by the definition of $\widehat{\theta}^{\mathsf{con}}$. Thus, $\|\xi - \theta^*\|_\infty \leq 5$. By Lemma 2.8.3, we get the lower bound

$$\ell_n(\widehat{\theta}^{\mathsf{con}}) \geq \ell_n(\widehat{\theta}_\lambda) + (\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda)^T \nabla \ell_n(\widehat{\theta}_\lambda) + c_1 np \|\widehat{\theta}^{\mathsf{con}} - \widehat{\theta}_\lambda\|^2, \tag{2.39}$$

for some constant $c_1 > 0$. By (2.38) and (2.39), we have

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

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

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

Finally, since

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

the minimizer of (2.37) is in the interior of the constraint. By the convexity of (2.37), we have $\widehat{\theta}^{\mathsf{con}} = \widehat{\theta}$, and thus the desired conclusion $\|\widehat{\theta} - \theta^*\|_\infty \le 5$ is obtained. □

Proof of Lemma 2.8.5. Our proof largely follows the arguments in [25] 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 [25] is not needed anymore. This requires a few minor changes in the proof of [25]. We still write down every step of the proof for the result to be self-contained.

Define a gradient descent sequence

$$\theta^{(t+1)} = \theta^{(t)} - \eta \left(\nabla \ell_n(\theta^{(t)}) + \lambda \theta^{(t)} \right). \tag{2.40}$$

We also need to introduce a leave-one-out gradient descent sequence. Define

$$\begin{aligned} \ell_n^{(m)}(\theta) \;=\; & \sum_{1 \le i < j \le n : i,j \ne m} A_{ij} \left[\bar{y}_{ij} \log \frac{1}{\psi(\theta_i - \theta_j)} + (1 - \bar{y}_{ij}) \log \frac{1}{1 - \psi(\theta_i - \theta_j)} \right] \\ & + \sum_{i \in [n] \setminus \{m\}} p \left[\psi(\theta_i^* - \theta_m^*) \log \frac{1}{\psi(\theta_i - \theta_m)} + \psi(\theta_m^* - \theta_i^*) \log \frac{1}{\psi(\theta_m - \theta_i)} \right]. \end{aligned}$$

With the objective $\ell_n^{(m)}(\theta)$, we define

$$\theta^{(t+1,m)} = \theta^{(t,m)} - \eta \left(\nabla \ell_n^{(m)}(\theta^{(t,m)}) + \lambda \theta^{(t,m)} \right). \tag{2.41}$$

We initialize both (2.40) and (2.41) by $\theta^{(0)} = \theta^{(0,m)} = \theta^*$ and use the same step size $\eta = \frac{1}{\lambda + np}$. Note that $\mathbb{1}_n^T \theta^* = 0$ implies $\mathbb{1}_n^T \theta^{(t)} = \mathbb{1}_n^T \theta^{(t,m)} = 0$ for all t. See Section 4.3 of [26]. We will establish the following bounds,

$$\max_{m \in [n]} \|\theta^{(t,m)} - \theta^{(t)}\| \leq 1, \tag{2.42}$$

$$\|\theta^{(t)} - \theta^*\| \leq \sqrt{\frac{n}{\log n}}, \tag{2.43}$$

$$\max_{m \in [n]} |\theta_m^{(t,m)} - \theta_m^*| \leq 1. \tag{2.44}$$

It is obvious that (2.42), (2.43) and (2.44) hold for $t = 0$. We use a mathematical induction argument to show (2.42), (2.43) and (2.44) for a general t. Let us suppose (2.42), (2.43) and (2.44) are true, and we need to show the same conclusions continue to hold for $t + 1$.

First, we have

$$\begin{aligned} \theta^{(t+1)} - \theta^{(t+1,m)} &= (1 - \eta\lambda)(\theta^{(t)} - \theta^{(t,m)}) - \eta(\nabla \ell_n(\theta^{(t)}) - \nabla \ell_n^{(m)}(\theta^{(t,m)})) \\ &= \left((1 - \eta\lambda) I_n - \eta H(\xi)\right)(\theta^{(t)} - \theta^{(t,m)}) - \eta \left(\nabla \ell_n(\theta^{(t,m)}) - \nabla \ell_n^{(m)}(\theta^{(t,m)}) \right), \end{aligned}$$

where ξ is a convex combination of $\theta^{(t)}$ and $\theta^{(t,m)}$. By (2.42) and (2.44), we have

$$\|\theta^{(t)} - \theta^*\|_\infty \leq \max_{m \in [n]} \|\theta^{(t,m)} - \theta^{(t)}\| + \max_{m \in [n]} |\theta_m^{(t,m)} - \theta_m^*| \leq 2, \tag{2.45}$$

and

$$\|\theta^{(t,m)} - \theta^*\|_\infty \leq \|\theta^{(t)} - \theta^*\|_\infty + \|\theta^{(t,m)} - \theta^{(t)}\| \leq 3. \tag{2.46}$$

We thus have $\|\xi - \theta^*\|_\infty \leq 3$, and we can apply Lemma 2.8.3 to obtain the bound

$$\| \left((1-\eta\lambda)I_n - \eta H(\xi)\right)(\theta^{(t)} - \theta^{(t,m)})\| \leq (1 - \eta\lambda - c_1\eta np)\|\theta^{(t)} - \theta^{(t,m)}\|, \tag{2.47}$$

for some constant $c_1 > 0$. We also note that

$$\begin{aligned}
& \|\nabla \ell_n(\theta^{(t,m)}) - \nabla \ell_n^{(m)}(\theta^{(t,m)})\|^2 \\
= \; & \Big(\sum_{j \in [n]\setminus\{m\}} A_{jm}(\bar{y}_{jm} - \psi(\theta_j^* - \theta_m^*)) \\
& - \sum_{j \in [n]\setminus\{m\}} (A_{jm} - p)(\psi(\theta_j^{(t,m)} - \theta_m^{(t,m)}) - \psi(\theta_j^* - \theta_m^*)) \Big)^2 \\
& + \sum_{j \in [n]\setminus\{m\}} \Big(A_{jm}(\bar{y}_{jm} - \psi(\theta_j^* - \theta_m^*)) - (A_{jm} - p)(\psi(\theta_j^{(t,m)} - \theta_m^{(t,m)}) - \psi(\theta_j^* - \theta_m^*)) \Big)^2 \\
\leq \; & C_1 \frac{np \log n}{L} + C_1 np \log n \|\theta^{(t,m)} - \theta^*\|_\infty^2,
\end{aligned} \tag{2.48}$$

for some constant $C_1 > 0$ by Lemma 2.8.2 and Lemma 2.8.4. We combine the two bounds (2.47) and (2.48), and obtain

$$\begin{aligned}
\|\theta^{(t+1)} - \theta^{(t+1,m)}\| &\leq (1 - \eta\lambda - c_1\eta np)\|\theta^{(t)} - \theta^{(t,m)}\| \\
&\quad + \eta\sqrt{C_1 np \log n \left(L^{-1} + \|\theta^{(t,m)} - \theta^*\|_\infty^2\right)} \\
&\leq (1 - c_1\eta np) + \eta\sqrt{C_1 np \log n \left(L^{-1} + 9\right)} \qquad (2.49) \\
&\leq 1 \qquad (2.50)
\end{aligned}$$

where the inequality (2.49) is by (2.42) and (2.46). The inequality (2.50) requires that $\sqrt{C_1 np \log n \left(L^{-1} + 9\right)} \leq c_1 np$, which is implied by the condition that $p \geq \frac{c_0 \log n}{n}$ for some sufficiently large $c_0 > 0$. We thus have proved (2.42) for $t + 1$.

Next, we have

$$\begin{aligned}\theta^{(t+1)} - \theta^* &= \theta^{(t)} - \theta^* - \eta\left(\nabla \ell_n(\theta^{(t)}) + \lambda\theta^{(t)}\right) \\ &= (1-\eta\lambda)(\theta^{(t)} - \theta^*) - \eta\left(\nabla \ell_n(\theta^{(t)}) - \nabla \ell_n(\theta^*)\right) - \eta\lambda\theta^* - \eta\nabla\ell_n(\theta^*) \\ &= ((1-\eta\lambda)I_n - \eta H(\xi))\,(\theta^{(t)} - \theta^*) - \eta\lambda\theta^* - \eta\nabla\ell_n(\theta^*),\end{aligned}$$

where ξ is abused for a vector that is a convex combination of $\theta^{(t)}$ and θ^*. Since by (2.45) we get $\|\xi - \theta^*\|_\infty \le \|\theta^{(t)} - \theta^*\|_\infty \le 2$, we can use Lemma 2.8.3 to obtain the bound

$$((1-\eta\lambda)I_n - \eta H(\xi))\,(\theta^{(t)} - \theta^*) \le (1 - \eta\lambda - c_2\eta np)\|\theta^{(t)} - \theta^*\|, \tag{2.51}$$

for some constant $c_2 > 0$. We also note that

$$\|\nabla\ell_n(\theta^*)\|^2 = \sum_{i=1}^n \left(\sum_{j\in[n]\setminus\{i\}} A_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))\right)^2 \le C_2\frac{n^2p}{L}, \tag{2.52}$$

for some constant $C_2 > 0$ with high probability by Lemma 2.8.4. Combine the bounds (2.51) and (2.52), and we obtain

$$\begin{aligned}\|\theta^{(t+1)} - \theta^*\| &\le (1 - \eta\lambda - c_2\eta np)\|\theta^{(t)} - \theta^*\| + \eta\sqrt{C_2\frac{n^2p}{L}} + \eta\lambda\|\theta^*\| \\ &\le (1 - c_2\eta np)\sqrt{\frac{n}{\log n}} + \eta\sqrt{C_2\frac{n^2p}{L}} + \eta\lambda\|\theta^*\| \\ &\le \sqrt{\frac{n}{\log n}},\end{aligned}$$

where the last inequality is due to $\eta\sqrt{C_2\frac{n^2p}{L}} + \eta\lambda\|\theta^*\| \lesssim \frac{1}{\sqrt{Lp}} + \frac{1}{n^{3/2}p} = o\left(\eta np\sqrt{\frac{n}{\log n}}\right)$ by the choice of η and λ. Hence, (2.43) holds for $t+1$.

Finally, we have

$$\begin{aligned}
\theta_m^{(t+1,m)} - \theta_m^* &= \theta_m^{(t,m)} - \theta_m^* + \eta p \sum_{j\in[n]\setminus\{m\}} \left(\psi(\theta_m^* - \theta_j^*) - \psi(\theta_m^{(t,m)} - \theta_j^{(t,m)})\right) - \lambda\eta\theta_m^{(t,m)} \\
&= \theta_m^{(t,m)} - \theta_m^* + \eta p \sum_{j\in[n]\setminus\{m\}} \psi'(\xi_j)(\theta_m^* - \theta_j^* - \theta_m^{(t,m)} + \theta_j^{(t,m)}) - \lambda\eta\theta_m^{(t,m)} \\
&= \left(1 - \eta\lambda - \eta p \sum_{j\in[n]\setminus\{m\}} \psi'(\xi_j)\right)(\theta_m^{(t,m)} - \theta_m^*) - \lambda\eta\theta_m^* \\
&\quad + \eta p \sum_{j\in[n]\setminus\{m\}} \psi'(\xi_j)(\theta_j^{(t,m)} - \theta_j^*),
\end{aligned}$$

where ξ_j is a scalar between $\theta_m^* - \theta_j^*$ and $\theta_m^{(t,m)} - \theta_j^{(t,m)}$. By (2.46), we have $|\xi_j - \theta_m^* + \theta_j^*| \le |\theta_m^* - \theta_j^* - \theta_m^{(t,m)} + \theta_j^{(t,m)}| \le 6$, which implies $\|\xi\|_\infty$ is bounded. We then have $\sum_{j\in[n]\setminus\{m\}} \psi'(\xi_j) \ge c_3 n$ for some constant $c_3 > 0$, and thus

$$\left|\left(1 - \eta\lambda - \eta p \sum_{j\in[n]\setminus\{m\}} \psi'(\xi_j)\right)(\theta_m^{(t,m)} - \theta_m^*)\right| \le (1 - \eta\lambda - c_3\eta n p)|\theta_m^{(t,m)} - \theta_m^*|. \quad (2.53)$$

We also have

$$\left|\sum_{j\in[n]\setminus\{m\}} \psi'(\xi_j)(\theta_j^{(t,m)} - \theta_j^*)\right| \le \|\theta^{(t,m)} - \theta^*\|_1 \le \sqrt{n}\|\theta^{(t,m)} - \theta^*\| \le \sqrt{n}\left(1 + \sqrt{\frac{n}{\log n}}\right), \quad (2.54)$$

where the last inequality is by (2.42) and (2.43). Combine the bounds (2.53) and (2.54), and we get

$$\begin{aligned}
|\theta_m^{(t+1,m)} - \theta_m^*| &\le (1 - \eta\lambda - c_3\eta n p)|\theta_m^{(t,m)} - \theta_m^*| + \eta p\sqrt{n}\left(1 + \sqrt{\frac{n}{\log n}}\right) + \lambda\eta|\theta_m^*| \\
&\le (1 - c_3\eta n p) + \eta p\sqrt{n} + \eta p\frac{n}{\sqrt{\log n}} + \lambda\eta|\theta_m^*| \\
&\le 1,
\end{aligned}$$

where the last inequality is because of $\eta p\sqrt{n} + \eta p\frac{n}{\sqrt{\log n}} + \lambda\eta|\theta^*_m| = o(\eta np)$ by the choice of η and λ. Hence, (2.44) holds for $t+1$.

To summarize, we have shown that (2.42), (2.43) and (2.44) hold for all $t \leq t^*$ with probability at least $1 - O(t^* n^{-10})$. The reason why we have the probability $1 - O(t^* n^{-10})$ is because we need to apply Lemma 2.8.2 with a different weight at each iteration to show (2.48). Note that the bound (2.45) holds for all $t \leq t^*$ as well and we thus have $\|\theta^{(t^*)} - \theta^*\|_\infty \leq 2$. With a standard optimization result for a strongly convex objective function, we have

$$\|\theta^{(t^*)} - \widehat{\theta}_\lambda\| \leq \left(1 - \frac{\lambda}{\lambda + np}\right)^{t^*} \|\widehat{\theta}_\lambda - \theta^*\|.$$

See Lemma 6.7 of [25]. By triangle inequality, we have

$$\|\widehat{\theta}_\lambda - \theta^*\|_\infty \leq \|\theta^{(t^*)} - \widehat{\theta}_\lambda\| + \|\theta^{(t^*)} - \theta^*\|_\infty \leq \left(1 - \frac{\lambda}{\lambda + np}\right)^{t^*} \sqrt{n}\|\widehat{\theta}_\lambda - \theta^*\|_\infty + 2.$$

Since $\left(1 - \frac{\lambda}{\lambda+np}\right) \leq 1 - \frac{1}{1+n^2}$, we can take $t^* = n^3$ in order that $\left(1 - \frac{\lambda}{\lambda+np}\right)^{t^*} \sqrt{n} \leq \frac{1}{2}$. This implies $\|\widehat{\theta}_\lambda - \theta^*\|_\infty \leq 4$ with probability at least $1 - O(n^{-7})$ as desired. □

2.8.4 Proof of Theorem 2.3.1

We give separate proofs for the conclusions (2.7) and (2.8) in this section.

Proof of (2.7) of Theorem 2.3.1. By the definition of $\widehat{\theta}$, we have $\ell_n(\theta^*) \geq \ell_n(\widehat{\theta})$. We then apply Taylor expansion and obtain

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

where ξ is a convex combination of $\widehat{\theta}$ and θ^*. By Proposition 2.8.1, we have $\|\widehat{\theta} - \theta^*\|_\infty \leq 5$, which implies $\|\xi - \theta^*\|_\infty \leq 5$. Thus, we can apply Lemma 2.8.3 and get $\frac{1}{2}(\widehat{\theta} - \theta^*)^T H(\xi)(\widehat{\theta} - \theta^*) \geq c_1 np\|\widehat{\theta} - \theta^*\|^2$ for some constant $c_1 > 0$. Together with $\ell_n(\theta^*) \geq \ell_n(\widehat{\theta})$ and a Cauchy-

Schwarz inequality, we have $\|\widehat{\theta} - \theta^*\|^2 \leq \frac{\|\nabla \ell_n(\theta^*)\|^2}{(c_1 np)^2}$. Use (2.52) and Lemma 2.8.4, we obtain the desired conclusion that $\|\widehat{\theta} - \theta^*\|^2 \lesssim \frac{1}{Lp}$. □

The proof of (2.8) is more involved. It is based on a leave-one-out argument that is very different from the one used in [25]. Let us decompose the objective function $\ell_n(\theta)$ as

$$\ell_n(\theta) = \ell_n^{(-m)}(\theta_{-m}) + \ell_n^{(m)}(\theta_m | \theta_{-m}), \tag{2.55}$$

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 (2.55) are defined as

$$\begin{aligned}
\ell_n^{(-m)}(\theta_{-m}) &= \sum_{1 \leq i < j \leq n : i, j \neq m} A_{ij} \left[\bar{y}_{ij} \log \frac{1}{\psi(\theta_i - \theta_j)} + (1 - \bar{y}_{ij}) \log \frac{1}{1 - \psi(\theta_i - \theta_j)} \right], \\
\ell_n^{(m)}(\theta_m | \theta_{-m}) &= \sum_{j \in [n] \setminus \{m\}} A_{mj} \left[\bar{y}_{mj} \log \frac{1}{\psi(\theta_m - \theta_j)} + (1 - \bar{y}_{mj}) \log \frac{1}{1 - \psi(\theta_m - \theta_j)} \right].
\end{aligned}$$

Define

$$\theta_{-m}^{(m)} = \operatorname*{argmin}_{\theta_{-m} : \|\theta_{-m} - \theta_{-m}^*\|_\infty \leq 5} \ell_n^{(-m)}(\theta_{-m}). \tag{2.56}$$

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 2.8.6. *Under the setting of Theorem 2.3.1, there exists some constant $C > 0$ such that*

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

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 (2.7), 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

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

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 $\|\xi - \theta_{-m}^*\|_\infty \leq \|\theta_{-m}^{(m)} - \theta_{-m}^*\|_\infty \leq 5$, we can apply Lemma 2.8.3 to the subset of the data, and obtain

$$\frac{1}{2}(\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1})^T H^{(-m)}(\xi)(\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1}) \geq c_1 np \|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1}\|^2,$$

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

$$\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1}\|^2 \leq \frac{\|\nabla \ell_n^{(-m)}(\theta_{-m}^*)\|^2}{(c_1 np)^2}.$$

Apply (2.52) and Lemma 2.8.4 to the subset of the data, and we obtain that $\|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbb{1}_{n-1}\|^2 \leq C\frac{1}{pL}$ with probability at least $1 - O(n^{-10})$. Finally, a union bound argument leads to the desired result. □

With the help of Lemma 2.8.6, we are ready to prove (2.8).

Proof of (2.8) of Theorem 2.3.1. By Proposition 2.8.1, we have $\|\widehat{\theta}_{-m} - \theta_{-m}^*\|_\infty \leq \|\widehat{\theta} - \theta^*\|_\infty \leq 5$, and thus $\widehat{\theta}_{-m}$ is feasible for the constraint of (2.56). By the definition of $\theta_{-m}^{(m)}$,

we have

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

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 (2.56), we must have $\|\xi - \theta_{-m}^*\|_\infty \leq 5$. Then, we can apply Lemma 2.8.3 to the subset of the data, and obtain

$$\frac{1}{2}(\theta_{-m}^{(m)} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1})^T H^{(-m)}(\xi)(\theta_{-m}^{(m)} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}) \geq c_1 np \|\theta_{-m}^{(m)} - \widehat{\theta}_{-m} - \bar{a}_m \mathbb{1}_{n-1}\|^2,$$

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

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

For each $i \in [n] \backslash \{m\}$, by the decomposition (2.55), 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

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

We therefore have the bound

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

where the last inequality is by Lemma 2.8.1. This implies

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

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

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

We then have

$$\begin{aligned}
\max_{m\in[n]} \|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - a_m \mathbb{1}_{n-1}\|^2 &\leq C_1 \frac{\max_{m\in[n]} \sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta^*_m - \theta^*_i))^2}{n^2 p^2} \\
&\quad + C_1 \frac{\|\widehat{\theta} - \theta^*\|_\infty^2}{np}, \qquad (2.57)
\end{aligned}$$

for some constant $C_1 > 0$.

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

$$\begin{aligned} f^{(m)}(\theta_m|\theta_{-m}) &= \frac{\partial}{\partial\theta_m}\ell_n^{(m)}(\theta_m|\theta_{-m}) = -\sum_{i\in[n]\backslash\{m\}} A_{mi}(\bar{y}_{mi}-\psi(\theta_m-\theta_i)), \\ g^{(m)}(\theta_m|\theta_{-m}) &= \frac{\partial^2}{\partial\theta_m^2}\ell_n^{(m)}(\theta_m|\theta_{-m}) = \sum_{i\in[n]\backslash\{m\}} A_{mi}\psi(\theta_m-\theta_i)\psi(\theta_i-\theta_m). \end{aligned}$$

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

$$\ell_n^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) + \ell_n^{(-m)}(\widehat{\theta}_{-m}) \geq \ell_n(\widehat{\theta}),$$

which implies

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

where ξ is a scalar between θ_m^* and $\widehat{\theta}_m$. By Proposition 2.8.1, $|\xi-\theta_m^*| \leq |\widehat{\theta}_m-\theta_m^*| \leq \|\widehat{\theta}-\theta^*\|_\infty \leq 5$. Therefore, for any $i\in[n]\backslash\{m\}$, $|\xi-\widehat{\theta}_i| \leq |\xi-\theta_m^*|+|\theta_m^*-\theta_i^*|+|\widehat{\theta}_i-\theta_i^*| \leq 10+\kappa$. This implies $\frac{1}{2}g^{(m)}(\xi|\widehat{\theta}_{-m}) \geq c_2 np$ for some constant $c_2 > 0$ with the help of Lemma 2.8.1. We then have the bound

$$(\widehat{\theta}_m-\theta_m^*)^2 \leq \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|^2}{(c_2np)^2}. \tag{2.58}$$

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

$$\begin{aligned}
|f^{(m)}(\theta^*_m|\widehat{\theta}_{-m})| &= \left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta^*_m - \widehat{\theta}_i))\right| \\
&\leq \left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta^*_m - \theta^*_i))\right| \qquad (2.59)\\
&\quad + \left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta^*_m - \theta^*_i) - \psi(\theta^*_m - \theta^{(m)}_i + a_m))\right| \qquad (2.60)\\
&\quad + \left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta^*_m - \theta^{(m)}_i + a_m) - \psi(\theta^*_m - \widehat{\theta}_i))\right|. \qquad (2.61)
\end{aligned}$$

We use Lemma 2.8.2 to bound (2.60). We have

$$\begin{aligned}
&\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta^*_m - \theta^*_i) - \psi(\theta^*_m - \theta^{(m)}_i + a_m))\right| \\
\leq\ & p\left|\sum_{i\in[n]\setminus\{m\}} (\psi(\theta^*_m - \theta^*_i) - \psi(\theta^*_m - \theta^{(m)}_i + a_m))\right| \qquad (2.62)\\
&+ \left|\sum_{i\in[n]\setminus\{m\}} (A_{mi} - p)(\psi(\theta^*_m - \theta^*_i) - \psi(\theta^*_m - \theta^{(m)}_i + a_m))\right| \qquad (2.63)\\
\leq\ & p\sqrt{n}\|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\| + C_2\log n\|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\|_\infty \\
&+C_2\sqrt{p\log n}\|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\| \\
\leq\ & (p\sqrt{n} + C_2\sqrt{p\log n})\|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\| + C_2\log n\|\widehat{\theta} - \theta^*\|_\infty \\
&+C_2\log n\|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - a_m\mathbb{1}_{n-1}\|.
\end{aligned}$$

With the help of 2.8.1, we can also bound (2.61), and we get

$$\begin{aligned}
& \left| \sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta^*_m - \theta_i^{(m)} + a_m) - \psi(\theta^*_m - \widehat{\theta}_i)) \right| \\
\leq\ & \sqrt{\sum_{i\in[n]\setminus\{m\}} A_{mi}} \|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - a_m \mathbb{1}_{n-1}\| \\
\leq\ & C_3\sqrt{np}\|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - a_m \mathbb{1}_{n-1}\|. \qquad (2.64)
\end{aligned}$$

Plug the bounds into (2.58), and we have

$$\begin{aligned}
\|\widehat{\theta} - \theta^*\|_\infty \leq\ & \frac{\max_{m\in[n]} \left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta^*_m - \theta^*_i))\right|}{c_2 np} \\
& + \frac{(p\sqrt{n} + C_2\sqrt{p\log n}) \max_{m\in[n]} \|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\|}{c_2 np} \\
& + \frac{(C_2\log n + C_3\sqrt{np})\|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - a_m\mathbb{1}_{n-1}\|}{c_2 np} + \frac{C_2\log n\|\widehat{\theta} - \theta^*\|_\infty}{c_2 np}.
\end{aligned}$$

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

$$\begin{aligned}
\|\widehat{\theta} - \theta^*\|_\infty \leq\ & C_4\frac{\max_{m\in[n]} \left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta^*_m - \theta^*_i))\right|}{np} \\
& + C_4\frac{p\sqrt{n}\max_{m\in[n]} \|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\|}{np} \\
& + C_4\frac{(\log n + \sqrt{np})\|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - a_m\mathbb{1}_{n-1}\|}{np}. \qquad (2.65)
\end{aligned}$$

Let us plug the above bound into (2.57). Then, after some rearrangement, we obtain

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

By Lemma 2.8.4 and Lemma 2.8.6, we have

$$\max_{m\in[n]} \|\theta^{(m)}_{-m} - \widehat{\theta}_{-m} - a_m \mathbb{1}_{n-1}\| \leq C_7 \sqrt{\frac{1}{npL}}. \tag{2.66}$$

Now we can plug the bound (2.66) back into (2.65), and together with Lemma 2.8.4 and Lemma 2.8.6, we have

$$\|\widehat{\theta} - \theta^*\|_\infty \leq C_8 \sqrt{\frac{\log n}{npL}}, \tag{2.67}$$

which is the desired conclusion. Tracking all the probabilistic events that we have used in the proof, we can conclude that both (2.66) and (2.67) hold with probability at least $1 - O(n^{-7})$. □

2.8.5 Proofs of Theorem 2.3.2 and Theorem 2.3.3

In the proof of (2.8), we have established the byproduct (2.66). This bound turns out to be extremely important for us to establish the result of Theorem 2.3.2. We therefore list it, together with its consequence, as a lemma.

Lemma 2.8.7. *Under the setting of Theorem 2.3.1, there exists some constant $C > 0$ such that*

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

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

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 (2.56).*

Proof. The first conclusion has been established in (2.66). 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 (2.67). □

Now we are ready to prove Theorem 2.3.2.

Proof of Theorem 2.3.2. 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

$$\eta = \frac{1}{2} - \frac{V(\kappa)}{(1-\bar{\delta})\Delta^2 npL}\log\frac{n-k}{k}, \tag{2.68}$$

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

$$\bar{\Delta}_i = \begin{cases} \min\left(\eta(\theta^*_k - \theta^*_{k+1}) + \theta^*_i - \theta^*_k, \left(\frac{\log n}{np}\right)^{1/4}\right), & 1 \le i \le k, \\ \min\left((1-\eta)(\theta^*_k - \theta^*_{k+1}) + \theta^*_{k+1} - \theta^*_i, \left(\frac{\log n}{np}\right)^{1/4}\right), & k+1 \le i \le n. \end{cases} \tag{2.69}$$

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 (2.8) of Theorem 2.3.1, we have shown that

$|f^{(m)}(\theta^*_m|\widehat{\theta}_{-m})|$ can be bounded by the sum of (2.59), (2.61), (2.62) and (2.63). For (2.59), we can use Hoeffding's inequality and Lemma 2.8.1 and obtain the bound

$$\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta^*_m - \theta^*_i))\right| \leq C_1\sqrt{\frac{x\sum_{i\in[n]\setminus\{m\}} A_{mi}}{L}} \leq C_2\sqrt{\frac{xnp}{L}}, \tag{2.70}$$

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

$$\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta^*_m - \theta^*_i))\right| \leq C_2\sqrt{\bar{\Delta}_m^{3/2}(np)^2}, \tag{2.71}$$

with probability at least $1 - O(n^{-10}) - e^{-\bar{\Delta}_m^{3/2}Lnp}$. Since we have already shown (2.61) can be bounded by (2.64) with probability at least $1 - O(n^{-10})$, an application of Lemma 2.8.7 implies that

$$\left|\sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta^*_m - \theta^{(m)}_i + a_m) - \psi(\theta^*_m - \widehat{\theta}_i))\right| \leq C_3\sqrt{\frac{1}{L}}, \tag{2.72}$$

with probability at least $1 - O(n^{-7})$. By Cauchy-Schwarz inequality, we can bound (2.62) by $p\sqrt{n}\|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\|$. With the help of Lemma 2.8.6, we have

$$p\left|\sum_{i\in[n]\setminus\{m\}} (\psi(\theta^*_m - \theta^*_i) - \psi(\theta^*_m - \theta^{(m)}_i + a_m))\right| \leq C_4\sqrt{\frac{np}{L}}, \tag{2.73}$$

with probability at least $1 - O(n^{-9})$. For (2.63), we use Bernstein's inequality, and we have

$$\begin{aligned}&\left|\sum_{i\in[n]\setminus\{m\}} (A_{mi} - p)(\psi(\theta^*_m - \theta^*_i) - \psi(\theta^*_m - \theta^{(m)}_i + a_m))\right| \\ &\leq C_5\sqrt{px}\|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\| + C_5x\|\theta^{(m)}_{-m} - \theta^*_{-m} - a_m\mathbb{1}_{n-1}\|_\infty,\end{aligned} \tag{2.74}$$

with probability at least $1-e^{-x}$. We choose $x = \min\left(\bar{\Delta}_m^2 Lnp\frac{np}{\log n}, 7\log n\right)$. Then, with the help of Lemma 2.8.6 and Lemma 2.8.7, we have

$$\begin{aligned}
&\left|\sum_{i\in[n]\setminus\{m\}} (A_{mi}-p)(\psi(\theta_m^*-\theta_i^*)-\psi(\theta_m^*-\theta_i^{(m)}+a_m))\right| \\
\leq\ & C_6\frac{1}{\sqrt{L}}\sqrt{\min\left(\bar{\Delta}_m^2 Lnp\frac{np}{\log n}, 7\log n\right)} + C_6\sqrt{\frac{\log n}{npL}}\min\left(\bar{\Delta}_m^2 Lnp\frac{np}{\log n}, 7\log n\right) \qquad (2.75)
\end{aligned}$$

with probability at least $1-O(n^{-7})-\exp\left(-\bar{\Delta}_m^2 npL\frac{np}{\log n}\right)$. Combining the bounds (2.71)-(2.75), 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 (2.58).

In the second step, we define

$$\bar{\theta}_m = \theta_m^* - \frac{f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})}{g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})}. \qquad (2.76)$$

We need to show $\bar{\theta}_m$ and $\widehat{\theta}_m$ are close. By Proposition 2.8.1, $\|\widehat{\theta}-\theta^*\|_\infty \leq 5$, and thus $g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) \geq c_1 np$ for some constant $c_1>0$, so that we have the bound $|\bar{\theta}_m-\theta_m^*| \leq \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}$. In fact, given the inequality (2.58), we can choose c_1 to be sufficiently small so that $|\widehat{\theta}_m-\theta_m^*| \leq \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}$ is also true. Therefore, we can express $\bar{\theta}_m$ and $\widehat{\theta}_m$ as

$$\begin{aligned}
\bar{\theta}_m &= \operatorname*{argmin}_{|\theta_m-\theta_m^*|\leq\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}} \bar{\ell}_n^{(m)}(\theta_m|\widehat{\theta}_{-m}), \\
\widehat{\theta}_m &= \operatorname*{argmin}_{|\theta_m-\theta_m^*|\leq\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}} \ell_n^{(m)}(\theta_m|\widehat{\theta}_{-m}),
\end{aligned}$$

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 (2.55) 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

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

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

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

where the last inequality uses Lemma 2.8.1. Therefore, for any θ_m that satisfies $|\theta_m - \theta_m^*| \leq \frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}$, the difference between the two objective functions can be bounded by

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

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,

$$\begin{aligned}
& \frac{1}{2}g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})(\widehat{\theta}_m - \bar{\theta}_m)^2 \\
= \;& \bar{\ell}_n^{(m)}(\widehat{\theta}_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) \\
\leq \;& \ell_n^{(m)}(\widehat{\theta}_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) + \frac{C_7 np}{2}\left(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}\right)^3 \\
\leq \;& \ell_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) - \bar{\ell}_n^{(m)}(\bar{\theta}_m|\widehat{\theta}_{-m}) + \frac{C_7 np}{2}\left(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}\right)^3 \\
\leq \;& 2\frac{C_7 np}{2}\left(\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|}{c_1 np}\right)^3.
\end{aligned}$$

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

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

Since $|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m})|$ has been shown to be bounded by the sum of (2.71)-(2.75), we have

$$|\widehat{\theta}_m - \bar{\theta}_m| \leq \delta\bar{\Delta}_m, \tag{2.77}$$

for some $\delta = o(1)$ with probability at least $1 - O(n^{-7}) - \exp(-\bar{\Delta}_m^{3/2}Lnp) - \exp\left(-\bar{\Delta}_m^2 npL\frac{np}{\log n}\right)$ 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 (2.72), (2.73) and (2.75). Given the fact that $g^{(m)}(\theta_m^*|\theta_{-m}^*) \gtrsim np$, we have

$$\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - f^{(m)}(\theta_m^*|\theta_{-m}^*)|}{g^{(m)}(\theta_m^*|\theta_{-m}^*)} \leq \delta\bar{\Delta}_m, \tag{2.78}$$

for some $\delta = o(1)$ with probability at least $1 - O(n^{-7}) - \exp\left(-\bar{\Delta}_m^2 npL\frac{np}{\log n}\right)$ 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 (2.77) and (2.78). We also need to give a sharp approximation to $g^{(m)}(\theta_m^*|\widehat{\theta}_{-m})$. We have

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

By Lemma 2.8.1, Lemma 2.8.6 and Lemma 2.8.7, 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 2.8.2, and then $\sum_{i\in[n]\setminus\{m\}}(A_{mi} - p)\left|\theta_i^{(m)} - a_m - \theta_i^*\right|$ 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 2.8.6 and Lemma 2.8.7, 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

$$\left|g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\theta_{-m}^*)\right| \lesssim \sqrt{\frac{np}{L}} + \frac{(\log n)^{3/2}}{\sqrt{npL}}. \tag{2.79}$$

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

$$\frac{\left|g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\theta_{-m}^*)\right|}{g^{(m)}(\theta_m^*|\theta_{-m}^*)} \leq \delta, \tag{2.80}$$

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 (2.77), (2.78) and (2.80).

In the last step, we will apply Lemma 2.3.1 with $t = (1-\eta)\theta_k^* + \eta\theta_{k+1}^*$ to finish the proof. Recall the definition of η in (2.68). For any $i \le k$, we have

$$
\begin{aligned}
& \mathbb{P}\left(\widehat{\theta}_i \le (1-\eta)\theta_k^* + \eta\theta_{k+1}^*)\right) \\
\le\ & \mathbb{P}\left(\widehat{\theta}_i - \theta_i^* \le -\eta(\theta_k^* - \theta_{k+1}^*) - (\theta_i^* - \theta_k^*)\right) \\
\le\ & \mathbb{P}\left(\bar{\theta}_i - \theta_i^* \le -(1-\delta)\bar{\Delta}_i\right) + \mathbb{P}\left(|\bar{\theta}_i - \widehat{\theta}_i| > \delta\bar{\Delta}_i\right) \\
\le\ & \mathbb{P}\left(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -(1+\delta^2-3\delta)\bar{\Delta}_i\right) + \mathbb{P}\left(|\bar{\theta}_i - \widehat{\theta}_i| > \delta\bar{\Delta}_i\right) \\
& + \mathbb{P}\left(\frac{\left|g^{(i)}(\theta_i^*|\widehat{\theta}_{-i}) - g^{(i)}(\theta_i^*|\theta_{-i}^*)\right|}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} > \delta\right) + \mathbb{P}\left(\frac{|f^{(i)}(\theta_i^*|\widehat{\theta}_{-i}) - f^{(i)}(\theta_i^*|\theta_{-i}^*)|}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} > \delta\bar{\Delta}_i\right) \\
\le\ & \mathbb{P}\left(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -(1-3\delta)\bar{\Delta}_i\right) + O(n^{-7}) \qquad (2.81) \\
& + \exp(-\bar{\Delta}_i^{3/2}Lnp) + \exp\left(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\right),
\end{aligned}
$$

where the last inequality is due to (2.77), (2.78) and (2.80). Define the event

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

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 (2.77), (2.78) and (2.80) are still true. We then have

$$
\begin{aligned}
& \mathbb{P}\left(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -(1-3\delta)\bar{\Delta}_i\right) \\
\le & \sup_{A\in\mathcal{A}_i} \mathbb{P}\left(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))}{\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi(\theta_i^*-\theta_j^*)\psi(\theta_j^*-\theta_i^*)} \le -(1-3\delta)\bar{\Delta}_i \Big| A\right) + \mathbb{P}(A\in\mathcal{A}_i^c) \\
\le & \sup_{A\in\mathcal{A}_i} \exp\left(-\frac{\frac{1}{2}(1-3\delta)^2\bar{\Delta}_i^2\left(L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^*-\theta_j^*)\right)^2}{L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^*-\theta_j^*) + \frac{1-3\delta}{3}\bar{\Delta}_i L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^*-\theta_j^*)}\right) \qquad (2.82) \\
& + O(n^{-7}) \\
= & \exp\left(-\frac{1+o(1)}{2}\bar{\Delta}_i^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^*-\theta_j^*)\right) + O(n^{-7}) \qquad (2.83) \\
\le & \exp\left(-\frac{1+o(1)}{2}(\eta(\theta_k^*-\theta_{k+1}^*) + (\theta_i^*-\theta_k^*))^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^*-\theta_j^*)\right) \qquad (2.84) \\
& + O(n^{-7}) \\
\le & \exp\left(-\frac{1+o(1)}{2}(\bar{\Delta} + (\theta_i^*-\theta_k^*))^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^*-\theta_j^*)\right) + O(n^{-7}). \qquad (2.85)
\end{aligned}
$$

The bound (2.82) is by Bernstein's inequality. We then use the definition of $\mathcal{A}_i$ to obtain the expression (2.83). To see why (2.84) is true, note that when $\bar{\Delta}_i^2 = \sqrt{\frac{\log n}{np}}$, the first term of (2.83) can be absorbed into $O(n^{-7})$. Finally, in (2.85), we have used the notation $\bar{\Delta} = \min\left(\eta(\theta_k^* - \theta_{k+1}^*), \left(\frac{\log n}{np}\right)^{1/4}\right)$. For each $j \in [n]$, define

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

The derivative of this function is

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

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 $\min_{t\in(c_2,\kappa]} h_j(t) \geq c_3$. With this fact, we can bound the exponent of (2.85) as

$$\begin{aligned}
& (\bar{\Delta} + (\theta_i^* - \theta_k^*))^2 Lp \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*) \\
\geq\ & Lp \sum_{j\in[n]\setminus\{i\}} \min\left(\bar{\Delta}^2\psi'(\theta_k^* - \theta_j^*), c_3\right) \\
\geq\ & Lp(k-1)\min\left(\bar{\Delta}^2\psi'(\theta_1^* - \theta_k^*), c_3\right) + Lp(n-k)\min\left(\bar{\Delta}^2\psi'(\theta_k^* - \theta_n^*), c_3\right) \\
=\ & Lp\bar{\Delta}^2\left((k-1)\psi'(\theta_1^* - \theta_k^*) + (n-k)\psi'(\theta_k^* - \theta_n^*)\right) \qquad (2.86) \\
\geq\ & (1+o(1))Lp\min\left(\eta^2\Delta^2, \sqrt{\frac{\log n}{np}}\right)\frac{n}{V(\kappa)}
\end{aligned}$$

where the equality (2.86) uses the fact that $\bar{\Delta} \to 0$. Therefore, we can further bound (2.85) as

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

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

$$\mathbb{P}\left(\widehat{\theta}_i \leq (1-\eta)\theta_k^* + \eta\theta_{k+1}^*)\right) \leq \exp\left(-\frac{(1-\delta')\eta^2\Delta^2 npL}{2V(\kappa)}\right) + O(n^{-7}), \qquad (2.87)$$

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

$$\mathbb{P}\left(\widehat{\theta}_i \geq (1-\eta)\theta_k^* + \eta\theta_{k+1}^*)\right) \leq \exp\left(-\frac{(1-\delta')(1-\eta)^2\Delta^2 npL}{2V(\kappa)}\right) + O(n^{-7}), \qquad (2.88)$$

for all all $i \geq k+1$. 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 (2.68) with $\bar{\delta} = \delta'$. By Lemma 2.3.1, we have

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

By Markov's inequality, the above bound implies

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

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 $\mathsf{H}_k(\widehat{r}, r^*) \in \{0, (2k)^{-1}, 2(2k)^{-1}, 3(2k)^{-1}, \cdots, 1\}$. Therefore, we always have

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

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

Proof of Theorem 2.3.3. With some rearrangements, the condition is equivalent to

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

Since ϵ is a constant, it implies

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

for any $\delta = o(1)$. Therefore, $\mathsf{H}_k(\widehat{r}, r^*) = o(k^{-1})$ when $k \to \infty$. Given the fact that $\mathsf{H}_k(\widehat{r}, r^*) \in \{0, (2k)^{-1}, 2(2k)^{-1}, 3(2k)^{-1}, \cdots, 1\}$, 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 $\left(\frac{1}{2} - \frac{V(\kappa)}{(1-\delta)npL\Delta^2}\log\frac{n-k}{k}\right)^2 > c_1$ for some constant $c_1 > 0$. Therefore, $\mathsf{H}_k(\widehat{r}, r^*) = o(1) = o(k^{-1})$, which implies $\mathsf{H}_k(\widehat{r}, r^*) = 0$. □

2.9 Analysis of the Spectral Method

We prove results for the spectral method in this section. This includes Theorem 2.4.1, Theorem 2.4.2 and Theorem 2.4.3. The proofs of Theorem 2.4.1 and Theorem 2.4.2 are given in Section 2.9.1, and then we prove Theorem 2.4.3 in Section 2.9.2.

2.9.1 Proofs of Theorem 2.4.1 and Theorem 2.4.2

The proof of Theorem 2.4.1 relies on a leave-one-out argument introduced by [25]. Without loss of generality, we consider $r_i^* = i$ so that $\theta^*_{r_i^*} = \theta_i^*$. Following [25], we define a transition matrix $P^{(m)}$ for each $m \in [n]$. For any $i \neq j$, $P_{ij}^{(m)} = P_{ij}$ if $i \neq m$ and $j \neq m$ and otherwise $P_{ij}^{(m)} = \frac{p}{d}\psi(\theta_i^* - \theta_j^*)$. For any $i \in [n]$, $P_{ii}^{(m)} = \sum_{j\in[n]\setminus\{i\}} P_{ij}^{(m)}$. Let $\pi^{(m)}$ be the stationary distribution of $P^{(m)}$. The following ℓ_2 norm bound has essentially been proved in [25].

Lemma 2.9.1. *Under the setting of Theorem 2.4.1, there exists a constant $C > 0$ such that*

$$\max_{m\in[n]} \|\pi^{(m)} - \widehat{\pi}\| \leq C\frac{1}{n}\sqrt{\frac{\log n}{npL}},$$

$$\max_{m\in[n]} \|\pi^{(m)} - \pi^*\|_\infty \leq C\frac{1}{n}\sqrt{\frac{\log n}{npL}},$$

$$\max_{m\in[n]} \|\pi^{(m)} - \pi^*\| \leq C\frac{1}{n}\sqrt{\frac{1}{pL}},$$

with probability at least $1 - O(n^{-4})$.

Proof. By Lemma 5.6 and Lemma 5.7 of [25], one can obtain $\|\pi^{(m)} - \widehat{\pi}\| \leq C_1\sqrt{\frac{\log n}{npL}}\|\pi^*\|_\infty + \|\widehat{\pi} - \pi^*\|_\infty$ for some constant $C_1 > 0$ with probability at least $1 - O(n^{-5})$. Theorem 2.6 of [25] gives the bound $\|\widehat{\pi} - \pi^*\|_\infty \leq C_2\sqrt{\frac{\log n}{npL}}\|\pi^*\|_\infty$ with probability at least $1 - O(n^{-5})$. A union bound argument together with the fact that $\|\pi^*\|_\infty \asymp n^{-1}$ leads to the first conclusion. The second conclusion is a consequence of triangle inequality. By Theorem 5.2 of [25], we have $\|\widehat{\pi} - \pi^*\| \leq C_3\frac{1}{n}\sqrt{\frac{1}{pL}}$ with probability at least $1 - O(n^{-1})$. Thus, we obtain the last conclusion by applying triangle inequality again. □

We also need a lemma that relates the asymptotic variance of $\widehat{\pi}_i$ to the function $\overline{V}(\kappa)$.

Lemma 2.9.2. *For any positive* $\kappa_1, \kappa_2 = O(1)$*, we have*

$$\begin{aligned}
&\min_{\substack{x_1,\ldots,x_k\in[0,\kappa_1]\\ x_{k+1},\ldots,x_n\in[0,\kappa_2]}} \frac{(\sum_{i=1}^k \psi(x_i) + \sum_{i=k+1}^n \psi(-x_i))^2}{\sum_{i=1}^k \psi'(x_i)(1+e^{x_i})^2 + \sum_{i=k+1}^n \psi'(x_i)(1+e^{-x_i})^2} \\
= \;& \frac{(k\psi(\kappa_1) + (n-k)\psi(-\kappa_2))^2}{k\psi'(\kappa_1)(1+e^{\kappa_1})^2 + (n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2},
\end{aligned}$$

for n *that is sufficiently large.*

Proof. The problem is equivalent to the solution of the following: the optimum of the problem

$$\min_{\substack{x_1,\ldots,x_k\in[1,M_1]\\ x_{k+1},\ldots,x_n\in[1,M_2]}} \frac{(\sum_{i=1}^k \frac{2x_i}{1+x_i} + \sum_{i=k+1}^n \frac{2}{1+x_i})^2}{\sum_{i=1}^k x_i + \sum_{i=k+1}^n \frac{1}{x_i}} = \min_{\substack{x_1,\ldots,x_k\in[1,M_1]\\ x_{k+1},\ldots,x_n\in[1,M_2]}} f(x_1,\cdots,x_n)$$

is obtained at $x_1 = ... = x_k = M_1, x_{k+1} = ... = x_n = M_2$. We will show that for any given $x_{k+1}, ..., x_n \in [1, M_2]$, the function is minimized at $x_1 = ... = x_k = M_1$. Moreover, for any

given $x_1, ..., x_k$, the function is minimized at $x_{k+1} = ... = x_n = M_2$. We only need to prove the former claim and the latter one can be proved similarly. Define

$$g(x_1, \cdots, x_k) = \frac{\left(\sum_{i=1}^k \frac{2x_i}{1+x_i} + \alpha\right)^2}{\sum_{i=1}^k x_i + \beta},$$

where $\alpha = \sum_{i=k+1}^n \frac{2}{1+x_i}, \beta = \sum_{i=k+1}^n \frac{1}{x_i}$. We first analyze the behavior of $g(x_1, \cdots, x_k)$ at each coordinate. By direct calculation, we have

$$\begin{aligned}
\frac{\partial \log g(x_1, \cdots, x_k)}{\partial x_1} &= \frac{4}{(1+x_1)^2(\sum_{i=1}^k \frac{2x_i}{1+x_i} + \alpha)} - \frac{1}{\sum_{i=1}^k x_i + \beta} \\
&= \frac{4(\sum_{i=1}^k x_i + \beta) - (1+x_1)^2(\sum_{i=1}^k \frac{2x_i}{1+x_i} + \alpha)}{(1+x_1)^2(\sum_{i=1}^k \frac{2x_i}{1+x_i} + \alpha)(\sum_{i=1}^k x_i + \beta)}.
\end{aligned}$$

The sign of the partial derivative is determined by its numerator

$$\begin{aligned}
& 4(\sum_{i=1}^k x_i + \beta) - (1+x_1)^2(\sum_{i=1}^k \frac{2x_i}{1+x_i} + \alpha) \\
= & -\left(\sum_{i=2}^k \frac{2x_i}{1+x_i} + \alpha + 2\right) x_1^2 - \left(\sum_{i=2}^k \frac{4x_i}{1+x_i} + 2\alpha - 2\right) x_1 \\
& +4(\sum_{i=2}^k x_i + \beta) - \left(\sum_{i=2}^k \frac{2x_i}{1+x_i} + \alpha\right),
\end{aligned}$$

which is a quadratic decreasing function of $x_1 \in [1, M_1]$. Therefore, $g(x_1, \cdots, x_k)$ is either monotone of $x_1 \in [1, M_1]$, or it is first increasing then decreasing. This implies that the optimum is achieved either at $x_1 = 1$ or $x_1 = M_1$. Since $g(x_1, \cdots, x_k)$ is symmetric, we therefore know that the optimizer must satisfy $(x_1, \cdots, x_k) \in \{1, M_1\}^k$. Using symmetry again, we can conclude that the value of $\min_{x_1, \cdots, x_k \in [1, M_1]} g(x_1, \cdots, x_k)$ is determined by the number of coordinates that take M_1. For $i \in [k]$, we define g_i to be the value of $g(x_1, \cdots, x_k)$ with $x_1 = \cdots = x_i = M_1$ and $x_{i+1} = \cdots = x_k = 1$. We now need to show g_i is nonincreasing

in $i \in [k]$. Note that

$$
\begin{aligned}
g_i \geq g_{i+1} \quad &\iff \quad \frac{(i\frac{2M_1}{M_1+1}+k-i+\alpha)^2}{iM_1+k-i+\beta} \geq \frac{(\frac{M_1-1}{M_1+1}+i\frac{2M_1}{M_1+1}+k-i+\alpha)^2}{M_1-1+iM_1+k-i+\beta} \\
&\iff \quad \frac{M_1-1}{iM_1+k-i+\beta} \geq \frac{(\frac{M_1-1}{M_1+1})^2}{(i\frac{2M_1}{M_1+1}+k-i+\alpha)^2} + \frac{2(\frac{M_1-1}{M_1+1})}{i\frac{2M_1}{M_1+1}+k-i+\alpha} \\
&\iff \quad \frac{(M_1+1)^2}{iM_1+k-i+\beta} - \frac{M_1-1}{(i\frac{2M_1}{M_1+1}+k-i+\alpha)^2} - \frac{2(M_1+1)}{i\frac{2M_1}{M_1+1}+k-i+\alpha} \geq 0 \\
&\iff \quad \frac{(i\frac{M_1-1}{M_1+1}+k+\alpha)(M_1+1)^2}{i(M_1-1)+k+\beta} - \frac{M_1-1}{i\frac{M_1-1}{M_1+1}+k+\alpha} - 2(M_1+1) \geq 0 \\
&\iff \quad \frac{i(M_1-1)+(k+\alpha)(M_1+1)}{i(M_1-1)+k+\beta} - \frac{M_1-1}{i(M_1-1)+(k+\alpha)(M_1+1)} - 2 \geq 0 \\
&\impliedby \quad \frac{i(M_1-1)+(k+\beta)(M_1+1)}{i(M_1-1)+k+\beta} - \frac{M_1-1}{i(M_1-1)+(k+\beta)(M_1+1)} - 2 \geq 0 \quad (2.89) \\
&\iff \quad \frac{-i(M_1-1)+(k+\beta)(M_1-1)}{i(M_1-1)+k+\beta} - \frac{M_1-1}{i(M_1-1)+(k+\beta)(M_1+1)} \geq 0 \\
&\impliedby \quad \frac{-i+(k+\beta)}{i(M_1-1)+k+\beta} - \frac{1}{i(M_1-1)+(k+\beta)(M_1+1)} \geq 0 \\
&\iff \quad (k+\beta)^2(M_1+1) \geq i(M_1-1)+i^2(M_1-1)+(2i+1)(k+\beta) \\
&\impliedby \quad (k+\beta)^2(M_1+1) \geq (k-1)^2(M_1-1)+(k-1)(M_1-1)+(2k-1)(k+\beta) \\
&\iff \quad k^2(M_1+1)+2\beta(M_1+1)k+\beta^2(M_1+1) \geq k^2(M_1+1)+(-M_1+2\beta)k-\beta \\
&\iff \quad (2\beta+1)M_1k+\beta^2(M_1+1)+\beta \geq 0
\end{aligned}
$$

where the last display is trivially true. We have used $\alpha \geq \beta$ for the step (2.89). Therefore, $\min_{x_1,\cdots,x_k \in [1,M_1]} g(x_1,\cdots,x_k) = g_k$, and the proof is complete. □

Now we are ready to prove Theorem 2.4.1.

Proof of Theorem 2.4.1. When the error exponent is of constant order, the bound is also a constant, and the result already holds since $\mathsf{H}_k(\widehat{r}, r^*) \leq 1$. Therefore, we only need to consider the case when the error exponent tends to infinity. We first introduce some

notation. Define

$$\eta = \frac{1}{2} - \frac{\overline{V}(\kappa)}{(1-\bar{\delta})\Delta^2 npL} \log \frac{n-k}{k}, \tag{2.90}$$

where $\bar{\delta} = o(1)$ is chosen so that $\eta > 0$ is satisfied. The specific choice of $\bar{\delta}$ will be determined later in the proof. We will continue to use the notation $\bar{\Delta}_i$ that is defined in (2.69). Since the diverging exponent implies $\overline{\mathsf{SNR}} \to \infty$, we have $\min_{i\in[n]} \bar{\Delta}_i^2 Lnp \to \infty$ and $\max_{i\in[n]} \bar{\Delta}_i \to 0$.

Since $\widehat{\pi}$ is the stationary distribution of P, we have $\hat{\pi}^T P = \hat{\pi}^T$. This implies that for any $m \in [n]$, we have $\sum_{j=1}^n P_{jm}\widehat{\pi}_j = \widehat{\pi}_m$. We can equivalently write this identity as

$$\widehat{\pi}_m = \frac{\sum_{j\in[n]\setminus\{m\}} P_{jm}\widehat{\pi}_j}{1-P_{mm}} = \frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{mj}\widehat{\pi}_j}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}}.$$

We approximate $\widehat{\pi}_m$ by

$$\bar{\pi}_m = \frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{mj}\pi^*_j}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}}. \tag{2.91}$$

The approximation error can be bounded by

$$\begin{aligned} |\widehat{\pi}_m - \bar{\pi}_m| \;\le\; & \left| \frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{mj}(\widehat{\pi}_j - \pi_j^{(m)})}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}} \right| && (2.92) \\ & + \left| \frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{mj}(\pi_j^{(m)} - \pi^*_j)}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}} \right|. && (2.93) \end{aligned}$$

The two terms (2.92) and (2.93) share a common denominator, which can be lower bounded by

$$\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm} \ge \sum_{j\in[n]\setminus\{m\}} A_{jm}\psi(\theta^*_j - \theta^*_m) - \left| \sum_{j\in[n]\setminus\{m\}} A_{jm}(\bar{y}_{jm} - \psi(\theta^*_j - \theta^*_m)) \right|. \tag{2.94}$$

By Lemma 2.8.1 and Lemma 2.8.4, we have $\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm} \ge c_1 np$ for some constant

$c_1 > 0$ with probability at least $1 - O(n^{-10})$. With this lower bound, we then bound (2.92) as

$$\begin{aligned}\left|\frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{mj}(\widehat{\pi}_j - \pi_j^{(m)})}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}}\right| &\leq \frac{\sqrt{\sum_{j\in[n]\setminus\{m\}} A_{1j}\bar{y}_{mj}^2}\|\widehat{\pi} - \pi^{(m)}\|}{c_1 np} \\ &\leq \frac{\sqrt{\sum_{j\in[n]\setminus\{m\}} A_{1j}}\|\widehat{\pi} - \pi^{(m)}\|}{c_1 np} \\ &\leq C_1\frac{1}{n}\sqrt{\frac{\log n}{(np)^2 L}},\end{aligned}$$

with probability at least $1 - O(n^{-4})$. In the last inequality, we have used Lemma 2.8.1 and Lemma 2.9.1. For (2.93), we can bound it as

$$\begin{aligned}&\left|\frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{mj}(\pi_j^{(m)} - \pi_j^*)}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}}\right| \\ \leq\ &\frac{\left|\sum_{j\in[n]\setminus\{m\}} A_{jm}(\bar{y}_{mj} - \psi(\theta_m^* - \theta_j^*))(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np} \\ &+\frac{p\left|\sum_{j\in[n]\setminus\{m\}} \psi(\theta_m^* - \theta_j^*)(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np} \\ &+\frac{\left|\sum_{j\in[n]\setminus\{m\}} (A_{jm} - p)\psi(\theta_m^* - \theta_j^*)(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np}.\end{aligned}$$

We bound the three terms above separately. For the first term, we use Hoeffding's inequality (Lemma 2.12.1), and get

$$\frac{\left|\sum_{j\in[n]\setminus\{m\}} A_{jm}(\bar{y}_{mj} - \psi(\theta_m^* - \theta_j^*))(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np} \leq C_2\frac{\sqrt{\frac{x}{L}\sum_{j\in[n]\setminus\{m\}} A_{jm}(\pi_j^{(m)} - \pi_j^*)^2}}{np}, \tag{2.95}$$

with probability at least $1 - e^{-x}$. By Lemma 2.8.1 and Lemma 2.9.1, we have

$$\sqrt{\sum_{j\in[n]\setminus\{m\}} A_{jm}(\pi_j^{(m)} - \pi_j^*)^2} \le \|\pi^{(m)} - \pi^*\|_\infty \sqrt{\sum_{j\in[n]\setminus\{m\}} A_{jm}} \le C_3 \frac{1}{n}\sqrt{\frac{\log n}{L}},$$

with probability at least $1 - O(n^{-4})$. Taking $x = \bar{\Delta}_m^2 npL\sqrt{\frac{npL}{\log n}}$, we have

$$\frac{\left|\sum_{j\in[n]\setminus\{m\}} A_{jm}(\bar{y}_{mj} - \psi(\theta_m^* - \theta_j^*))(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np} \le C_4 \frac{1}{n}\bar{\Delta}_m \left(\frac{\log n}{Lnp}\right)^{1/4},$$

with probability at least $1 - O(n^{-4}) - \exp\left(-\bar{\Delta}_m^2 npL\sqrt{\frac{npL}{\log n}}\right)$. Next, for the second term, we apply Lemma 2.9.1 and get

$$\frac{p\left|\sum_{j\in[n]\setminus\{m\}} \psi(\theta_m^* - \theta_j^*)(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np} \le \frac{\|\pi^{(m)} - \pi^*\|}{c_1\sqrt{n}} \le C_5 \frac{1}{n}\sqrt{\frac{1}{npL}},$$

with probability at least $1 - O(n^{-4})$. For the third term, we use Bernstein's inequality (Lemma 2.12.2), and get

$$\frac{\left|\sum_{j\in[n]\setminus\{m\}} (A_{jm} - p)\psi(\theta_m^* - \theta_j^*)(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np} \le C_6 \frac{\sqrt{px}\|\pi^{(m)} - \pi^*\|}{np} + C_6 \frac{x\|\pi^{(m)} - \pi^*\|_\infty}{np}, \tag{2.96}$$

with probability at least $1 - e^{-x}$. We choose $x = \min\left(\bar{\Delta}_m^2 Lnp\frac{np}{\log n}, 4\log n\right)$. Then, with the help of Lemma 2.9.1, we have

$$\begin{aligned}
&\frac{\left|\sum_{j\in[n]\setminus\{m\}} (A_{jm} - p)\psi(\theta_m^* - \theta_j^*)(\pi_j^{(m)} - \pi_j^*)\right|}{c_1 np} \\
&\le C_7 \frac{1}{n}\frac{1}{np\sqrt{L}}\sqrt{\min\left(\bar{\Delta}_m^2 Lnp\frac{np}{\log n}, \log n\right)} + C_7 \frac{1}{n}\frac{1}{np}\sqrt{\frac{\log n}{npL}}\min\left(\bar{\Delta}_m^2 Lnp\frac{np}{\log n}, \log n\right),
\end{aligned} \tag{2.97}$$

with probability at least $1 - O(n^{-4}) - \exp\left(-\bar{\Delta}_m^2 npL\frac{np}{\log n}\right)$.

To summarize, we have proved that

$$\frac{|\widehat{\pi}_m - \bar{\pi}_m|}{\pi_m^*} \leq \delta(1 - e^{-\bar{\Delta}_m}), \tag{2.98}$$

for some $\delta = o(1)$ with probability at least

$$1 - O(n^{-4}) - \exp\left(-\bar{\Delta}_m^2 npL\frac{np}{\log n}\right) - \exp\left(-\bar{\Delta}_m^2 npL\sqrt{\frac{npL}{\log n}}\right)$$

under the assumption that $\bar{\Delta}_m = o(1)$, $npL\bar{\Delta}_m^2 \to \infty$ and $\frac{np}{\log n} \to \infty$.

Next, we note that by the definition of $\bar{\pi}_m$, we have

$$\bar{\pi}_m - \pi_m^* = \frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}(\bar{y}_{mj} - \psi(\theta_m^* - \theta_j^*))(\pi_j^* + \pi_m^*)}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}}. \tag{2.99}$$

By Lemma 2.8.4 and the inequality (2.94), the denominator of (2.99) satisfies

$$\left|\frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\psi(\theta_j^* - \theta_m^*)} - 1\right| \leq \delta, \tag{2.100}$$

for some $\delta = o(1)$ with probability at least $1 - O(n^{-10})$. Note that we can choose the same δ to accommodate both bounds (2.98) and (2.100).

We will apply Lemma 2.3.1 with

$$t = \frac{e^{(1-\eta)\theta_k^* + \eta\theta_{k+1}^*}}{\sum_{j=1}^n e^{\theta_j^*}} \tag{2.101}$$

to finish the proof. Recall the definition of η in (2.90). For $i \le k$, we have

$$
\begin{aligned}
& \mathbb{P}\left(\widehat{\pi}_i \le \frac{e^{(1-\eta)\theta_k^* + \eta\theta_{k+1}^*}}{\sum_{j=1}^n e^{\theta_j^*}}\right) \\
= \;& \mathbb{P}\left(\frac{\widehat{\pi}_i - \pi_i^*}{\pi_i^*} \le e^{(1-\eta)\theta_k^* + \eta\theta_{k+1}^* - \theta_i^*} - 1\right) \\
\le \;& \mathbb{P}\left(\frac{\widehat{\pi}_i - \pi_i^*}{\pi_i^*} \le e^{-\bar{\Delta}_i} - 1\right) \\
\le \;& \mathbb{P}\left(\frac{\bar{\pi}_i - \pi_i^*}{\pi_i^*} \le -(1-\delta)(1-e^{-\bar{\Delta}_i})\right) + \mathbb{P}\left(\frac{|\bar{\pi}_i - \widehat{\pi}_i|}{\pi_i^*} > \delta(1-e^{-\bar{\Delta}_i})\right) \\
\le \;& \mathbb{P}\left(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^* - \theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \le -(1-\delta)^2(1-e^{-\bar{\Delta}_i})\right) \\
& + \mathbb{P}\left(\frac{|\bar{\pi}_i - \widehat{\pi}_i|}{\pi_i^*} > \delta(1-e^{-\bar{\Delta}_i})\right) + \mathbb{P}\left(\left|\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}\bar{y}_{ji}}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} - 1\right| > \delta\right) \\
\le \;& \mathbb{P}\left(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^* - \theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \le -(1-\delta)^2(1-e^{-\bar{\Delta}_i})\right) \\
& + O(n^{-4}) + \exp\left(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\right) + \exp\left(-\bar{\Delta}_i^2 npL\sqrt{\frac{npL}{\log n}}\right), \qquad (2.102)
\end{aligned}
$$

where the last inequality is by (2.98) and (2.100). Define the event

$$
\begin{aligned}
\mathcal{A}_i = \{A : & \left|\frac{\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^* - \theta_j^*)\left(1+e^{\theta_j^* - \theta_i^*}\right)^2}{p\sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*)\left(1+e^{\theta_j^* - \theta_i^*}\right)^2} - 1\right| \le \delta, \\
& \left|\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)}{p\sum_{j\in[n]\setminus\{i\}} \psi(\theta_j^* - \theta_i^*)} - 1\right| \le \delta\}. \qquad (2.103)
\end{aligned}
$$

Then, by Bernstein's inequality, we have

$$
\begin{aligned}
& \mathbb{P}\left(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^*-\theta_i^*)} \le -(1-\delta)^2(1-e^{-\bar{\Delta}_i})\right) \\
\le & \sup_{A\in\mathcal{A}_i} \mathbb{P}\left(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^*-\theta_i^*)} \le -(1-\delta)^2(1-e^{-\bar{\Delta}_i})\Big| A\right) \\
& +\mathbb{P}(A\in\mathcal{A}_i^c) \\
\le & \exp\left(-\frac{(1-o(1))Lp\bar{\Delta}_i^2\left(\sum_{j\in[n]\setminus\{i\}}\psi(\theta_j^*-\theta_i^*)\right)^2}{2\sum_{j\in[n]\setminus\{i\}}\psi'(\theta_i^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_i^*}\right)^2}\right)+O(n^{-4}) \qquad (2.104) \\
\le & \exp\left(-\frac{(1-o(1))Lp(\bar{\Delta}+\theta_i^*-\theta_k^*)^2\left(\sum_{j\in[n]\setminus\{i\}}\psi(\theta_j^*-\theta_i^*)\right)^2}{2\sum_{j\in[n]\setminus\{i\}}\psi'(\theta_i^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_i^*}\right)^2}\right)+O(n^{-4}). \qquad (2.105)
\end{aligned}
$$

The inequality (2.105) is by the same argument that leads to (2.84) and (2.85). We use the notation $\bar{\Delta}=\min\left(\eta(\theta_k^*-\theta_{k+1}^*),\left(\frac{\log n}{np}\right)^{1/4}\right)$ in (2.105). Define

$$
h_i(t)=\frac{(\bar{\Delta}+t)^2\left(\sum_{j\in[n]\setminus\{i\}}\psi(\theta_j^*-\theta_k^*-t)\right)^2}{\sum_{j\in[n]\setminus\{i\}}\psi'(t+\theta_k^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_k^*-t}\right)^2}, \qquad \text{for all } t\ge 0.
$$

Though $h_i(t)$ is a complicated function, by the fact that $\bar{\Delta}=o(1)$ and $\max_{j,k}|\theta_j^*-\theta_k^*|\le\kappa=O(1)$, one can directly analyze the derivative of $h_i(t)$ to conclude that there exists some small constant $c_2>0$ such that $h_i(t)$ is increasing on $[0,c_2]$. Moreover, there also exists a

small constant $c_3 > 0$ such that $\min_{t\in[c_2,\kappa]} h_i(t) \geq c_3 n$. This implies

$$\begin{aligned}
& \frac{Lp(\bar{\Delta}+\theta_i^*-\theta_k^*)^2\left(\sum_{j\in[n]\backslash\{i\}}\psi(\theta_j^*-\theta_i^*)\right)^2}{2\sum_{j\in[n]\backslash\{i\}}\psi'(\theta_i^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_i^*}\right)^2} \\
\geq & \frac{Lp\bar{\Delta}^2\left(\sum_{j\in[n]\backslash\{i\}}\psi(\theta_j^*-\theta_k^*)\right)^2}{2\sum_{j\in[n]\backslash\{i\}}\psi'(\theta_k^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_k^*}\right)^2} \wedge \frac{c_3 npL}{2} \\
= & \frac{Lp\bar{\Delta}^2\left(\sum_{j\in[n]\backslash\{i\}}\psi(\theta_j^*-\theta_k^*)\right)^2}{2\sum_{j\in[n]\backslash\{i\}}\psi'(\theta_k^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_k^*}\right)^2},
\end{aligned}$$

where the last inequality is due to the fact that $\bar{\Delta} = o(1)$. We further bound the above exponent by

$$\begin{aligned}
& \frac{Lp\bar{\Delta}^2\left(\sum_{j\in[n]\backslash\{i\}}\psi(\theta_j^*-\theta_k^*)\right)^2}{2\sum_{j\in[n]\backslash\{i\}}\psi'(\theta_k^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_k^*}\right)^2} \\
= & (1-o(1))\frac{Lp\bar{\Delta}^2\left(\sum_{j=1}^n\psi(\theta_j^*-\theta_k^*)\right)^2}{2\sum_{j=1}^n\psi'(\theta_k^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_k^*}\right)^2} \\
\geq & (1-o(1))\frac{Lp\bar{\Delta}^2}{2} \\
& \min_{\substack{\kappa_1+\kappa_2\leq\kappa\\ \kappa_1,\kappa_2\geq 0}} \min_{\substack{x_1,\cdots,x_k\in[0,\kappa_1]\\ x_{k+1},\cdots,x_n\in[0,\kappa_2]}} \frac{\left(\sum_{j=1}^k\psi(x_j)+\sum_{j=k+1}^n\psi(-x_j)\right)^2}{\sum_{j=1}^k\psi'(x_j)(1+e^{x_j})^2+\sum_{j=k+1}^n\psi'(x_j)(1+e^{-x_j})^2} \\
= & (1-o(1))\frac{Lp\bar{\Delta}^2}{2}\min_{\substack{\kappa_1+\kappa_2\leq\kappa\\ \kappa_1,\kappa_2\geq 0}}\frac{(k\psi(\kappa_1)+(n-k)\psi(-\kappa_2))^2}{k\psi'(\kappa_1)(1+e^{\kappa_1})^2+(n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2} \qquad (2.106) \\
= & (1-o(1))\frac{Lpn\bar{\Delta}^2}{2\overline{V}(\kappa)}.
\end{aligned}$$

The equality (2.106) is due to Lemma 2.9.2. With the above analysis of the error exponent,

we can further bound (2.105) as

$$\begin{aligned} & \exp\left(-\frac{1-o(1)}{2}Lp\min\left(\eta^2\Delta^2,\sqrt{\frac{\log n}{np}}\right)\frac{n}{\overline{V}(\kappa)}\right)+O(n^{-4}) \\ \leq\ & \exp\left(-\frac{(1-o(1))\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)+O(n^{-4}). \end{aligned}$$

The last inequality holds because when $\min\left(\eta^2\Delta^2,\sqrt{\frac{\log n}{np}}\right)=\sqrt{\frac{\log n}{np}}$, the first term becomes $\exp\left(-\frac{(1-o(1))L\sqrt{np\log n}}{2\overline{V}(\kappa)}\right)$, which can be absorbed by $O(n^{-4})$. Since

$$\exp\left(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\right)+\exp\left(-\bar{\Delta}_i^2 npL\sqrt{\frac{npL}{\log n}}\right)\leq\exp\left(-\frac{(1-o(1))\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)+O(n^{-4})$$

, we have

$$\mathbb{P}\left(\widehat{\pi}_i\leq\frac{e^{(1-\eta)\theta_k^*+\eta\theta_{k+1}^*}}{\sum_{j=1}^n e^{\theta_j^*}}\right)\leq\exp\left(-\frac{(1-\delta_1)\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)+O(n^{-4}), \tag{2.107}$$

with some $\delta_1=o(1)$ for all $i\leq k$. With a similar argument, we also have

$$\mathbb{P}\left(\widehat{\pi}_i\geq\frac{e^{(1-\eta)\theta_k^*+\eta\theta_{k+1}^*}}{\sum_{j=1}^n e^{\theta_j^*}}\right)\leq\exp\left(-\frac{(1-\delta_1)(1-\eta)^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)+O(n^{-4}), \tag{2.108}$$

for all all $i\geq k+1$. It can be checked that the δ_1 above can be set independent of the $\bar{\delta}$ in the definition of η. Now we choose η as in (2.90) with $\bar{\delta}=\delta_1$. By Lemma 2.3.1, we have

$$\begin{aligned} \mathbb{E}\mathsf{H}_k(\widehat{r},r^*) &\leq \exp\left(-\frac{(1-\bar{\delta})\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)+\frac{n-k}{k}\exp\left(-\frac{(1-\bar{\delta})(1-\eta)^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)+O(n^{-4}) \\ &\leq 2\exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1-\bar{\delta})\overline{\mathsf{SNR}}}}{2}-\frac{1}{\sqrt{(1-\bar{\delta})\overline{\mathsf{SNR}}}}\log\frac{n-k}{k}\right)^2\right)+O(n^{-4}). \end{aligned}$$

By Markov's inequality, the above bound implies

$$\mathsf{H}_k(\widehat{r}, r^*) \leq \exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1-\delta')\overline{\mathsf{SNR}}}}{2} - \frac{1}{\sqrt{(1-\delta')\overline{\mathsf{SNR}}}}\log\frac{n-k}{k}\right)^2\right) + O(n^{-3}),$$

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

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

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

$$\mathsf{H}_k(\widehat{r}, r^*) \leq 2\exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1-\delta')\overline{\mathsf{SNR}}}}{2} - \frac{1}{\sqrt{(1-\delta')\overline{\mathsf{SNR}}}}\log\frac{n-k}{k}\right)^2\right),$$

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

Proof of Theorem 2.4.2. The proof is the same as that of Theorem 2.3.3. □

2.9.2 Proof of Theorem 2.4.3

To prove Theorem 2.4.3, we need two additional lemmas. The first lemma can be viewed as a reverse version of the inequality in Lemma 2.3.1.

Lemma 2.9.3. *Suppose $\widehat{r}$ is a rank vector induced by $\widehat{\theta}$, we then have*

$$\mathsf{H}_k(\widehat{r}, r^*) \geq \frac{1}{k}\max_{t\in\mathbb{R}}\min\left(\sum_{i:r_i^*\leq k}\mathbb{I}\left\{\widehat{\theta}_i < t\right\}, \sum_{i:r_i^*>k}\mathbb{I}\left\{\widehat{\theta}_i > t\right\}\right).$$

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

Proof. Following the proof of Lemma 2.3.1, we have

$$\begin{aligned} 2k\mathsf{H}_k(\widehat{r}, r^*) &= 2\max\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{r}_i > k\right\}, \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{r}_i \le k\right\}\right) \\ &\ge 2\max\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{\theta}_i < \widehat{\theta}_{(k)}\right\}, \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{\theta}_i > \widehat{\theta}_{(k+1)}\right\}\right) \\ &\ge 2\min_t \max\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{\theta}_i < t\right\}, \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{\theta}_i > t\right\}\right) \qquad (2.109) \\ &= 2\max_t \min\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{\theta}_i < t\right\}, \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{\theta}_i > t\right\}\right). \qquad (2.110) \end{aligned}$$

where (2.109) and (2.110) follow the same argument that leads to (2.232) and (2.233). □

Proof of Theorem 2.4.3. We first note that condition (2.19) necessarily implies $\Delta = o(1)$. Throughout the proof, we assume $\kappa = \Omega(1)$ and there exists some $\delta_1 = o(1)$ such that

$$\frac{\sqrt{(1+\delta_1)\overline{\mathsf{SNR}}}}{2} - \frac{1}{\sqrt{(1+\delta_1)\overline{\mathsf{SNR}}}}\log\frac{n-k}{k} \to \infty. \qquad (2.111)$$

The case with $\kappa = o(1)$ or $\overline{\mathsf{SNR}}$ not satisfying (2.111) will be addressed at the end of the proof.

Choose $\kappa_1, \kappa_2 \ge 0$ such that we have both $\kappa_1 + \kappa_2 \le \kappa$ and

$$\frac{k\psi'(\kappa_1)(1+e^{\kappa_1})^2 + (n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2}{(k\psi(\kappa_1) + (n-k)\psi(-\kappa_2))^2/n} = \overline{V}(\kappa).$$

Let $\rho = o(1)$ be a vanishing number that will be specified later. Since $k \to \infty$ and $\kappa = \Omega(1)$, one can easily check that $\kappa_2 = \Omega(1)$. Define $\theta_i^* = \kappa_1$ 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 $k + \rho(n-k) < i \le n$. For the simplicity of proof, we choose ρ so that both ρk and $\rho(n-k)$ are integers. Define r^*

to be $r_i^* = i, \forall i \in [n]$. Then we have

$$\sup_{\substack{r \in \mathfrak{S}_n \\ \theta \in \Theta(k,\Delta,\kappa)}} \mathbb{E}_{(\theta,r)} \mathsf{H}_k(\widehat{r}, r) \geq \mathbb{E}_{(\theta^*,r^*)} \mathsf{H}_k(\widehat{r}, r^*).$$

We will utilize several results established in the proof of Theorem 2.4.1. Define

$$\eta = \frac{1}{2} - \frac{\overline{V}(\kappa)}{(1+\bar{\delta})\Delta^2 npL} \log \frac{n-k}{k}, \tag{2.112}$$

for $\bar{\delta} = o(1)$. The specific choice of $\bar{\delta}$ will be specified later in the proof. Also define $t = \frac{e^{(1-\eta)\theta_k^* + \eta\theta_{k+1}^*}}{\sum_{j=1}^n e^{\theta_j^*}} = \frac{e^{-\eta\Delta}}{\sum_{j=1}^n e^{\theta_j^*}}$. Then, by Lemma 2.9.3, we have

$$\begin{aligned}
\mathsf{H}_k(\widehat{r}, r^*) &\geq \frac{1}{k} \min \left(\sum_{i=1}^k \mathbb{I}\left\{\widehat{\pi}_i < t\right\}, \sum_{i=k+1}^n \mathbb{I}\left\{\widehat{\pi}_i > t\right\} \right) \\
&\geq \frac{1}{k} \min \left(\sum_{k-\rho k < i \leq k} \mathbb{I}\left\{\widehat{\pi}_i < t\right\}, \sum_{k < i \leq k+\rho(n-k)} \mathbb{I}\left\{\widehat{\pi}_i > t\right\} \right).
\end{aligned}$$

For any $\delta > 0$, define the function $\phi(\delta) = \frac{\sqrt{(1+\delta)\overline{\mathsf{SNR}}}}{2} - \frac{1}{\sqrt{(1+\delta)\overline{\mathsf{SNR}}}} \log \frac{n-k}{k}$. It suffices to show there exists some constant $C > 0$ such that

$$\mathbb{P}_{(\theta^*,r^*)} \left(\sum_{k-\rho k < i \leq k} \mathbb{I}\left\{\widehat{\pi}_i < t\right\} \geq Ck \exp\left(-\frac{\phi(\bar{\delta})^2}{2}\right) \right) \geq 1 - o(1), \tag{2.113}$$

$$\text{and } \mathbb{P}_{(\theta^*,r^*)} \left(\sum_{k < i \leq k+\rho(n-k)} \mathbb{I}\left\{\widehat{\pi}_i > t\right\} \geq Ck \exp\left(-\frac{\phi(\bar{\delta})^2}{2}\right) \right) \geq 1 - o(1). \tag{2.114}$$

Suppose both inequalities hold, we have

$$\mathbb{P}_{(\theta^*,r^*)} \left(\mathsf{H}_k(\widehat{r}, r^*) > 0 \right) \geq 1 - o(1).$$

By Markov's inequality, we also have

$$\begin{aligned}\mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r},r^*) &\geq C\exp\left(-\frac{\phi(\bar{\delta})^2}{2}\right)\mathbb{P}_{(\theta^*,r^*)}\left(\mathsf{H}_k(\widehat{r},r^*)\geq C\exp\left(-\frac{\phi(\bar{\delta})^2}{2}\right)\right)\\ &\geq \frac{C}{2}\exp\left(-\frac{\phi(\bar{\delta})^2}{2}\right).\end{aligned}$$

Therefore, we obtain the desired conclusions.

In the rest of the proof, we are going to establish (2.113). Recall the definition of $\bar{\pi}$ in (2.91). For any $k-\rho k<i\leq k$, define the event $\mathcal{F}$ as

$$\mathcal{F}_i=\left\{\frac{|\widehat{\pi}_i-\bar{\pi}_i|}{\pi_i^*}\leq\delta_0(1-e^{-\eta\Delta}) \text{ and } \left|\frac{\sum_{j\in[n]\backslash\{i\}}A_{ji}\bar{y}_{ji}}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}-1\right|\leq\delta_0\right\}.$$

Using a similar argument that leads to (2.98) and (2.100), we can show that there exists some $\delta_0=o(1)$ not dependent on $\bar{\delta}$, such that

$$\mathbb{P}_{(\theta^*,r^*)}(\mathcal{F}_i)\geq 1-\left(O(n^{-4})+\exp\left(-\eta^2\Delta^2 npL\frac{np}{\log n}\right)+\exp\left(-\eta^2\Delta^2 npL\sqrt{\frac{npL}{\log n}}\right)\right). \tag{2.115}$$

Suppose $\mathcal{F}_i$ holds, we then have

$$\begin{aligned}
\mathbb{I}\left\{\widehat{\pi}_i < t\right\} &= \mathbb{I}\left\{\widehat{\pi}_i < \frac{e^{(1-\eta)\theta_k^* + \eta\theta_{k+1}^*}}{\sum_{j=1}^n e^{\theta_j^*}}\right\} \\
&= \mathbb{I}\left\{\frac{\widehat{\pi}_i - \pi_i^*}{\pi_i^*} \leq e^{(1-\eta)\theta_k^* + \eta\theta_{k+1}^* - \theta_i^*} - 1\right\} \\
&= \mathbb{I}\left\{\frac{\widehat{\pi}_i - \pi_i^*}{\pi_i^*} \leq e^{-\eta\Delta} - 1\right\} \\
&\geq \mathbb{I}\left\{\frac{\bar{\pi}_i - \pi_i^*}{\pi_i^*} \leq -(1+\delta_0)(1-e^{-\eta\Delta})\right\} \\
&\geq \mathbb{I}\left\{\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^* - \theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \leq -(1+\delta_0)^2(1-e^{-\eta\Delta})\right\} \\
&\geq \mathbb{I}\left\{\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^* - \theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \leq -(1+\delta_0)^2\eta\Delta\right\}. \quad (2.116)
\end{aligned}$$

We use the notation L_i for the indicator function on the right hand side of (2.116). In other words, we have shown that

$$\begin{aligned}
\sum_{k-\rho k<i\leq k} \mathbb{I}\left\{\widehat{\pi}_i < t\right\} &\geq \sum_{k-\rho k<i\leq k} L_i \mathbb{I}_{\mathcal{F}_i} \\
&\geq \sum_{k-\rho k<i\leq k} L_i - \sum_{k-\rho k<i\leq k} \mathbb{I}_{\mathcal{F}_i^c}.
\end{aligned}$$

By (2.115), we have

$$\mathbb{E}\left(\sum_{k-\rho k<i\leq k} \mathbb{I}_{\mathcal{F}_i^c}\right) \leq O(n^{-3}) + \rho k \exp\left(-\eta^2\Delta^2 npL\frac{np}{\log n}\right) + \rho k \exp\left(-\eta^2\Delta^2 npL\sqrt{\frac{npL}{\log n}}\right).$$

Since the above bounds is of smaller order than $k\exp\left(-\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\left(\frac{np}{\log n}\right)^{1/4}\right)$, we can use

Markov's inequality and obtain

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{k-\rho k<i\le k}\mathbb{I}_{\mathcal{F}_i^c}\le k\exp\left(-\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\left(\frac{np}{\log n}\right)^{1/4}\right)\right)\ge 1-o(1). \tag{2.117}$$

To lower bound $\sum_{k-\rho k<i\le k}L_i$, we define

$$\mathcal{A}=\Bigg\{A:\forall k-\rho k<i\le k,\left|\frac{\sum_{j\in[n]\setminus\{i\}}A_{ij}\psi'(\theta_i^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_i^*}\right)^2}{p\sum_{j\in[n]\setminus\{i\}}\psi'(\theta_i^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_i^*}\right)^2}-1\right|\le\delta_0, \tag{2.118}$$

$$\left|\frac{\sum_{j\in[n]\setminus\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}{p\sum_{j\in[n]\setminus\{i\}}\psi(\theta_j^*-\theta_i^*)}-1\right|\le\delta_0, \tag{2.119}$$

$$\left|\sum_{k-\rho k<j<k}A_{ji}\psi'(\theta_i^*-\theta_j^*)(1+e^{\theta_j^*-\theta_i^*})^2\right|\le 2\rho kp+10\log n\Bigg\}. \tag{2.120}$$

By Bernstein's inequality and union bound, we have $\mathbb{P}(A\in\mathcal{A})\ge 1-O(n^{-3})$. From now on, we use the notation $\mathbb{P}_A$ for the conditional probability $\mathbb{P}_{(\theta^*,r^*)}(\cdot|A)$ given A. For any $s>0$,

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{k-\rho k<i\le k}L_i\ge s\right)\ge\mathbb{P}(A\in\mathcal{A})\inf_{A\in\mathcal{A}}\mathbb{P}_A\left(\sum_{k-\rho k<i\le k}L_i\ge s\right). \tag{2.121}$$

To study $\mathbb{P}_A\left(\sum_{k-\rho k<i\le k}L_i\ge s\right)$, we define the set $S=\{i\in[n]:i\le k-\rho k \text{ or } i>k\}$.

Note that for each $k-\rho k<i\le k$, we have $L_i\ge L_{i,1}-L_{i,2}-L_{i,3}$, where

$$L_{i,1}=\mathbb{I}\left\{\frac{\sum_{j\in S}A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}\le-(1+2\delta')(1+\delta_0)^2\eta\Delta\right\}$$
$$L_{i,2}=\mathbb{I}\left\{\frac{\sum_{k-\rho k<j<i}A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}\ge\delta'(1+\delta_0)^2\eta\Delta\right\}$$
$$L_{i,3}=\mathbb{I}\left\{\frac{\sum_{i<j\le k}A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}\ge\delta'(1+\delta_0)^2\eta\Delta\right\},$$

for some $\delta'=o(1)$ whose value will be determined later. We are going to control each term separately.

(1). Analysis of $L_{i,1}$. Note that conditional on A, $\{L_{i,1}\}_{k-\rho k<i\le k}$ are all independent Bernoulli random variables. We have $L_{i,1}\sim\text{Bernoulli}(p_i)$, where $p_i=\mathbb{E}_{(\theta^*,r^*)}(L_{i,1}|A)$. By Chebyshev's inequality, we have

$$\mathbb{P}_A\left(\sum_{k-\rho k<i\le k}L_{i,1}\ge\frac{1}{2}\sum_{k-\rho k<i\le k}p_i\right)\ge1-\frac{4}{\sum_{k-\rho k<i\le k}p_i}.$$

By Lemma 2.9.4 stated and proved at the end of the section, we can lower bound each p_i by

$$\begin{aligned}p_i&=\mathbb{P}_A\left(\frac{\sum_{j\in S}A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}\le-(1+2\delta')(1+\delta_0)^2\eta\Delta\right)\\&\ge C_1\exp\left(-\frac{(1+\delta_2)\eta^2\Delta^2npL}{2\overline{V}(\kappa)}-C_1'\eta\sqrt{\frac{\Delta^2npL}{\overline{V}(\kappa)}}\right),\end{aligned}$$

for some constants $C_1,C_1'>0$ and some $\delta_2=o(1)$ that are not dependent on η. By (2.111), there exists some $\delta_3=o(1)$ such that

$$\sum_{k-\rho k<i\le k}p_i\ge C_1k\exp\left(-\frac{(1+\delta_3)\eta^2\Delta^2npL}{2\overline{V}(\kappa)}\right).\tag{2.122}$$

To obtain (2.226), we need to set ρ that tends to zero sufficiently slow so that it can be absorbed into the exponent. Note that condition (2.19) is equivalent to

$$\frac{(1+\epsilon)\overline{\mathsf{SNR}}}{2}\left(\frac{1}{2}-\frac{1}{(1+\epsilon)\overline{\mathsf{SNR}}}\log\frac{n-k}{k}\right)^2<\log k.$$

Since ϵ is a constant, it implies

$$\frac{\overline{\mathsf{SNR}}}{2}\left(\frac{1}{2}-\frac{1}{(1+\bar{\delta})\overline{\mathsf{SNR}}}\log\frac{n-k}{k}\right)^2<(1-\epsilon')^{-1}\log k,$$

for some constant $\epsilon'>0$. As a result, under the condition that $k\to\infty$, we have

$$\sum_{k-\rho k<i\le k}p_i\ge\sum_{k-\rho k<i\le k}C_1\exp\left(-(1+\delta_3)(1-\epsilon')\log k\right)\ge k^{\frac{\epsilon'}{2}}\to\infty.$$

Hence, we have proved

$$\inf_{A\in\mathcal{A}}\mathbb{P}_A\left(\sum_{k-\rho k<i\le k}L_{i,1}\ge\frac{1}{2}C_1k\exp\left(-\frac{(1+\delta_2)\eta^2\Delta^2npL}{2\overline{V}(\kappa)}-C_1'\eta\sqrt{\frac{\Delta^2npL}{\overline{V}(\kappa)}}\right)\right)\ge1-o(1).$$

(2). Analysis of $L_{i,2}$. By (2.221)-(2.223) and Bernstein's inequality, we can bound $\mathbb{E}(L_{i,2}|A)$ by

$$\begin{aligned}&e^{\left(-\frac{\left(\delta'(1+\delta_0)^2\eta\Delta L\sum_{j\in[n]\setminus\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)\right)^2}{2\left(L\sum_{k-\rho k<j<i}A_{ji}\psi'(\theta_i^*-\theta_j^*)(1+e^{\theta_j^*-\theta_i^*})^2+\frac{1}{3}\delta'(1+\delta_0)^2\eta\Delta L\sum_{j\in[n]\setminus\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)\right)}\right)}\\&\le\exp\left(-\frac{\left(\delta'(1+\delta_0)^2\eta\Delta L\sum_{j\in[n]\setminus\{i\}}p\psi(\theta_j^*-\theta_i^*)\right)^2}{4\left(2L\rho kp+10\log n+\frac{1}{3}\delta'(1+\delta_0)^2\eta\Delta L\sum_{j\in[n]\setminus\{i\}}p\psi(\theta_j^*-\theta_i^*)\right)}\right).\end{aligned}$$

Now we set $\delta'=\max\{\rho^{\frac{1}{2}},\Delta^{\frac{4}{3}},\left(\frac{\log n}{np}\right)^{\frac{1}{2}}\}$. Then, there exists some constant $C_2,C_3>0$ such

that

$$\mathbb{E}(L_{i,2}|A) \leq \exp\left(-C_2\rho^{-\frac{1}{2}}npL\eta^2\Delta^2\right) \leq \exp\left(-C_3\rho^{-1/2}\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right).$$

Then,

$$\mathbb{E}\left(\sum_{k-\rho k<i\leq k} L_{i,2}\middle| A\right) \leq \rho k \exp\left(-C_3\rho^{-1/2}\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right).$$

By Markov inequality, we have

$$\inf_{A\in\mathcal{A}} \mathbb{P}_A\left(\sum_{k-\rho k<i\leq k} L_{i,2} \geq \rho k \exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)\right) \leq \exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right). \tag{2.123}$$

(3). Analysis of $L_{i,3}$. By a similar argument, we also have

$$\inf_{A\in\mathcal{A}} \mathbb{P}_A\left(\sum_{k-\rho k<i\leq k} L_{i,3} \geq \rho k \exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)\right) \leq \exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\eta^2\Delta^2 npL}{2\overline{V}(\kappa)}\right). \tag{2.124}$$

Now we can combine the above analyses of $L_{i,1}$, $L_{i,2}$ and $L_{i,3}$. Since $\rho = o(1)$, the bounds (2.227) and (2.228) are of smaller order than (2.226). We have

$$\inf_{A\in\mathcal{A}} \mathbb{P}_A\left(\sum_{k-\rho k<i\leq k} L_i \geq C_4 k \exp\left(-\frac{(1+\delta_2)\eta^2\Delta^2 npL}{2\overline{V}(\kappa)} - C_1'\eta\sqrt{\frac{\Delta^2 npL}{\overline{V}(\kappa)}}\right)\right) \geq 1 - o(1), \tag{2.125}$$

for some constant $C_4 > 0$. Then (2.220) and (2.224) lead to

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{k-\rho k<i\leq k} \mathbb{I}\left\{\widehat{\pi}_i < t\right\} \geq C_4 k \exp\left(-\frac{(1+\delta_2)\eta^2\Delta^2 npL}{2\overline{V}(\kappa)} - C_1'\eta\sqrt{\frac{\Delta^2 npL}{\overline{V}(\kappa)}}\right)\right) \geq 1 - o(1). \tag{2.126}$$

We are going to show it leads to (2.113) by selecting an appropriate $\bar{\delta}$ as follows. We write $\eta = \eta_{\bar{\delta}} = \frac{1}{2} - \frac{\overline{V}(\kappa)}{(1+\bar{\delta})\Delta^2 npL} \log \frac{n-k}{k}$ to make the dependence on $\bar{\delta}$ explicit. Recall that δ_2 and C_1' are independent of the $\bar{\delta}$ in the definition of $\eta_{\bar{\delta}}$. First we can let $\bar{\delta} > \delta_1$, then we have

$$\begin{aligned}
\frac{(1+\delta_2)\eta_{\bar{\delta}}^2 \Delta^2 npL}{2\overline{V}(\kappa)} + C_1' \eta_{\bar{\delta}} \sqrt{\frac{\Delta^2 npL}{\overline{V}(\kappa)}} &\leq \left(1 + \delta_2 + 2C_1' \left(\eta_{\bar{\delta}} \frac{\Delta^2 npL}{\overline{V}(\kappa)}\right)^{-\frac{1}{2}}\right) \frac{\eta_{\bar{\delta}}^2 \Delta^2 npL}{2\overline{V}(\kappa)} \\
&\leq \left(1 + \delta_2 + 2C_1' \left(\eta_{\delta_1} \frac{\Delta^2 npL}{\overline{V}(\kappa)}\right)^{-\frac{1}{2}}\right) \frac{\eta_{\bar{\delta}}^2 \Delta^2 npL}{2\overline{V}(\kappa)} \\
&\leq (1+\delta_4) \frac{\eta_{\bar{\delta}}^2 \Delta^2 npL}{2\overline{V}(\kappa)},
\end{aligned}$$

for some $\delta_4 = o(1)$ not dependent on $\bar{\delta}$. Here the second inequality is due to the fact that η_δ is in increasing function of δ, and the last inequality is due to (2.111). Then we can let $\bar{\delta} \geq \delta_4$ to have the above expression to be upper bounded by $\left(1+\bar{\delta}\right) \frac{\eta_{\bar{\delta}}^2 \Delta^2 npL}{2\overline{V}(\kappa)}$. Hence, (2.230) leads to

$$\mathbb{P}_{(\theta^*, r^*)} \left(\sum_{k-\rho k < i \leq k} \mathbb{I}\left\{\widehat{\pi}_i < t\right\} \geq C_4 k \exp\left(-\frac{(1+\bar{\delta})\eta_{\bar{\delta}}^2 \Delta^2 npL}{2\overline{V}(\kappa)}\right)\right) \geq 1 - o(1), \qquad (2.127)$$

witch establishes (2.113).

Similar to (2.230), we can establish

$$\mathbb{P}_{(\theta^*, r^*)} \left(\sum_{k < i \leq k + \rho(n-k)} \mathbb{I}\left\{\widehat{\pi}_i > t\right\} \geq C_4 (n-k) e^{\left(-\frac{(1+\delta_2)(1-\eta_{\bar{\delta}})^2 \Delta^2 npL}{2\overline{V}(\kappa)} - C_1'(1-\eta_{\bar{\delta}})\sqrt{\frac{\Delta^2 npL}{\overline{V}(\kappa)}}\right)} \right)$$

$$\geq 1 - o(1).$$

Due to (2.111), we have $(1-\eta_{\bar{\delta}}) \in [0,1]$, then

$$\begin{aligned}\frac{(1+\delta_2)(1-\eta_{\bar{\delta}})^2\Delta^2 npL}{2\overline{V}(\kappa)} + C_1'(1-\eta_{\bar{\delta}})\sqrt{\frac{\Delta^2 npL}{\overline{V}(\kappa)}} &\leq \frac{(1+\delta_2)(1-\eta_{\bar{\delta}})^2\Delta^2 npL}{2\overline{V}(\kappa)} + C_1'\sqrt{\frac{\Delta^2 npL}{\overline{V}(\kappa)}} \\ &\leq (1+\delta_5)\,\frac{(1-\eta_{\bar{\delta}})^2\Delta^2 npL}{2\overline{V}(\kappa)},\end{aligned}$$

for some $\delta_5 = o(1)$ not dependent on $\bar{\delta}$. Since $(1-\eta_{\bar{\delta}})^2\Delta^2 npL/(2\overline{V}(\kappa)) = \eta_{\bar{\delta}}^2\Delta^2 npL/(2\overline{V}(\kappa)) + 2\log\frac{n-k}{k}/(1+\bar{\delta})$, we have

$$\begin{aligned}&(n-k)\exp\left(-\frac{(1+\delta_2)(1-\eta_{\bar{\delta}})^2\Delta^2 npL}{2\overline{V}(\kappa)} - C_1'\eta_{\bar{\delta}}\sqrt{\frac{\Delta^2 npL}{\overline{V}(\kappa)}}\right) \\ &\geq k\exp\left(\log\frac{n-k}{k} - (1+\delta_5)\,\frac{(1-\eta_{\bar{\delta}})^2\Delta^2 npL}{2\overline{V}(\kappa)}\right) \\ &= k\exp\left(\frac{\bar{\delta}-\delta_5}{1+\bar{\delta}}\log\frac{n-k}{k} - (1+\delta_5)\,\frac{\eta_{\bar{\delta}}^2\Delta^2 npL}{2\overline{V}(\kappa)}\right).\end{aligned}$$

By letting $\bar{\delta} \geq \delta_5$ and using the same argument as in obtaining (2.127), we have

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{k<i\leq k+\rho(n-k)} \mathbb{I}\left\{\widehat{\pi}_i > t\right\} \geq C_4 k\exp\left(-\frac{(1+\bar{\delta})\eta_{\bar{\delta}}^2\Delta^2 npL}{2\overline{V}(\kappa)}\right)\right) \geq 1-o(1), \quad (2.128)$$

which establishes (2.114). To sum up, we can choose $\bar{\delta} = \max\{\delta_1, \delta_4, \delta_5\}$ to establish (2.113) and (2.114).

The above proof assumes that $\kappa = \Omega(1)$ and $\overline{\mathsf{SNR}}$ satisfies (2.111). When these two conditions do not hold, we need to slightly modify the argument. When (2.111) is not satisfied, there must exist some small constant $\bar{\epsilon} > 0$ such that $\frac{\sqrt{(1+\bar{\epsilon})\overline{\mathsf{SNR}}}}{2} - \frac{1}{\sqrt{(1+\bar{\epsilon})\overline{\mathsf{SNR}}}}\log\frac{n-k}{k} = O(1)$. We can then take ρ to be a sufficiently small constant, and the proof will go through with some slight modification. When $\kappa = o(1)$, we can simply construct θ^* by $\theta_i^* = 0$ for $1 \leq i \leq k$ and $\theta_i^* = -\Delta$ for $k+1 \leq i \leq n$. □

Finally, we state and prove Lemma 2.9.4 to close this section.

Lemma 2.9.4. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, $\rho = o(1)$, $k \to \infty$ *and (2.19) holds for some arbitrarily small constant* $\epsilon > 0$. *Choose* $\kappa_1, \kappa_2 \geq 0$ *such that we have both* $\kappa_1 + \kappa_2 \leq \kappa$ *and*

$$\frac{k\psi'(\kappa_1)(1+e^{\kappa_1})^2 + (n-k)\psi'(\kappa_2)(1+e^{-\kappa_2})^2}{(k\psi(\kappa_1) + (n-k)\psi(-\kappa_2))^2/n} = \overline{V}(\kappa).$$

Define $\theta_i^* = \kappa_1$ *for all* $1 \leq i \leq k - \rho k$, $\theta_i^* = 0$ *for* $k - \rho k < i \leq k$, $\theta_i^* = -\Delta$ *for* $k+1 \leq i \leq k + \rho(n-k)$ *and* $\theta_i^* = -\kappa_2$ *for* $k + \rho(n-k) < i \leq n$ *and* $S = \{i \in [n] : i \leq k - \rho k \text{ or } i > k\}$. *There exists some constants* $C_1 > 0$ *such that for any* $\tilde{\delta} = o(1)$, *there exists* $C_2 > 0$ *and* $\delta_1 = o(1)$ *such that for any* $\eta < 1/2$ *and any* $A \in \mathcal{A}$ *where* $\mathcal{A}$ *is defined in (2.221)-(2.223), we have*

$$\begin{aligned} &\mathbb{P}\left(\frac{\sum_{j\in S} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^* - \theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \leq -(1+\tilde{\delta})\eta\Delta \middle| A \right) \\ &\geq C_1 \exp\left(-\frac{1+\delta_1}{2}\eta_+^2 \overline{SNR} - C_2\eta_+\sqrt{\overline{SNR}} \right). \end{aligned} \tag{2.129}$$

for any $k - \rho k < i \leq k$.

Proof. We suggest readers to go through the proof of Lemma 2.10.3 in Section 2.10.2 first. The proof of Lemma 2.9.4 basically follows that of Lemma 2.10.3. We will omit repeated details in the proof of Lemma 2.10.3 and only present key steps and calculations specific to this Lemma 2.9.4.

We denote $q_j = \psi(\theta_i - \theta_j)$. Then $1 + e^{\theta_j^* - \theta_i^*} = 1/q_j$ and $\psi(\theta_j - \theta_i) = 1 - q_j$. Then what we need to lower bound can be written as

$$\mathbb{P}_A\left(\sum_{\ell\in[L]} \sum_{j\in S} A_{ji} \frac{q_j - y_{ij\ell}}{q_j} \geq Lt' \right),$$

where $t' = (1+\delta')\eta\Delta \sum_{j\in[n]\setminus\{i\}} p(1-q_j)$ for some $\delta' = o(1)$ due to (2.221)-(2.223), and

$\mathbb{P}_A$ is the conditional probability given A. Note that δ' can be chosen independent of η. We remark that

$$\overline{\mathsf{SNR}} = (1+\delta'')\frac{L\Delta^2(\sum_{j\in[n]\setminus\{i\}} p(1-q_j))^2}{\sum_{j\in S} p\frac{1-q_j}{q_j}}$$

due to $\rho = o(1)$ for some $\delta'' = o(1)$ independent of η. We still first consider the regime when

$$\eta\sqrt{\overline{\mathsf{SNR}}} \to \infty, \tag{2.130}$$

This implies $\eta \in (0, 1/2)$.

The conditional cumulant of $\sum_{j\in S} A_{ji}\frac{q_j - y_{ijl}}{q_j}$ for each $l \in [L]$ is

$$\nu(u) = \sum_{j\in S} A_{ji}\log\left(q_j e^{\frac{u(q_j-1)}{q_j}} + (1-q_j)e^u\right) = \sum_{j\in S} A_{ji}\left[-u\frac{1-q_j}{q_j} + \log((1-q_j)e^{u/q_j} + q_j)\right].$$

The function $\nu(u)$ acts as the same role as $K(u)$ in the proof of Lemma 2.10.3. Define

$$u^* = \arg\min_{u\geq 0}\left(L\nu(u) - uLt'\right).$$

Its first derivative is

$$\nu'(u) = \sum_{j\in S} A_{ji}\left[\frac{\frac{(1-q_j)}{q_j}e^{u/q_j}}{(1-q_j)e^{u/q_j} + q_j} - \frac{1-q_j}{q_j}\right].$$

Following the same argument in the proof of Lemma 2.10.3, we need to pin down a range for u^*. First due to (2.130) and $\nu'(0) = 0$, we have $t' > 0$ and thus $\nu'(0) - t' < 0$. Now for $u = o(1)$, we can approximate $\nu'(u)$ by Taylor expansion and obtain

$$1 - \delta_2 \leq \frac{\nu'(u)}{\overline{\nu'}(u)} \leq 1 + \delta_2, \tag{2.131}$$

for some $0 < \delta_2 = o(1)$, where $\overline{\nu'}(u) = \sum_{j\in S} p\frac{1-q_i}{q_i}u$. Note that we can replace A_{ji} by p because of the condition $A \in \mathcal{A}$. Then we consider $\tilde{u} = \frac{2t'}{\sum_{j\in S} p\frac{1-q_i}{q_i}}$, which is $o(1)$ since $\Delta = o(1)$ and $\rho = o(1)$. Therefore,

$$\nu'(\tilde{u}) - t' \geq (1-\delta_2)\overline{\nu'}(\tilde{u}) - t' = (1-\delta_2)t' > 0.$$

This implies that $u^* \in \left(0, \frac{2t'}{\sum_{j\in S} p\frac{1-q_i}{q_i}}\right)$. Thus $u^* = o(1)$.

When $u = o(1)$, $\nu(u)$ also follows a second order Taylor expansion such that:

$$1 - \delta_3 \leq \frac{\nu(u)}{\bar{\nu}(u)} \leq 1 + \delta_3,$$

where $\bar{\nu}(u) = \frac{1}{2}\sum_{j\in S} p\frac{1-q_j}{q_j}u^2$ and $\delta_3 = o(1)$ due to (2.221)-(2.223).

Following the change-of-measure argument in the proof of Lemma 2.10.3, the probability of interest can be lower bounded by

$$\exp\left(-u^*T + L\nu(u^*) - Lu^*t'\right)\mathbb{Q}_A\left(0 \leq \sum_{l=1}^{L}\sum_{j\in S} Z_{jl} - Lt' \leq T\right),$$

where $\mathbb{Q}_A$ is a measure under which Z_{jl} are all independent given A and follow

$$\mathbb{Q}_A(Z_{jl} = s) = e^{A_{ji}u^*s - A_{ji}\nu_j(u^*)}\mathbb{P}_A\left(A_{ji}\frac{q_j - y_{ijl}}{q_j} = s\right)$$

and $\nu_j(u) = -u\frac{1-q_j}{q_j} + \log((1-q_j)e^{u/q_j} + q_j)$. Then for each Z_{jl} such that $A_{ij} = 1$, its second and 4th moment under $\mathbb{Q}_A$ can be analyzed:

$$\mathbb{Q}_A((Z_{jl} - \mathbb{Q}_A(Z_{jl}))^2) = \nu_j''(u^*) = \frac{1-q_j}{q_j}\frac{e^{u^*/q_j}}{[(1-q_j)e^{u^*/q_j} + q_j]^2} \in (C_1', C_2'), \qquad (2.132)$$

$$\mathbb{Q}_A((Z_{jl} - \mathbb{Q}_A(Z_{jl}))^4) = \nu_j''''(u^*) + 3\nu_j''(u^*) \leq (3 + C_4')\nu_j''(u^*) \leq C_3', \tag{2.133}$$

where (2.133) comes from

$$\nu_j''''(u) = \frac{1-q_j}{q_j^3} e^{u/q_j} \frac{(1-q_j)^3 e^{3u/q_j} - 3(1-q_j)^2 q_j e^{2u/q_j} - 3(1-q_j) q_j^2 e^{2u/q_j} + q_j^3}{[(1-q_j)e^{u/q_j} + q_j]^5}$$
$$\leq \max_{j \in S} 1/q_j^2 \nu''(u) \leq C_4' \nu''(u).$$

Now, to lower bound $L\nu(u^*) - Lu^*t'$:

$$\begin{aligned} L\nu(u^*) - Lu^*t' &\geq L(1-\delta_3)\frac{1}{2}\sum_{j \in S} p\frac{1-q_j}{q_j} u^{*2} - Lu^*t' \\ &\geq L \min_{u \in (0,1)} \left[(1-\delta_3)\frac{1}{2}\sum_{j \in S} p\frac{1-q_j}{q_j} u^{*2} - u^*t' \right] \\ &\geq -\frac{1}{2}\frac{Lt'^2}{(1-\delta_3)\sum_{j \in S} p\frac{1-q_j}{q_j}} \\ &\geq -\frac{1+\delta_4}{2}\eta^2 \overline{\mathsf{SNR}}, \end{aligned} \tag{2.134}$$

where (2.134) is achieved at $u = \frac{t'}{(1-\delta_3)\sum_{j \in S} p\frac{1-q_j}{q_j}}$ and $\delta_4 = o(1)$ since $\rho = o(1)$. This gives us the desired exponent. We remark that δ_4 is independent of η.

To choose T, observe that

$$\mathsf{Var}_{\mathbb{Q}_A}\left(\sum_{l \in [L]} \sum_{j \in S} Z_{jl}\right) \leq \tilde{C}_1 npL,$$

for some constant $\tilde{C}_1 > 0$ using (2.221) - (2.223) , (2.132) and $\rho = o(1)$. Thus we choose $T = \sqrt{\tilde{C}_1 npL}$, which leads to a term $C_2\eta\sqrt{\overline{\mathsf{SNR}}}$ in the exponent for some $C_2 > 0$ independent of η.

Finally, to lower bound the $\mathbb{Q}_A$ measure, we only need to verify the vanishing property

of the 4th moment approximation bound in Lemma 3.7.4:

$$\sqrt{L\sum_{j\in S}A_{ji}\left(\frac{\mathbb{Q}_A((Z_{j1}-\mathbb{Q}_A(Z_{j1})^4)}{(L\sum_{j\in S}A_{ji}\mathbb{Q}_A((Z_{j1}-\mathbb{Q}_A(Z_{j1})^2))^2}\right)^{3/4}}$$
$$\leq \tilde{C}_2(npL)^{-1/4} \tag{2.135}$$

where (2.135) is by (2.132), (2.133) and $\rho = o(1)$. To summarize, we have proved

$$\mathbb{P}_A\left(\sum_{l\in[L]}\sum_{j\in S}A_{ji}\frac{q_j-y_{ijl}}{q_j}\geq Lt'\right)\geq C_1\exp\left(-\frac{1+\delta_5}{2}\eta^2\overline{\mathsf{SNR}}-C_2\eta\sqrt{\overline{\mathsf{SNR}}}\right)$$

for some constant $C_1, C_2 > 0$ and $\delta_5 = o(1)$, when (2.130) holds. This δ_5 can be used as the δ_1 in (2.129). We remark that C_1, C_2, δ_5 are all independent of η.

Finally, when

$$\eta\sqrt{\overline{\mathsf{SNR}}}\leq C_3$$

for some constant $C_3 > 0$. This condition, together with (2.221)-(2.223) and $\rho = o(1)$, implies that

$$Lt'\leq C_4\sqrt{L\sum_{j\in S}A_{ji}\frac{1-q_i}{q_i}}.$$

Therefore,

$$\mathbb{P}_A\left(\sum_{l\in[L]}\sum_{j\in S}A_{ji}\frac{q_j-y_{ijl}}{q_j}\geq Lt'\right)\geq\mathbb{P}_A\left(\sum_{l\in[L]}\sum_{j\in S}A_{ji}\frac{q_j-y_{ijl}}{q_j}\geq C_5\sqrt{L\sum_{j\in S}A_{ji}\frac{1-q_i}{q_i}}\right)$$
$$\geq c_1-o(1) \tag{2.136}$$

where (2.136) comes from Lemma 3.7.4. The 4th moment approximation can be checked to be of order $(npL)^{-1/4}$ similarly as in (2.135) using (2.221)-(2.223) and $\rho = o(1)$ since the

second and fourth moment of $\frac{q_j - y_{ijl}}{q_j}$ are at the constant order under measure $\mathbb{P}_A$, which completes the proof.

□

2.10 Proofs of Lower Bounds

This section collects the proofs of lower bound results of the paper. The lower bound for exact recovery is proved in Section 2.10.1, and the partial recovery lower bound is proved in Section 2.10.2.

2.10.1 Proof of Theorem 2.3.4

The key mathematical argument in the proof of Theorem 2.3.4 is to characterize the maximum of dependent binomial random variables. For this purpose, we need a high-dimensional central limit theorem result by [27]. The following lemma is adapted from [27] for our purpose.

Lemma 2.10.1. *Consider independent random vectors $X_1, \cdots, X_n \in \mathbb{R}^d$ with mean zero. Assume there exist constants $c_1, c_2, C_1, C_2 > 0$ such that*

$$\begin{aligned}
&\min_{i,j} \mathbb{E} X_{ij}^2 \geq c_1, \\
&\max_{i,j} \mathbb{E} \exp(|X_{ij}|/C_1) \leq 2, \\
&(\log(nd))^7 \leq C_2 n^{-(1+c_2)}.
\end{aligned}$$

Then, there exist independent Gaussian vectors $Z_1, \cdots, Z_n$ satisfying $\mathbb{E} Z_i = 0$ and $\mathsf{Cov}(Z_i) = \mathsf{Cov}(X_i)$, such that

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\max_{j \in [d]} \sum_{i=1}^n X_{ij} \leq t \right) - \mathbb{P}\left(\max_{j \in [d]} \sum_{i=1}^n Z_{ij} \leq t \right) \right| \leq C n^{-c},$$

for some constants $c, C > 0$ only depending on c_1, c_2, C_1, C_2.

With the above Gaussian approximation, we only need to analyze the maximum of dependent Gaussian random variables. The following lemma can be found in [57].

Lemma 2.10.2. *Consider $Z = (Z_1, \cdots, Z_n)^T \sim N(0, \Sigma)$. Then, for any $\alpha \in (0,1)$, there exists some constant $C_\alpha > 0$ such that for all $n \geq \sqrt{2\pi}e^3 \log 1/\alpha$,*

$$\mathbb{P}\left(\max_{i\in[n]} Z_i > \lambda^{1/2}\sqrt{2\log n - \log\log n - C_\alpha} - \Lambda^{1/2}\Phi^{-1}(1-\alpha)\right) \geq 1 - 2\alpha,$$

where $\lambda = \min_{i\in[n]} \Sigma_{ii} - \frac{\max_{i\in[n]} \sum_{j\in[n]\setminus\{i\}} \Sigma_{ij}^2}{\lambda_{\min}(\Sigma)}$ and $\Lambda = \max_{i\in[n]} \Sigma_{ii}$.

Now we are ready to prove Theorem 2.3.4.

Proof of Theorem 2.3.4. We first note that the condition (2.15) implies that $\Delta = o(1)$. Choose $\kappa_1, \kappa_2 \geq 0$ such that we have both $\kappa_1 + \kappa_2 \leq \kappa$ and

$$\frac{n}{k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2)} = V(\kappa).$$

We first consider the case $k \to \infty$ and $\kappa = \Omega(1)$. In this case, one can easily check that $\kappa_2 = \Omega(1)$. Our least favorable $\theta^* \in \Theta(k, \Delta, \kappa)$ is constructed as follows. Let $\rho = o(1)$ be a vanishing number that will be specified later. Define $\theta_i^* = \kappa_1$ for all $1 \leq i \leq k - \rho k$, $\theta_i^* = 0$ for $k - \rho k < i \leq k$, $\theta_i^* = -\Delta$ for $k < i \leq k + \rho(n-k)$ and $\theta_i^* = -\kappa_2$ for $k + \rho(n-k) < i \leq n$. For the simplicity of proof, we choose ρ so that both ρk and $\rho(n-k)$ are integers. Consider a subset $\mathcal{R}_{k,\rho} \subset \mathfrak{S}_n$ that is defined by

$$\mathcal{R}_{k,\rho} = \{r \in \mathfrak{S}_n : r_i = i \text{ for all } i \leq k - \rho k \text{ or } i > k + \rho(n-k)\}. \tag{2.137}$$

We then have the lower bound

$$\inf_{\widehat{r}} \sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta(k,\Delta,\kappa)}} \mathbb{P}_{(\theta^*, r^*)} \left(\mathsf{H}_k(\widehat{r}, r^*) > 0 \right) \geq \inf_{\widehat{r}} \sup_{r^* \in \mathcal{R}_{k,\rho}} \mathbb{P}_{(\theta^*, r^*)} \left(\mathsf{H}_k(\widehat{r}, r^*) > 0 \right).$$

For each $z = \{z_i\}_{k-\rho k < i \leq k+\rho(n-k)} \in \{0,1\}^{\rho n}$, we define $\mathbb{Q}_z$ as a joint probability of the observations $\{A_{ij}\}$ and $\{y_{ijl}\}$. To sample data from $\mathbb{Q}_z$, we first sample $A \sim \mathcal{G}(n,p)$, and then for any (i,j) such that $A_{ij} = 1$, sample $y_{ijl} \sim \text{Bernoulli}(\psi(\mu_i(z) - \mu_j(z)))$ independently for $l \in [L]$. The vector $\mu(z)$ is defined by $\mu_i(z) = \theta_i^*$ for all $i \leq k - \rho k$ or $i > \rho(n-k)$ and $\mu_i(z) = \Delta \mathbb{I}\{z_i = 1\}$ for all $k - \rho k < i \leq k + \rho(n-k)$. Then, we have

$$\begin{aligned} \inf_{\widehat{r}} \sup_{r^* \in \mathcal{R}_{k,\rho}} \mathbb{P}_{(\theta^*, r^*)} \left(\mathsf{H}_k(\widehat{r}, r^*) > 0 \right) &\geq \inf_{\widehat{z}} \sup_{z^* \in \mathcal{Z}_k} \mathbb{Q}_{z^*} \left(\widehat{z} \neq z^* \right) \\ &\geq \inf_{\widehat{z}} \frac{1}{|\mathcal{Z}_k|} \sum_{z^* \in \mathcal{Z}_k} \mathbb{Q}_{z^*}(\widehat{z} \neq z^*), \end{aligned}$$

where

$$\mathcal{Z}_k = \left\{ z = \{z_i\}_{k-\rho k < i \leq k+\rho(n-k)} \in \{0,1\}^{\rho n} : \sum_i z_i = \rho k \right\}.$$

The Bayes risk $\frac{1}{|\mathcal{Z}_k|} \sum_{z^* \in \mathcal{Z}_k} \mathbb{Q}_{z^*}(\widehat{z} \neq z^*)$ is minimized by

$$\widehat{z} = \operatorname*{argmin}_{z \in \mathcal{Z}_k} \ell_n(\mu(z)), \tag{2.138}$$

where

$$\ell_n(\mu(z)) = \sum_{1 \leq i < j \leq n} A_{ij} \left[\bar{y}_{ij} \log \frac{1}{\psi(\mu_i(z) - \mu_j(z))} + (1 - \bar{y}_{ij}) \log \frac{1}{1 - \psi(\mu_i(z) - \mu_j(z))} \right].$$

It suffices to lower bound the probability $\mathbb{Q}_{z^*}(\widehat{z} \neq z^*)$ for the estimator (2.138) and for each $z^* \in \mathcal{Z}_k$. By symmetry, the value of $\mathbb{Q}_{z^*}(\widehat{z} \neq z^*)$ is the same for any $z^* \in \mathcal{Z}_k$. We therefore

can set $z_i^* = \mathbb{I}\{i \leq k\}$ without loss of generality. Define

$$\mathcal{N}(z^*) = \left\{ z \in \mathcal{Z}_k : \sum_i \mathbb{I}\{z_i \neq z_i^*\} = 2 \right\}.$$

Then, we have

$$\mathbb{Q}_{z^*}(\widehat{z} \neq z^*) \geq \mathbb{Q}_{z^*}\left(\min_{z \in \mathcal{N}(z^*)} \ell_n(\mu(z)) < \ell_n(\mu(z^*)) \right).$$

By direct calculation, we have

$$\begin{aligned} & \ell_n(\mu(z)) - \ell_n(\mu(z^*)) \\ = & \sum_{1 \leq i < j \leq n} A_{ij}(\bar{y}_{ij} - \psi(\mu_i(z^*) - \mu_j(z^*)))(\mu_i(z^*) - \mu_j(z^*) - \mu_i(z) + \mu_j(z)) \\ & + \sum_{1 \leq i < j \leq n} A_{ij} D\left(\psi(\mu_i(z^*) - \mu_j(z^*)) \| \psi(\mu_i(z) - \mu_j(z))\right). \end{aligned}$$

For any $z \in \mathcal{N}(z^*)$, there exists some $k - \rho k < a \leq k$ and some $k < b \leq k + \rho(n-k)$ such

that $z_a = 0$, $z_b = 1$ and $z_i = z_i^*$ for all other i's. Then,

$$\begin{aligned}
&\sum_{1\le i<j\le n} A_{ij} D\left(\psi(\mu_i(z^*) - \mu_j(z^*)) \| \psi(\mu_i(z) - \mu_j(z))\right) \\
\le\ & \sum_{i=1}^{k-\rho k} A_{ia} D(\psi(\kappa_1) \| \psi(\kappa_1 + \Delta)) + \sum_{i=k+\rho(n-k)+1}^{n} A_{ia} D(\psi(-\kappa_2) \| \psi(-\kappa_2 + \Delta)) \\
&+ \sum_{i=1}^{k-\rho k} A_{ib} D(\psi(\kappa_1 + \Delta) \| \psi(\kappa_1)) + \sum_{i=k+\rho(n-k)+1}^{n} A_{ib} D(\psi(-\kappa_2 + \Delta) \| \psi(-\kappa_2)) \\
&+ \sum_{i=k-\rho k+1}^{k} A_{ia} D(\psi(0) \| \psi(\Delta)) + \sum_{i=k+1}^{k+\rho(n-k)} A_{ia} D(\psi(-\Delta) \| \psi(0)) \\
&+ \sum_{i=k-\rho k+1}^{k} A_{ib} D(\psi(\Delta) \| \psi(0)) + \sum_{i=k+1}^{k+\rho(n-k)} A_{ib} D(\psi(0) \| \psi(-\Delta)) + A_{ab} D(\psi(\Delta) \| \psi(-\Delta)) \\
\le\ & (1+\delta)(1-\rho)p\left[kD(\psi(\kappa_1) \| \psi(\kappa_1 + \Delta)) + (n-k)D(\psi(-\kappa_2) \| \psi(-\kappa_2 + \Delta))\right] \qquad (2.139) \\
&+(1+\delta)(1-\rho)p\left[kD(\psi(\kappa_1 + \Delta) \| \psi(\kappa_1)) + (n-k)D(\psi(-\kappa_2 + \Delta) \| \psi(-\kappa_2))\right] \\
&+(1+\delta)\rho p\left[kD(\psi(0) \| \psi(\Delta)) + (n-k)D(\psi(-\Delta) \| \psi(0))\right] \\
&+(1+\delta)\rho p\left[kD(\psi(\Delta) \| \psi(0)) + (n-k)D(\psi(0) \| \psi(-\Delta))\right] + (1+\delta)pD(\psi(\Delta) \| \psi(-\Delta)) \\
\le\ & (1+\delta)^2(1-\rho)p\Delta^2\left[k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2)\right] + (1+\delta)^2\rho p\Delta^2\frac{n}{4} \qquad (2.140) \\
\le\ & (1+\delta)^3 p\Delta^2\frac{n}{V(\kappa)}. \qquad (2.141)
\end{aligned}$$

The inequality (2.139) holds with probability at least $1 - O(n^{-10})$ by Bernstein's inequality. The inequality (2.140) is a Taylor expansion argument with the help of $\Delta = o(1)$. We obtain (2.141) by the choice that $\rho = o(1)$. Note that we can choose some $\delta = o(1)$ to make all of (2.139), (2.140) and (2.141) hold. We also have

$$\begin{aligned}
&\sum_{1\le i<j\le n} A_{ij}(\bar{y}_{ij} - \psi(\mu_i(z^*) - \mu_j(z^*)))(\mu_i(z^*) - \mu_j(z^*) - \mu_i(z) + \mu_j(z)) \\
=\ & -\Delta \sum_{i\in[n]\setminus\{a\}} A_{ia}(\bar{y}_{ia} - \mathbb{E}\bar{y}_{ia}) + \Delta \sum_{i\in[n]\setminus\{b\}} A_{ib}(\bar{y}_{ib} - \mathbb{E}y_{ib}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\min_{z\in\mathcal{N}(z^*)} \ell_n(\mu(z)) - \ell_n(\mu(z^*)) \\
\leq\ & -\max_{(1-\rho)k<a\leq k} \Delta \sum_{i\in[n]\setminus\{a\}} A_{ia}(\bar{y}_{ia} - \mathbb{E}\bar{y}_{ia}) + \Delta \min_{k<b\leq k+\rho(n-k)} \sum_{i\in[n]\setminus\{b\}} A_{ib}(\bar{y}_{ib} - \mathbb{E}y_{ib}) \\
& +(1+\delta)^3 p\Delta^2 \frac{n}{V(\kappa)},
\end{aligned}$$

with probability at least $1 - O(n^{-10})$. This leads to the bound

$$\begin{aligned}
& \mathbb{Q}_{z^*}\left(\min_{z\in\mathcal{N}(z^*)} \ell_n(\mu(z)) < \ell_n(\mu(z^*))\right) \\
\geq\ & \mathbb{Q}_{z^*}\Bigg(\max_{(1-\rho)k<a\leq k} \sum_{i\in[n]\setminus\{a\}} A_{ia}(\bar{y}_{ia} - \mathbb{E}\bar{y}_{ia}) \\
& - \min_{k<b\leq k+\rho(n-k)} \sum_{i\in[n]\setminus\{b\}} A_{ib}(\bar{y}_{ib} - \mathbb{E}y_{ib}) > (1+\delta)^3 p\Delta \frac{n}{V(\kappa)}\Bigg) - O(n^{-10}) \\
\geq\ & \mathbb{Q}_{z^*}\Bigg(\max_{(1-\rho)k<a\leq k} \sum_{i\in[n]\setminus\{a\}} A_{ia}(\bar{y}_{ia} - \mathbb{E}\bar{y}_{ia}) - \min_{k<b\leq k+\rho(n-k)} \sum_{i\in[n]\setminus\{b\}} A_{ib}(\bar{y}_{ib} - \mathbb{E}y_{ib}) \qquad (2.142) \\
& > \sqrt{2(1-\epsilon/2)}\sqrt{\frac{np}{LV(\kappa)}}\left(\sqrt{\log k} + \sqrt{\log(n-k)}\right)\Bigg) - O(n^{-10}) \\
\geq\ & \mathbb{Q}_{z^*}\left(\max_{(1-\rho)k<a\leq k} \sum_{i\in[n]\setminus\{a\}} A_{ia}(\bar{y}_{ia} - \mathbb{E}\bar{y}_{ia}) > \sqrt{2(1-\epsilon/2)}\sqrt{\frac{np}{LV(\kappa)}}\sqrt{\log k}\right) \qquad (2.143) \\
& +\mathbb{Q}_{z^*}\left(-\min_{k<b\leq k+\rho(n-k)} \sum_{i\in[n]\setminus\{b\}} A_{ib}(\bar{y}_{ib} - \mathbb{E}y_{ib}) > \sqrt{2(1-\epsilon/2)}\sqrt{\frac{np}{LV(\kappa)}}\sqrt{\log(n-k)}\right) \\
& -1 - O(n^{-10}),
\end{aligned}$$

where we have used the condition of the theorem to derive (2.142). The last inequality (2.143) is by union bound $\mathbb{P}(A\cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$. To lower bound (2.143), we introduce the

notation

$$T_a = \sum_{i\in[n]\setminus\{a\}} A_{ia}(\bar{y}_{ia} - \mathbb{E}\bar{y}_{ia}), \quad (1-\rho)k < a \le k.$$

The covariance structure of $\{T_a\}_{(1-\rho)k<a\le k}$ can be quantified by the matrix $\Sigma \in \mathbb{R}^{(\rho k)\times(\rho k)}$, which is defined by $\Sigma_{ab} = \mathsf{Cov}(T_a, T_b|A)$. We then construct a vector $S = \{S_a\}_{(1-\rho)k<a\le k}$ that is jointly Gaussian conditioning on A. The conditional covariance of S is also Σ. By Lemma 2.10.1, we have

$$\mathbb{Q}_{z^*}\left(\max_{(1-\rho)k<a\le k}\sum_{i\in[n]\setminus\{a\}} A_{ia}(\bar{y}_{ia} - \mathbb{E}\bar{y}_{ia}) > \sqrt{2(1-\epsilon/2)}\sqrt{\frac{np}{LV(\kappa)}}\sqrt{\log k}\right) \quad (2.144)$$

$$\ge \mathbb{P}\left(\max_{(1-\rho)k<a\le k} S_a > \sqrt{2(1-\epsilon/2)}\sqrt{\frac{np}{LV(\kappa)}}\sqrt{\log k}\right) - O\left(\frac{1}{(\log n)^c}\right). \quad (2.145)$$

To see how Lemma 2.10.1 implies (2.145), we can take $X_{la} = \frac{1}{\sqrt{np}}\sum_{i\in[n]\setminus\{a\}} A_{ia}(y_{ial} - \mathbb{E}y_{ial})$. Conditioning on A, we observe that $\{X_{la}\}$ is independent across $l \in [L]$. The conditional variance of X_{la} given A is bounded away from zero with high probability by Lemma 2.8.1. Moreover, one can find a constant $C > 0$, such that $\mathbb{E}\left[\exp(|X_{la}|/C)\middle|A\right] \le 2$ by Hoeffding's inequality. Then, we can apply Lemma 2.10.1 for a given A and obtain (2.145) under the condition $L > (\log n)^8$. We need Lemma 2.10.2 to lower bound the probability in (2.145). For each a,

$$\begin{aligned}
\Sigma_{aa} &= \mathsf{Var}(T_a|A) \\
&= \frac{1}{L}\sum_{i\in[n]\setminus\{a\}} A_{ia}\psi'(\mu_i(z^*) - \mu_a(z^*)) \\
&= \frac{\psi'(\kappa_1)}{L}\sum_{i=1}^{k-\rho k} A_{ia} + \frac{1}{4L}\sum_{i=k-\rho k+1}^{k} A_{ia} + \frac{\psi'(\kappa_2)}{L}\sum_{i=k+1}^{k+\rho(n-k)} A_{ia} + \frac{\psi'(\Delta)}{L}\sum_{i=k+\rho(n-k)+1}^{n} A_{ia}.
\end{aligned}$$

By Lemma 2.8.1, we have

$$\max_{(1-\rho)k<a\le k} \Sigma_{aa} \le \frac{1}{4L} \sum_{i\in[n]\setminus\{a\}} A_{ia} \le \frac{np}{2L}, \tag{2.146}$$

with probability at least $1 - O(n^{-10})$. Similar to the proof of Lemma 2.8.1, we can use Bernstein's inequality and a union bound argument to obtain that

$$\begin{aligned} \min_{(1-\rho)k<a\le k} \Sigma_{aa} &\ge \min_{(1-\rho)k<a\le k} \left[\frac{\psi'(\kappa_1)}{L} \sum_{i=1}^{k-\rho k} A_{ia} + \frac{\psi'(\kappa_2)}{L} \sum_{i=k+1}^{k+\rho(n-k)} A_{ia} \right] \\ &\ge \frac{(1-\delta)(1-\rho)p}{L} \big(k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2) \big) \\ &= \frac{(1-\delta)(1-\rho)pn}{LV(\kappa)}, \end{aligned} \tag{2.147}$$

for some $\delta = o(1)$ with probability at least $1 - O(n^{-10})$. For each $a \neq b$,

$$\Sigma_{ab} = \mathsf{Cov}(T_a, T_b | A) = A_{ab} \frac{\psi'(\mu_a(z^*) - \mu_b(z^*))}{L}.$$

Then, Bernstein's inequality and a union bound argument, we have

$$\max_a \sum_{b:b\neq a} \Sigma_{ab}^2 \le \frac{1}{16L^2} \max_{(1-\rho)k<a\le k} \sum_{b:b\neq a} A_{ab} \le C_1 \frac{\rho kp + \log n}{L^2}, \tag{2.148}$$

with probability at least $1-O(n^{-10})$. We can also obtain a similar bound for $\max_a \sum_{b:b\neq a} \Sigma_{ab}$. This allows us to give a lower bound on $\lambda_{\min}(\Sigma)$:

$$\lambda_{\min}(\Sigma) \ge \min_{(1-\rho)k<a\le k} \Sigma_{aa} - \max_a \sum_{b:b\neq a} \Sigma_{ab} \ge \frac{(1-\delta)(1-\rho)pn}{LV(\kappa)} - C_2 \frac{\rho kp + \log n}{L} \ge c_1 \frac{pn}{L}. \tag{2.149}$$

To apply Lemma 2.10.2, we shall choose ρ that satisfies both $\log(\rho k) = (1 + o(1)) \log k$ and $\rho = o(1)$. The existence of such ρ is guaranteed by $k \to \infty$. With the bounds (2.146)-(2.149),

we can apply Lemma 2.10.2, and obtain

$$\mathbb{P}\left(\max_{(1-\rho)k<a\leq k} S_a > \sqrt{2(1-\epsilon/2)}\sqrt{\frac{np}{LV(\kappa)}}\sqrt{\log k}\right) \geq 0.98 - O(n^{-1}).$$

We then obtain the desired lower bound for (2.144). A similar argument also leads to

$$\begin{aligned}
&\mathbb{Q}_{z^*}\left(-\min_{k<b\leq k+\rho(n-k)} \sum_{i\in[n]\setminus\{b\}} A_{ib}(\bar{y}_{ib} - \mathbb{E}y_{ib}) > \sqrt{2(1-\epsilon/2)}\sqrt{\frac{np}{LV(\kappa)}}\sqrt{\log(n-k)}\right) \\
\geq\ & 0.99 - O\left(\frac{1}{(\log n)^c}\right).
\end{aligned}$$

Therefore, $\mathbb{Q}_{z^*}(\widehat{z} \neq z^*) \geq 0.95$ and we obtain the desired conclusion.

The above proof assumes that $k \to \infty$ and $\kappa = \Omega(1)$. When these two conditions do not hold, we need to slightly modify the argument. Let us briefly discuss two cases. In the first case, $k = O(1)$ and $\kappa = \Omega(1)$. In this case, we can construct θ^* by $\theta_i^* = 0$ for $1 \leq i \leq k$, $\theta_i^* = -\Delta$ for $k < i \leq k + \rho(n-k)$ and $\theta_i^* = -\kappa$ for $k + \rho(n-k) < i \leq n$. In the second case, $\kappa = o(1)$, and then we can take θ^* with $\theta_i^* = 0$ for $1 \leq i \leq k$ and $\theta_i^* = -\Delta$ for $k < i \leq n$. The remaining part of the proof will go through with similar arguments, and we will omit the details. □

2.10.2 Proof of Theorem 2.6.1

We first establish a lemma that lower bounds the error of a critical testing problem.

Lemma 2.10.3. *Assume $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$, $\rho = o(1)$, $k \to \infty$ and (2.15) holds for some arbitrarily small constant $\epsilon > 0$. Choose $\kappa_1, \kappa_2 \geq 0$ such that we have both $\kappa_1 + \kappa_2 \leq \kappa$ and*

$$\frac{n}{k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2)} = V(\kappa).$$

Define $\theta_i = \kappa_1$ for $1 \leq i \leq k - \rho k$, $\theta_i = 0$ for $k - \rho k < i \leq k$, $\theta_i = -\Delta$ for $k + 2 \leq$

$i \le k + \rho(n-k)$ and $\theta_i = -\kappa_2$ for $k + \rho(n-k) < i \le n$. Suppose we have independent $A_i \sim \text{Bernoulli}(p)$ and $z_{il} \sim \text{Bernoulli}(\psi(\theta_i))$ for all $i \in [n]\backslash\{k+1\}$ and $l \in [L]$. Then, there exists some $\delta = o(1)$ such that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{l=1}^{L} \sum_{i\in[n]\backslash\{k+1\}} A_i \left[z_{il} \log \frac{\psi(\theta_i+\Delta)}{\psi(\theta_i)} + (1-z_{il}) \log \frac{1-\psi(\theta_i+\Delta)}{1-\psi(\theta_i)}\right] \ge \log \frac{k}{n-k-1}\right) \\
\ge\ & C \exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1+\delta)\mathit{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta)\mathit{SNR}}} \log \frac{n-k}{k}\right)_+^2\right),
\end{aligned}$$

for some constant $C > 0$.

Proof. We first consider the case

$$\frac{\sqrt{(1+\delta)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta)\mathsf{SNR}}} \log \frac{n-k}{k} \to \infty, \tag{2.150}$$

for some $\delta = o(1)$ to be specified later. Throughout the proof, we use $\mathbb{P}_A$ for the conditional distribution $\mathbb{P}(\cdot|A)$. We use the notation

$$Z_l = \sum_{i\in[n]\backslash\{k+1\}} A_i \left[z_{il} \log \frac{\psi(\theta_i+\Delta)}{\psi(\theta_i)} + (1-z_{il}) \log \frac{1-\psi(\theta_i+\Delta)}{1-\psi(\theta_i)}\right].$$

Its conditional cumulant generating function is

$$K(u) = \sum_{i\in[n]\backslash\{k+1\}} A_i \log\left(\psi(\theta_i)^{1-u}\psi(\theta_i+\Delta)^u + (1-\psi(\theta_i))^{1-u}(1-\psi(\theta_i+\Delta))^u\right).$$

Define

$$u^* = \operatorname*{argmin}_{u\ge 0}\left(LK(u) - u \log \frac{k}{n-k-1}\right).$$

By direct calculation, we have

$$\begin{aligned} K'(0) &= -\sum_{i\in[n]\setminus\{k+1\}} A_i D(\psi(\theta_i)\|\psi(\theta_i+\Delta)). \\ K'(1) &= \sum_{i\in[n]\setminus\{k+1\}} A_i D(\psi(\theta_i+\Delta)\|\psi(\theta_i)). \end{aligned}$$

By Bernstein's inequality,

$$K'(0) \leq -(1-\delta_1)p \sum_{i\in[n]\setminus\{k+1\}} D(\psi(\theta_i)\|\psi(\theta_i+\Delta)), \tag{2.151}$$

$$K'(1) \geq (1-\delta_1)p \sum_{i\in[n]\setminus\{k+1\}} D(\psi(\theta_i+\Delta)\|\psi(\theta_i)), \tag{2.152}$$

with some $\delta_1 = o(1)$ for probability at least $1 - O(n^{-1})$. Given that $\Delta = o(1)$, which is implied by (2.15), and $\rho = o(1)$, we have $\sum_{i\in[n]\setminus\{k+1\}} D(\psi(\theta_i)\|\psi(\theta_i+\Delta)) = (1+o(1))\frac{n\Delta^2}{2V(\kappa)}$ and $\sum_{i\in[n]\setminus\{k+1\}} D(\psi(\theta_i+\Delta)\|\psi(\theta_i)) = (1+o(1))\frac{n\Delta^2}{2V(\kappa)}$. With the condition (2.150), we know that $LK'(0) - \log\frac{k}{n-k-1} < 0$ and $LK'(1) - \log\frac{k}{n-k-1} > 0$. Thus, we must have $u^* \in (0,1)$. In fact, the range of u^* can be further narrowed down. We apply a Taylor expansion of $K'(u)$ as a function of Δ near 0, and we obtain

$$K'(u) = \sum_{i\in[n]\setminus\{k+1\}} A_i \left[-\frac{1}{2}\psi'(\theta_i)\Delta^2 + \psi'(\theta_i)u\Delta^2 + O(|\Delta|^3)\right].$$

Note that the remainder term $O(|\Delta|^3)$ can be bounded by $|\Delta|^3$ up to some constant uniformly for all $u \in (0,1)$. By Bernstein's inequality, we have

$$K'(u) \geq -(1+\delta_1)\left(\frac{1}{2}-u\right)\frac{np\Delta^2}{V(\kappa)}, \tag{2.153}$$

for all $u \in (0, 1/2)$ with probability at least $1 - O(n^{-1})$. By (2.153), there exists $\delta' = o(1)$

such that

$$K'\left(\frac{1}{2}-\frac{1}{(1+\delta')\mathsf{SNR}}\log\frac{n-k}{k}\right)>0,$$

and therefore, we must have

$$u^*\in\left(0,\frac{1}{2}-\frac{1}{(1+\delta')\mathsf{SNR}}\log\frac{n-k}{k}\right). \tag{2.154}$$

We also introduce a quadratic approximation for $K(u)$, which is

$$\overline{K}(u)=\frac{np\Delta^2}{2V(\kappa)}(u^2-u).$$

It can be shown that

$$1-\delta_2\le\frac{K(u)}{\overline{K}(u)}\le 1+\delta_2, \tag{2.155}$$

uniformly over all $u\in(0,1)$ for some $\delta_2=o(1)$ with probability at least $1-O(n^{-1})$. The inequality (2.155) can be obtained by a Taylor expansion argument followed by Bernstein's inequality, similar to the approximation obtained in (2.153).

Define a probability distribution $\mathbb{Q}_A$, under which $Z_1,\cdots,Z_L$ are i.i.d. given A and follow

$$\mathbb{Q}_A(Z_l=s)=\mathbb{P}_A(Z_l=s)e^{u^*s-K(u^*)},$$

for any s. It fact, each Z_l, under the measure $\mathbb{Q}_A$ can be written as the sum of several independent random variables, i.e. $Z_l=\sum_{i\in[n]\setminus\{k+1\}}Z_{il}$ where

$$\mathbb{Q}_A(Z_{il}=s)=e^{A_iu^*s-A_iK_i(u^*)}\mathbb{P}_A\left(A_i\left[z_{il}\log\frac{\psi(\theta_i+\Delta)}{\psi(\theta_i)}+(1-z_{il})\log\frac{1-\psi(\theta_i+\Delta)}{1-\psi(\theta_i)}\right]=s\right),$$

and $K_i(u)=\log\left(\psi(\theta_i)^{1-u}\psi(\theta_i+\Delta)^u+(1-\psi(\theta_i))^{1-u}(1-\psi(\theta_i+\Delta))^u\right)$. Then for each

Z_{il} such that $A_i = 1$, we can compute its second and 4th moment as

$$\mathbb{Q}_A((Z_{il} - \mathbb{Q}_A(Z_{il}))^2) = K_i''(u^*) = \psi'(\theta_i)\Delta^2 \frac{e^{u^*\Delta}}{(1 - \psi(\theta_i) + e^{u^*\Delta}\psi(\theta_i))^2} \in (C_1'\Delta^2, C_2'\Delta^2), \tag{2.156}$$

$$\mathbb{Q}_A((Z_{il} - \mathbb{Q}_A(Z_{il}))^4) = K_i''''(u^*) + 3K_i''(u^*)^2 \leq \Delta^2 K_i''(u^*) + 3K_i''(u^*)^2, \tag{2.157}$$

where $C_1', C_2' > 0$ in (2.156) are some constants and we have used

$$\begin{aligned} K_i''''(u^*) &= \psi'(\theta_i)\Delta^4 e^{u^*\Delta} \\ &\quad \times \frac{\psi(\theta_i)^3 e^{3u^*\Delta} - 3\psi(\theta_i)\psi'(\theta_i)e^{2u^*\Delta} - 3\psi'(\theta_i)(1 - \psi(\theta^*))e^{u^*\Delta} + (1 - \psi(\theta_i))^3}{(1 - \psi(\theta_i) + \psi(\theta_i)e^{u^*\Delta})^5} \\ &\leq \psi'(\theta_i)\Delta^4 e^{u^*\Delta} \frac{1}{(1 - \psi(\theta_i) + \psi(\theta_i)e^{u^*\Delta})^2} = \Delta^2 K_i''(u^*) \end{aligned}$$

in (2.157).

Define $\mathcal{A}$ to be the event of A that (2.151), (2.152), (2.153), (2.155) and

$$\frac{1}{2}np \leq \sum_{i \in [n] \setminus \{k+1\}} A_i \leq 2np, \tag{2.158}$$

all hold. We know that $\mathbb{P}(A \in \mathcal{A}) \geq 1 - O(n^{-1})$.

With the above preparations, we can lower bound $\mathbb{P}\left(\sum_{l=1}^{L} Z_l \geq \log \frac{k}{n-k-1}\right)$ by

$$\inf_{A \in \mathcal{A}} \mathbb{P}_A \left(\sum_{l=1}^{L} Z_l \geq \log \frac{k}{n-k-1} \right) \mathbb{P}(A \in \mathcal{A}) \geq \frac{1}{2} \inf_{A \in \mathcal{A}} \mathbb{P}_A \left(\sum_{l=1}^{L} Z_l \geq \log \frac{k}{n-k-1} \right).$$

For any $A \in \mathcal{A}$, a change-of-measure argument leads to the lower bound

$$\begin{aligned}
& \mathbb{P}_A \left(\sum_{l=1}^{L} Z_l \geq \log \frac{k}{n-k-1} \right) \\
= \; & \exp\left(LK(u^*) - u^* \frac{k}{n-k-1} \right) \\
& \times \mathbb{Q}_A \left[\mathbb{I}\left\{ \sum_{l=1}^{L} Z_l - \log \frac{k}{n-k-1} \geq 0 \right\} \exp\left(-u^* (\sum_{l=1}^{L} Z_l - \log \frac{k}{n-k-1}) \right) \right] \\
\geq \; & \exp\left(-u^* T + LK(u^*) - u^* \log \frac{k}{n-k-1} \right) \mathbb{Q}_A \left(0 \leq \sum_{l=1}^{L} Z_l - \log \frac{k}{n-k-1} \leq T \right),
\end{aligned}$$

for any $T > 0$ to be specified. We first lower bound the exponent $LK(u^*) - u^* \log \frac{k}{n-k-1}$ by

$$\begin{aligned}
LK(u^*) - u^* \log \frac{k}{n-k-1} &= \min_{u \in (0,1)} \left(LK(u) - u \log \frac{k}{n-k-1} \right) \\
&\geq \min_{u \in (0,1)} \left(L(1+\delta_2)\overline{K}(u) - u \log \frac{k}{n-k-1} \right) \\
&\geq -\frac{1}{2} \left(\frac{\sqrt{(1+\delta_3)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta_3)\mathsf{SNR}}} \log \frac{n-k}{k} \right)^2,
\end{aligned}$$

for some $\delta_3 = o(1)$. We then need to choose an appropriate T so that the probability $\mathbb{Q}_A \left(0 \leq \sum_{l=1}^{L} Z_l - \log \frac{k}{n-k-1} \leq T \right)$ can be bounded below by some constant. To achieve this purpose, we note that

$$\mathsf{Var}_{\mathbb{Q}_A} \left(\sum_{l=1}^{L} Z_l \right) = L \sum_{i \in [n] \setminus \{k+1\}} A_i K_i''(u^*) \leq C_1 \Delta^2 L \sum_{i \in [n] \setminus \{k+1\}} A_i \leq 2 C_1 \Delta^2 L n p,$$

for some constant $C_1 > 0$ due to (2.156), where $\mathsf{Var}_{\mathbb{Q}_A}$ is the variance operator under the measure $\mathbb{Q}_A$. Thus, we set $T = \sqrt{2C_1 \Delta^2 L n p}$. With this choice, and by (2.154), we have

$$u^* T \leq \sqrt{2C_1 \Delta^2 L n p} \left(\frac{1}{2} - \frac{1}{(1+\delta')\mathsf{SNR}} \log \frac{n-k}{k} \right).$$

Therefore, u^*T is at most the order of the square-root of the desired exponent, and thus it is negligible.

Finally, we need to show $\mathbb{Q}_A\left(0 \leq \sum_{l=1}^{L} Z_l - \log\frac{k}{n-k-1} \leq T\right)$ is lower bounded by some constant. Note that the definition of u^* implies that $\sum_{l=1}^{L} Z_l - \log\frac{k}{n-k-1}$ has mean zero under $\mathbb{Q}_A$. By the definition of T, we have

$$\begin{aligned}
&\mathbb{Q}_A\left(0 \leq \sum_{l=1}^{L} Z_l - \log\frac{k}{n-k-1} \leq T\right) \\
\geq\ &\mathbb{Q}_A\left(0 \leq \sum_{l=1}^{L} Z_l - \log\frac{k}{n-k-1} \leq \sqrt{\mathsf{Var}\left(\sum_{l=1}^{L} Z_l \middle| A\right)}\right) \\
=\ &\mathbb{Q}_A\left(0 \leq \sum_{l=1}^{L}\sum_{i\in[n]\setminus\{k+1\}} Z_{il} - \log\frac{k}{n-k-1} \leq \sqrt{\mathsf{Var}\left(\sum_{l=1}^{L}\sum_{i\in[n]\setminus\{k+1\}} Z_{il} \middle| A\right)}\right).
\end{aligned}$$

We apply the central limit theorem in Lemma 3.7.4 to bound the above probability. The 4th moment approximation bound in Lemma 3.7.4 is

$$\begin{aligned}
&\sqrt{L\sum_{i\in[n]\setminus\{k+1\}} A_i\left(\frac{K_i''''(u^*)+3K_i''(u^*)^2}{(L\sum_{i\in[n]\setminus\{k+1\}} A_iK_i''(u^*))^2}\right)^{3/4}} \\
&\leq \sqrt{L\sum_{i\in[n]\setminus\{k+1\}} A_i\left(\frac{\Delta^2K_i''(u^*)+3K_i''(u^*)^2}{(L\sum_{i\in[n]\setminus\{k+1\}} A_iK_i''(u^*))^2}\right)^{3/4}} && (2.159) \\
&\leq \sqrt{L\sum_{i\in[n]\setminus\{k+1\}} A_i\left(\frac{C_2'+3C_2'^2}{(L\sum_{i\in[n]\setminus\{k+1\}} A_iC_1')^2}\right)^{3/4}} && (2.160) \\
&\leq C_2\left(L\sum_{i\in[n]\setminus\{k+1\}} A_i\right)^{-1/4} && (2.161)
\end{aligned}$$

which tends to zero by (2.158). We have used (2.157) in (2.159), (2.156) in (2.160). We thus

have

$$\mathbb{Q}_A\left(0 \le \sum_{l=1}^{L} Z_l - \log\frac{k}{n-k-1} \le T\right) \ge \mathbb{P}\left(0 \le N(0,1) \le 1\right) - o(1),$$

which is bounded below by a constant. To summarize, we have shown that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{l=1}^{L} Z_l \ge \log\frac{k}{n-k-1}\right) \\
&\ge C_3 \exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1+\delta_4)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta_4)\mathsf{SNR}}}\log\frac{n-k}{k}\right)^2\right)
\end{aligned}$$

for some $\delta_4 = o(1)$ and some constant $C_3 > 0$ when (2.150) holds with $\delta = \delta_4$.

To close the proof, we need a different argument when

$$\frac{\sqrt{(1+\delta_4)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta_4)\mathsf{SNR}}}\log\frac{n-k}{k} \le C_4,$$

for some constant $C_4 > 0$. This condition, together with Bernstein's inequality, implies that

$$\sum_{l=1}^{L}\mathbb{E}(Z_l|A) - \log\frac{k}{n-k-1} \ge -C_5\sqrt{Lnp\Delta^2}, \tag{2.162}$$

with probability at least $1 - O(n^{-1})$. Define $\overline{\mathcal{A}}$ to be an event of A such that both (2.158) and (2.162) hold. It is clear that $\mathbb{P}(\overline{\mathcal{A}}) \ge 1 - O(n^{-1})$. We then have

$$\begin{aligned}
\mathbb{P}\left(\sum_{l=1}^{L} Z_l \ge \log\frac{k}{n-k-1}\right) &\ge \frac{1}{2}\inf_{A\in\overline{\mathcal{A}}}\mathbb{P}_A\left(\sum_{l=1}^{L} Z_l \ge \log\frac{k}{n-k-1}\right) \\
&\ge \frac{1}{2}\inf_{A\in\overline{\mathcal{A}}}\mathbb{P}_A\left(\sum_{l=1}^{L}(Z_l - \mathbb{E}(Z_l|A)) \ge C_5\sqrt{Lnp\Delta^2}\right) \qquad (2.163) \\
&\ge c_1 - o(1), \qquad (2.164)
\end{aligned}$$

for some constant $c_1 > 0$. The inequality (2.163) is by (2.162). For (2.164), we use the

Gaussian approximation in Lemma 3.7.4, and the 4th moment approximation bound is of order $\left(L\sum_{i\in[n]\backslash\{k+1\}} A_i\right)^{-1/4}$ by similar calculation as in (2.161) under measure $\mathbb{P}_A$, which tends to zero by (2.158). The proof is complete. □

Proof of Theorem 2.6.1. We first note that the condition (2.15) implies that $\Delta = o(1)$. Choose $\kappa_1, \kappa_2 \geq 0$ such that we have both $\kappa_1 + \kappa_2 \leq \kappa$ and

$$\frac{n}{k\psi'(\kappa_1) + (n-k)\psi'(\kappa_2)} = V(\kappa).$$

We first consider the case $k \to \infty$ and $\kappa = \Omega(1)$. In this case, one can easily check that $\kappa_2 = \Omega(1)$. Our least favorable $\theta', \theta'' \in \Theta'(k, \Delta, \kappa)$ is constructed as follows. Let $\rho = o(1)$ be a vanishing number that will be specified later. Define $\theta'_i = \kappa_1$ for all $1 \leq i \leq k - \rho k$, $\theta'_i = 0$ for $k - \rho k < i \leq k$, $\theta'_i = -\Delta$ for $k < i \leq k + \rho(n-k)$ and $\theta'_i = -\kappa_2$ for $k + \rho(n-k) < i \leq n$. For the simplicity of proof, we choose ρ so that both ρk and $\rho(n-k)$ are integers. For θ'', we set $\theta''_i = \theta'_i$ for all $i \in [n]\backslash\{k+1\}$ and $\theta''_{k+1} = 0$. Recall the definition of the subset $\mathcal{R}_{k,\rho} \subset \mathfrak{S}_n$ in (2.137). We then have

$$\begin{aligned}
\inf_{\widehat{r}} \sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta'(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) &\geq \inf_{\widehat{r}} \sup_{\substack{r^* \in \mathcal{R}_{k,\rho} \\ \theta^* \in \{\theta', \theta''\}}} \mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \\
&\geq \inf_{\widehat{r}} \frac{1}{2} \sum_{\theta^* \in \{\theta', \theta''\}} \frac{1}{|\mathcal{R}_{k,\rho}|} \sum_{r^* \in \mathcal{R}_{k,\rho}} \mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*).
\end{aligned}$$

That is, we first lower bound the minimax risk by the Bayes risk. Since

$$\mathsf{H}_k(\widehat{r}, r^*) \geq \frac{1}{2k} \sum_{k - \rho k < i \leq k + \rho(n-k)} \left(\mathbb{I}\left\{\widehat{r}_i > k, r^*_i \leq k\right\} + \mathbb{I}\left\{\widehat{r}_i \leq k, r^*_i > k\right\} \right),$$

we have

$$
\begin{aligned}
&\inf_{\widehat{r}} \sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta'(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*, r^*)} \mathsf{H}_k(\widehat{r}, r^*) \\
&\geq \inf_{\widehat{r}} \frac{1}{2} \sum_{\theta^* \in \{\theta', \theta''\}} \frac{1}{\left|\mathcal{R}_{k,\rho}\right|} \sum_{r^* \in \mathcal{R}_{k,\rho}} \mathbb{E}_{(\theta^*, r^*)} \frac{1}{2k} \\
&\qquad \sum_{k - \rho k < i \leq k + \rho(n-k)} \left(\mathbb{I}\left\{\widehat{r}_i > k, r_i^* \leq k\right\} + \mathbb{I}\left\{\widehat{r}_i \leq k, r_i^* > k\right\} \right) \\
&\geq \frac{1}{4k\left|\mathcal{R}_{k,\rho}\right|} \sum_{k - \rho k < i \leq k + \rho(n-k)} \\
&\qquad \inf_{\widehat{r}} \sum_{\theta^* \in \{\theta', \theta''\}} \left(\sum_{\substack{r^* \in \mathcal{R}_{k,\rho} \\ r_i^* \leq k}} \mathbb{P}_{(\theta^*, r^*)}(\widehat{r}_i > k) + \sum_{\substack{r^* \in \mathcal{R}_{k,\rho} \\ r_i^* \geq k+2}} \mathbb{P}_{(\theta^*, r^*)}(\widehat{r}_i \leq k) \right) \\
&\geq \frac{1}{4k\left|\mathcal{R}_{k,\rho}\right|} \sum_{k - \rho k < i \leq k + \rho(n-k)} \inf_{\widehat{r}} \left(\sum_{\substack{r^* \in \mathcal{R}_{k,\rho} \\ r_i^* \leq k}} \mathbb{P}_{(\theta'', r^*)}(\widehat{r}_i > k) + \sum_{\substack{r^* \in \mathcal{R}_{k,\rho} \\ r_i^* \geq k+2}} \mathbb{P}_{(\theta', r^*)}(\widehat{r}_i \leq k) \right).
\end{aligned}
$$

At this point, we need to introduce some extra notation. For any $r, r' \in \mathfrak{S}_n$, we define the Hamming distance without normalization as $\mathcal{H}(r, r') = \sum_{i=1}^n \mathbb{I}\left\{r_i \neq r_i'\right\}$. For each $k - \rho k < i \leq k + \rho(n-k)$, we can partition the set $\mathcal{R}_{k,\rho}$ into three disjoint subsets. Define

$$
\begin{aligned}
\mathcal{R}_{k,\rho}^{(1)} &= \left\{r \in \mathcal{R}_{k,\rho} : r_i \leq k\right\}, \\
\mathcal{R}_{k,\rho}^{(2)} &= \left\{r \in \mathcal{R}_{k,\rho} : r_i = k+1\right\}, \\
\mathcal{R}_{k,\rho}^{(3)} &= \left\{r \in \mathcal{R}_{k,\rho} : r_i \geq k+2\right\}.
\end{aligned}
$$

It is easy to see that $\mathcal{R}_{k,\rho} = \cup_{j=1}^3 \mathcal{R}_{k,\rho}^{(j)}$. We note that the three subsets all depend on the index i, but we shall suppress this dependence to avoid notational clutter. For any $r \in \mathcal{R}_{k,\rho}^{(2)}$,

define

$$\begin{aligned}\mathcal{N}_{2\to 1}(r) &= \left\{r'' \in \mathcal{R}^{(1)}_{k,\rho} : \mathcal{H}(r, r'') = 2\right\},\\ \mathcal{N}_{2\to 3}(r) &= \left\{r' \in \mathcal{R}^{(3)}_{k,\rho} : \mathcal{H}(r, r') = 2\right\}.\end{aligned}$$

Since for any different permutations, the smallest Hamming distance between them is 2, $\mathcal{N}_{2\to 1}(r)$ and $\mathcal{N}_{2\to 3}(r)$ can be understood as neighborhoods r within $\mathcal{R}^{(1)}_{k,\rho}$ and $\mathcal{R}^{(3)}_{k,\rho}$, respectively. It is easy to check that $\{\mathcal{N}_{2\to 1}(r)\}_{r\in\mathcal{R}^{(2)}_{k,\rho}}$ are disjoint subsets, and they form a partition of $\mathcal{R}^{(1)}_{k,\rho}$. Similarly, $\{\mathcal{N}_{2\to 3}(r)\}_{r\in\mathcal{R}^{(2)}_{k,\rho}}$ are disjoint subsets, and form a partition of $\mathcal{R}^{(3)}_{k,\rho}$. With these notation, we have

$$\begin{aligned}&\inf_{\widehat{r}} \sup_{\substack{r^*\in\mathfrak{S}_n\\ \theta^*\in\Theta'(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r}, r^*)\\
&\geq \frac{1}{4k\left|\mathcal{R}_{k,\rho}\right|}\sum_{k-\rho k<i\leq k+\rho(n-k)}\\
&\qquad\qquad \inf_{\widehat{r}}\sum_{r\in\mathcal{R}^{(2)}_{k,\rho}}\left(\sum_{r''\in\mathcal{N}_{2\to 1}(r)}\mathbb{P}_{(\theta'',r'')}(\widehat{r}_i > k) + \sum_{r'\in\mathcal{N}_{2\to 3}(r)}\mathbb{P}_{(\theta',r')}(\widehat{r}_i \leq k)\right)\\
&= \frac{1}{4k\left|\mathcal{R}_{k,\rho}\right|}\sum_{k-\rho k<i\leq k+\rho(n-k)}\\
&\qquad\qquad \inf_{\widehat{r}}\sum_{r\in\mathcal{R}^{(2)}_{k,\rho}}\sum_{\substack{r''\in\mathcal{N}_{2\to 1}(r)\\ r'\in\mathcal{N}_{2\to 3}(r)}}\left(\frac{1}{n-k-1}\mathbb{P}_{(\theta'',r'')}(\widehat{r}_i > k) + \frac{1}{k}\mathbb{P}_{(\theta',r')}(\widehat{r}_i \leq k)\right)\\
&\geq \frac{1}{4k(n-k-1)\left|\mathcal{R}_{k,\rho}\right|}\sum_{k-\rho k<i\leq k+\rho(n-k)}\\
&\qquad\qquad \sum_{r\in\mathcal{R}^{(2)}_{k,\rho}}\sum_{\substack{r''\in\mathcal{N}_{2\to 1}(r)\\ r'\in\mathcal{N}_{2\to 3}(r)}}\inf_{0\leq\phi\leq 1}\left[\mathbb{E}_{(\theta'',r'')}\phi + \frac{n-k-1}{k}\mathbb{E}_{(\theta',r')}(1-\phi)\right],\end{aligned}$$

where we have used the fact $|\mathcal{N}_{2\to 1}(r)| = k$ and $|\mathcal{N}_{2\to 3}(r)| = n-k-1$ to obtain the equality

in the above display. To this end, it suffices to give a lower bound for the testing problem

$$\inf_{0\leq\phi\leq 1}\left[\mathbb{E}_{(\theta'',r'')}\phi+\frac{n-k-1}{k}\mathbb{E}_{(\theta',r')}(1-\phi)\right], \tag{2.165}$$

for any $r''\in\mathcal{N}_{2\to 1}(r)$ and any $r'\in\mathcal{N}_{2\to 3}(r)$ with any $r\in\mathcal{R}^{(2)}_{k,\rho}$ and any $k-\rho k<i\leq k+\rho(n-k)$.

For the two probability distributions in (2.165), the probability $\mathbb{P}_{(\theta'',r'')}$ is the BTL model with parameter $\{\theta''_{r''_i}\}_{i\in[n]}$ and the probability $\mathbb{P}_{(\theta',r')}$ is the BTL model with parameter $\{\theta'_{r'_i}\}_{i\in[n]}$. It turns out the two vectors $\{\theta''_{r''_i}\}_{i\in[n]}$ and $\{\theta'_{r'_i}\}_{i\in[n]}$ only differ by one entry. To see this, let i and j' be the two coordinates that r and r' differ and let i and j'' be the two coordinates that r and r'' differ. Then, r' and r'' differ at the ith, the j'th and the j''th coordinates. This immediately implies $\theta'_{r'_l}=\theta''_{r''_l}$ for all $l\in[n]\backslash\{i,j',j''\}$. By the definitions of $\mathcal{N}_{2\to 1}$ and $\mathcal{N}_{2\to 3}$, we have $r'_i=r_{j'}, r'_{j'}=k+1, r'_{j''}=r_{j''}$ and $r''_i=r_{j''}, r''_{j'}=r_{j'}, r''_{j''}=k+1$. Moreover, we also have $r_{j'}\geq k+2$ and $r_{j''}\leq k$. We remind the readers that all the three coordinates are in the interval $[k-\rho k+1, k+\rho(n-k)]$. According to the definitions of θ' and θ'', we then have $\theta'_{r'_{j''}}=\theta''_{r''_{j''}}=0$ and $\theta'_{r'_{j'}}=\theta''_{r''_{j'}}=-\Delta$. For the only different coordinate, we have $\theta'_{r'_i}=-\Delta$ and $\theta''_{r''_i}=0$.

Since $\{\theta''_{r''_i}\}_{i\in[n]}$ and $\{\theta'_{r'_i}\}_{i\in[n]}$ only differ by a single coordinate, the testing problem (2.165) is equivalent to

$$\inf_{0\leq\phi\leq 1}\left[\mathbb{E}_{(\theta'',\bar{r})}\phi+\frac{n-k-1}{k}\mathbb{E}_{(\theta',\bar{r})}(1-\phi)\right], \tag{2.166}$$

where $\bar{r}_i=i$ for all $i\in[n]$. The equivalence between (2.165) and (2.166) can be obtained by the existence of a simultaneous permutation that maps the two vectors $\{\theta''_{r''_i}\}_{i\in[n]}$ and

$\{\theta'_{r'_i}\}_{i\in[n]}$ to θ'' and θ'. By Neyman-Pearson lemma, we can lower bound (2.166) by

$$\mathbb{P}_{(\theta'',\bar{r})}\left(\frac{d\mathbb{P}_{(\theta',\bar{r})}}{d\mathbb{P}_{(\theta'',\bar{r})}} \geq \frac{k}{n-k-1}\right). \tag{2.167}$$

This probability can be lower bounded by

$$C\exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1+\delta)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta)\mathsf{SNR}}}\log\frac{n-k}{k}\right)_+^2\right),$$

with some constant $C>0$ and some $\delta = o(1)$ according to Lemma 2.10.3. Since

$$|\mathcal{R}^{(2)}_{k,\rho}|/|\mathcal{R}_{(k,\rho)}| = (1-\rho)n,$$

we have

$$\begin{aligned}
&\inf_{\widehat{r}} \sup_{\substack{r^*\in\mathfrak{S}_n \\ \theta^*\in\Theta'(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r},r^*) \\
\geq\ & C_1\rho\exp\left(-\frac{1}{2}\left(\frac{\sqrt{(1+\delta)\mathsf{SNR}}}{2} - \frac{1}{\sqrt{(1+\delta)\mathsf{SNR}}}\log\frac{n-k}{k}\right)_+^2\right),
\end{aligned}$$

for some constant $C_1>0$. When the exponent diverges, we can choose ρ that tends to zero sufficiently slow so that it can be absorbed into the exponent. Otherwise, we can simply set ρ to be a sufficiently small constant, and the above proof will still go through. One can use a similar argument as Lemma 2.10.3 to show (2.167) is bounded below by some constant. In this case, we have $\inf_{\widehat{r}}\sup_{\substack{r^*\in\mathfrak{S}_n \\ \theta^*\in\Theta'(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r},r^*)$ bounded below by some constant as desired.

Finally, we briefly discuss how to modify the proof when either $k\to\infty$ or $\kappa = \Omega(1)$ does not hold. When $k\to\infty$ and $\kappa = o(1)$, we can take $\theta'_i = 0$ for $1\leq i\leq k$ and $\theta'_i = -\Delta$ for $k<i\leq n$. The vector θ'' is still defined according to $\theta''_i = \theta'_i$ for all $i\in[n]\backslash\{k+1\}$ and

$\theta''_{k+1} = 0$. The proof will go through with some slight modifcation. When $k = O(1)$, the condition (2.15) is equivalent to $\mathsf{SNR} < (1-\epsilon)2\log n$ for some constant $\epsilon > 0$, and we only need to prove a constant minimax lower bound. This is obviously true becasue

$$\begin{aligned}\inf_{\widehat{r}} \sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta'(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*,r^*)} \mathsf{H}_k(\widehat{r}, r^*) &\geq \inf_{\widehat{r}} \sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta(k,\Delta,\kappa)}} \mathbb{E}_{(\theta^*,r^*)} \mathsf{H}_k(\widehat{r}, r^*) \\ &\geq \inf_{\widehat{r}} \sup_{\substack{r^* \in \mathfrak{S}_n \\ \theta^* \in \Theta(k,\Delta,\kappa)}} \frac{1}{2k} \mathbb{P}_{(\theta^*,r^*)} \left(\mathsf{H}_k(\widehat{r}, r^*) > 0\right),\end{aligned}$$

which is lower bounded by a constant by Theorem 2.3.4 and the condition that $k = O(1)$. □

2.11 Proofs of Local Error Rates

In this section, we prove Theorem 2.7.1 and Theorem 2.7.2.

2.11.1 Proof of Theorem 2.7.1

We first give Lemma 2.11.1 to characterize entrywise tail behaviors of the MLE (2.5) which is crucial to the upper bound in Theorem 2.7.1.

Lemma 2.11.1. *Assume $\frac{np}{\log n} \to \infty$ and $\kappa = O(1)$. Then, for the rank vector $\widehat{r}$ that is induced by the MLE (2.5), for any small constant $0.1 > \delta > 0$, there exists some constant $C > 0$, such that for any $t \in \mathbb{R}$, any $\theta^* \in \Theta(k, 0, \kappa)$, $r^* \in \mathfrak{S}_n$, we have*

$$\mathbb{P}_{(\theta^*,r^*)} \left(\widehat{\theta}_i \leq t\right) \leq C \exp\left(-\frac{(1-\delta)(\theta^*_{r_i^*} - t)_+^2 npL}{2V_{r_i^*}(\theta^*)}\right) + Cn^{-7}, r_i^* \leq k; \tag{2.168}$$

$$\mathbb{P}_{(\theta^*,r^*)} \left(\widehat{\theta}_i \geq t\right) \leq C \exp\left(-\frac{(1-\delta)(t - \theta^*_{r_i^*})_+^2 npL}{2V_{r_i^*}(\theta^*)}\right) + Cn^{-7}, r_i^* \geq k+1 \tag{2.169}$$

Proof. The proof follows the proof of Theorem 2.3.2 with slight modifications. Without loss

of generality, we can assume $r_i^* = i$ for all $\in [n]$. Let

$$\bar{\Delta}_i = \begin{cases} \min\left((\theta_i^* - t)_+, \left(\frac{\log n}{np}\right)^{1/4}\right), & 1 \le i \le k, \\ \min\left((t - \theta_i^*)_+, \left(\frac{\log n}{np}\right)^{1/4}\right), & k+1 \le i \le n. \end{cases} \tag{2.170}$$

We only need to prove (2.168) since (2.169) can be proved similarly.

Consider any $m \in [k]$. When $(\theta_m^* - t)_+^2 npL \le c'$ for some large enough constant to be specified later, we can directly bound the probability using the trivial bound 1. Thus, we only need to consider the regime when $(\theta_m^* - t)_+^2 npL > c'$.

Following the proof of Theorem 2.3.2, we have (2.71)-(2.77) and (2.79) hold. Note that we now have $\bar{\Delta}_m^2 Lnp > c'$ instead of $\bar{\Delta}_m^2 Lnp \to \infty$ which is needed in the proof of Theorem 2.3.2. As a consequence, we now have (2.78) and (2.80) hold with $\delta = 4C_4 e^{\kappa}/\sqrt{c'}$ instead of some $o(1)$ as in the proof of Theorem 2.3.2. To sum up, with this δ, we have

$$|\widehat{\theta}_m - \bar{\theta}_m| \le \delta \bar{\Delta}_m, \tag{2.171}$$

$$\frac{|f^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - f^{(m)}(\theta_m^*|\theta_{-m}^*)|}{g^{(m)}(\theta_m^*|\theta_{-m}^*)} \le \delta \bar{\Delta}_m, \tag{2.172}$$

$$\frac{\left|g^{(m)}(\theta_m^*|\widehat{\theta}_{-m}) - g^{(m)}(\theta_m^*|\theta_{-m}^*)\right|}{g^{(m)}(\theta_m^*|\theta_{-m}^*)} \le \delta, \tag{2.173}$$

hold with probability at least $1 - O(n^{-7}) - \exp(-\bar{\Delta}_m^{3/2} Lnp) - \exp\left(-\bar{\Delta}_m^2 npL \frac{np}{\log n}\right)$. We can make δ to be an arbitrarily small constant by setting c' large as $\kappa = O(1)$.

Then for any $i \le k$, by the same argument as in the proof of Theorem 2.3.2, we have

$$\begin{aligned}
& \mathbb{P}\left(\widehat{\theta}_i \le t\right) \\
\le\ & \mathbb{P}\left(\widehat{\theta}_i - \theta_i^* \le -(\theta_i^* - t)\right) \\
\le\ & \mathbb{P}\left(\bar{\theta}_i - \theta_i^* \le -(1-\delta)\bar{\Delta}_i\right) + \mathbb{P}\left(|\bar{\theta}_i - \widehat{\theta}_i| > \delta\bar{\Delta}_i\right) \\
\le\ & \mathbb{P}\left(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -(1-3\delta)\bar{\Delta}_i\right) + O(n^{-7}) \qquad (2.174) \\
& + \exp(-\bar{\Delta}_i^{3/2} Lnp) + \exp\left(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\right),
\end{aligned}$$

which has the same upper bound as in (2.81). We then have the same (2.82) and the event $\mathcal{A}_i$ as in the proof of Theorem 2.3.2. As a result,

$$\begin{aligned}
& \mathbb{P}\left(-\frac{f^{(i)}(\theta_i^*|\theta_{-i}^*)}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \le -(1-3\delta)\bar{\Delta}_i\right) \\
\le\ & \sup_{A\in\mathcal{A}_i} \exp\left(-\frac{\frac{1}{2}(1-3\delta)^2\bar{\Delta}_i^2\left(L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^*-\theta_j^*)\right)^2}{L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^*-\theta_j^*) + \frac{1-3\delta}{3}\bar{\Delta}_i L\sum_{j\in[n]\setminus\{i\}} A_{ij}\psi'(\theta_i^*-\theta_j^*)}\right) \\
& + O(n^{-7}) \\
=\ & \exp\left(-\frac{1-\delta'}{2}\bar{\Delta}_i^2 Lp\sum_{j\in[n]\setminus\{i\}}\psi'(\theta_i^*-\theta_j^*)\right) + O(n^{-7}) \qquad (2.175) \\
\le\ & \exp\left(-\frac{1-\delta'}{2}(\theta_i^*-t)^2 Lp\sum_{j\in[n]\setminus\{i\}}\psi'(\theta_i^*-\theta_j^*)\right) + O(n^{-7}) \qquad (2.176) \\
=\ & \exp\left(-\frac{1-\delta''}{2V_i(\theta^*)}(\theta_i^*-t)^2 npL\right) + O(n^{-7}) \qquad (2.177)
\end{aligned}$$

where δ', δ'' are able to be any small constant (by adjusting c'). We use the definition of $\mathcal{A}_i$ to obtain the expression (2.175). To see why (2.176) is true, note that when $\bar{\Delta}_i^2 = \sqrt{\frac{\log n}{np}}$, the first term of (2.175) can be absorbed into $O(n^{-7})$. (2.177) comes from $\frac{\sum_{j\in[n]\setminus\{i\}}\psi'(\theta_i^*-\theta_j^*)}{\sum_{j\in[n]}\psi'(\theta_i^*-\theta_j^*)} = 1 + o(1)$.

Since $\exp(-\bar{\Delta}_i^{3/2}Lnp) + \exp\left(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\right) \leq \exp\left(-\frac{1+o(1)}{2V_i(\theta^*)}(\theta_i^* - t)^2 npL\right) + O(n^{-7})$, we have for any small constant $\delta > 0$, there exists some constant $C > 0$, such that

$$\mathbb{P}\left(\widehat{\theta}_i \leq t\right) \leq C\exp\left(-\frac{1-\delta}{2V_i(\theta^*)}(\theta_i^* - t)^2 npL\right) + Cn^{-7}, \tag{2.178}$$

for all $i \leq k$ which completes the proof. □

Proof of (2.28) of Theorem 2.7.1. The upper bound (2.28) is a straightforward consequence of Lemma 2.3.1 and Lemma 2.11.1. We have

$$\begin{aligned}
&\mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r}, r^*) \\
&\leq C\frac{1}{k}\left[\sum_{i=1}^{k}\exp\left(-\frac{(1-\delta)(\theta_i^* - t)_+^2 npL}{2V_i(\theta^*)}\right) + \sum_{i=k+1}^{n}\exp\left(-\frac{(1-\delta)(t - \theta_i^*)_+^2 npL}{2V_i(\theta^*)}\right)\right] + Cn^{-6}.
\end{aligned}$$

□

The rest of the section focuses on the lower bound (2.29). The proof follows the proof of Theorem 2.3.4 with some modification. We include it below for completeness.

Proof of (2.29) of Theorem 2.7.1. We are going to prove

$$\mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r}, r^*) \gtrsim \frac{R_1([k], \theta^*, t^*, -\delta) + R_2([n]\backslash[k], \theta^*, t^*, -\delta)}{k} \tag{2.179}$$

where t^* is the unique solution such that $R_1([k], \theta^*, t^*, -\delta) = R_2([n]\backslash[k], \theta^*, t^*, -\delta)$. We first show the existence and uniqueness of t^*. Note that $R_1([k], \theta^*, t, -\delta)$ increases with t while $R_2([n]\backslash[k], \theta^*, t, -\delta)$ decreases with t. Moreover, since $\lim_{t\to-\infty} R_1([k], \theta^*, t, -\delta) = \lim_{t\to+\infty} R_2([n]\backslash[k], \theta^*, t, -\delta) = 0$, such t^* must exist due to continuity. The uniqueness comes from $R_1([k], \theta^*, t, -\delta)$, as a function of t, is strictly increasing on $(-\infty, \theta_1^*]$ and $R_2([n]\backslash[k], \theta^*, t, -\delta)$, as a function of t, is strictly decreasing on $[\theta_n^*, +\infty)$ and $\theta_1^* \geq \theta_n^*$.

Define

$$S_1(t) = \left\{ i \in [n] : i \le k, (\theta_i^* - t)_+ \le (\log n/np)^{1/4} \right\}, \tag{2.180}$$
$$S_2(t) = \left\{ i \in [n] : i \ge k+1, (t - \theta_i^*)_+ \le (\log n/np)^{1/4} \right\}.$$

Since we assume $\inf_t(R_1([k], \theta^*, t, -\delta) + R_2([n]\backslash[k], \theta^*, t, -\delta)) \to \infty$, we must have

$$R_1([k], \theta^*, t^*, -\delta) \to \infty \tag{2.181}$$

and hence,

$$\frac{R_1(S_1(t^*), \theta^*, t^*, -\delta)}{R_1([k], \theta^*, t^*, -\delta)} \ge \frac{1}{2}, \frac{R_1(S_2(t^*), \theta^*, t^*, -\delta)}{R_1([n]\backslash[k], \theta^*, t^*, -\delta)} \ge \frac{1}{2}. \tag{2.182}$$

This is because $R_1([k], \theta^*, t^*, -\delta) - R_1(S_1(t^*), \theta^*, t^*, -\delta) \le n^{-6}$ and $R_2([n]\backslash[k], \theta^*, t^*, -\delta) - R_2(S_2(t^*), \theta^*, t^*, -\delta) \le n^{-6}$ by the definition of $S_1(t^*)$, $S_2(t^*)$ and $np/\log n \to \infty$.

Now by Lemma 2.9.3, we have

$$\begin{aligned} \mathsf{H}_k(\widehat{r}, r^*) &\ge \frac{1}{k} \min \left(\sum_{i=1}^{k} \mathbb{I}\left\{ \widehat{\theta}_i < t^* \right\}, \sum_{i=k+1}^{n} \mathbb{I}\left\{ \widehat{\theta}_i > t^* \right\} \right) \\ &\ge \frac{1}{k} \min \left(\sum_{i \in S_1(t^*)} \mathbb{I}\left\{ \widehat{\theta}_i < t^* \right\}, \sum_{i \in S_2(t^*)} \mathbb{I}\left\{ \widehat{\theta}_i > t^* \right\} \right). \end{aligned} \tag{2.183}$$

It suffices to show there exists some constant $C > 0$ such that

$$\mathbb{P}_{(\theta^*, r^*)} \left(\frac{1}{k} \sum_{i \in S_1(t^*)} \mathbb{I}\left\{ \widehat{\theta}_i < t^* \right\} \ge \frac{4C}{k} R_1(S_1(t^*), \theta^*, t^*, -\delta) \right) \ge 3/4 \tag{2.184}$$

$$\text{and } \mathbb{P}_{(\theta^*, r^*)} \left(\frac{1}{k} \sum_{i \in S_2(t^*)} \mathbb{I}\left\{ \widehat{\theta}_i > t \right\} \ge \frac{4C}{k} R_2(S_2(t^*), \theta^*, t^*, -\delta) \right) \ge 3/4. \tag{2.185}$$

This is because

$$
\begin{aligned}
&\mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r},r^*)\\
&\geq C\frac{R_1([k],\theta^*,t^*,-\delta)+R_2([n]\backslash[k],\theta^*,t^*,-\delta)}{k}\\
&\qquad\times\mathbb{P}_{(\theta^*,r^*)}\left(\mathsf{H}_k(\widehat{r},r^*)\geq C\frac{R_1([k],\theta^*,t^*,-\delta)+R_2([n]\backslash[k],\theta^*,t^*,-\delta)}{k}\right) &(2.186)\\
&\geq C\frac{R_1([k],\theta^*,t^*,-\delta)+R_2([n]\backslash[k],\theta^*,t^*,-\delta)}{k}\\
&\qquad\times\mathbb{P}_{(\theta^*,r^*)}\left(\begin{array}{c}\frac{\sum_{i\in S_1(t^*)}\mathbb{I}\left\{\widehat{\theta}_i<t^*\right\}}{k}\geq\frac{2C}{k}R_1([k],\theta^*,t^*,-\delta)\text{ and}\\ \frac{\sum_{i\in S_2(t^*)}\mathbb{I}\left\{\widehat{\theta}_i>t\right\}}{k}\geq\frac{2C}{k}R_2([n]\backslash[k],\theta^*,t^*,-\delta)\end{array}\right) &(2.187)\\
&\geq C\frac{R_1([k],\theta^*,t^*,-\delta)+R_2([n]\backslash[k],\theta^*,t^*,-\delta)}{k}\\
&\qquad\times\mathbb{P}_{(\theta^*,r^*)}\left(\begin{array}{c}\frac{\sum_{i\in S_1(t^*)}\mathbb{I}\left\{\widehat{\theta}_i<t^*\right\}}{k}\geq\frac{4C}{k}R_1(S_1(t^*),\theta^*,t^*,-\delta)\text{ and}\\ \frac{\sum_{i\in S_2(t^*)}\mathbb{I}\left\{\widehat{\theta}_i>t\right\}}{k}\geq\frac{4C}{k}R_2(S_2(t^*),\theta^*,t^*,-\delta)\end{array}\right) &(2.188)\\
&\geq\frac{C}{2}\frac{R_1([k],\theta^*,t^*,-\delta)+R_2([n]\backslash[k],\theta^*,t^*,-\delta)}{k}. &(2.189)
\end{aligned}
$$

Therefore, we obtain the desired conclusion. (2.186) is a consequence of Markov inequality; (2.187) comes from (2.183) and the choice of t^*; (2.188) is due to (2.182); (2.184) and (2.185) lead to (2.189).

In the rest of the proof, we are going to establish (2.184) and then (2.185) can be proved similarly. Define

$$
S_1'(\rho,t^*)=\left\{i\in S_1(t^*):\rho\left|S_1(t^*)\right|\text{ indices in }S_1(t^*)\text{ with the smallest }\frac{(\theta_i^*-t^*)_+^2}{V_i(\theta^*)}\right\} \tag{2.190}
$$

for some small enough constant $\rho>0$ to be specified later. That is, $S_1'(\rho,t^*)$ is a subset of $S_1(t^*)$ of size $\rho\left|S_1(t^*)\right|$ with the smallest $\frac{(\theta_i^*-t^*)_+^2}{V_i(\theta^*)}$ values. We remark that condition (2.181) and (2.182) necessarily imply $\left|S_1'(\rho,t^*)\right|\to\infty$ when ρ is a constant. We shall also assume

$\rho \left|S_1'(\rho, t^*)\right|$ is an integer. Furthermore, note that the definition of $S_1'(\rho, t^*)$ implies:

$$R_1(S_1(t^*), \theta^*, t^*, -\delta) \geq R_1(S_1'(\rho, t^*), \theta^*, t^*, -\delta) \geq \rho R_1(S_1(t^*), \theta^*, t^*, -\delta). \tag{2.191}$$

Therefore, to establish (2.184), we only need to show

$$\mathbb{P}_{(\theta^*, r^*)} \left(\sum_{i \in S_1'(\rho, t^*)} \mathbb{I}\left\{\widehat{\theta}_i < t^*\right\} \geq C' R_1(S_1'(\rho, t^*), \theta^*, t^*, -\delta) \right) \geq 3/4. \tag{2.192}$$

for some constant $C' > 0$. The remaining proof is then devoted to proving (2.192).

Recall the definition of $\bar{\theta}$ in (2.76). Define $\tilde{\Delta}_i = (\theta_i^* - t^*)_+ \vee \alpha\sqrt{\frac{1}{npL}}$ where α is some large enough constant to be determined later. Define the event $\mathcal{F}_i$ as

$$\begin{aligned}\mathcal{F}_i = \{&|\widehat{\theta}_i - \bar{\theta}_i| \leq \frac{\delta_0}{3}\tilde{\Delta}_i, \frac{|f^{(i)}(\theta_i^*|\widehat{\theta}_{-i}) - f^{(i)}(\theta_i^*|\theta_{-i}^*)|}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \leq \frac{\delta_0}{3}\tilde{\Delta}_i, \\ &\frac{\left|g^{(i)}(\theta_i^*|\widehat{\theta}_{-i}) - g^{(i)}(\theta_i^*|\theta_{-i}^*)\right|}{g^{(i)}(\theta_i^*|\theta_{-i}^*)} \leq \frac{\delta_0}{3}\}.\end{aligned}$$

When $(\theta_i^* - t^*)_+^2 npL > \alpha$, using a similar argument that leads to (2.171)-(2.173), we can show that there exists some constant $\delta_0 > 0$, such that

$$\mathbb{P}_{(\theta^*, r^*)}(\mathcal{F}_i) \geq 1 - \left(O(n^{-7}) + \exp\left(-\tilde{\Delta}_i^2 npL \frac{np}{\log n}\right) + \exp\left(-\tilde{\Delta}_i^{3/2} npL\right)\right). \tag{2.193}$$

When $(\theta_i^* - t^*)_+^2 npL \leq \alpha$, we can show

$$\mathbb{P}_{(\theta^*, r^*)}(\mathcal{F}_i) \geq 1 - \left(O(n^{-7}) + e^{-(npL)^{1/4}} + e^{-\sqrt{\log n}}\right). \tag{2.194}$$

instead. To establish it, we can choose $x = (npL)^{1/4}$ in (2.70) and $x = \sqrt{\log n}$ in (2.75) and then follow the same proof of (2.77), (2.78), and (2.80) as in the proof of Theorem 2.3.2. In

both cases, this δ_0 can be made arbitrarily small by setting α large.

Assuming $\mathcal{F}_i$ is true, we can use arguments similar to the establishment of (2.81) to have

$$\mathbb{I}\left\{\widehat{\theta}_i < t^*\right\} \geq \mathbb{I}\left\{\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi'(\theta_j^* - \theta_i^*)} \leq -(1+\delta_0)(\theta_i^* - t^*)_+\right\}.$$

Define the RHS of the above display as L_i. Then we have shown that

$$\sum_{i\in S_1'(\rho,t^*)} \mathbb{I}\left\{\widehat{\theta}_i < t^*\right\} \geq \sum_{i\in S_1'(\rho,t^*)} L_i\mathbb{I}_{\mathcal{F}_i} \geq \sum_{i\in S_1'(\rho,t^*)} L_i - \sum_{i\in S_1'(\rho,t^*)} \mathbb{I}_{\mathcal{F}_i^c}. \tag{2.195}$$

By (2.193) and (2.194), we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i\in S_1'(\rho,t^*)} \mathbb{I}_{\mathcal{F}_i^c}\right) \\
&\leq O(n^{-6}) + \sum_{i:i\in S_1'(\rho,t^*),(\theta_i^*-t^*)_+^2 npL>\alpha} \left(\exp\left(-\tilde{\Delta}^2 npL\frac{np}{\log n}\right) + \exp\left(-\tilde{\Delta}^{3/2}npL\right)\right) \\
&\quad + \sum_{i:i\in S_1'(\rho,t^*),(\theta_i^*-t^*)_+^2 npL\leq\alpha} \left(\exp\left(-(npL)^{1/4}\right) + \exp\left(-\sqrt{\log n}\right)\right).
\end{aligned}$$

Using $\theta_i^* - t^* \leq (\log n/np)^{1/4}$ for $i \in S_1(t^*)$ and $np/\log n \to \infty$, we see that the above bound is of smaller order than

$$n^{-5.9} + \sum_{i\in S_1'(\rho,t^*)} \exp\left[-\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\left(\left(\frac{np}{\log n}\right)^{1/9} \wedge (\log n)^{1/5}\right)\right],$$

and we can use Markov's inequality and obtain

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{i\in S_i'(t^*)} \mathbb{I}_{\mathcal{F}_i^c} \leq n^{-5.9} + \sum_{i\in S_1'(\rho,t^*)} e^{-\frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)}\left(\left(\frac{np}{\log n}\right)^{1/9}\wedge(\log n)^{1/5}\right)}\right) \geq 1 - o(1). \tag{2.196}$$

Now to lower bound $\sum_{i\in S_1'(\rho,t^*)} L_i$, we define

$$\mathcal{A} = \Bigg\{ A : \forall i \in S_1(t^*), \left| \frac{\sum_{j\in[n]\backslash\{i\}} A_{ij}\psi'(\theta_i^* - \theta_j^*)}{p\sum_{j\in[n]\backslash\{i\}} \psi'(\theta_i^* - \theta_j^*)} - 1 \right| \le \delta_0, \tag{2.197}$$

$$\left| \sum_{j\in S_1'(\rho,t^*)} A_{ji}\psi'(\theta_i^* - \theta_j^*) \right| \le 2\rho kp + 10\log n \Bigg\}. \tag{2.198}$$

By Bernstein's inequality and union bound, we have $\mathbb{P}(A \in \mathcal{A}) \ge 1 - O(n^{-10})$. From now on, we use the notation $\mathbb{P}_A$ for the conditional probability $\mathbb{P}_{(\theta^*,r^*)}(\cdot|A)$ given A. For any $s > 0$,

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{i\in S_1'(\rho,t^*)} L_i \ge s \right) \ge \mathbb{P}(A \in \mathcal{A}) \inf_{A\in\mathcal{A}} \mathbb{P}_A \left(\sum_{i\in S_1'(\rho,t^*)} L_i \ge s \right). \tag{2.199}$$

Now we study $\mathbb{P}_A\left(\sum_{i\in S_1'(\rho,t^*)} L_i \ge s\right)$. Define $S = [n]\backslash S_1'(\rho,t^*)$. Note that for each $i \in S_1'(\rho,t^*)$, we have $L_i \ge L_{i,1} - L_{i,2} - L_{i,3}$, where

$$L_{i,1} = \mathbb{I}\left\{ \frac{\sum_{j\in S} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))}{\sum_{j\in[n]\backslash\{i\}} A_{ji}\psi'(\theta_j^* - \theta_i^*)} \le -(1+2\delta')(1+\delta_0)\tilde{\Delta}_i \right\},$$

$$L_{i,2} = \mathbb{I}\left\{ \frac{\sum_{j\in S_1'(\rho,t^*):j<i} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))}{\sum_{j\in[n]\backslash\{i\}} A_{ji}\psi'(\theta_j^* - \theta_i^*)} \ge \delta'(1+\delta_0)\tilde{\Delta}_i \right\},$$

$$L_{i,3} = \mathbb{I}\left\{ \frac{\sum_{j\in S_1'(\rho,t^*):i<j} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))}{\sum_{j\in[n]\backslash\{i\}} A_{ji}\psi'(\theta_j^* - \theta_i^*)} \ge \delta'(1+\delta_0)\tilde{\Delta}_i \right\}$$

for some small constant $\delta' > 0$ whose value will be determined later. We are going to control each term separately.

(1). Analysis of $L_{i,1}$. Note that conditional on A, $\{L_{i,1}\}_{i\in S_1'(\rho,t^*)}$ are all independent Bernoulli random variables. We have $L_{i,1} \sim \text{Bernoulli}(p_i)$, where $p_i = \mathbb{E}_{(\theta^*,r^*)}(L_{i,1}|A)$. By

Chebyshev's inequality, we have

$$\mathbb{P}_A\left(\sum_{i\in S_1'(\rho,t^*)} L_{i,1} \geq \frac{1}{2}\sum_{i\in S_1'(\rho,t^*)} p_i\right) \geq 1 - \frac{4}{\sum_{i\in S_1'(\rho,t^*)} p_i}.$$

By Lemma 2.11.2, we can lower bound each p_i by

$$\begin{aligned} p_i &= \mathbb{P}_A\left(\frac{\sum_{j\in S} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \leq -(1+2\delta')(1+\delta_0)^2\tilde{\Delta}_i\right) \\ &\geq C_1 \exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)} - C_1'\sqrt{\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)}}\right), \end{aligned}$$

for some constants $C_1, C_1' > 0$ and some small constant $\delta_2 > 0$. Note that δ_2 can be an arbitrarily small constant by making δ' and ρ small as well as making α large. Thus we can choose δ', ρ small enough and α large enough to let $\delta_2 < \delta/2$. Then we have

$$\begin{aligned} \sum_{i\in S_1'(\rho,t^*)} p_i &\geq C_1 \sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)} - C_1'\sqrt{\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)}}\right) \\ &\geq C_1 R_1(S_1'(\rho,t^*), \theta^*, t^*, -\delta) \qquad (2.200) \\ &\geq C_1\rho R_1(S_1(t^*), \theta^*, t^*, -\delta). \qquad (2.201) \end{aligned}$$

where (2.200) can be achieved by setting α large and (2.201) comes from (2.191). As a result, under the condition (2.181), we have $\sum_{i\in S_1'(\rho,t^*)} p_i \to \infty$.

Hence, we have proved

$$\inf_{A\in\mathcal{A}} \mathbb{P}_A\left(\sum_{i\in S_1'(\rho,t^*)} L_{i,1} \geq \frac{1}{2}C_1 \sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)} - C_1'\sqrt{\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)}}\right)\right) \geq 1 - o(1).$$

(2). Analysis of $L_{i,2}$. By (2.197)-(2.198) and Bernstein's inequality, we can bound

$\mathbb{E}(L_{i,2}|A)$ by

$$\exp\left(-\frac{\left(\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}A_{ji}\psi'(\theta_j^*-\theta_i^*)\right)^2}{2\left(L\sum_{j\in S_1'(\rho,t^*):j<i}A_{ji}\psi'(\theta_i^*-\theta_j^*)+\frac{1}{3}\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}A_{ji}\psi'(\theta_j^*-\theta_i^*)\right)}\right)$$
$$\leq \exp\left(-\frac{\left(\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}p\psi'(\theta_j^*-\theta_i^*)\right)^2}{4\left(2L\rho kp+10\log n+\frac{1}{3}\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}p\psi'(\theta_j^*-\theta_i^*)\right)}\right).$$

Now we set $\delta'=\rho^{1/8}$, and make ρ small enough to ensure (2.201). Then, there exists some constants $C_2, C_3>0$ such that

$$\mathbb{E}(L_{i,2}|A)\leq \exp\left(-C_2\rho^{-\frac{1}{2}}npL\tilde{\Delta}_i^2\right)\leq \exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)}\right).$$

due to $\tilde{\Delta}_i=o(1)$ and $np/\log n\to\infty$. Then,

$$\mathbb{E}\left(\sum_{i\in S_1'(\rho,t^*)}L_{i,2}\middle|A\right)\leq\sum_{i\in S_1'(\rho,t^*)}\exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)}\right).$$

By Markov inequality, we have

$$\begin{aligned}&\inf_{A\in\mathcal{A}}\mathbb{P}_A\left(\sum_{i\in S_1'(\rho,t^*)}L_{i,2}\geq\sum_{i\in S_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)}\right)\right)\\&\leq\frac{\sum_{i\in S_1'(\rho,t^*)}\exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)}\right)}{\sum_{i\in S_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)}\right)}.\end{aligned}\tag{2.202}$$

(3). Analysis of $L_{i,3}$. By a similar argument, we also have

$$\begin{aligned}
&\inf_{A\in\mathcal{A}} \mathbb{P}_A \left(\sum_{i\in S_1'(\rho,t^*)} L_{i,3} \geq \sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1}{2} C_3 \rho^{-1/2} \frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)} \right) \right) \\
&\leq \frac{\sum_{i\in S_1'(\rho,t^*)} \exp\left(-C_3 \rho^{-1/2} \frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)} \right)}{\sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1}{2} C_3 \rho^{-1/2} \frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)} \right)}.
\end{aligned} \tag{2.203}$$

Now we can combine the above analyses of $L_{i,1}$, $L_{i,2}$ and $L_{i,3}$. Since we are allowed to choose ρ to be an arbitrarily small constant, we shall make

$$\sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1}{2} C_3 \rho^{-1/2} \frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)} \right) \leq \frac{1}{8} C_1 \sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1+\delta_2}{2} \frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)} - C_1' \sqrt{\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)}} \right)$$

and

$$\frac{\sum_{i\in S_1'(\rho,t^*)} \exp\left(-C_3 \rho^{-1/2} \frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)} \right)}{\sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1}{2} C_3 \rho^{-1/2} \frac{\tilde{\Delta}_i^2 npL}{2V_i(\theta^*)} \right)} \leq \frac{1}{16}.$$

Thus, we have

$$\inf_{A\in\mathcal{A}} \mathbb{P}_A \left(\sum_{i\in S_1'(\rho,t^*)} L_i \geq C_4 \sum_{i\in S_1'(\rho,t^*)} \exp\left(-\frac{1+\delta_2}{2} \frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)} - C_1' \sqrt{\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)}} \right) \right) \geq \frac{7}{8} - o(1), \tag{2.204}$$

for some constant $C_4 > 0$. Then (2.195), (2.196), (2.199) together with (2.181) lead to

$$\mathbb{P}_{(\theta^*,r^*)} \left(\sum_{i\in S_1'(\rho,t^*)} \mathbb{I}\left\{ \widehat{\theta}_i < t^* \right\} \geq \frac{C_4}{2} \sum_{i\in S_1'(\rho,t^*)} e^{-\frac{1+\delta_2}{2} \frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)} - C_1' \sqrt{\frac{\tilde{\Delta}_i^2 npL}{V_i(\theta^*)}}} \right) \geq \frac{7}{8} - o(1). \tag{2.205}$$

Finally, (2.192) follows from (2.201) which completes the proof. □

We state Lemma 2.11.2 to close this section. Its proof is essentially the same as the proof of Lemma 2.9.4 and hence is omitted here.

Lemma 2.11.2. *Assume* $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$. *Recall the definition of* $S_1'(\rho, t^*)$ *in (2.190),* $S = [n] \backslash S_1'(\rho, t^*)$ *and* $\tilde{\Delta}_i = (\theta_i^* - t^*)_+ \vee \alpha\sqrt{\frac{1}{npL}}$. *There exists some constants* $C_1, C_2 > 0$ *such that for any small constant* $0.1 > \tilde{\delta} > 0$, *there exists constant* $\delta_1 > 0$ *such that for any constant* $\alpha > 0$, $i \in S_1'(\rho, t^*)$, *any* $A \in \mathcal{A}$ *where* $\mathcal{A}$ *is defined in (2.197)-(2.198), any* $\theta^* \in \Theta(k, 0, \kappa)$ *and any* $r^* \in \mathfrak{S}_n$, *we have*

$$\begin{aligned}
&\mathbb{P}_{(\theta^*, r^*)}\left(\frac{\sum_{j \in S} A_{ji}(\bar{y}_{ij} - \psi(\theta^*_{r_i^*} - \theta^*_{r_j^*}))}{\sum_{j \in [n]\backslash\{i\}} A_{ji}\psi'(\theta^*_{r_j^*} - \theta^*_{r_i^*})} \le -(1+\tilde{\delta})\tilde{\Delta}_i \middle| A \right) \\
&\ge C_1 \exp\left(-\frac{1+\delta_1}{2}\frac{\tilde{\Delta}_i^2 npL}{V_{r_i^*}(\theta^*)} - C_2\sqrt{\frac{\tilde{\Delta}_i^2 npL}{V_{r_i^*}(\theta^*)}} \right). \qquad (2.206)
\end{aligned}$$

Moreover, δ_1 *is able to be arbitrarily small if* $\tilde{\delta}$ *and* ρ *are small enough.*

2.11.2 Proof of Theorem 2.7.2

We first give Lemma 2.11.3 to characterize entrywise tail behaviors of the spectral method (3.5) which is crucial to the upper bound in Theorem 2.7.2.

Lemma 2.11.3. *Assume* $\frac{np}{\log n} \to \infty$ *and* $\kappa = O(1)$. *Then, for the rank vector* $\widehat{r}$ *that is induced by the stationary distribution of the Markov chain (3.5), for any small constant* $0.1 > \delta > 0$, *there exists some constant* $C > 0$, *such that for any* $t \in \mathbb{R}$, *any* $\theta^* \in \Theta(k, 0, \kappa)$, $r^* \in \mathfrak{S}_n$, *we have*

$$\mathbb{P}_{(\theta^*, r^*)}\left(\widehat{\pi}_i \le \frac{e^t}{\sum_{j \in [n]} e^{\theta_j^*}} \right) \le C \exp\left(-\frac{(1-\delta)(\theta^*_{r_i^*} - t)_+^2 npL}{2\overline{V}_{r_i^*}(\theta^*)} \right) + Cn^{-4}, r_i^* \le k; \quad (2.207)$$

$$\mathbb{P}_{(\theta^*,r^*)}\left(\widehat{\pi}_i \geq \frac{e^t}{\sum_{j\in[n]} e^{\theta_j^*}}\right) \leq C\exp\left(-\frac{(1-\delta)(t-\theta^*_{r_i^*})_+^2 npL}{2\overline{V}_{r_i^*}(\theta^*)}\right) + Cn^{-4}, r_i^* \geq k+1 \tag{2.208}$$

Proof. The proof follows the proof of Theorem 2.4.1 with slight modifications. Without loss of generality, we can assume $r_i^* = i$ for all $\in [n]$. Define $\bar{\Delta}_i$ as in (2.170). We only need to prove (2.207) since (2.208) can be proved similarly.

Consider any $m \in [k]$. When $(\theta_m^* - t)_+^2 npL \leq c'$ for some large enough constant to be specified later, we can directly bound the probability using the trivial bound 1. Thus, we only need to consider the regime when $(\theta_m^* - t)_+^2 npL > c'$.

Following the proof of Theorem 2.4.1, we have (2.91)-(2.102) and (2.100) hold. Note that we now have $\bar{\Delta}_m^2 Lnp > c'$ instead of $\bar{\Delta}_m^2 Lnp \to \infty$ which is needed in the proof of Theorem 2.4.1. As a consequence, we now have (2.98) hold with $\delta = 4C_4 e^{\kappa}/\sqrt{c'}$ instead of some $o(1)$ as in the proof of Theorem 2.4.1. To sum up, with this δ, we have

$$\frac{|\widehat{\pi}_m - \bar{\pi}_m|}{\pi_m^*} \leq \delta(1 - e^{-\bar{\Delta}_m}), \tag{2.209}$$

$$\left|\frac{\sum_{j\in[n]\setminus\{m\}} A_{jm}\bar{y}_{jm}}{\sum_{j\in[n]\setminus\{m\}} A_{jm}\psi(\theta_j^* - \theta_m^*)} - 1\right| \leq \delta, \tag{2.210}$$

hold with probability at least $1 - O(n^{-4}) - \exp\left(-\bar{\Delta}_m^2 npL\frac{np}{\log n}\right) - \exp\left(-\bar{\Delta}_m^2 npL\sqrt{\frac{npL}{\log n}}\right)$. We can make δ to be an arbitrarily small constant by setting c' large as $\kappa = O(1)$.

Then for any $i \le k$, by the same argument as in the proof of Theorem 2.4.1, we have

$$\begin{aligned}
& \mathbb{P}\left(\widehat{\pi}_i \le \frac{e^t}{\sum_{j=1}^n e^{\theta_j^*}}\right) \\
= \;& \mathbb{P}\left(\frac{\widehat{\pi}_i - \pi_i^*}{\pi_i^*} \le e^{-(\theta_i^* - t)} - 1\right) \\
\le \;& \mathbb{P}\left(\frac{\widehat{\pi}_i - \pi_i^*}{\pi_i^*} \le e^{-\bar{\Delta}_i} - 1\right) \\
\le \;& \mathbb{P}\left(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1 + e^{\theta_j^* - \theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \le -(1-\delta)^2(1 - e^{-\bar{\Delta}_i})\right) \\
& + O(n^{-4}) + \exp\left(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\right) + \exp\left(-\bar{\Delta}_i^2 npL\sqrt{\frac{npL}{\log n}}\right),
\end{aligned}$$

which has the same upper bound as in (2.102). We then have the same (2.104) as in the proof of Theorem 2.4.1 which leads to

$$\begin{aligned}
& \mathbb{P}\left(\frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1 + e^{\theta_j^* - \theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \le -(1-\delta)^2(1 - e^{-\bar{\Delta}_i})\right) \\
\le \;& \exp\left(-\frac{(1-o(1))Lp\bar{\Delta}_i^2\left(\sum_{j\in[n]\setminus\{i\}}\psi(\theta_j^* - \theta_i^*)\right)^2}{2\sum_{j\in[n]\setminus\{i\}}\psi'(\theta_i^* - \theta_j^*)\left(1 + e^{\theta_j^* - \theta_i^*}\right)^2}\right) + O(n^{-4}) \\
= \;& \exp\left(-\frac{(1-\delta_2)npL\bar{\Delta}_i^2}{2\overline{V}_i(\theta^*)}\right) + O(n^{-4}) \\
\le \;& \exp\left(-\frac{(1-\delta_2)npL(\theta_i^* - t)_+^2}{2\overline{V}_i(\theta^*)}\right) + O(n^{-4})
\end{aligned}$$

with $\delta_1, \delta_2 > 0$ being some constant that can be arbitrarily small. The last inequality holds because when $\min\left((\theta_i^* - t)_+^2, \sqrt{\frac{\log n}{np}}\right) = \sqrt{\frac{\log n}{np}}$, then the first term becomes $\exp\left(-\frac{(1-\delta_2)L\sqrt{np\log n}}{2\overline{V}_i(\theta^*)}\right)$, which can be absorbed by $O(n^{-4})$. Since $\exp\left(-\bar{\Delta}_i^2 npL\frac{np}{\log n}\right) +$

$\exp\left(-\bar{\Delta}_i^2 npL\sqrt{\frac{npL}{\log n}}\right) \le \exp\left(-\frac{(1-\delta_2)(\theta_i^*-t)_+^2 npL}{2\overline{V}_i(\theta^*)}\right) + O(n^{-4})$, we have

$$\mathbb{P}\left(\widehat{\pi}_i \le \frac{e^t}{\sum_{j=1}^n e^{\theta_j^*}}\right) \le 2\exp\left(-\frac{(1-\delta_2)(\theta_i^*-t)_+^2 npL}{2\overline{V}_i(\theta^*)}\right) + O(n^{-4}), \tag{2.211}$$

for all $i \le k$. The proof is complete. □

Proof of (2.31) of Theorem 2.7.2. The upper bound (2.31) is a straightforward consequence of Lemma 2.11.3 in the same way as the proof of (2.28) of Theorem 2.7.1, and hence is omitted here. □

The rest of the section focuses on the lower bound (2.29). The proof follows the proof of Theorem 2.3.4 with some modification and is also very similar to the proof of (2.29) of Theorem 2.7.1. We include it below for completeness.

Proof of (2.32) of Theorem 2.7.2. To prove the lower bound (2.32), we are going to show

$$\mathbb{E}_{(\theta^*,r^*)}\mathsf{H}_k(\widehat{r},r^*) \gtrsim \frac{\overline{R}_1([k],\theta^*,t^*,-\delta) + \overline{R}_2([n]\backslash[k],\theta^*,t^*,-\delta)}{k} \tag{2.212}$$

where t^* is the unique solution such that $\overline{R}_1([k],\theta^*,t^*,-\delta) = \overline{R}_2([n]\backslash[k],\theta^*,t^*,-\delta)$. The existence and uniqueness of t^* follow the same argument as in the proof of (2.29) of Theorem 2.7.1. Recall the definition of $S_1(t)$ in (2.180). Since we assume $\inf_t(\overline{R}_1([k],\theta^*,t,-\delta) + \overline{R}_2([n]\backslash[k],\theta^*,t,-\delta)) \to \infty$, we have

$$\overline{R}_1([k],\theta^*,t^*,-\delta) \to \infty. \tag{2.213}$$

The proof of (2.212) follows the proof of Theorem 2.4.3. We will omit repeated details

and only present the differences. Define

$$\overline{S}_1'(\rho, t^*) = \left\{ i \in S_1(t^*) : \rho \, |S_1(t^*)| \text{ indices in } S_1(t^*) \text{ with the smallest } \frac{(\theta_i^* - t^*)_+^2}{\overline{V}_i(\theta^*)} \right\} \tag{2.214}$$

for some small enough constant $\rho > 0$ to be specified later. Following the same argument as in the proof of (2.29) of Theorem 2.7.1, we only need to show

$$\mathbb{P}_{(\theta^*, r^*)} \left(\sum_{i \in \overline{S}_1'(\rho, t^*)} \mathbb{I}\left\{ \widehat{\pi}_i < t \right\} \geq C' \overline{R}_1(\overline{S}_1'(\rho, t^*), \theta^*, t^*, -\delta) \right) \geq 3/4. \tag{2.215}$$

for some constant $C' > 0$. The remaining proof is then devoted to proving (2.215).

Recall the definition of $\bar{\pi}$ in (2.91). Define $\tilde{\Delta}_i = (\theta_i^* - t^*)_+ \vee \alpha \sqrt{\frac{1}{npL}}$ where α is some large enough constant to be determined later. Define the event $\overline{\mathcal{F}}_i$ as

$$\overline{\mathcal{F}}_i = \left\{ \frac{|\widehat{\pi}_i - \bar{\pi}_i|}{\pi_i^*} \leq \delta_0 (1 - e^{-\tilde{\Delta}_i}) \text{ and } \left| \frac{\sum_{j \in [n] \setminus \{i\}} A_{ji} \bar{y}_{ji}}{\sum_{j \in [n] \setminus \{i\}} A_{ji} \psi(\theta_j^* - \theta_i^*)} - 1 \right| \leq \delta_0 \right\}.$$

When $(\theta_i^* - t^*)_+^2 npL > \alpha$, using a similar argument that leads to (2.209)-(2.210), we can show that there exists some constant $\delta_0 > 0$, such that

$$\mathbb{P}_{(\theta^*, r^*)}(\overline{\mathcal{F}}_i) \geq 1 - \left(O(n^{-4}) + \exp\left(-\tilde{\Delta}_i^2 npL \frac{np}{\log n} \right) + \exp\left(-\tilde{\Delta}_i^2 npL \sqrt{\frac{npL}{\log n}} \right) \right). \tag{2.216}$$

When $(\theta_i^* - t^*)_+^2 npL \leq \alpha$, we can show

$$\mathbb{P}_{(\theta^*, r^*)}(\overline{\mathcal{F}}_i) \geq 1 - \left(O(n^{-4}) + e^{-(np/\log n)^{1/2}} + e^{-\sqrt{\log n}} \right). \tag{2.217}$$

instead. To establish it, we can choose $x = (np/\log n)^{1/2}$ in (2.95) and $x = \sqrt{\log n}$ in (2.96) and then follow the same proof of (2.98) and (2.100) as in the proof of Theorem 2.3.2. In both cases, this δ_0 can be made arbitrarily small by setting α large.

Assuming $\overline{\mathcal{F}}_i$ is true, we can use arguments similar to the establishment of (2.116) to have

$$\mathbb{I}\left\{\widehat{\pi}_i < t\right\} \geq \mathbb{I}\left\{ \frac{\sum_{j\in[n]\setminus\{i\}} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^* - \theta_i^*)} \leq -(1+\delta_0)^2\alpha\sqrt{\frac{1}{npL}} \right\}. \tag{2.218}$$

Define the RHS of the above display as $\overline{L}_i$.

$$\sum_{i\in\overline{S}_1'(\rho,t^*)} \mathbb{I}\left\{\widehat{\pi}_i < t\right\} \;\geq\; \sum_{i\in\overline{S}_1'(\rho,t^*)} \overline{L}_i \mathbb{I}_{\overline{\mathcal{F}}_i} \geq \sum_{i\in\overline{S}_1'(\rho,t^*)} \overline{L}_i - \sum_{i\in\overline{S}_1'(\rho,t^*)} \mathbb{I}_{\overline{\mathcal{F}}_i^c}. \tag{2.219}$$

By (2.216) and (2.217), we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i\in\overline{S}_1'(\rho,t^*)} \mathbb{I}_{\overline{\mathcal{F}}_i^c}\right) \\
&\leq O(n^{-3}) + \sum_{i:i\in\overline{S}_1'(\rho,t^*),(\theta_i^*-t^*)_+^2 npL>\alpha} \exp\left(-\tilde{\Delta}_i^2 npL\frac{np}{\log n}\right) + \exp\left(-\tilde{\Delta}_i^2 npL\sqrt{\frac{npL}{\log n}}\right) \\
&\quad + \sum_{i:i\in\overline{S}_1'(\rho,t^*),(\theta_i^*-t^*)_+^2 npL\leq\alpha} \exp\left(-(np/\log n)^{1/2}\right) + \exp\left(-\sqrt{\log n}\right).
\end{aligned}$$

Since the above bound is of smaller order than

$$n^{-2.9} + \sum_{i\in\overline{S}_1'(\rho,t^*)} \exp\left[-\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\left(\left(\frac{np}{\log n}\right)^{1/4} \wedge (\log n)^{1/4}\right)\right],$$

we can use Markov's inequality and obtain

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{i\in S_i'(t^*)} \mathbb{I}_{\overline{\mathcal{F}}_i^c} \leq n^{-2.9} + \sum_{i\in\overline{S}_1'(\rho,t^*)} e^{-\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\left(\left(\frac{np}{\log n}\right)^{1/4}\wedge(\log n)^{1/4}\right)}\right) \geq 1 - o(1). \tag{2.220}$$

Now to lower bound $\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_i$, we define

$$\overline{\mathcal{A}}=\Bigg\{A:\forall i\in S_1(t^*),\left|\frac{\sum_{j\in[n]\backslash\{i\}}A_{ij}\psi'(\theta_i^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_i^*}\right)^2}{p\sum_{j\in[n]\backslash\{i\}}\psi'(\theta_i^*-\theta_j^*)\left(1+e^{\theta_j^*-\theta_i^*}\right)^2}-1\right|\le\delta_0, \tag{2.221}$$

$$\left|\frac{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}{p\sum_{j\in[n]\backslash\{i\}}\psi(\theta_j^*-\theta_i^*)}-1\right|\le\delta_0, \tag{2.222}$$

$$\left|\sum_{j\in\overline{S}_1'(\rho,t^*)}A_{ji}\psi'(\theta_i^*-\theta_j^*)(1+e^{\theta_j^*-\theta_i^*})^2\right|\le 2\rho kp+10\log n\Bigg\}. \tag{2.223}$$

By Bernstein's inequality and union bound, we have $\mathbb{P}(A\in\overline{\mathcal{A}})\ge 1-O(n^{-3})$. From now on, we use the notation $\mathbb{P}_A$ for the conditional probability $\mathbb{P}_{(\theta^*,r^*)}(\cdot|A)$ given A. For any $s>0$,

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_i\ge s\right)\ge\mathbb{P}(A\in\overline{\mathcal{A}})\inf_{A\in\overline{\mathcal{A}}}\mathbb{P}_A\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_i\ge s\right). \tag{2.224}$$

Now we study $\mathbb{P}_A\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}L_i\ge s\right)$. Define $S=[n]\backslash\overline{S}_1'(\rho,t^*)$. Note that for each $i\in\overline{S}_1'(\rho,t^*)$, we have $L_i\ge L_{i,1}-L_{i,2}-L_{i,3}$, where

$$\overline{L}_{i,1}=\mathbb{I}\left\{\frac{\sum_{j\in S}A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}\le-(1+2\delta')(1+\delta_0)^2\tilde{\Delta}_i\right\},$$

$$\overline{L}_{i,2}=\mathbb{I}\left\{\frac{\sum_{j\in\overline{S}_1'(\rho,t^*):j<i}A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}\ge\delta'(1+\delta_0)^2\tilde{\Delta}_i\right\},$$

$$\overline{L}_{i,3}=\mathbb{I}\left\{\frac{\sum_{j\in\overline{S}_1'(\rho,t^*):i<j}A_{ji}(\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\backslash\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)}\ge\delta'(1+\delta_0)^2\tilde{\Delta}_i\right\}$$

for some small constant $\delta'>0$ whose value will be determined later. We are going to control

each term separately.

(1). Analysis of $\overline{L}_{i,1}$. Note that conditional on A, $\{\overline{L}_{i,1}\}_{i\in\overline{S}_1'(\rho,t^*)}$ are all independent Bernoulli random variables. We have $\overline{L}_{i,1} \sim \text{Bernoulli}(p_i)$, where $p_i = \mathbb{E}_{(\theta^*,r^*)}(\overline{L}_{i,1}|A)$. By Chebyshev's inequality, we have

$$\mathbb{P}_A\left(\sum_{i\in\overline{S}_1'(\rho,t^*)} \overline{L}_{i,1} \geq \frac{1}{2}\sum_{i\in\overline{S}_1'(\rho,t^*)} p_i\right) \geq 1 - \frac{4}{\sum_{i\in\overline{S}_1'(\rho,t^*)} p_i}.$$

By Lemma 2.11.4, we can lower bound each p_i by

$$\begin{aligned} p_i &= \mathbb{P}_A\left(\frac{\sum_{j\in S} A_{ji}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))(1+e^{\theta_j^*-\theta_i^*})}{\sum_{j\in[n]\setminus\{i\}} A_{ji}\psi(\theta_j^*-\theta_i^*)} \leq -(1+2\delta')(1+\delta_0)^2\tilde{\Delta}_i\right) \\ &\geq C_1\exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2 npL}{\overline{V}_i(\theta^*)} - C_1'\sqrt{\frac{\tilde{\Delta}_i^2 npL}{\overline{V}_i(\theta^*)}}\right), \end{aligned}$$

for some constants $C_1, C_1' > 0$ and some small constant $\delta_2 > 0$. Note that δ_2 can be an arbitrarily small constant by making δ' and ρ small as well as making α large. Thus we can choose δ', ρ small enough and α large enough to let $\delta_2 < \delta/2$. Then we have

$$\begin{aligned} &\sum_{i\in\overline{S}_1'(\rho,t^*)} p_i \geq C_1 \sum_{i\in\overline{S}_1'(\rho,t^*)} \exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2 npL}{\overline{V}_i(\theta^*)} - C_1'\sqrt{\frac{\tilde{\Delta}_i^2 npL}{\overline{V}_i(\theta^*)}}\right) \\ &\geq C_1\overline{R}_1(\overline{S}_1'(\rho,t^*),\theta^*,t^*,-\delta) \qquad (2.225) \\ &\geq C_1\rho\overline{R}_1(S_1(t^*),\theta^*,t^*,-\delta). \qquad (2.226) \end{aligned}$$

by the same argument as in the proof of (2.29) of Theorem 2.7.1. As a result, under the condition (2.213), we have $\sum_{i\in\overline{S}_1'(\rho,t^*)} p_i \to \infty$.

Hence, we have proved

$$\inf_{A\in\overline{\mathcal{A}}}\mathbb{P}_A\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_{i,1}\geq\frac{1}{2}C_1\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2 npL}{\overline{V}_i(\theta^*)}-C_1'\sqrt{\frac{\tilde{\Delta}_i^2 npL}{\overline{V}_i(\theta^*)}}\right)\right)\geq 1-o(1).$$

(2). Analysis of $\overline{L}_{i,2}$. By (2.221)-(2.223) and Bernstein's inequality, we can bound $\mathbb{E}(\overline{L}_{i,2}|A)$ by

$$\begin{aligned}
&e^{-\frac{\left(\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)\right)^2}{2\left(L\sum_{j\in\overline{S}_1'(\rho,t^*):j<i}A_{ji}\psi'(\theta_i^*-\theta_j^*)(1+e^{\theta_j^*-\theta_i^*})^2+\frac{1}{3}\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}A_{ji}\psi(\theta_j^*-\theta_i^*)\right)}}\\
&\leq\exp\left(-\frac{\left(\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}p\psi(\theta_j^*-\theta_i^*)\right)^2}{4\left(2L\rho kp+10\log n+\frac{1}{3}\delta'(1+\delta_0)^2\tilde{\Delta}_i L\sum_{j\in[n]\setminus\{i\}}p\psi(\theta_j^*-\theta_i^*)\right)}\right).
\end{aligned}$$

Now we set $\delta'=\rho^{1/8}$, and make ρ small enough to ensure (2.226). Then, there exists some constants $C_2,C_3>0$ such that

$$\mathbb{E}(\overline{L}_{i,2}|A)\leq\exp\left(-C_2\rho^{-\frac{1}{2}}npL\tilde{\Delta}_i^2\right)\leq\exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\right).$$

Then,

$$\mathbb{E}\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_{i,2}\middle|A\right)\leq\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\right).$$

By Markov inequality, we have

$$\begin{aligned}
&\inf_{A\in\overline{\mathcal{A}}}\mathbb{P}_A\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_{i,2}\geq\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\right)\right)\\
&\leq\frac{\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\right)}{\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2 npL}{2\overline{V}_i(\theta^*)}\right)}.
\end{aligned}\tag{2.227}$$

(3). Analysis of $\overline{L}_{i,3}$. By a similar argument, we also have

$$\begin{aligned}
&\inf_{A\in\overline{\mathcal{A}}}\mathbb{P}_A\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_{i,3}\geq\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2npL}{2\overline{V}_i(\theta^*)}\right)\right)\\
&\leq\frac{\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2npL}{2\overline{V}_i(\theta^*)}\right)}{\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2npL}{2\overline{V}_i(\theta^*)}\right)}.
\end{aligned}\tag{2.228}$$

Now we can combine the above analyses of $\overline{L}_{i,1}$, $\overline{L}_{i,2}$ and $\overline{L}_{i,3}$. Since we are allowed to choose ρ to be an arbitrarily small constant, we shall make

$$\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2npL}{2\overline{V}_i(\theta^*)}\right)\leq\frac{1}{8}C_1\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2npL}{\overline{V}_i(\theta^*)}-C_1'\sqrt{\frac{\tilde{\Delta}_i^2npL}{\overline{V}_i(\theta^*)}}\right)$$

and

$$\frac{\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2npL}{2\overline{V}_i(\theta^*)}\right)}{\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1}{2}C_3\rho^{-1/2}\frac{\tilde{\Delta}_i^2npL}{2\overline{V}_i(\theta^*)}\right)}\leq\frac{1}{16}.$$

Thus, we have

$$\inf_{A\in\overline{\mathcal{A}}}\mathbb{P}_A\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\overline{L}_i\geq C_4\sum_{i\in\overline{S}_1'(\rho,t^*)}\exp\left(-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2npL}{\overline{V}_i(\theta^*)}-C_1'\sqrt{\frac{\tilde{\Delta}_i^2npL}{\overline{V}_i(\theta^*)}}\right)\right)\geq\frac{7}{8}-o(1),\tag{2.229}$$

for some constant $C_4>0$. Then (2.219), (2.220), (2.224) together with (2.213) lead to

$$\mathbb{P}_{(\theta^*,r^*)}\left(\sum_{i\in\overline{S}_1'(\rho,t^*)}\mathbb{I}\left\{\widehat{\pi}_i<t\right\}\geq\frac{C_4}{2}\sum_{i\in\overline{S}_1'(\rho,t^*)}e^{-\frac{1+\delta_2}{2}\frac{\tilde{\Delta}_i^2npL}{\overline{V}_i(\theta^*)}-C_1'\sqrt{\frac{\tilde{\Delta}_i^2npL}{\overline{V}_i(\theta^*)}}}\right)\geq\frac{7}{8}-o(1).\tag{2.230}$$

Finally, (2.215) follows from (2.226) which completes the proof. □

We state Lemma 2.11.4 to close this section. Its proof is essentially the same as the proof of Lemma 2.9.4 and hence is omitted here.

Lemma 2.11.4. *Assume $\frac{np}{\log n} \to \infty$, $\kappa = O(1)$. Recall the definition of $\overline{S}_1'(\rho, t^*)$ in (2.214), $S = [n] \backslash \overline{S}_1'(\rho, t^*)$ and $\tilde{\Delta}_i = (\theta_i^* - t^*)_+ \vee \alpha\sqrt{\frac{1}{npL}}$. There exists some constants $C_1, C_2 > 0$ such that for any small constant $0.1 > \tilde{\delta} > 0$, there exists constant $\delta_1 > 0$ such that for any constant $\alpha > 0$, $i \in S_1'(t^*)$, any $A \in \overline{\mathcal{A}}$ where $\overline{\mathcal{A}}$ is defined in (2.221)-(2.223), any $\theta^* \in \Theta(k, 0, \kappa)$ and any $r^* \in \mathfrak{S}_n$, we have*

$$\mathbb{P}_{(\theta^*, r^*)} \left(\frac{\sum_{j \in S} A_{ji}(\bar{y}_{ij} - \psi(\theta^*_{r_i^*} - \theta^*_{r_j^*}))(1 + e^{\theta^*_{r_j^*} - \theta^*_{r_i^*}})}{\sum_{j \in [n] \backslash \{i\}} A_{ji} \psi(\theta^*_{r_j^*} - \theta^*_{r_i^*})} \leq -(1 + \tilde{\delta})\tilde{\Delta}_i \middle| A \right)$$
$$\geq C_1 \exp\left(-\frac{1 + \delta_1}{2} \frac{\tilde{\Delta}_i^2 npL}{\overline{V}_{r_i^*}(\theta^*)} - C_2 \sqrt{\frac{\tilde{\Delta}_i^2 npL}{\overline{V}_{r_i^*}(\theta^*)}} \right). \tag{2.231}$$

Moreover, δ_1 is able to be arbitrarily small if $\tilde{\delta}$ and ρ are small enough.

2.12 Proofs of Technical Lemmas

In this section, we prove Lemma 2.3.1, Lemma 2.8.1, Lemma 2.8.2, Lemma 2.8.3 and Lemma 2.8.4. We first list some additional technical results that will be needed in the proofs.

Lemma 2.12.1 (Hoeffding's inequality)**.** *For independent random variables $X_1, \cdots, X_n$ that satisfy $a_i \leq X_i \leq b_i$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t \right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right),$$

for any $t > 0$.

Lemma 2.12.2 (Bernstein's inequality)**.** *For independent random variables $X_1, \cdots, X_n$ that satisfy $|X_i| \leq M$ and $\mathbb{E}X_i = 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^{n}\mathbb{E}X_i^2 + \frac{1}{3}Mt}\right),$$

for any $t > 0$.

Lemma 2.12.3 (Central limit theorem, Theorem 2.20 of [91])**.** *If $Z \sim N(0,1)$ and $W = \sum_{i=1}^{n} X_i$ where X_i are independent mean 0 and $\mathsf{Var}(W) = 1$, then*

$$\sup_t |\mathbb{P}(W \leq t) - \mathbb{P}(Z \leq t)| \leq 2\sqrt{3\sum_{i=1}^{n}\left(\mathbb{E}X_i^4\right)^{3/4}}.$$

Proof of Lemma 2.3.1. Without loss of generality, we consider $r_i^* = i$ so that $\theta_1^* \geq \cdots \geq \theta_n^*$. Then, we can write the loss as $2k\mathsf{H}_k(\widehat{r}, r^*) = \sum_{i=1}^{k}\mathbb{I}\{\widehat{r}_i > k\} + \sum_{i=k+1}^{n}\mathbb{I}\{\widehat{r}_i \leq k\}$. Since $\widehat{r} \in \mathfrak{S}_n$, we must have $\sum_{i=1}^{k}\mathbb{I}\{\widehat{r}_i > k\} = \sum_{i=k+1}^{n}\mathbb{I}\{\widehat{r}_i \leq k\}$. This implies

$$\begin{aligned}
2k\mathsf{H}_k(\widehat{r}, r^*) &= 2\min\left(\sum_{i=1}^{k}\mathbb{I}\{\widehat{r}_i > k\}, \sum_{i=k+1}^{n}\mathbb{I}\{\widehat{r}_i \leq k\}\right) \\
&\leq 2\min\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{\theta}_i \leq \widehat{\theta}_{(k+1)}\right\}, \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{\theta}_i \geq \widehat{\theta}_{(k)}\right\}\right) \\
&\leq 2\max_t\min\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{\theta}_i \leq t\right\}, \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{\theta}_i \geq t\right\}\right) \qquad (2.232) \\
&= 2\min_t\max\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{\theta}_i \leq t\right\}, \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{\theta}_i \geq t\right\}\right) \qquad (2.233) \\
&\leq 2\min_t\left(\sum_{i=1}^{k}\mathbb{I}\left\{\widehat{\theta}_i \leq t\right\} + \sum_{i=k+1}^{n}\mathbb{I}\left\{\widehat{\theta}_i \geq t\right\}\right).
\end{aligned}$$

The inequality (2.232) uses the fact that $\widehat{\theta}_{(k)} \geq \widehat{\theta}_{(k+1)}$ where $\{\theta_{(i)}\}_{i=1}^{n}$ are the order statistics

with $\widehat{\theta}_{(1)}$ being the largest and $\widehat{\theta}_{(n)}$ being the smallest. The equality (2.233) holds since $\sum_{i=1}^{k} \mathbb{I}\left\{\widehat{\theta}_i \leq t\right\}$ is a nondecreasing function of t and $\sum_{i=k+1}^{n} \mathbb{I}\left\{\widehat{\theta}_i \geq t\right\}$ is a nonincreasing function of t. □

Proof of Lemma 2.8.1. The first conclusion is a direct consequence of Bernstein's inequality and a union bound argument. The second and third conclusion is a standard property of random graph Laplacian [103]. □

Proof of Lemma 2.8.2. To see the first conclusion, we note that $\mathbb{E}(A_{ij} - p)^2 \leq p$ and $\mathsf{Var}((A_{ij} - p)^2) \lesssim p$, and thus we can apply Bernstein's inequality followed by a union bound argument to obtain the desired result. The second conclusion is a direct consequence of Bernstein's inequality and a union bound argument. □

Proof of Lemma 2.8.3. For any $u \in \mathbb{R}^n$ such that $\mathbb{1}_n^T u = 0$,

$$u^T H(\theta) u = \sum_{1 \leq i < j \leq n} A_{ij} \psi(\theta_i - \theta_j) \psi(\theta_j - \theta_i)(u_i - u_j)^2.$$

Since $\psi(\theta_i - \theta_j)\psi(\theta_j - \theta_i) \geq \frac{1}{4} e^{-M}$, we have $\lambda_{\min,\perp}(H(\theta)) \geq \frac{1}{4} e^{-M} \lambda_{\min,\perp}(\mathcal{L}_A)$. By Lemma 2.8.1, we obtain the desired result. □

Proof of Lemma 2.8.4. Let $\mathcal{U} = \left\{u \in \mathbb{R}^n : \sum_{i \in [n]} u_i^2 \leq 1\right\}$ be the unit ball in $\mathbb{R}^n$. Then there exists a subset of $\mathcal{V} \subset \mathcal{U}$ such that for any $u \in \mathcal{U}$, there is a $v \in \mathcal{V}$ satisfying $\|u - v\| \leq 1/2$. Moreover, we also have $\log |\mathcal{V}| \leq C' n$ for some constant C'. See Lemma 5.2

of [105]. Then for any $u \in \mathcal{U}$, with the corresponding $v \in \mathcal{V}$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} u_i \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right) \\
&= \sum_{i=1}^{n} v_i \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right) + \sum_{i=1}^{n} (u_i - v_i) \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right) \\
&\leq \sum_{i=1}^{n} v_i \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right) + \frac{1}{2} \sqrt{\sum_{i=1}^{n} \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right)^2}.
\end{aligned}$$

Maximize u and v on both sides of the inequality, after rearrangement, we have

$$\begin{aligned}
&\sqrt{\sum_{i=1}^{n} \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right)^2} \\
&\leq 2 \max_{v \in \mathcal{V}} \sum_{i=1}^{n} v_i \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right) \\
&= 2 \max_{v \in \mathcal{V}} \sum_{i<j} A_{ij} (v_i - v_j)(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)).
\end{aligned}$$

Conditional on A, applying Hoeffding's inequality and union bound on the last line, we have

$$\begin{aligned}
\sum_{i=1}^{n} \left(\sum_{j \in [n] \setminus \{i\}} A_{ij} (\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)) \right)^2 &\leq C'' \frac{(\log n + n) \max_{v \in \mathcal{V}} \sum_{i<j} A_{ij} (v_i - v_j)^2}{L} \\
&\leq C'' \frac{(\log n + n) \lambda_{\max}(\mathcal{L}_A)}{L}
\end{aligned}$$

with probability at least $1 - O(n^{-10})$. By Lemma 2.8.1, we obtain the desired bound for the first conclusion.

The second conclusion is a direct application of Hoeffding's inequality and a union bound argument.

The proof of the third conclusion is similar to that of the first one. Define $\mathcal{U}_i = \left\{u \in \mathbb{R}^{n-1} : \sum_{j\in[n]\setminus\{i\}} A_{ij}u_j^2 \le 1\right\}$. Conditioning on A, one can think of $\mathcal{U}_i$ as a unit ball with dimension $\sum_{j\in[n]\setminus\{i\}} A_{ij} - 1$. Then, there exists a subset $\mathcal{V}_i \subset \mathcal{U}_i$ such that for any $u \in \mathcal{U}_i$, there is a $v \in \mathcal{V}_i$ that satisfies $\|u - v\| \le \frac{1}{2}$. Moreover, we also have $\log|\mathcal{V}_i| \le 2\sum_{j\in[n]\setminus\{i\}} A_{ij}$ by Lemma 5.2 of [105]. For any $u \in \mathcal{U}_i$, with the corresponding $v \in \mathcal{V}_i$, following a similar argument of the proof of the first conclusion, we have

$$\sqrt{\sum_{j\in[n]\setminus\{i\}} A_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))^2} \le 2\max_{v\in\mathcal{V}_i} \sum_{j\in[n]\setminus\{i\}} A_{ij}v_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)),$$

which implies

$$\sqrt{\max_{i\in[n]} \sum_{j\in[n]\setminus\{i\}} A_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))^2} \le 2\max_{i\in[n]}\max_{v\in\mathcal{V}_i} \sum_{j\in[n]\setminus\{i\}} A_{ij}v_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)).$$

Applying Hoeffding's inequality and union bound, we have

$$\max_{i\in[n]} \sum_{j\in[n]\setminus\{i\}} A_{ij}(\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*))^2 \le C_1 \frac{\log n + \max_{i\in[n]} \sum_{j\in[n]\setminus\{i\}} A_{ij}}{L},$$

with probability at least $1 - O(n^{-10})$. Finally, applying Lemma 2.8.1, we obtain the desired bound for the third conclusion, which concludes the proof. □

CHAPTER 3

OPTIMAL FULL RANKING FROM PAIRWISE COMPARISONS

3.1 Introduction

In this chapter, we study the problem of *full ranking*, the estimation of the entire rank vector r^*. To the best of our knowledge, theoretical analysis of full ranking under the BTL model has not been considered in the literature yet. We rigorously formulate the full ranking problem from a decision-theoretic perspective, and derive the minimax rate with respect to a loss function that measures the difference between two permutation vectors. To be specific, our main result of the paper shows that

$$\inf_{\widehat{r}\in\mathfrak{S}_n}\sup_{r^*\in\mathfrak{S}_n}\mathbb{E}\mathsf{K}(\widehat{r},r^*)\asymp\begin{cases}\exp\left(-\Theta(Lp\beta)\right), & Lp\beta>1,\\ n\wedge\sqrt{\frac{1}{Lp\beta}}, & Lp\beta\le 1,\end{cases} \tag{3.1}$$

where $\mathfrak{S}_n$ is the set of all rank vectors of size n, $\mathsf{K}(\widehat{r},r^*)$ is the *Kendall's tau distance* that counts the number of inversions between two ranks, and β is the minimal gap between skill parameters of different players. The precise definitions of these quantities will be given in Section 3.2. The minimax rate (3.1) exhibits a transition between an exponential rate and a polynomial rate. This is a unique phenomenon in the estimation of a full rank vector. In contrast, under the same BTL model, the minimax rate of estimating the skill parameters is always polynomial [86, 25], and the minimax rate of top-k ranking is always exponential [21]. Whether (3.1) is exponential or polynomial depends on the value of $Lp\beta$ that plays the role of signal-to-noise ratio. When $Lp\beta > 1$, the exponential minimax rate is a consequence of the discreteness of a rank vector. On the other hand, when $Lp\beta \le 1$, the discrete nature of ranking is blurred by the noise, and thus estimating the rank vector is effectively estimating a continuous parameter, which leads to a polynomial rate. A more detailed statement of the minimax rate (3.1) with an explicit exponent in the regime of exponential rate will be given

in Section 3.3.

Achieving the minimax rate (3.1) is a nontrivial problem. To this end, we propose a *divide-and-conquer* algorithm that first partitions the n players into several leagues and then computes a local MLE using games in each league. Finally, a full rank vector is obtained by aggregating local ranking results from all leagues. The divide-and-conquer technique is the basis of efficient algorithms for all kinds of sorting problems [94, 65, 68]. Our adaption of this classical technique in the optimal full ranking is motivated by both information-theoretic and computational considerations. From an information-theoretic perspective, games between players whose skill parameters are significantly different from each other have little effect on the final ranking result. This phenomenon can be revealed by a simple local Fisher information calculation of each player. The league partition step groups players with similar skill parameters together, thus maximizing information in the follow-up step of local MLE. From a computational perspective, the local MLE computed within each league involves an objective function whose Hessian matrix is well conditioned, a property that is crucial for efficient convex optimization. The description and the analysis of our algorithm are given in Section 3.4.

Before the end of the introduction section, let us also remark that the more general problem of permutation estimation has also been considered in various other settings in the literature [12, 13, 30, 31, 87, 47, 80, 48, 88]. For instance, in the problem of noisy sorting [13, 80], one assumes a data generating process that satisfies $\mathbb{P}(y_{ijl} = 1) > \frac{1}{2} + \gamma$ when $r_i^* < r_j^*$. In the feature matching problem [31, 30], it is assumed that $X_i - Y_{r_i^*} \sim \mathcal{N}(0, \sigma_i^2)$ for some permutation r^*, and the goal is to match the two data sequences X and Y by recovering the unknown permutation. An extension of this problem, called shuffled regression, assumes that the response variable y_i and regression function $x_{r_i^*}^T \beta$ are linked by an unknown permutation. Estimation of the unknown permutation in shuffled regression has been considered by [87].

The rest of the paper is organized as follows. We introduce the problem setting in

Section 3.2. The minimax rate of the full ranking is presented in Section 3.3. In Section 3.4, we introduce and analyze a divide-and-conquer algorithm that achieves the minimax rate. Numerical studies of the algorithm are given in Section 3.5. In Section 3.6, we discuss a few extensions and future projects that are related to the paper. Finally, Section 3.7 collects technical proofs of the results of the paper.

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 $\{1, 2, ..., d\}$. Given two numbers $a, b \in \mathbb{R}$, we use $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For any $x \in \mathbb{R}$, $\lfloor x \rfloor$ stands for the largest integer that is no greater than x and $\lceil x \rceil$ is the smallest integer that is no less than x. For two positive sequences $\{a_n\}, \{b_n\}$, $a_n \lesssim b_n$ or $a_n = O(b_n)$ means $a_n \leq C b_n$ for some constant $C > 0$ independent of n, $a_n = \Omega(b_n)$ means $b_n = O(a_n)$, and we use $a_n \asymp b_n$ or $a_n = \Theta(b_n)$ when both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. 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. We use the notation $S = S_1 \uplus S_2$ to denote a partition of S such that $S_1 \cap S_2 = \varnothing$ and $S = S_1 \cup S_2$. 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 $\|v\|_\infty = \max_{1 \leq i \leq d} |v_i|$. For a matrix $A \in \mathbb{R}^{n \times m}$, we use $\|A\|_{\mathrm{op}}$ for its operator norm, which is the largest singular value. The notation $\mathbb{1}_d$ means a d-dimensional column vector of all ones. 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.

3.2 A Decision-Theoretic Framework of Full Ranking

The BTL Model. Consider n players, each associated with a positive latent skill parameter w_i^* for $i \in [n]$. The games played among the n players are modeled by an Erdős-Rényi random graph $A \sim \mathcal{G}(n, p)$. To be specific, we have $A_{ij} \overset{iid}{\sim} \text{Bernoulli}(p)$ for all $1 \leq i < j \leq n$.

For any pair (i, j) such that $A_{ij} = 1$, we observe the outcomes of L games played between i and j, modeled by the Bradley-Terry-Luce (BTL) model (1.1). Our goal is to estimate the ranks of the n players.

To formulate the problem of full ranking from a decision-theoretic perspective, we can reparametrize the BTL model (1.1) by a sorted vector θ^* and a rank vector r^*. A sorted vector θ^* satisfies $\theta_1^* \geq \theta_2^* \geq \cdots \geq \theta_n^*$, and a rank vector r^* is an element of the permutation set $\mathfrak{S}_n$. We have

$$y_{ijl} \overset{ind}{\sim} \text{Bernoulli}(\psi(\theta^*_{r_i^*} - \theta^*_{r_j^*})), \quad l = 1, \cdots, L, \tag{3.2}$$

where $\psi(\cdot)$ is the sigmoid function $\psi(t) = \frac{1}{1+e^{-t}}$. In the original representation (1.1), we have $w_i^* = \exp(\theta^*_{r_i^*})$ for all $i \in [n]$. With (3.2), the full ranking problem is to estimate the rank vector r^* from the random comparison data.

Loss Function for Full Ranking. To measure the difference between an estimator $\widehat{r} \in \mathfrak{S}_n$ and the true $r^* \in \mathfrak{S}_n$, we introduce the *Kendall's tau distance*, defined by

$$\mathsf{K}(\widehat{r}, r^*) = \frac{1}{n} \sum_{1 \leq i < j \leq n} \mathbb{I}\Big\{\text{sign}(\widehat{r}_i - \widehat{r}_j)\text{sign}(r_i^* - r_j^*) < 0\Big\}, \tag{3.3}$$

where $\text{sign}(x)$ represents the sign of x and $n\mathsf{K}(\widehat{r}, r^*)$ counts the number of inversions between $\widehat{r}$ and r^*. Another distance is the normalized ℓ_1 loss, defined as

$$\mathsf{F}(\widehat{r}, r^*) = \frac{1}{n} \sum_{i=1}^{n} |\widehat{r}_i - r_i^*|, \tag{3.4}$$

also known as the *Spearman's footrule*. The two loss functions can be related by the following inequality,

$$\frac{1}{2}\mathsf{F}(\widehat{r}, r^*) \leq \mathsf{K}(\widehat{r}, r^*) \leq \mathsf{F}(\widehat{r}, r^*). \tag{3.5}$$

See [37] for the derivation of (3.5). The inequality (3.5) establishes an equivalence between the estimation of the vector r^* and that of the matrix of pairwise relation $\mathbb{I}\{r_i^* < r_j^*\}$, a key fact that we will explore in constructing an optimal algorithm.

Recall the normalized Hamming loss in top-k ranking in Chapter 2,

$$\mathsf{H}_k(\widehat{r}, r^*) = \frac{1}{2k}\left(\sum_{i=1}^{n} \mathbb{I}\left\{\widehat{r}_i > k, r_i^* \le k\right\} + \sum_{i=1}^{n} \mathbb{I}\left\{\widehat{r}_i \le k, r_i^* > k\right\}\right). \tag{3.6}$$

The comparison between (3.3) and (3.6) reveals the key difference between the two problems. While top-k ranking only requires a correct classification of the two groups, the quality of the full ranking depends on the accuracy of each individual $|\widehat{r}_i - r_i^*|$. It is easy to see that $\mathsf{K}(\widehat{r}, r) = 0$ implies $\mathsf{H}_k(\widehat{r}, r^*) = 0$, but the opposite direction is not true.

Regularity of Skill Parameters. For the nuisance parameter θ^* of the model (3.2), it is necessary that the skill parameters of neighboring players θ_i^* and θ_{i+1}^* are separated so that the identification of the ranks is possible. We introduce a parameter space that serves for this purpose. For any $\beta > 0$ and any $C_0 \ge 1$, define

$$\Theta_n(\beta, C_0) = \left\{\theta \in \mathbb{R}^n : \theta_1 \ge \cdots \ge \theta_n, 1 \le \frac{|\theta_i - \theta_j|}{\beta|i-j|} \le C_0 \text{ for any } i \ne j\right\}.$$

In other words, neighboring θ_i^* and θ_{i+1}^* are required to be separated by at least β. The magnitude of β then characterizes the difficulty of full ranking. The number C_0 characterizes the regularity of the space of sorted vectors $\Theta_n(\beta, C_0)$. The special case $\Theta_n(\beta, 1)$ only consists of fully regular θ's that can be written as $\theta_i = \alpha - \beta i$. Throughout the paper, we assume that $C_0 \ge 1$ is an absolute constant, but allow β to be a function of the sample size n, with the possibility that $\beta \to 0$.

The assumption $\theta^* \in \Theta_n(\beta, C_0)$ implies that the numbers $\theta_1^*, \cdots, \theta_n^*$ to be roughly evenly spaced. This assumption, which can be certainly relaxed, allows us to obtain relatively clean

formulas of the minimax rate of full ranking. By restricting our focus to the space $\Theta_n(\beta, C_0)$, we will develop a clear but nontrivial understanding of the full ranking problem in this paper. The extension of our results beyond $\theta^* \in \Theta_n(\beta, C_0)$ will be briefly discussed in Section 3.6.

3.3 Minimax Rates of Full Ranking

In this section, we present the minimax rate of full ranking under the BTL model. To better understand the results, we first derive the minimax rate of full ranking under a Gaussian pairwise comparison model in Section 3.3.1. This allows us to highlight some of the unique and nontrivial features of the BTL model by comparing the minimax rates of the two different distributions. Readers who are already familiar with the BTL model can directly start with Section 3.3.2.

3.3.1 Results for a Gaussian Model

Consider the same comparison scheme modeled by the Erdős-Rényi random graph $A \sim \mathcal{G}(n, p)$. For any pair (i, j) such that $A_{ij} = 1$, we independently observe

$$y_{ij} \sim \mathcal{N}(\theta^*_{r^*_i} - \theta^*_{r^*_j}, \sigma^2). \tag{3.7}$$

The joint distribution of $\{A_{ij}\}$ and $\{y_{ij}\}$, under the above generating process, is denoted by $\mathbb{P}_{(\theta^*, \sigma^2, r^*)}$. Estimation of the rank vector $r^* \in \mathfrak{S}_n$ under the Gaussian model (3.7) is much less complicated than the same problem under (3.2), because of the separate parametrization of mean and variance.

Theorem 3.3.1. *Assume $\theta^* \in \Theta_n(\beta, C_0)$ for some constant $C_0 \geq 1$ and $\frac{np}{\log n} \to \infty$. Then,*

for any constant δ that can be arbitrarily small, we have

$$\inf_{\widehat{r}\in\mathfrak{S}_n}\sup_{r^*\in\mathfrak{S}_n}\mathbb{E}_{(\theta^*,\sigma^2,r^*)}\mathsf{K}(\widehat{r},r^*)\gtrsim\begin{cases}\frac{1}{n-1}\sum_{i=1}^{n-1}\exp\left(-\frac{(1+\delta)np(\theta_i^*-\theta_{i+1}^*)^2}{4\sigma^2}\right), & \frac{np\beta^2}{\sigma^2}>1,\\ n\wedge\sqrt{\frac{\sigma^2}{np\beta^2}}, & \frac{np\beta^2}{\sigma^2}\le 1.\end{cases}$$

Moreover, let $\widehat{r}$ be the rank obtained by sorting the MLE $\widehat{\theta}$, and then

$$\sup_{r^*\in\mathfrak{S}_n}\mathbb{E}_{(\theta^*,\sigma^2,r^*)}\mathsf{K}(\widehat{r},r^*)\lesssim\begin{cases}\frac{1}{n-1}\sum_{i=1}^{n-1}\exp\left(-\frac{(1-\delta)np(\theta_i^*-\theta_{i+1}^*)^2}{4\sigma^2}\right)+n^{-5}, & \frac{np\beta^2}{\sigma^2}>1,\\ n\wedge\sqrt{\frac{\sigma^2}{np\beta^2}}, & \frac{np\beta^2}{\sigma^2}\le 1.\end{cases}$$

Both inequalities are up to constant factors only depending on C_0 and δ.

Theorem 3.3.1 characterizes the statistical fundamental limit of full ranking under the Gaussian comparison model. The result holds for each individual $\theta^*\in\Theta_n(\beta,C_0)$. It is interesting to note that the minimax rate exhibits a transition between an exponential rate and a polynomial rate. By scrutinizing the proof, the constant δ can be replaced by some sequence $\delta_n=o(1)$. Therefore, consider a special example $\theta^*\in\Theta_n(\beta,1)$, and the minimax rate (ignoring the n^{-5} term) can be simplified as

$$\begin{cases}\exp\left(-\frac{(1+o(1))np\beta^2}{4\sigma^2}\right), & \frac{np\beta^2}{\sigma^2}>1,\\ n\wedge\sqrt{\frac{\sigma^2}{np\beta^2}}, & \frac{np\beta^2}{\sigma^2}\le 1.\end{cases}\tag{3.8}$$

The behavior of (3.8) is illustrated in Figure 3.1. The quantity $\frac{np\beta^2}{\sigma^2}$ plays the role of the signal-to-noise ratio of the ranking problem. In the high SNR regime $\frac{np\beta^2}{\sigma^2}>1$, the difficulty of the ranking problem is dominated by whether the data can distinguish each r_i^* from its neighboring values. Therefore, ranking is essentially a *hypothesis testing* problem, which leads to an exponential rate. In the low SNR regime $\frac{np\beta^2}{\sigma^2}\le 1$, the discrete nature of ranking is absent because of the noise level. The recovery of r^* is equivalent to the estimation of a

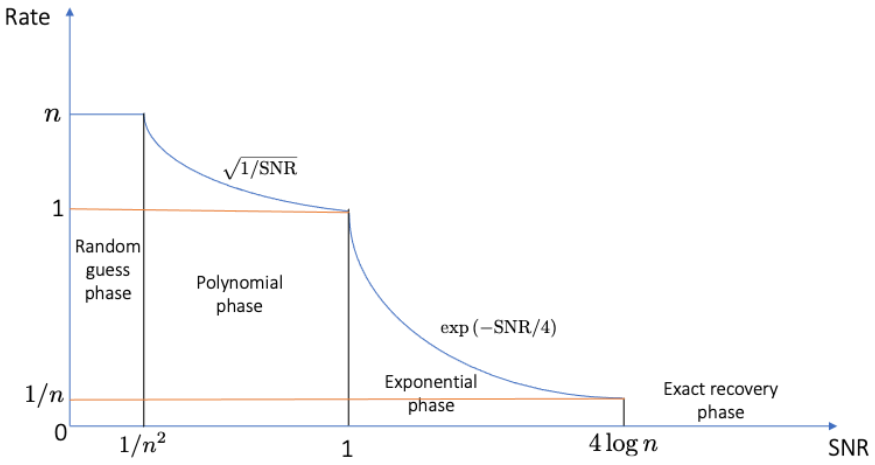

Figure 3.1: *Illustration of the the minimax rate of full ranking.*

continuous vector in $\mathbb{R}^n$, which is essentially a *parameter estimation* problem. The polynomial rate $n \wedge \sqrt{\frac{\sigma^2}{np\beta^2}}$ is the usual minimax rate for estimating an n-dimensional parameter under the ℓ_1 loss. It is also worthing noting that the rate (3.8) implies that the rank vector can be exactly recovered when $\frac{np\beta^2}{\sigma^2} > C \log n$ for any constant $C > 4$. This is because in this regime, we have $\mathsf{K}(\widehat{r}, r^*) = o(n^{-1})$ with high probability by a direct application of Markov's inequality. According to the definition of $\mathsf{K}(\widehat{r}, r^*)$, we know $\mathsf{K}(\widehat{r}, r^*) = o(n^{-1})$ implies $\mathsf{K}(\widehat{r}, r^*) = 0$.

The upper bound of Theorem 3.3.1 involves an extra n^{-5} term in the high SNR regime. According to the proof, the number 5 in the exponent can actually be replaced by an arbitrarily large constant. The n^{-5} term does not contribute to the high-probability bound. By a direct application of Markov's inequality, when $\frac{np\beta^2}{\sigma^2} \to \infty$, we have

$$\mathsf{K}(\widehat{r}, r^*) \lesssim \frac{1}{n-1} \sum_{i=1}^{n-1} \exp\left(-\frac{(1-\delta) np (\theta_i^* - \theta_{i+1}^*)^2}{4\sigma^2} \right), \tag{3.9}$$

with probability $1 - o(1)$. Notice that the high-probability bound (3.9) does not involve the n^{-5}. This is because when $\mathsf{K}(\widehat{r}, r^*)$ is nonzero, it must be at least n^{-1} by the definition of the loss function. Therefore, n^{-5} can always be absorbed into the other term of the upper

bound.

We also remark that the condition $\frac{np}{\log n} \to \infty$ guarantees that the random graph A is connected with high probability. It is well known that when $p \leq c\frac{\log n}{n}$ for some sufficiently small constant $c > 0$, the random graph has several disjoint components, which makes the comparisons between different components impossible. The condition $\frac{np}{\log n} \to \infty$ can be slightly relaxed to $\frac{np}{\log n} > C$ for some sufficiently large constant that depends on δ by carefully tracking the dependence among all constants in the proof, but we will just assume $\frac{np}{\log n} \to \infty$ throughout the paper to avoid this lengthy exercise of constants tracking.

An optimal estimator that achieves the minimax rate is the rank vector induced by the MLE, which is defined by

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta} \sum_{1 \leq i < j \leq n} A_{ij} \Big(y_{ij} - (\theta_i - \theta_j) \Big)^2. \tag{3.10}$$

We note that the parameter θ^* in (3.7) is identifiable up to a global shift. We may put an extra constraint $\mathbb{1}_n^T \theta = 0$ in the least-squares estimator above, so that $\widehat{\theta}$ is uniquely defined. However, this constraint is actually not essential, since even without it, the rank vector $\widehat{r}$ induced by $\widehat{\theta}$ is still uniquely defined. To study the property of $\widehat{\theta}$, we introduce a diagonal matrix $D \in \mathbb{R}^{n \times n}$ whose entries are given by $D_{ii} = \sum_{j \in [n] \setminus \{i\}} A_{ij}$. Then, $\mathcal{L}_A = D - A$ is the graph Laplacian of A. A standard least-squares analysis of (3.10) leads to the fact that up to some global shift,

$$\widehat{\theta} \sim \mathcal{N}\left(\theta^*, \sigma^2 \mathcal{L}_A^\dagger \right), \tag{3.11}$$

where $\mathcal{L}_A^\dagger$ is the generalized inverse of $\mathcal{L}_A$. The covariance matrix of (3.11) is optimal by achieving the intrinsic Cramér-Rao lower bound of the problem [10]. Without loss of generality, we can assume $r_i^* = i$ for each $i \in [n]$. Then, by the definition of the loss function

(3.3), we have

$$\mathbb{E}\mathsf{K}(\widehat{r}, r^*) = \frac{1}{n} \sum_{1 \le i < j \le n} \mathbb{P}\left(\widehat{r}_i > \widehat{r}_j\right) = \frac{1}{n} \sum_{1 \le i < j \le n} \mathbb{P}\left(\widehat{\theta}_i > \widehat{\theta}_j\right),$$

and each $\mathbb{P}\left(\widehat{\theta}_i > \widehat{\theta}_j\right)$ can be accurately estimated by a Gaussian tail bound under the distribution (3.11), which then leads to the upper bound result of Theorem 3.3.1. A detailed proof of Theorem 3.3.1, including a lower bound analysis, is given in Section 3.7.1.

3.3.2 Some Intuitions for the BTL Model

Before stating the minimax rate for the BTL model, we discuss a few key differences that one can expect from the result. Without loss of generality, we assume $r_i^* = i$ for all $i \in [n]$ throughout the discussion to simplify the notation. Let us consider a problem of oracle estimation of the skill parameter of the first player θ_1^*. To be specific, we would like to estimate θ_1^* by assuming that $\theta_2^*, \cdots, \theta_n^*$ are known. The Fisher information of this problem can be shown as

$$I^{\text{oracle}}(\theta_1^*) = Lp \sum_{j=2}^{n} \psi'(\theta_1^* - \theta_j^*). \tag{3.12}$$

The formula (3.12) characterizes the individual contribution of each player to the overall information in estimating θ_1^*. That is, the information from the games between 1 and j is quantified by $Lp\psi'(\theta_1^* - \theta_j^*)$. Since $\psi'(t) = \frac{e^t}{(1+e^t)^2} \le e^{-|t|}$, we have

$$\psi'(\theta_1^* - \theta_j^*) \le \exp\left(-|\theta_1^* - \theta_j^*|\right).$$

In other words, $\psi'(\theta_1^* - \theta_j^*)$ is an exponentially small function of the skill difference $|\theta_1^* - \theta_j^*|$. This means for players whose skills are significantly different from θ_1^*, their games with Player 1 offers little information in the inference of θ_1^*.

This phenomenon can be intuitively understood from the following simple example illus-

trated in Figure 3.2. Consider four players with specific skill parameters $(\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) =$

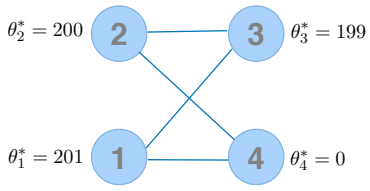

Figure 3.2: *A comparison graph of four players.*

$(201, 200, 199, 0)$, and we would like to compare the first two players. With the direct link between 1 and 2 missing, the only way to compare Players 1 and 2 is through their performances against Players 3 and 4. Since both $\theta_1^* - \theta_4^* = 201$ and $\theta_2^* - \theta_4^* = 200$ are very large numbers, it is very likely that Player 4 will lose all games against Players 1 and 2. On the other hand, we have $\theta_1^* - \theta_3^* = 2$ and $\theta_2^* - \theta_3^* = 1$, and thus Player 3 is likely to lose more games against Player 1 than against Player 2. Therefore, we can conclude that Player 1 is stronger than Player 2 based on their performances against Player 3, and the games against Player 4 offer essentially no information for this purpose. This example clearly illustrates that closer opponents are more informative.

Mathematically, for any $\theta^* \in \Theta_n(\beta, C_0)$ and any $M > 0$, it can be easily shown that

$$I^{\text{oracle}}(\theta_1^*) \leq (1 + O(e^{-M}))Lp \sum_{j \leq M/\beta} \psi'(\theta_1^* - \theta_j^*). \tag{3.13}$$

Therefore, (3.12) and (3.13) imply that

$$I^{\text{oracle}}(\theta_1^*) = (1 + O(e^{-M}))Lp \sum_{j \leq M/\beta} \psi'(\theta_1^* - \theta_j^*). \tag{3.14}$$

There is no need to consider the games against players with $j > M/\beta$. Moreover, we also observe from (3.14) that the parameter β plays two different roles in the BTL model:

1. The parameter β is the minimal gap between different players, and it quantifies the

signal strength of the BTL model.

2. The number $1/\beta$ quantifies the number of close opponents of each player, and thus p/β can be understood as the effective sample size of the BTL model.

While the first role is also shared by the β in the Gaussian comparison model (3.7), the second role dramatically distinguishes the BTL model from its Gaussian counterpart. The effective sample size of the Gaussian model is np, compared with p/β of the BTL model. This critical difference is a consequence of the nonlinearity of the logistic function. Increasing β magnifies the signal but reduces the effective sample size at the same time. The precise role of β in full ranking under the BTL model will be clarified by the formula of the minimax rate.

3.3.3 Results for the BTL Model

To present the minimax rate of full ranking under the BTL model, we first introduce some new quantities. For any $i \in [n]$, define

$$V_i(\theta^*) = \frac{n}{\sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*)}. \tag{3.15}$$

The quantity (3.15) is interpreted as the variance function of the ith best player. With a slight abuse of notation, the expectation associated with the BTL model is denoted as $\mathbb{E}_{(\theta^*,r^*)}$.

Theorem 3.3.2. *Assume $\theta^* \in \Theta_n(\beta, C_0)$ for some constant $C_0 \geq 1$ and $\frac{p}{(\beta\vee n^{-1})\log n} \to \infty$. Then, for any constant δ that can be arbitrarily small, we have*

$$\inf_{\widehat{r}\in\mathfrak{S}_n} \sup_{r^*\in\mathfrak{S}_n} \mathbb{E}_{(\theta^*,r^*)} \mathsf{K}(\widehat{r}, r^*) \gtrsim \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n-1} \exp\left(-\frac{(1+\delta)npL(\theta_i^*-\theta_{i+1}^*)^2}{4V_i(\theta^*)}\right), & \frac{Lp\beta^2}{\beta\vee n^{-1}} > 1, \\ n \wedge \sqrt{\frac{\beta\vee n^{-1}}{Lp\beta^2}}, & \frac{Lp\beta^2}{\beta\vee n^{-1}} \leq 1. \end{cases}$$

Moreover, let $\widehat{r}$ be the rank computed by Algorithm 2, and then if additionally $\frac{L}{\log n} \to \infty$, we have

$$\sup_{r^* \in \mathfrak{S}_n} \mathbb{E}_{(\theta^*, r^*)} \mathsf{K}(\widehat{r}, r^*) \lesssim \begin{cases} \frac{1}{n-1} \sum_{i=1}^{n-1} \exp\left(-\frac{(1-\delta) npL(\theta_i^* - \theta_{i+1}^*)^2}{4V_i(\theta^*)} \right) + n^{-5}, & \frac{Lp\beta^2}{\beta \vee n^{-1}} > 1, \\ n \wedge \sqrt{\frac{\beta \vee n^{-1}}{Lp\beta^2}}, & \frac{Lp\beta^2}{\beta \vee n^{-1}} \le 1. \end{cases}$$

Both inequalities are up to constant factors only depending on C_0 and δ.

Similar to Theorem 3.3.1, the result of Theorem 3.3.2 holds for each individual $\theta^* \in \Theta_n(\beta, C_0)$, and the minimax rate also exhibits a transition between an exponential rate and a polynomial rate. To better understand the minimax rate formula, we use Lemma 3.7.6 to quantify the order of the variance function $V_i(\theta^*)$. There exist constants $C_1, C_2 > 0$, such that

$$C_1 \left(\beta \vee \frac{1}{n} \right) \le \frac{V_i(\theta^*)}{n} \le C_2 \left(\beta \vee \frac{1}{n} \right).$$

Therefore, when $\beta \lesssim n^{-1}$, the minimax rate (ignoring the n^{-5} term) can be simplified as

$$\begin{cases} \exp\left(-\Theta(nLp\beta^2) \right), & nLp\beta^2 > 1, \\ n \wedge \sqrt{\frac{1}{nLp\beta^2}}, & nLp\beta^2 \le 1. \end{cases} \tag{3.16}$$

The formula (3.16) also exhibits a transition between a polynomial rate and an exponential rate. Its behavior can be illustrated by Figure 3.1 with SNR being $\Theta(nLP\beta^2)$. It is worth noting that the condition $\frac{L}{\log n} \to \infty$ is not needed when $\beta \lesssim n^{-1}$, and the minimax rate can be achieved by ranking the MLE,[1]

$$\widehat{\theta} = \operatorname*{argmax}_{\theta} \left[\bar{y}_{ij} \log \frac{1}{\psi(\theta_i - \theta_j)} + (1 - \bar{y}_{ij}) \log \frac{1}{1 - \psi(\theta_i - \theta_j)} \right], \tag{3.17}$$

where $\bar{y}_{ij} = \frac{1}{L} \sum_{l=1}^{L} y_{ijl}$.

1. The error rate (3.16) for the MLE (3.17) is an immediate consequence of Lemma 3.4.3.

In comparison, when $\beta \gtrsim n^{-1}$, the minimax rate (ignoring the n^{-5} term) is simplified into

$$\begin{cases} \exp\left(-\Theta(Lp\beta)\right), & Lp\beta > 1, \\ n \wedge \sqrt{\frac{1}{Lp\beta}}, & Lp\beta \leq 1. \end{cases} \tag{3.18}$$

Compared with the minimax rate (3.8) for the Gaussian comparison model, the dependence of (3.18) on β is weaker. This is a consequence of the dual roles of β discussed in Section 3.3.2. In fact, by writing

$$Lp\beta = L\beta^{-1}p\beta^2,$$

we can directly observe the effects of $\beta^{-1}p$ and β^2 as the effective sample size and the signal strength, respectively. On the other hand, the number of total players n has very little effect on the minimax rate formula.

The condition $\frac{p}{(\beta \vee n^{-1})\log n} \to \infty$ required by Theorem 3.3.2 can be equivalently written as $\frac{np}{\log n} \to \infty$ and $\frac{p}{\beta \log n} \to \infty$. Compared with the setting of Theorem 3.3.1, an additional condition $\frac{p}{\beta \log n} \to \infty$ is assumed for the BTL model. This condition can be seen as a consequence of the Fisher information formula (3.14) that statistical inference on the skill parameter of each player only depends on the player's close opponents. In other words, for each θ_i^*, the information is available in the games on the local graph

$$\mathcal{A}_i = \left\{ A_{jk} : |r_j^* - r_i^*| \leq \frac{M}{\beta}, |r_k^* - r_i^*| \leq \frac{M}{\beta} \right\}. \tag{3.19}$$

All the other games have little information in the statistical inference of θ_i^*. Therefore, it is required that the local graph $\mathcal{A}_i$ is connected. The condition $\frac{p}{\beta \log n} \to \infty$ guarantees the connectivity of $\mathcal{A}_i$ for all $i \in [n]$. Note that the size of the local graph is $O(\beta^{-1})$, which again justifies that the effective sample size of the BTL model is p/β instead of pn in the Gaussian case. Since the local graph $\mathcal{A}_i$ is unknown, the additional $\frac{L}{\log n} \to \infty$ assumption is needed in the upper bound to estimate it or its surrogate.

3.4 A Divide-and-Conquer Algorithm

We introduce a fully adaptive and computationally efficient algorithm for ranking under the BTL model in this section. We first outline the main idea in Section 3.4.1. Details of the algorithm are presented in Section 3.4.2, and the statistical properties are analyzed in Section 3.4.3.

3.4.1 An Overview

In the Gaussian comparison model, we first compute the global MLE for the skill parameters via the least-squares optimization (3.10), and then rank the players according to the estimators of the skills. This simple idea does not generalize to the BTL model, since the statistical information of each player concentrates on its close opponents, a phenomenon that is discussed in Section 3.3.2. Therefore, instead of using the global MLE, we should maximize likelihood functions that are only defined by players whose abilities are close. This modification not only addresses the information-theoretic issue of the BTL model that we just mentioned, but it also leads to Hessian matrices that are well conditioned, a property that is critical for efficient convex optimization.

For Player i, the set of close opponents that are sufficient for optimal statistical inference is given by $\mathcal{A}_i$ defined in (3.19). Suppose the knowledge of $\mathcal{A}_i$ was available, we could compute the local MLE using games only against players in $\mathcal{A}_i$. This idea is roughly correct, but there are several nontrivial issues that we need to solve before making it actually work. The first issue lies in the identifiability of the BTL model that θ_i^* can only be estimated up to a translation, which makes the comparison between $\widehat{\theta}_i$ obtained from $\mathcal{A}_i$ and $\widehat{\theta}_j$ obtained from $\mathcal{A}_j$ meaningless. The second issue is that the set $\mathcal{A}_i$ is unknown, and we need a data-driven procedure to identify the close opponents of each player.

We propose an algorithm that first partitions the n players into several leagues and then use local MLE to compare the skills of players within the same league. The league partition

is data-driven, and serves as a surrogate for the local graphs $\mathcal{A}_i$'s. Moreover, for two players i and j in the same league, the MLEs of their skill parameters are computed using the same set of opponents, and thus $\widehat{\theta}_i - \widehat{\theta}_j$ is a well-defined estimator of $\theta_i^* - \theta_j^*$.

Another key idea we use in our proposed algorithm is that the estimation of r^* is closely related to the estimation of the *pairwise relation matrix* R^* defined as

$$R_{ij}^* = \mathbb{I}\{r_i^* < r_j^*\} \quad \text{for all } 1 \le i \ne j \le n. \tag{3.20}$$

For any estimator of R^*, it can be converted into an estimator of the rank vector r^* according to Lemma 3.4.1. As a result, we shall focus on constructing a good estimator for all the pairwise relations $\{\mathbb{I}\{r_i^* < r_j^*\}\}_{i<j}$.

This divide-and-conquer algorithm, which will be described in Section 3.4.2, resembles typical strategies adopted in professional sports such as European football leagues where different teams are put into different leagues according to their skill level and teams only need to compete with other teams within the same league, which not only saves resources, but also leads to more accurate ranking of the teams. It is computationally efficient and we will show the algorithm achieves the minimax rate of full ranking.

3.4.2 Details of The Proposed Algorithm

We first decompose the set $[L]$ by $\{1, \cdots, L_1\}$ and $\{L_1 + 1, \cdots, L\}$. Games in the first set are used as preliminary games for league partition, and games in the second set are used for computing the MLE. Under the condition $\frac{L}{\log n} \to \infty$, we can set the number L_1 as $L_1 = \lceil \sqrt{L \log n} \rceil$. Define

$$\bar{y}_{ij}^{(1)} = \frac{1}{L_1} \sum_{l=1}^{L_1} y_{ijl} \quad \text{and} \quad \bar{y}_{ij}^{(2)} = \frac{1}{L - L_1} \sum_{l=L_1+1}^{L} y_{ijl}$$

as the summary statistics in $\{1, \cdots, L_1\}$ and $\{L_1 + 1, \cdots, L\}$, respectively.

The proposed algorithm consists of four steps, which we describe in detail below before presenting the whole procedure in Algorithm 2.

Step 1: League Partition. For each $i \in [n]$, we define

$$w_i^{(1)} = \sum_{j \in [n]} A_{ij} \mathbb{I}\{\bar{y}_{ij}^{(1)} \leq \psi(-2M)\}, \tag{3.21}$$

where M is some sufficiently large constant. The indicator $\mathbb{I}\{\bar{y}_{ij}^{(1)} \leq \psi(-2M)\}$ describes the event that Player i is completely dominated by Player j in the preliminary games. The quantity $w_i^{(1)}$ then counts the number of players who have dominated Player i. If $w_i^{(1)}$ is sufficiently small, Player i should belong to the top league since only few or no players could dominate Player i. Indeed, the first league is defined by

$$S_1 = \left\{ i \in [n] : w_i^{(1)} \leq h \right\}, \tag{3.22}$$

where h is chosen as $h = \frac{pM}{\beta}$. A data-driven h will be described in the Section 3.4.5. Similarly, $w_i^{(2)}$ and the second league S_2 can be defined by replacing $[n]$ with $[n] \backslash \{S_1\}$ in (3.21) and (3.22). Sequentially, we compute $w_i^{(k+1)}$ and S_{k+1} based on players in $[n] \backslash (S_1 \cup \cdots \cup S_k)$ for all $k \geq 1$. This procedure will terminate as soon as the number of the players who are yet to be classified is small enough, at which point all of the remaining players will be grouped together into the last league. The entire procedure of league partition is described in Algorithm 1.

Step 2: Local MLEs and Within-League Pairwise Relation Estimation. Having obtained the league partition $S_1, \cdots, S_K$, we need to compare players in the same league in the next step. Given the ambiguity between neighboring leagues, we shall also compare

Algorithm 1: A league partition algorithm

Input : $\{A_{ij}\bar{y}_{ij}^{(1)}\}_{1\le i<j\le n}$ and $\{A_{ij}\}_{1\le i<j\le n}$; M and h

Output: A partition of $[n]$: $S_1, \cdots, S_K$ such that $[n] = \uplus_{k=1}^{K} S_k$

1 For i in $[n]$, compute $w_i^{(1)} \leftarrow \sum_{j\in[n]} A_{ij}\mathbb{I}\{\bar{y}_{ij}^{(1)} \le \psi(-2M)\}$.
Set $S_1 \leftarrow \left\{i \in [n] : w_i^{(1)} \le h\right\}$ and $k = 1$.

2 While $n - (|S_1| + \cdots + |S_k|) > |S_k|/2$,
For each $i \in [n]\backslash(S_1 \cup \cdots \cup S_k)$,
compute $w_i^{(k+1)} \leftarrow \sum_{j\in[n]\backslash(S_1\cup\cdots\cup S_k)} A_{ij}\mathbb{I}\{\bar{y}_{ij}^{(1)} \le \psi(-2M)\}$.
Set $S_{k+1} \leftarrow \left\{i \in [n]\backslash(S_1 \cup \cdots \cup S_k) : w_i^{(k+1)} \le h\right\}$ and $k \leftarrow k+1$.

3 Set $K \leftarrow k-1$ and $S_K \leftarrow S_K \cup ([n]\backslash(S_1 \cup \cdots \cup S_{K-1}))$.

players if the leagues they belong to are next to each other. Therefore, for each $k \in [K-1]$, we need to compute the MLE for $\{\theta^*_{r_i^*}\}_{i\in S_k\cup S_{k+1}}$. This leads to the comparison between any two players in $S_k \cup S_{k+1}$. Define

$$\mathcal{E} = \left\{(i,j) : 1 \le i < j \le n, \psi(-M) \le \bar{y}_{ij}^{(1)} \le \psi(M)\right\}. \tag{3.23}$$

For each $k \in [K-1]$, the local negative log likelihood function is given by

$$\ell^{(k)}(\theta) = \sum_{\substack{(i,j)\in\mathcal{E} \\ i,j\in S_{k-1}\cup S_k\cup S_{k+1}\cup S_{k+2}}} A_{ij}\left[\bar{y}_{ij}^{(2)}\log\frac{1}{\psi(\theta_i-\theta_j)} + (1-\bar{y}_{ij}^{(2)})\log\frac{1}{1-\psi(\theta_i-\theta_j)}\right]. \tag{3.24}$$

When $k = 1$ or $k = K-1$, we use the notation $S_0 = S_{K+1} = \varnothing$. Note that the negative log likelihood function is only defined for edges in $\mathcal{E}$. In other words, only games between close opponents are considered. Moreover, some of the top players in S_k may have close opponents in the previous league S_{k-1}, and some of the bottom players in S_{k+1} may have close opponents in the next league S_{k+2}. The likelihood should include these games as well

for optimal inference of the parameters $\{\theta^*_{r^*_i}\}_{i\in S_k\cup S_{k+1}}$. The MLE is defined by

$$\widehat{\theta}^{(k)} \in \operatorname{argmin} \ell^{(k)}(\theta), \tag{3.25}$$

which is any vector that minimizes $\ell^{(k)}(\theta)$. Then, for any $i \in S_k$ and any $j \in S_k \cup S_{k+1}$, set

$$R_{ij} = \mathbb{I}\{\widehat{\theta}^{(k)}_i > \widehat{\theta}^{(k)}_j\}.$$

Note that $\{\widehat{\theta}^{(k)}_i\}_{i\in S_k\cup S_{k+1}}$ is defined only up to a common translation, but even with such ambiguity, the comparison indicator R_{ij} is uniquely defined.

We also remark that the computation of the MLE (3.25) is a straightforward convex optimization. It can be shown that the Hessian matrix of the objective function is well conditioned (Lemma 3.7.14), and thus a standard gradient descent algorithm converges to the optimum with a linear rate [25, 21].

Step 3: Cross-League Pairwise Relation Estimation. Consider i and j that belong to S_k and S_l respectively with $|k-l| \geq 2$. This is a pair of players that are separated by at least an entire league between them. For all such pairs, we set

$$R_{ij} = \mathbb{I}\{k < l\}.$$

Combined with the entries that are computed in Step 2, all upper triangular entries of the matrix R have been filled. The remaining entries of R can be filled according to the rule $R_{ij} + R_{ji} = 1$.

Step 2 and Step 3 together serve the purpose of estimating the pairwise relation matrix R^* defined in (3.20). Illustrated in Figure 3.3, the matrix R^* can be decomposed into blocks $\{R^*_{S_k\times S_l}\}_{k<l}$ according to the league partition $\{S_k\}_{k\in[K]}$. The yellow blocks close to the

diagonal are estimated by the procedure described in Step 2. In Figure 3.3, the data used in the two local MLEs ($k = 1$ and $k = 4$) are marked by different patterns for illustration. For example, when $k = 4$, we obtain estimators for $R^*_{S_4 \times S_4}$ and $R^*_{S_4 \times S_5}$ based on the local MLE that involves observations from $\{(i, j) \in \mathcal{E} : i, j \in S_3 \cup S_4 \cup S_5 \cup S_6\}$. The blue blocks are away from the diagonal and are estimated in Step 3. The remaining blocks in the lower triangular part are estimated according to $R_{ij} + R_{ji} = 1$.

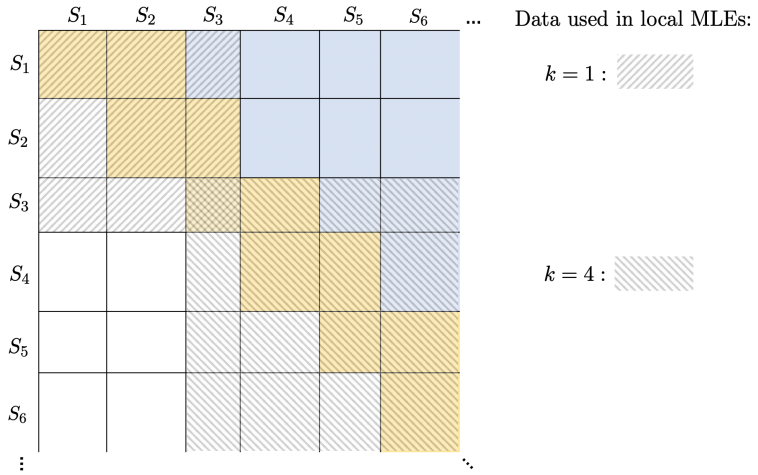

Figure 3.3: *Illustration of Step 2 and Step 3.*

Step 4: Full Rank Estimation. In the last step, we convert the pairwise relations estimator R into a rank estimator. First, compute the score for the ith player by

$$s_i = \sum_{j \in [n] \setminus \{i\}} R_{ij}.$$

Then, the rank estimator $\widehat{r}$ is obtained by sorting the scores $\{s_i\}_{i \in [n]}$.

The whole procedure of full ranking is summarized as Algorithm 2.

Algorithm 2: A divide-and-conquer full ranking algorithm

Input : $\{A_{ij}\bar{y}_{ij}^{(1)}\}_{1\le i<j\le n}$, $\{A_{ij}\bar{y}_{ij}^{(2)}\}_{1\le i<j\le n}$ and $\{A_{ij}\}_{1\le i<j\le n}$; M and h

Output: A rank vector $\widehat{r}\in\mathfrak{S}_n$

1 Run Algorithm 1 and obtain the partition $[n]=\uplus_{k=1}^{K}S_k$.
Set $S_0=S_{K+1}=\varnothing$.

2 For $k\in[K-1]$,
compute the local MLE $\widehat{\theta}^{(k)}$ according to (3.25).
For $i\in S_k$ and $j\in S_k\cup S_{k+1}$,
set $R_{ij}\leftarrow\mathbb{I}\{\widehat{\theta}_i^{(k)}>\widehat{\theta}_j^{(k)}\}$.

3 For $k\in[K-2]$ and $l\in[k+2:K]$,
For $(i,j)\in S_k\times S_l$,
set $R_{ij}\leftarrow 1$.
For $i\in[n]$ and $j\in[i+1:n]$,
set $R_{ji}\leftarrow 1-R_{ij}$.

4 For $i\in[n]$,
compute $s_i\leftarrow\sum_{j\in[n]\setminus\{i\}}R_{ij}$.
Sort $\{s_i\}_{i\in[n]}$ from high to low and obtain a full rank vector $\widehat{r}$.

3.4.3 Statistical Properties of Each Step

The purpose of this section is to prove the upper bound result of Theorem 3.3.2 by analyzing the statistical properties of Algorithm 2. The four components of the algorithm will be analyzed separately. We will first analyze Step 4 in Section 3.4.3, then Step 1 in Section 3.4.3, followed by Step 3 in Section 3.4.3 and finally Step 2 in Section 3.4.3. The results of these individual components will be combined to derive the minimax optimality of Algorithm 2, presented in Section 3.4.4.

From Pairwise Relations to Full Ranking (Step 4).

We first establish a result that clarifies the role of Step 4 of Algorithm 2. Consider any matrix $R\in\{0,1\}^{n\times n}$ that satisfies $R_{ij}+R_{ji}=1$ for any $i\neq j$. Let $\widehat{r}$ be the rank vector obtained by sorting $\{\sum_{j\in[n]\setminus\{i\}}R_{ij}\}_{i\in[n]}$ from high to low. The error of $\widehat{r}$ is controlled by the following lemma.

Lemma 3.4.1. *For any $r^* \in \mathfrak{S}_n$, define its pairwise relation matrix R^* such that $R^*_{ij} = \mathbb{I}\{r^*_i < r^*_j\}$. Then, we have*

$$\mathsf{K}(\widehat{r}, r^*) \le \frac{4}{n} \sum_{1 \le i \ne j \le n} \mathbb{I}\{R_{ij} \ne R^*_{ij}\}.$$

Lemma 3.4.1 is a deterministic inequality that bounds the error of the rank estimation by the estimation error of pairwise relations. It implies that to accurately rank n players, it is sufficient to accurately estimate the pairwise relations between all pairs.

Statistical Properties of League Partition (Step 1).

The partition output by Algorithm 1 satisfies several nice properties that are stated by the following theorem.

Theorem 3.4.1. *Assume $\theta^* \in \Theta_n(\beta, C_0)$ for some constant $C_0 \ge 1$, $\frac{L}{\log n} \to \infty$ and $\frac{p}{(\beta \vee n^{-1}) \log n} \to \infty$. Let $\{S_k\}_{k \in [K]}$ be the output of Algorithm 1 with $L_1 = \lceil \sqrt{L \log n} \rceil$, $1 \le M = O(1)$ and $h = \frac{pM}{\beta}$. Then, there exist some constants $C_1, C_2, C_3 > 0$ only depending on C_0 such that the following conclusions hold with probability at least $1 - O(n^{-9})$:*

1. Boundedness: *For any $k \in [K]$ and any $i, j \in S_{k-1} \cup S_k \cup S_{k+1}$, we have $|\theta^*_{r^*_i} - \theta^*_{r^*_j}| \le C_1 M$. Recall the convention that $S_0 = S_{K+1} = \varnothing$;*

2. Inclusiveness: *For any $k \in [K]$ and any $i \in S_k$, we have $\left\{ j \in [n] : |r^*_i - r^*_j| \le \frac{C_2 M}{\beta} \right\} \subset S_{k-1} \cup S_k \cup S_{k+1}$;*

3. Separation: *For any $i \in S_k$ and $j \in S_l$ such that $l - k \ge 2$, we have $\theta^*_{r^*_i} > \theta^*_{r^*_j}$;*

4. Independence: *For any $k \in [K]$, we have $S_k = \check{S}_k$. Here, $\{\check{S}_k\}_{k \in [K]}$ is a partition that is measurable with respect to the σ-algebra generated by $\{(A_{ij}, \bar{y}^{(1)}_{ij}) : |\theta^*_{r^*_i} - \theta^*_{r^*_j}| > 1.9M\}$;*

5. Continuity: *For any* $k \in [K-1]$ *and any* $i \in S_{k-1} \cup S_k \cup S_{k+1} \cup S_{k+2}$, *we have* $\left|\left\{j \in [n] : |\theta^*_{r_i^*} - \theta^*_{r_j^*}| \le \frac{M}{2}\right\} \cap (S_{k-1} \cup S_k \cup S_{k+1} \cup S_{k+2})\right| \ge C_3\left(\frac{M}{\beta} \wedge n\right)$.

We give some remarks on each conclusion of Theorem 3.4.1. The first conclusion asserts that the skill parameters of players from the neighboring leagues are close to each other. This property is complemented by the second conclusion that the close opponents of each player are either from the same league, the previous league, or the next league. In other words, for any $k \in [K]$ and any $i \in S_k$, the local graph $\{A_{jk} : j, k \in S_{k-1} \cup S_k \cup S_{k+1}\}$ can be viewed as a data-driven surrogate of $\mathcal{A}_i$ defined in (3.19). Moreover, the second conclusion also implies that $|S_{k-1} \cup S_k \cup S_{k+1}| \gtrsim \frac{1}{\beta} \wedge n$, from which we can deduce the bound $K = O(n\beta \vee 1)$ that controls the number of iterations Algorithm 1 needs before it is terminated.[2] Conclusion 3 implies that the partition $\{S_k\}_{k\in[K]}$ is roughly correlated with the true rank in the sense that it correctly identifies the comparisons between players who do not belong to neighboring leagues. Conclusion 4 shows that almost all of the randomness of the partition is from that of $\{(A_{ij}, \bar{y}_{ij}^{(1)}) : \theta^*_{r_i^*} - \theta^*_{r_j^*} \le -1.9M\}$. This fact leads to a crucial independence property in the later analysis of the local MLE. Conclusions 1, 2, 4, and 5 are crucial in the analysis of Step 2 in Section 3.4.3, while Conclusion 3 will be used in the analysis of Step 3 in Section 3.4.3.

The proof of Theorem 3.4.1 is a delicate mathematical induction argument that iteratively explores the asymptotic independence between consecutive constructions of leagues. To be specific, the random variable

$$w_i^{(k+1)} = \sum_{j \in [n] \setminus (S_1 \cup \cdots \cup S_k)} A_{ij} \mathbb{I}\{\bar{y}_{ij}^{(1)} \le \psi(-2M)\}$$

2. We can in fact prove a stronger result that $\frac{1}{\beta} \wedge n \lesssim |S_k| \lesssim \frac{1}{\beta} \wedge n$ uniformly for all $k \in [K]$ with probability at least $1 - O(n^{-9})$.

can be sandwiched between $\underline{w}_i^{(k+1)}$ and $\overline{w}_i^{(k+1)}$[3]. We show that both $\underline{w}_i^{(k+1)}$ and $\overline{w}_i^{(k+1)}$, when conditioning on the previous leagues $S_1, \cdots, S_k$, approximately follow Binomial distributions. Essentially, the A_{ij}'s that contribute to the summation of $w_i^{(k+1)}$ are disjoint from the A_{ij}'s that lead to the constructions of $S_1, \cdots, S_k$, which then implies an asymptotic independence property between $\left(\underline{w}_i^{(k+1)}, \overline{w}_i^{(k+1)}\right)$ and $S_1, \cdots, S_k$. This phenomenon

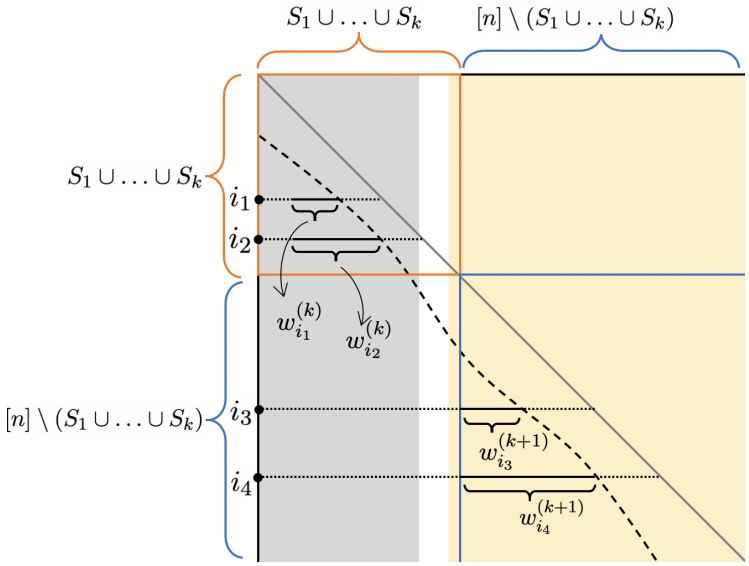

Figure 3.4: *Illustration of the independence property of Algorithm 1.*

is illustrated in Figure 3.4. In the picture, we use the orange block to denote $S_1 \cup \cdots \cup S_k$, the set that has already been partitioned. The next step of the algorithm is to construct the $(k+1)$th league from $[n] \setminus (S_1 \cup \cdots \cup S_k)$, which is the blue block. From the positions of $w_i^{(k+1)}$'s, we observe that the construction of S_{k+1} depends on A_{ij}'s that are in the yellow area. On the other hand, since the area on the left hand side of the dashed curve satisfies $\bar{y}_{ij}^{(1)} \leq \psi(-2M)$, the construction of the first k leagues only depends on A_{ij}'s that are in the grey area. The independence property can be easily seen from the separation between

3. Here $\underline{w}_i^{(k+1)}$ is defined as $\sum_{j \in [n] \setminus (S_1' \cup \cdots \cup S_k')} A_{ij} \mathbb{I}\{\theta^*_{r_j^*} \geq \theta^*_{r_i^*} + 2M + \delta_1\}$, for some quantity δ_1 such that $\mathbb{I}\{\theta^*_{r_j^*} \geq \theta^*_{r_i^*} + 2M + \delta_1\}$ is smaller than $\mathbb{I}\{\bar{y}_{ij}^{(1)} \leq \psi(-2M)\}$ for all pairs (i,j) with high probability, and $S_1' \cup \cdots \cup S_k'$ is another partition of $[n]$ that is equal to $S_1 \cup \cdots \cup S_k$ with high probability. $\overline{w}_i^{(k+1)}$ is defined similarly. See proof of Theorem 3.4.1 for details.

the grey and the yellow areas. A rigorous proof of Theorem 3.4.1, which is based on this argument, will be given in Section 3.7.3.

Statistical Properties of Cross-League Estimation (Step 3).

The analysis of Step 3 is quite straightforward following the results from the league partition. Assume the Conclusion 3 of Theorem 3.4.1 holds. Then for any $i \in S_k$ and $j \in S_l$ such that $l - k \geq 2$, we have $R^*_{ij} = 1$. Since $R_{ij} = 1$ for all such pairs, we have $\sum_{k\in[K-2]} \sum_{l\in[k+2:K]} \mathbb{I}\left\{R_{ij} \neq R^*_{ij}, i \in S_k, j \in S_l\right\} = 0$.

Statistical Properties of Local MLEs (Step 2).

The main challenge of analyzing the local MLE is the dependence between the partition $\{S_k\}_{k\in[K]}$ and the likelihood (3.24). We are going to use Conclusion 4 of Theorem 3.4.1 to resolve this issue. Define

$$\check{A}_{ij} = A_{ij}\mathbb{I}\{|\theta^*_{r^*_i} - \theta^*_{r^*_j}| \leq M/2\} + A_{ij}\mathbb{I}\left\{(i,j) \in \mathcal{E}, M/2 < |\theta^*_{r^*_i} - \theta^*_{r^*_j}| < 1.1M\right\},$$

and

$$\check{\ell}^{(k)}(\theta) = \sum_{i,j\in\check{S}_{k-1}\cup\check{S}_k\cup\check{S}_{k+1}\cup\check{S}_{k+2}} \check{A}_{ij}\left[\bar{y}^{(2)}_{ij} \log\frac{1}{\psi(\theta_i - \theta_j)} + (1 - \bar{y}^{(2)}_{ij})\frac{1}{1 - \psi(\theta_i - \theta_j)}\right].$$

The maximizer of $\check{\ell}^{(k)}(\theta)$ is denoted by

$$\check{\theta}^{(k)} \in \operatorname{argmin} \check{\ell}^{(k)}(\theta). \tag{3.26}$$

The introduction of $\check{\ell}^{(k)}(\theta)$ and $\check{\theta}^{(k)}$ is to disentangle the dependence of the MLE on the league partition. By Theorem 3.4.1, we know that $S_k = \check{S}_k$ for all $k \in [K]$. The concentration

of $\{\bar{y}_{ij}^{(1)}\}$ implies that $\{|\theta^*_{r_i^*} - \theta^*_{r_j^*}| \le M/2\} \subset \{(i,j) \in \mathcal{E}\} \subset \{|\theta^*_{r_i^*} - \theta^*_{r_j^*}| \le 1.1M\}$ for all $1 \le i < j \le n$. Therefore, we have

$$\mathbb{I}\{(i,j) \in \mathcal{E}\} = \mathbb{I}\{|\theta^*_{r_i^*} - \theta^*_{r_j^*}| \le M/2\} + \mathbb{I}\left\{(i,j) \in \mathcal{E}, M/2 < |\theta^*_{r_i^*} - \theta^*_{r_j^*}| < 1.1M\right\}.$$

We can thus conclude that $\ell^{(k)}(\theta) = \check{\ell}^{(k)}(\theta)$ for all θ with high probability. The result is formally stated below.

Lemma 3.4.2. *Assume $\theta^* \in \Theta_n(\beta, C_0)$ for some constant $C_0 \ge 1$, $\frac{L}{\log n} \to \infty$ and $\frac{p}{(\beta \vee n^{-1}) \log n} \to \infty$. Let $\{S_k\}_{k\in[K]}$ be the output of Algorithm 1 with $L_1 = \lceil \sqrt{L \log n} \rceil$, $1 \le M = O(1)$ and $h = \frac{pM}{\beta}$. Then, with probability at least $1 - O(n^{-8})$, we have $\ell^{(k)}(\theta) = \check{\ell}^{(k)}(\theta)$ for all θ and for all $k \in [K]$. As a consequence $\{\widehat{\theta}_i^{(k)}\}_{i \in S_k \cup S_{k+1}}$ and $\{\check{\theta}_i^{(k)}\}_{i \in S_k \cup S_{k+1}}$ are equivalent up to a common shift.*

With Lemma 3.4.2, it suffices to study (3.26) for the statistical property of the MLE. Note that $\{\check{A}_{ij}\}$ is measurable with respect to the σ-algebra generated by $\{(A_{ij}, \bar{y}_{ij}^{(1)}) : |\theta^*_{r_i^*} - \theta^*_{r_j^*}| < 1.1M\}$. Theorem 3.4.1 shows that $\{\check{S}_k\}$ is measurable with respect to the σ-algebra generated by $\{(A_{ij}, \bar{y}_{ij}^{(1)}) : |\theta^*_{r_i^*} - \theta^*_{r_j^*}| > 1.9M\}$. We then reach a very important conclusion that $\{\check{A}_{ij}\}$, $\{\bar{y}_{ij}^{(2)}\}$ and $\{\check{S}_k\}$ are mutually independent, and therefore we can analyze $\check{\theta}^{(k)}$ by conditioning on the partition $\{\check{S}_k\}$. To be more specific, for any $i, j \in \check{S}_k \cup \check{S}_{k+1}$ such that $\theta^*_{r_i^*} > \theta^*_{r_j^*}$, since $R_{ij} = \mathbb{I}\left\{\check{\theta}_i^{(k)} > \check{\theta}_j^{(k)}\right\}$, we will provide an upper bound for $\mathbb{P}\left(\check{\theta}_i^{(k)} < \check{\theta}_j^{(k)} \middle| \{\check{S}_k\}_{k\in[K]}\right)$.

To this end, we state a result that characterizes the performance of the MLE under a BTL model with bounded skill parameters. Consider a random graph with independent edges $B_{ij} \sim \text{Bernoulli}(p_{ij})$ for $1 \le i < j \le m$. For each $B_{ij} = 1$, observe i.i.d. $y_{ijl} \sim \text{Bernoulli}(\psi(\eta_i^* - \eta_j^*))$ for $l = 1, \cdots, L$. Let $\bar{y}_{ij} = \frac{1}{L}\sum_{l=1}^{L} y_{ijl}$, and we define the MLE by

$$\widehat{\eta} \in \operatorname{argmin} \sum_{1 \le i < j \le m} B_{ij} \left[\bar{y}_{ij} \log \frac{1}{\psi(\eta_i - \eta_j)} + (1 - \bar{y}_{ij}) \log \frac{1}{1 - \psi(\eta_i - \eta_j)}\right]. \tag{3.27}$$

Lemma 3.4.3. *Assume $\eta_1^* > \cdots > \eta_m^*$ and $\eta_1^* - \eta_m^* \le \kappa$. There exists some constant $c \in (0,1)$ such that $p_{ij} = p$ for all $|i-j| \le cm$ and $p_{ij} \le p$ otherwise. As long as $\frac{mp}{\log(m+n)} \to \infty$ and $\kappa = O(1)$, then for any $\delta > 0$ that is sufficiently small, there exists a constant $C > 0$ such that*

$$\mathbb{P}\left(\widehat{\eta}_i < \widehat{\eta}_j\right) \le C\left[\exp\left(-\frac{(1-\delta)L(\eta_i^* - \eta_j^*)^2}{2(W_i(\eta^*) + W_j(\eta^*))}\right) + n^{-7}\right],$$

for all $1 \le i < j \le m$, where $W_i(\eta^*) = \frac{1}{\sum_{j\in[m]\setminus\{i\}} p_{ij}\psi'(\eta_i^* - \eta_j^*)}$ for all $i \in [m]$.

The proof of Lemma 3.4.3, which relies on a recently developed leave-one-out technique in the analysis of the BTL model [25, 21], will be given in Section 3.7.4.

By conditioning on $\{\check{S}_k\}$, the statistical property of (3.26) is a direct consequence of Lemma 3.4.3. Note that $\mathbb{P}\left(\check{\theta}_i^{(k)} < \check{\theta}_j^{(k)} \middle| \{\check{S}_k\}_{k\in[K]}\right)$ is a function of $\{\check{S}_k\}_{k\in[K]}$, and we will establish a uniform upper bound for this conditional probability for any partition $\{\check{S}_k\}_{k\in[K]}$ satisfying the following conditions:

(i) For any $k \in [K]$ and any $i, j \in \check{S}_{k-1} \cup \check{S}_k \cup \check{S}_{k+1}$, we have $|\theta_{r_i^*}^* - \theta_{r_j^*}^*| \le C_1 M$;

(ii) For any $k \in [K]$ and any $i \in \check{S}_k$, we have $\left\{j \in [n] : |r_i^* - r_j^*| \le \frac{C_2 M}{\beta}\right\} \subset \check{S}_{k-1} \cup \check{S}_k \cup \check{S}_{k+1}$;

(iii) For any $k \in [K-1]$ and any $i \in \check{S}_{k-1} \cup \check{S}_k \cup \check{S}_{k+1} \cup \check{S}_{k+2}$, we have the relation $\left|\left\{j \in [n] : |\theta_{r_i^*}^* - \theta_{r_j^*}^*| \le \frac{M}{2}\right\} \cap (\check{S}_{k-1} \cup \check{S}_k \cup \check{S}_{k+1} \cup \check{S}_{k+2})\right| \ge C_3\left(\frac{M}{\beta} \wedge n\right)$.

Note that we use the convention $\check{S}_0 = \check{S}_{K+1} = \varnothing$ and C_1, C_2, C_3 are the same constants in Theorem 3.4.1. Consider any partition $\{\check{S}_k\}_{k\in[K]}$ satisfying the three conditions above. When applying Lemma 3.4.3, by Conditions (i) and (ii), we have $\kappa = 2C_1 M$ and $m = |\check{S}_{k-1} \cup \check{S}_k \cup \check{S}_{k+1} \cup \check{S}_{k+2}| \asymp \frac{1}{\beta} \wedge n$. We also know that for any $i, j \in \check{S}_{k-1} \cup \check{S}_k \cup \check{S}_{k+1} \cup \check{S}_{k+2}$ such that $|\theta_{r_i^*}^* - \theta_{r_j^*}^*| \le \frac{M}{2}$, we have $\check{A}_{ij} = A_{ij} \sim \text{Bernoulli}(p)$. Then, Condition (iii) implies the existence of a band in $\{(r_i^*, r_j^*) : i, j \in \check{S}_{k-1} \cup \check{S}_k \cup \check{S}_{k+1} \cup \check{S}_{k+2}\}$ with width at least

cm for some constant $c > 0$, such that $\check{A}_{ij} \sim \text{Bernoulli}(p)$ for all pairs in the band. For any other (i, j), we have $\check{A}_{ij} \sim \text{Bernoulli}(p_{ij})$ with $p_{ij} \leq p$. Having checked the conditions of Lemma 3.4.3, we obtain the following result for the local MLE (3.26),

$$\mathbb{P}\left(\check{\theta}_i^{(k)} < \check{\theta}_j^{(k)} \middle| \{\check{S}_k\}_{k\in[K]}\right) \leq C\left[\exp\left(-\frac{(1-\delta)npL(\theta^*_{r_i^*} - \theta^*_{r_j^*})^2}{2(V_{r_i^*}(\theta^*) + V_{r_j^*}(\theta^*))}\right) + n^{-7}\right], \tag{3.28}$$

for any $i, j \in \check{S}_k \cup \check{S}_{k+1}$ such that $\theta^*_{r_i^*} > \theta^*_{r_j^*}$. Recall the definition of $V_i(\theta^*)$ in (3.15). The constant δ in (3.28) can be made arbitrarily small with a sufficiently large M. To derive (3.28) from Lemma 3.4.3, we only need to show

$$p \sum_{j\in[n]\setminus\{i\}} \psi'(\theta_i^* - \theta_j^*) \leq \left(1 + O(e^{-C_2 M})\right) \sum_{j\in(\check{S}_{k-1}\cup\check{S}_k\cup\check{S}_{k+1}\cup\check{S}_{k+2})\setminus\{i\}} p_{ij}\psi'(\theta^*_{r_i^*} - \theta^*_{r_j^*}),$$

for all $i \in \check{S}_k \cup \check{S}_{k+1}$. This is true by a similar argument that leads to (3.14), together with Condition (ii). Finally, by Theorem 3.4.1, Conditions (i)-(iii) hold for $\{\check{S}_k\}_{k\in[K]}$ with high probability, and thus (3.28) is a high-probability bound. A similar bound to (3.28) also holds for (3.25) by the conclusion of Lemma 3.4.2.

3.4.4 Analysis of Algorithm 2

With the help of Lemma 3.4.1, Theorem 3.4.1, Lemma 3.4.2 and Lemma 3.4.3, we are ready to prove that Algorithm 2 achieves the minimax rate of full ranking.

Proof of Theorem 3.3.2 (upper bound). Let $\mathcal{G}$ be the event that the conclusions of Theorem 3.4.1 and Lemma 3.4.2 hold. We have $\mathbb{P}(\mathcal{G}^c) = O(n^{-8})$. In addition, we use the notation $\check{\mathcal{S}}$ for the event that $\{\check{S}_k\}_{k\in[K]}$ satisfies Conditions (i)-(iii) listed in Section 3.4.3. It is clear that $\mathcal{G} \subset \check{\mathcal{S}}$.

By Lemma 3.4.1, we have

$$\mathbb{E}\mathsf{K}(\widehat{r}, r^*) \leq \frac{4}{n} \sum_{1 \leq i \neq j \leq n} \mathbb{P}(R_{ij} \neq R^*_{ij}).$$

It suffices to give a bound for $\mathbb{P}(R_{ij} \neq R^*_{ij})$ for every pair $i \neq j$. Note that we have $\mathbb{P}(R_{ij} \neq R^*_{ij}) \leq \mathbb{P}\left(R_{ij} \neq R^*_{ij}, \mathcal{G}\right) + \mathbb{P}(\mathcal{G}^c)$. Then.

$$\begin{aligned}
\mathbb{P}(R_{ij} \neq R^*_{ij}, \mathcal{G}) &= \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbb{P}(R_{ij} \neq R^*_{ij}, \mathcal{G}, i \in S_k, j \in S_l) \\
&= \sum_{(k,l) \in [K]^2 : |k-l| \leq 1} \mathbb{P}(R_{ij} \neq R^*_{ij}, \mathcal{G}, i \in S_k, j \in S_l) \\
&\quad + \sum_{(k,l) \in [K]^2 : |k-l| \geq 2} \mathbb{P}(R_{ij} \neq R^*_{ij}, \mathcal{G}, i \in S_k, j \in S_l).
\end{aligned}$$

The second term above is zero. This is due to the analysis of Step 3 in Section 3.4.3 which shows $\sum_{(k,l) \in [K]^2 : |k-l| \geq 2} \mathbb{I}\{R_{ij} \neq R^*_{ij}, i \in S_k, j \in S_l\} = 0$ under the event $\mathcal{G}$. Hence, we only need to study the first term. Without loss of generality, consider $\theta_{r_i^*} > \theta^*_{r_j^*}$. Then, the event $\{R_{ij} \neq R^*_{ij}, \mathcal{G}, i \in S_k, j \in S_k\}$ is equivalent to $\{\widehat{\theta}_i^{(k)} < \widehat{\theta}_j^{(k)}, \mathcal{G}, i \in S_k, j \in S_k\}$, which

is further equivalent to $\{\check{\theta}_i^{(k)} < \check{\theta}_j^{(k)}, \mathcal{G}, i \in \check{S}_k, j \in \check{S}_k\}$ by the definition of $\mathcal{G}$. We thus have

$$\begin{aligned}
\mathbb{P}(R_{ij} \neq R_{ij}^*, \mathcal{G}) &= \sum_{(k,l)\in[K]^2:|k-l|\leq 1} \mathbb{P}(\check{\theta}_i^{(k)} < \check{\theta}_j^{(k)}, \mathcal{G}, i \in \check{S}_k, j \in \check{S}_l) \\
&\leq \sum_{(k,l)\in[K]^2:|k-l|\leq 1} \mathbb{P}(\check{\theta}_i^{(k)} < \check{\theta}_j^{(k)}, \check{\mathcal{S}}, i \in \check{S}_k, j \in \check{S}_l) \\
&= \sum_{(k,l)\in[K]^2:|k-l|\leq 1} \mathbb{P}\left(\check{\theta}_i^{(k)} < \check{\theta}_j^{(k)} \middle| \check{\mathcal{S}}, i \in \check{S}_k, j \in \check{S}_l\right) \mathbb{P}\left(\check{\mathcal{S}}, i \in \check{S}_k, j \in \check{S}_l\right) \\
&\leq C \left[e^{-\frac{(1-\delta)npL(\theta^*_{r_i^*} - \theta^*_{r_j^*})^2}{2(V_{r_i^*}(\theta^*) + V_{r_j^*}(\theta^*))}} + n^{-7} \right] \sum_{(k,l)\in[K]^2:|k-l|\leq 1} \mathbb{P}\left(\check{\mathcal{S}}, i \in \check{S}_k, j \in \check{S}_l\right) \\
&\leq C \left[\exp\left(-\frac{(1-\delta)npL(\theta^*_{r_i^*} - \theta^*_{r_j^*})^2}{2(V_{r_i^*}(\theta^*) + V_{r_j^*}(\theta^*))} \right) + n^{-7} \right],
\end{aligned}$$

for some constant $C > 0$ and some $\delta > 0$ that is arbitrarily small. The second last inequality above is by Lemma 3.4.3, or more specifically, (3.28), as we show (3.28) holds for any $\{\check{S}_k\}_{k\in[K]}$ satisfying Conditions (i)-(iii) listed in Section 3.4.3. Since $\mathbb{P}(\mathcal{G}^c) = O(n^{-8})$, we obtain the bound

$$\mathbb{P}(R_{ij} \neq R_{ij}^*) \leq 2C \left[\exp\left(-\frac{(1-\delta)npL(\theta^*_{r_i^*} - \theta^*_{r_j^*})^2}{2(V_{r_i^*}(\theta^*) + V_{r_j^*}(\theta^*))} \right) + n^{-7} \right], \tag{3.29}$$

for all $i \neq j$.

Summing the bound (3.29) over all $i \neq j$, we have

$$\begin{aligned}
\mathbb{E}\mathsf{K}(\widehat{r}, r^*) &\leq \frac{8C}{n} \sum_{1\leq i\neq j\leq n} \exp\left(-\frac{(1-\delta)npL(\theta^*_{r_i^*} - \theta^*_{r_j^*})^2}{2(V_{r_i^*}(\theta^*) + V_{r_j^*}(\theta^*))} \right) + 8Cn^{-6} \\
&= \frac{8C}{n} \sum_{1\leq i\neq j\leq n} \exp\left(-\frac{(1-\delta)npL(\theta^*_i - \theta^*_j)^2}{2(V_i(\theta^*) + V_j(\theta^*))} \right) + 8Cn^{-6}.
\end{aligned} \tag{3.30}$$

Now it is just a matter of simplifying the expression (3.30). We consider the following two cases: $\frac{Lp\beta^2}{\beta\vee n^{-1}} \le 1$ and $\frac{Lp\beta^2}{\beta\vee n^{-1}} > 1$.

First, we consider the case $\frac{Lp\beta^2}{\beta\vee n^{-1}} \le 1$. By Lemma 3.7.6 proved in Section 3.7.2, there exist constants $c_1, c_2 > 0$, such that

$$c_1\left(\beta \vee \frac{1}{n}\right) \le \frac{V_i(\theta^*)}{n} \le c_2\left(\beta \vee \frac{1}{n}\right), \tag{3.31}$$

for all $\theta^* \in \Theta_n(\beta, C_0)$ and all $i \in [n]$. Then, for each $i \in [n]$,

$$\begin{aligned}
\sum_{j\in[n]\setminus\{i\}} \exp\left(-\frac{(1-\delta)npL(\theta_i^*-\theta_j^*)^2}{2(V_i(\theta^*)+V_j(\theta^*))}\right) &\le \sum_{j\in[n]\setminus\{i\}} \exp\left(-\frac{1}{3c_2}(i-j)^2\frac{Lp\beta^2}{\beta\vee n^{-1}}\right) \\
&\le \int_0^\infty \exp\left(-\frac{1}{3c_2}x^2\frac{Lp\beta^2}{\beta\vee n^{-1}}\right)dx \\
&= \sqrt{\frac{3\pi c_2}{4}}\sqrt{\frac{\beta\vee n^{-1}}{Lp\beta^2}},
\end{aligned}$$

and we have $\mathbb{E}\mathsf{K}(\widehat{r}, r^*) \lesssim \sqrt{\frac{\beta\vee n^{-1}}{Lp\beta^2}}$. The definition of the loss function implies $\mathbb{E}\mathsf{K}(\widehat{r}, r^*) \le n$, and thus we obtain the rate $n \wedge \sqrt{\frac{\beta\vee n^{-1}}{Lp\beta^2}}$ when $\frac{Lp\beta^2}{\beta\vee n^{-1}} \le 1$.

Next, we consider the case $\frac{Lp\beta^2}{\beta\vee n^{-1}} > 1$. For any $|i-j| \le C_0\sqrt{c_2/c_1}$, we have $V_j(\theta^*) \le (1+\delta')V_i(\theta^*)$ for some $\delta' = o(1)$. This is by the definition of the variance function and the fact that $\sup_x \left|\frac{\psi'(x+\Delta)}{\psi'(x)} - 1\right| \lesssim |\Delta|$ for $\Delta = o(1)$. Therefore, we have

$$\begin{aligned}
&\sum_{1\le i\ne j\le n:|i-j|\le C_0\sqrt{c_2/c_1}} \exp\left(-\frac{(1-\delta)npL(\theta_i^*-\theta_j^*)^2}{2(V_i(\theta^*)+V_j(\theta^*))}\right) \\
\lesssim\ &\sum_{i=1}^{n-1} \exp\left(-\frac{(1-2\delta)npL(\theta_i^*-\theta_{i+1}^*)^2}{4V_i(\theta^*)}\right),
\end{aligned}$$

By (3.31), we also have

$$\begin{aligned}
& \sum_{1\le i\neq j\le n:|i-j|>C_0\sqrt{c_2/c_1}} \exp\left(-\frac{(1-2\delta)npL(\theta_i^*-\theta_j^*)^2}{2(V_i(\theta^*)+V_j(\theta^*))}\right) \\
\lesssim & \sum_{1\le i\neq j\le n:|i-j|>C_0\sqrt{c_2/c_1}} \exp\left(-\frac{(1-2\delta)pL\beta^2(i-j)^2}{2c_2(\beta\vee n^{-1})}\right) \\
\lesssim & \; n\exp\left(-\frac{(1-2\delta)pL\beta^2C_0^2}{2c_1(\beta\vee n^{-1})}\right) \\
\lesssim & \sum_{i=1}^{n-1}\exp\left(-\frac{(1-2\delta)npL(\theta_i^*-\theta_{i+1}^*)^2}{4V_i(\theta^*)}\right).
\end{aligned}$$

The desired bound for $\mathbb{E}\mathsf{K}(\widehat{r},r^*)$ immediately follows by summing up the above bounds. □

3.4.5 A Data-Driven h.

Our proposed algorithm relies on a tuning parameter $h=\frac{pM}{\beta}$ that is unknown in practice. This quantity can be replaced by a data-driven version, defined as

$$\widehat{h}=\frac{1}{n}\sum_{1\le i<j\le n}A_{ij}\mathbb{I}\{1.2M\le|\psi^{-1}(\bar{y}_{ij}^{(1)})|\le 1.8M\}. \tag{3.32}$$

A standard concentration result implies that $\widehat{h}\asymp\frac{pM}{\beta}$ with high probability. Moreover, by defining

$$\check{h}=\frac{1}{n}\sum_{\substack{1\le i<j\le n\\ 1.1M<|\theta^*_{r_i^*}-\theta^*_{r_j^*}|<1.9M}}A_{ij}\mathbb{I}\{1.2M\le|\psi^{-1}(\bar{y}_{ij}^{(1)})|\le 1.8M\},$$

it can be shown that $\widehat{h}=\check{h}$ with high probability. Since $\check{h}$ is measurable with respect to the σ-algebra generated by $\{(A_{ij},\bar{y}_{ij}^{(1)}):1.1M<|\theta^*_{r_i^*}-\theta^*_{r_j^*}|<1.9M\}$, we still have the asymptotic independence property between the league partition and local MLE after h being replaced by $\widehat{h}$ in Algorithm 1. Therefore, with a data-driven $\widehat{h}$ being used in the proposed

algorithm, the upper bound conclusion of Theorem 3.3.2 still holds.

3.5 Numerical Results

In this section, we conduct numerical experiments to study the statistical and computational properties of Algorithm 2.

Simulation Setting. In our experiment, we consider $\theta^* \in \mathbb{R}^n$ with $n = 1000$. In particular, we set $\theta_i^* = -\beta i$ for all $i \in [n]$ with some $\beta \in [0.001, 0.05]$. The range of β implies that the dynamic range $\theta_1^* - \theta_{1000}^*$ takes value in $[0.999, 49.95]$. We assume the true rank is the identity permutation, i.e., $r_i^* = i$ for all $i \in [n]$. We also consider three different (L, L_1) pairs: (50, 10), (75, 15), (100, 20) in Algorithm 2.

Implementation. In the implementation of Algorithm 2, we set $M = 5$. For the choice of h, though the recommended data-driven estimator (3.32) works for the theoretical purpose, it may not be a sensible choice for a data set with a moderate size. Note that with $M = 5$, we have $\psi(1.2M) = 0.9975274$ and $\psi(1.8M) = 0.9998766$, respectively, and thus the indicator $\mathbb{I}\{1.2M \le |\psi^{-1}(\bar{y}_{ij}^{(1)})| \le 1.8M\}$ is usually zero in (3.32). To address this issue, we set h by

$$h = 0.4 \times \frac{1}{n} \sum_{1 \le i < j \le n} A_{ij} \mathbb{I}\Big\{\psi(-M) \le \bar{y}_{ij}^{(1)} \le \psi(M)\Big\}.$$

The computation of the local MLE (3.25) is implemented by the MM algorithm [60]. All simulations are implemented in Python (along with NumPy package, whose backend is written in C) using a 2019 MacBook Pro, 15-inch, 2.6GHz 6-core Intel Core i7.

Accuracy of League Partition. We first study Algorithm 1, which is Step 1 of Algorithm 2. The purpose of Algorithm 1 is to divide all players into K leagues. The average value of K from 50 independent experiments is reported in Figure 3.5. This number increases with

β linearly, which agrees with our theoretical bound $K = O_{\mathbb{P}}(n\beta \vee 1)$.

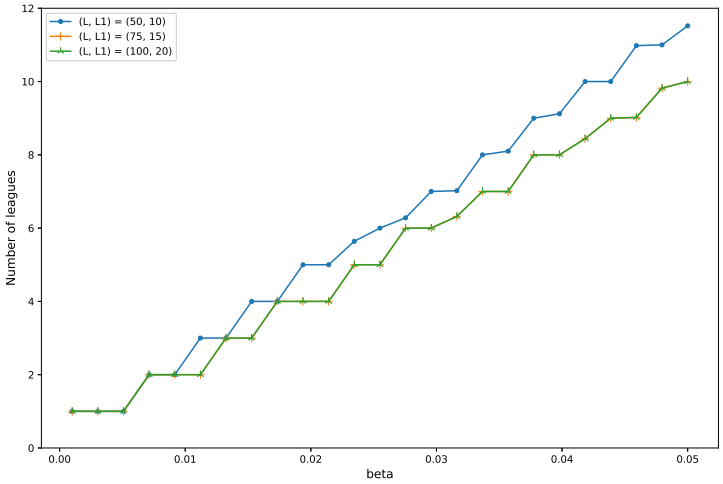

Figure 3.5: *The number of leagues obtained by Algorithm 1. The orange curve is mostly overlapped by the green curve.*

To quantify the accuracy of Algorithm 1, we define the following metric,

$$E_{partition} = \begin{cases} \frac{1}{K-2}\sum_{k=2}^{K-1} \mathbb{I}\left\{\max\left\{r_i^* : i \in \cup_{k'<k} S_{k'}\right\} > \min\left\{r_i^* : i \in \cup_{k'>k} S_{k'}\right\}\right\}, & K \geq 3, \\ 0, & K < 3. \end{cases}$$

The quantity $E_{partition}$ is essentially designed to verify the Conclusion 3 in Theorem 3.4.1, and we expect that $E_{partition}$ should be 0 with high probability. Note that Conclusion 3 of Theorem 3.4.1 guarantees the correctness of the cross-league pairwise relation estimation, which is Step 3 of Algorithm 2. For each combination of (β, L, L_1), we generate independent data and repeat the experiments 50 times. It turns out that $E_{partition}$ is always 0, which agrees with the theoretical property of the league partition.

Statistical Error. Next, we study the ranking error of the proposed divide-and-conquer algorithm (Algorithm 2) under the Kendall's tau distance defined by (3.3). For comparison, we also implement the global MLE and the spectral method. The MLE outputs the rank of

the entries of $\widehat{\theta}$ that maximizes the negative log-likelihood function

$$\sum_{1\le i<j\le n} A_{ij}\left[\bar{y}_{ij}\log\frac{1}{\psi(\theta_i-\theta_j)}+(1-\bar{y}_{ij})\log\frac{1}{1-\psi(\theta_i-\theta_j)}\right], \tag{3.33}$$

where $\bar{y}_{ij}=\frac{1}{L}\sum_{l=1}^{L} y_{ijl}$. The spectral method, also known as Rank Centrality, is a ranking algorithm proposed by [86]. Define a matrix $P\in\mathbb{R}^{n\times n}$ by

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

where d is set to be twice the maximum degree of the random graph A. Note that P is the transition matrix of a Markov chain. Let $\widehat{\pi}$ be the stationary distribution of this Markov chain, and the spectral method outputs the rank of the entries of the vector $\widehat{\pi}$.

Both the MLE and the spectral method have been studied for parameter estimation [86, 25] and top-k ranking [25, 21] under the BTL model. However, to the best of our knowledge, the statistical properties of the two methods for full ranking have not been studied in the literature. The recent work [25] has established the estimation errors of the skill parameter for both the MLE and the spectral method. Their results involve a factor of $e^{O(n\beta)}$ in the estimation error under an ℓ_∞ loss, which suggests that the MLE and the spectral method may not perform well when the dynamic range $n\beta$ diverges.

We implement the MLE, the spectral method, and the divide-and-conquer algorithm for various combinations of β and L. The results of each setting are computed by averaging across 50 independent experiments. As shown in Figure 3.6, the spectral method is significantly worse than the MLE and the divide-and-conquer algorithm. The performance of the spectral method may be explained by the $e^{O(n\beta)}$ factor in the ℓ_∞ norm error bound obtained by [25], though the exact relation between the ℓ_∞ error and the full ranking error is not clear to us. On the other hand, the error curves of the MLE and the divide-and-

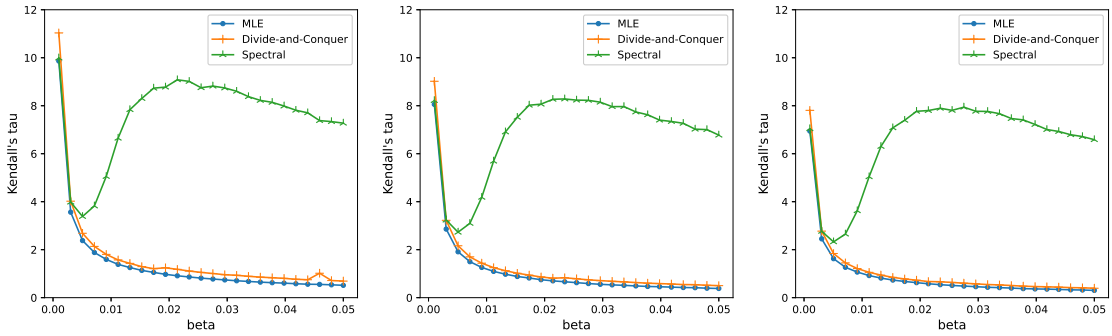

Figure 3.6: *Statistical error under Kendall's tau. Left:* $(L, L_1) = (50, 10)$*; Middle:* $(L, L_1) = (75, 15)$*; Right:* $(L, L_1) = (100, 20)$.

conquer algorithm are very close. Since the divide-and-conquer algorithm has been proved to be minimax optimal, the simulation results suggest that the MLE may also enjoy such statistical optimality.

The current analysis of the MLE [25, 21] crucially depends on the spectral property of the Hessian matrix $H(\theta^*)$ of the objective (3.33). It is known that the condition number of $H(\theta^*)$ on the subspace orthogonal to $\mathbb{1}_n$ is of order $e^{O(n\beta)}$, which explains the $e^{O(n\beta)}$ factor in the ℓ_∞ estimation error of the MLE [25]. However, our simulation study reveals that the error bound of [25] can be potentially loose. The definition of the Kendall's tau distance suggests that a sharp analysis of the MLE requires a careful study of the random variable $\widehat{\theta}_{r_i^*} - \widehat{\theta}_{r_j^*}$. We conjecture that the variance of $\widehat{\theta}_{r_i^*} - \widehat{\theta}_{r_j^*}$ should be approximately proportional to $(e_{r_i^*} - e_{r_j^*})^T H(\theta^*)^\dagger (e_{r_i^*} - e_{r_j^*})$, where e_j is the jth canonical vector with all entries being 0 except that the jth entry is 1. Since $H(\theta^*)$ can be viewed as the graph Laplacian of some random weighted graph, there may exist random matrix tools to study $(e_{r_i^*} - e_{r_j^*})^T H(\theta^*)^\dagger (e_{r_i^*} - e_{r_j^*})$ directly without using the naive condition number bound, and we leave this interesting direction as a future project.

In comparison, our divide-and-conquer algorithm does not need to solve the global MLE. Since the objective function of each local MLE is well conditioned (Lemma 3.7.14), Algorithm 2 is provably optimal in addition to its good performance in simulation.

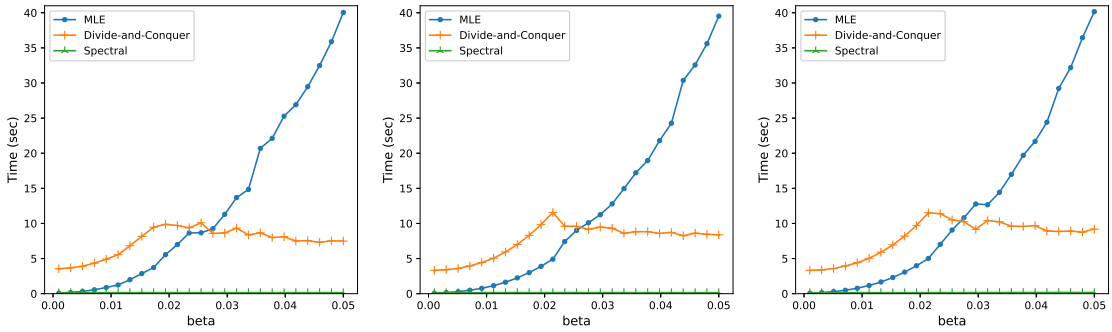

Figure 3.7: *Running time comparison. Left:* $(L, L_1) = (50, 10)$*; Middle:* $(L, L_1) = (75, 15)$*; Right:* $(L, L_1) = (100, 20)$.

Computational Cost. Finally, we compare the computational costs of the three methods. The average time needed to run the three algorithms is given in Figure 3.7. The spectral method, though suffers from its unsatisfactory statistical error, is the fastest, partly because finding the stationary distribution is just a single line of code using a NumPy function whose backend is C. The running time of the MLE grows rapidly as β increases. This can be explained by the growing condition number of the Hessian matrix $H(\theta^*)$. While the condition number may not affect the statistical error of the MLE, it does have a rather strong effect on its computational cost. On the other hand, the running time for the divide-and-conquer method (Algorithm 2) first increases with β, and then stabilizes. This is the effect of Algorithm 1, which divides a large difficult problem into many small sub-problems, and after that each small sub-problem can be conquered efficiently. In fact, we can further improve the computational efficiency by solving the sub-problems in parallel. The initial increase of the running time of Algorithm 2 is because of the additional league partition step. Recall that the league partition step divides the players into $K = O_{\mathbb{P}}(n\beta \vee 1)$ subsets. When β is small, we have a very small K. According to the formula (3.24), the local MLE is as difficult as the global MLE whenever $K \leq 4$. In this regime, the divide-and-conquer method is more time consuming because of the additional league partition step. On the other hand, as β grows, the computational advantage of the divide-and-conquer strategy

becomes significant. This makes our proposed algorithm scalable to large data sets, while preserving the statistical optimality, which concludes the divide-and-conquer algorithm as the best overall method among the three.

3.6 Discussion

In this paper, the problem of ranking n players from partial comparison data under the BTL model has been investigated. We have derived the minimax rate with respect to the Kendall's tau distance. A divide-and-conquer algorithm is proposed and is proved to achieve the minimax rate. In this section, we discuss a few directions along which the results of the paper can be extended.

The first extension one can consider is to assume that $A_{ij} \sim \text{Bernoulli}(p_{ij})$ independently for all $1 \le i < j \le n$. For this more general comparison graph, as long as we assume that all p_{ij}'s are of the same order in the sense that $\max_{ij} p_{ij} \le C \min_{ij} p_{ij}$ for some constant $C > 0$, Theorem 3.3.2 continues to hold with $\frac{V_i(\theta^*)}{np}$ replaced by

$$\frac{1}{\sum_{j \in [n] \setminus \{i\}} p_{ij} \psi'(\theta_i^* - \theta_j^*)},$$

and all the technical arguments in the proofs will still go through.

Another important condition that we impose throughout the paper is the regularity of the skill parameters $\theta^* \in \Theta_n(\beta, C_0)$. It assumes that $|\theta_i^* - \theta_j^*| \asymp \beta |i - j|$, which roughly describes that players with different skills are evenly distributed in the population. Without this condition, we conjecture that the minimax rate under the Kendall's tau loss should be

$$\inf_{\widehat{r} \in \mathfrak{S}_n} \sup_{r^* \in \mathfrak{S}_n} \mathbb{E}_{(\theta^*, r^*)} \mathsf{K}(\widehat{r}, r^*) \asymp \frac{1}{n} \sum_{1 \le i < j \le n} \exp\left(- \frac{(1 + o(1)) npL(\theta_i^* - \theta_j^*)^2}{2(V_i(\theta^*) + V_j(\theta^*))} \right).$$

In fact, this formula has already appeared in the upper bound analysis (3.30) and can be

simplified to the result of Theorem 3.3.2 when $\theta^* \in \Theta_n(\beta, C_0)$. Extending the result of Theorem 3.3.2 beyond the condition $\theta^* \in \Theta_n(\beta, C_0)$ is possible by some necessary modifications of the league partition step described in Algorithm 1. Without $|\theta_i^* - \theta_j^*| \asymp \beta|i-j|$, the partition formula $S_k = \{i \in [n] \backslash (S_1 \cup \cdots \cup S_{k-1}) : w_i^{(k)} \leq h\}$ should be replaced by $S_k = \{i \in [n] \backslash (S_1 \cup \cdots \cup S_{k-1}) : w_i^{(k)} \leq h_k\}$ for some sequence $\{h_k\}$ to account for the non-regularity of θ^*. Intuitively, the size of each $|S_k|$ should adaptively depend on the local density of the skill parameters in the neighborhood from which it is selected. Then, the major difficulty is to find a data-driven $\{\widehat{h}_k\}$ that estimates the local density. When $|\theta_i^* - \theta_j^*| \asymp \beta|i-j|$, we can just use the global estimator (3.32). Without this assumption, estimating $\{h_k\}$ is a much harder problem. In [61], it is assumed that the skill parameters $\theta_1^*, \cdots, \theta_n^*$ are i.i.d drawn from some distribution F instead of being fixed parameters, and the authors have studied the problem of estimating F, which is called the skill distribution, from the partial pairwise comparison data. Under this formulation, the estimation of the parameters $\{h_k\}$ can be linked to the problem of local bandwidth selection in kernel density estimation [64]. We leave this direction of research as one of our future projects.

A restriction of the BTL model is that it can only deal with pairwise comparison. One extension from pairwise comparison to multiple comparison is the popular Plackett-Luce model [89, 77]. Suppose there is a subset of J players $S = \{i_1, i_2, \cdots, i_J\}$. Under the Plackett-Luce model, the probability that j is selected among S is given by the formula $\frac{\exp(\theta_j)}{\sum_{i \in S} \exp(\theta_i)}$. Statistical analysis of ranking under the Plackett-Luce model is a problem that has been rarely explored. Both the minimax rate and the construction of optimal algorithms are important open problems.

The ranking problem has also been studied under nonparametric comparison models. For example, a nonparametric stochastically transitive model was proposed by [96, 97] and the problems of estimating the mean matrix and top-k ranking have been investigated. However, full ranking is still a problem that has not been well studied under nonparametric models.

One of the few works that we are aware of is [80] that assumes $\mathbb{P}(y_{ijl} = 1) > \frac{1}{2} + \gamma$ when $r_i^* < r_j^*$. An investigation of full ranking under more general nonparametric settings is another direction to be explored.

3.7 Proofs

3.7.1 Proof of Theorem 3.3.1

We prove Theorem 3.3.1 in this section. We first state and prove a few lemmas.

Lemma 3.7.1. *Assume $p \geq \frac{c_0 \log n}{n}$ for some sufficiently large $c_0 > 0$. Then, we have*

$$\|A - \mathbb{E}(A)\|_{\text{op}} \leq C\sqrt{np}, \tag{3.34}$$

$$\|D - \mathbb{E}(D)\|_{\text{op}} \leq C\sqrt{np \log n} \tag{3.35}$$

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

Proof. Bound (3.34) is a direct consequence of Theorem 5.2 in [70] and Bound (3.35) is from standard concentration of sums of i.i.d. Bernoulli random variables. □

Lemma 3.7.2. *Assume $p \geq \frac{c_0 \log n}{n}$ for some sufficiently large $c_0 > 0$. Then, we have*

$$np - 2C\sqrt{np \log n} \leq \lambda_{\min,\perp}(\mathcal{L}_A) = \min_{u \neq 0, \mathbb{1}_n^T u = 0} \frac{u^T \mathcal{L}_A u}{\|u\|},$$

$$np + 2C\sqrt{np \log n} \geq \lambda_{\max,\perp}(\mathcal{L}_A) = \max_{u \neq 0, \mathbb{1}_n^T u = 0} \frac{u^T \mathcal{L}_A u}{\|u\|}$$

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

Proof. Note the decomposition

$$\mathcal{L}_A = \mathbb{E}\mathcal{L}_A + D - \mathbb{E}D - (A - \mathbb{E}A)$$

and $\lambda_{\min,\perp}(\mathbb{E}\mathcal{L}_A) = \lambda_{\max,\perp}(\mathbb{E}\mathcal{L}_A) = np$. By Lemma 3.7.1, we have

$$\|D - \mathbb{E}D - (A - \mathbb{E}A)\|_{\text{op}} \leq 2C\sqrt{np\log n}$$

with probability at least $1 - O(n^{-10})$ for some $C > 0$. The Lemma can be seen immediately by Weyl's inequality. □

We introduce another notation $r^{*(i,j)} \in \mathfrak{S}_n$ to be the element in $\mathfrak{S}_n$ having

$$r_k^{*(i,j)} = \begin{cases} r_k^*, \text{ if } k \neq i, j \\ r_j^*, \text{ if } k = i \\ r_i^*, \text{ if } k = j \end{cases} . \tag{3.36}$$

That is, $r^{*(i,j)}$ is a permutation by swapping the i, jth position in r^* while keeping other positions fixed.

Lemma 3.7.3. *Assume $\frac{np}{\log n} \to \infty$. There exists $\delta = o(1)$, such that for any $\theta^* \in \Theta_n(\beta, C_0)$, any $r^* \in \mathfrak{S}_n$, any $i, j \in [n], i \neq j$, we have*

$$\begin{aligned} &\inf_{\widehat{r}} \frac{\mathbb{P}_{(\theta^*,\sigma^2,r^*)}\left(\widehat{r} \neq r^*\right) + \mathbb{P}_{(\theta^*,\sigma^2,r^{*(i,j)})}\left(\widehat{r} \neq r^{*(i,j)}\right)}{2} \\ &\gtrsim \min\left\{1, \sqrt{\frac{\sigma^2}{np(\theta^*_{r_i^*} - \theta^*_{r_j^*})^2}} \exp\left(-\frac{(1+\delta)np(\theta^*_{r_i^*} - \theta^*_{r_j^*})^2}{4\sigma^2}\right)\right\} \end{aligned}$$

Proof. Assume $r_i^* = a < r_j^* = b$ and thus $\theta_a^* \geq \theta_b^*$. Let $\mathcal{F}$ be the event about A on which Lemma 3.7.1 holds. We have $\mathbb{P}(\mathcal{F}) > 1/2$. To simplify notation, let $\mathbb{P}_A(\cdot) = \mathbb{P}_{(\theta^*,\sigma^2,r^*)}(\cdot|A)$ be the conditional probability. For any A, by Neyman-Pearson Lemma, the optimal proce-

dure is given by the likelihood ratio test. Then

$$
\begin{aligned}
&\inf_{\widehat{r}} \frac{\mathbb{P}_{(\theta^*,\sigma^2,r^*)}\left(\widehat{r} \neq r^*\right) + \mathbb{P}_{(\theta^*,\sigma^2,r^{*(i,j)})}\left(\widehat{r} \neq r^{*(i,j)}\right)}{2} \\
&\geq \mathbb{P}\left(\mathcal{F}\right) \inf_{A\in\mathcal{F}} \mathbb{P}_A\left(\frac{d\mathbb{P}_{(\theta^*,\sigma^2,r^{*(i,j)})}}{d\mathbb{P}_{(\theta^*,\sigma^2,r^*)}} \geq 1\right) \\
&\gtrsim \inf_{A\in\mathcal{F}} \mathbb{P}_A(-4A_{ij}(\theta_a^* - \theta_b^* + w_{ij}) + \sum_{k\neq i,j} -A_{ik}(\theta_a^* - \theta_b^* + 2w_{ik}) \\
&\qquad\qquad + \sum_{k\neq i,j} A_{jk}(\theta_b^* - \theta_a^* + 2w_{jk}) \geq 0) \\
&= \inf_{A\in\mathcal{F}} \mathbb{P}_A\left(\mathcal{N}(0, \frac{\sigma^2}{D_{ii}+D_{jj}+2A_{ij}}) \geq \frac{|\theta_a - \theta_b|}{2}\right) \\
&\gtrsim \min\left\{1, \sqrt{\frac{\sigma^2}{np(\theta_a^* - \theta_b^*)^2}} \exp\left(-\frac{(1+\delta)np(\theta_a^* - \theta_b^*)^2}{4\sigma^2}\right)\right\} \qquad (3.37)
\end{aligned}
$$

for some $\delta = o(1)$, where (3.37) comes from standard Gaussian tail bound and Lemma 3.7.1. □

Now we are ready to state the proof of Theorem 3.3.1.

Proof of Theorem 3.3.1. We prove the theorem for any $\theta^* \in \Theta_n(\beta, C_0)$. Note that conditional on A, the solution of the least squares problem (3.10) can be written as

$$\widehat{\theta} = c\mathbb{1}_n + \theta_{r^*}^* + Z,$$

where $\theta_{r^*}^* = (\theta_{r_1^*}^*, ..., \theta_{r_n^*}^*)^T$, $Z \sim \mathcal{N}(0, \sigma^2\mathcal{L}_A^\dagger)$ and $c\mathbb{1}_n$ is a global shift of the skill parameters. Let $x_{ij} = e_i - e_j$ where $\{e_1, ..., e_n\}$ are the standard basis of $\mathbb{R}^n$. Let $\mathcal{F}$ be the event about

A when Lemma 3.7.2 holds. Then

$$
\begin{aligned}
&\mathbb{E}_{(\theta^*,\sigma^2,r^*)}\left[\mathsf{K}(\widehat{r},r^*)\right] = \frac{1}{n}\sum_{1\le i<j\le n}\mathbb{P}_{(\theta^*,\sigma^2,r^*)}\left(\mathrm{sign}(\widehat{r}_i-\widehat{r}_j)\mathrm{sign}(r_i^*-r_j^*)<0\right)\\
&=\frac{1}{n}\sum_{1\le i<j\le n}\mathbb{P}_{(\theta^*,\sigma^2,r^*)}\left(\mathrm{sign}(\widehat{\theta}_i-\widehat{\theta}_j)\mathrm{sign}(r_i^*-r_j^*)>0\right)\\
&\le\frac{1}{n}\sum_{1\le i<j\le n}\sup_{A\in\mathcal{F}}\mathbb{P}\left(\mathcal{N}(0,\sigma^2x_{ij}^T\mathcal{L}_A^\dagger x_{ij})>\left|\theta^*_{r_i^*}-\theta^*_{r_j^*}\right||A\right)+O(n^{-9})\\
&\le\frac{1}{n}\sum_{1\le i<j\le n}\sup_{A\in\mathcal{F}}\min\left\{1,\sqrt{\frac{\sigma^2x_{ij}^T\mathcal{L}_A^\dagger x_{ij}}{2\pi(\theta^*_{r_i^*}-\theta^*_{r_j^*})^2}}\exp\left(-\frac{(\theta^*_{r_i^*}-\theta^*_{r_j^*})^2}{2\sigma^2x_{ij}^T\mathcal{L}_A^\dagger x_{ij}}\right)\right\}+O(n^{-9})\\
&\le\frac{1}{n}\sum_{1\le i<j\le n}\min\left\{1,\sqrt{\frac{\sigma^2(np-2C\sqrt{np\log n})^{-1}}{\pi(\theta^*_{r_i^*}-\theta^*_{r_j^*})^2}}\exp\left(-\frac{(np-2C\sqrt{np\log n})(\theta^*_{r_i^*}-\theta^*_{r_j^*})^2}{4\sigma^2}\right)\right\}\qquad(3.38)\\
&\quad+O(n^{-9})\\
&\lesssim\frac{1}{n}\sum_{1\le i<j\le n}\min\left\{1,\sqrt{\frac{\sigma^2}{np(\theta_i^*-\theta_j^*)^2}}\exp\left(-\frac{(1-\delta_1')np(\theta_i^*-\theta_j^*)^2}{4\sigma^2}\right)\right\}+n^{-9}
\end{aligned}
$$

for some $\delta_1'=o(1)$ independent of θ^*, σ^2 and r^*, where (3.38) is due to Lemma 3.7.2.

We first consider the high signal-to-noise ratio regime, where $\frac{np\beta^2}{\sigma^2} > 1$. In this scenario,

$$
\begin{aligned}
&\sum_{1\le i<j\le n} \min\left\{1, \sqrt{\frac{\sigma^2}{np(\theta_i^*-\theta_j^*)^2}}\exp\left(-\frac{(1-\delta_1')np(\theta_i^*-\theta_j^*)^2}{4\sigma^2}\right)\right\} \\
&\le \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \exp\left(-\frac{(1-\delta_1')np(\theta_i^*-\theta_j^*)^2}{4\sigma^2}\right) \\
&\le \sum_{i=1}^{n-1} \exp\left(-\frac{(1-\delta_1')np(\theta_i^*-\theta_{i+1}^*)^2}{4\sigma^2}\right) \sum_{j=i+1}^{n} \exp\left(-\frac{(1-\delta_1')np[(\theta_i^*-\theta_j^*)^2-(\theta_i^*-\theta_{i+1}^*)^2]}{4\sigma^2}\right) \\
&\le \sum_{i=1}^{n-1} \exp\left(-\frac{(1-\delta_1')np(\theta_i^*-\theta_{i+1}^*)^2}{4\sigma^2}\right) \sum_{j=i+1}^{n} \exp\left(-\frac{(1-\delta_1')np(j-i-1)\beta^2}{4\sigma^2}\right) \\
&\lesssim \sum_{i=1}^{n-1} \exp\left(-\frac{(1-\delta_1')np(\theta_i^*-\theta_{i+1}^*)^2}{4\sigma^2}\right)
\end{aligned}
$$

where the last inequality is due to summation of an exponentially decaying series. This gives the exponential rate in high signal-to-noise ratio regime.

Now, when $\frac{np\beta^2}{\sigma^2} \le 1$,

$$
\begin{aligned}
&\sum_{1\le i<j\le n} \min\left\{1, \sqrt{\frac{\sigma^2}{np(\theta_i^*-\theta_j^*)^2}}\exp\left(-\frac{(1-\delta_1')np(\theta_i^*-\theta_j^*)^2}{4\sigma^2}\right)\right\} \\
&\le \sum_{i=1}^{n-1}\sum_{k\ge 1} \sum_{\substack{j>i \\ (k-1)\sqrt{\frac{\sigma^2}{np\beta^2}}<j-i\le k\sqrt{\frac{\sigma^2}{np\beta^2}}}} \min\left\{1, \sqrt{\frac{\sigma^2}{np(\theta_i^*-\theta_j^*)^2}}\exp\left(-\frac{(1-\delta')np(\theta_i^*-\theta_j^*)^2}{4\sigma^2}\right)\right\} \\
&\lesssim \sqrt{\frac{\sigma^2}{np\beta^2}} \sum_{i=1}^{n-1}\left(\sum_{k\ge 0}\exp\left(-\frac{(1-\delta')k^2}{4}\right)\right) \lesssim n\sqrt{\frac{\sigma^2}{np\beta^2}} \wedge n^2
\end{aligned}
$$

where the last inequality also comes from summing an exponentially decaying series and n^2 is a trivial upper bound. This finishes the proof of the upper bound.

Now we look at the lower bound. For any $r^* \in \mathfrak{S}_n$, we have $r^{*(i,j)} \in \mathfrak{S}_n$ defined as in

(3.36). Then for any $\theta^* \in \Theta_n(\beta, C_0)$,

$$
\begin{aligned}
&\inf_{\widehat{r}} \sup_{r^* \in \mathfrak{S}_n} \mathbb{E}_{(\theta^*, \sigma^2, r^*)} \left[\mathsf{K}(\widehat{r}, r^*) \right] \\
&\geq \inf_{\widehat{r}} \frac{1}{n} \sum_{1 \leq i < j \leq n} \frac{1}{n!} \sum_{r^* \in \mathfrak{S}_n} \mathbb{P}_{(\theta^*, \sigma^2, r^*)} \left(\mathrm{sign}(\widehat{r}_i - \widehat{r}_j) \mathrm{sign}(r_i^* - r_j^*) < 0 \right) \\
&= \inf_{\widehat{r}} \frac{1}{n} \sum_{1 \leq i < j \leq n} \frac{1}{n!} \sum_{1 \leq a < b \leq n} \sum_{r^* : \{r_i^*, r_j^*\} = \{a, b\}} \mathbb{P}_{(\theta^*, \sigma^2, r^*)} \left(\mathrm{sign}(\widehat{r}_i - \widehat{r}_j) \mathrm{sign}(r_i^* - r_j^*) < 0 \right) \\
&\geq \frac{1}{n} \sum_{1 \leq i < j \leq n} \frac{2}{n(n-1)} \sum_{1 \leq a < b \leq n} \frac{1}{(n-2)!} \\
&\qquad \sum_{r^* : r_i^* = a, r_j^* = b} \inf_{\widehat{r}} \frac{\mathbb{P}_{(\theta^*, \sigma^2, r^*)} (\widehat{r}_i \neq a) + \mathbb{P}_{(\theta^*, \sigma^2, r^{*(i,j)})} (\widehat{r}_i \neq b)}{2} \\
&\gtrsim \frac{1}{n} \sum_{1 \leq i < j \leq n} \frac{2}{n(n-1)} \sum_{1 \leq a < b \leq n} \\
&\qquad \frac{1}{(n-2)!} \sum_{r^* : r_i^* = a, r_j^* = b} \min \left\{ 1, \sqrt{\frac{\sigma^2}{np(\theta_a^* - \theta_b^*)^2}} \exp \left(- \frac{(1+\delta') np (\theta_a^* - \theta_b^*)^2}{4\sigma^2} \right) \right\} \qquad (3.39) \\
&= \frac{1}{n} \sum_{1 \leq a < b \leq n} \min \left\{ 1, \sqrt{\frac{\sigma^2}{np(\theta_a^* - \theta_b^*)^2}} \exp \left(- \frac{(1+\delta') np (\theta_a^* - \theta_b^*)^2}{4\sigma^2} \right) \right\}
\end{aligned}
$$

for some $\delta' = o(1)$, where (3.39) comes from Lemma 3.7.3.

We still consider the high signal-to-noise ratio case first.

$$
\begin{aligned}
&\sum_{1 \leq i < j \leq n} \min \left\{ 1, \sqrt{\frac{\sigma^2}{np(\theta_i^* - \theta_j^*)^2}} \exp \left(- \frac{(1+\delta') np (\theta_i^* - \theta_j^*)^2}{4\sigma^2} \right) \right\} \\
&\geq \sum_{i=1}^{n-1} \sqrt{\frac{\sigma^2}{np(\theta_i^* - \theta_{i+1}^*)^2}} \exp \left(- \frac{(1+\delta') np (\theta_i^* - \theta_{i+1}^*)^2}{4\sigma^2} \right) \\
&\gtrsim \sum_{i=1}^{n-1} \exp \left(- \frac{(1+\delta) np (\theta_i^* - \theta_{i+1}^*)^2}{4\sigma^2} \right) \qquad (3.40)
\end{aligned}
$$

where δ in (3.40) can be chosen arbitrarily small when $np\beta^2/\sigma^2 > 1$, which concludes the

exponential lower bound.

For the polynomial lower bound when signal-to-noise ratio is small,

$$\begin{aligned}
&\sum_{1\le i<j\le n} \min\left\{1, \sqrt{\frac{\sigma^2}{np(\theta_i^*-\theta_j^*)^2}}\exp\left(-\frac{(1+\delta')np(\theta_i^*-\theta_j^*)^2}{4\sigma^2}\right)\right\}\\
&\gtrsim \sum_{i=1}^{n} \sum_{\substack{j\neq i\\ |j-i|\le\sqrt{\frac{\sigma^2}{np\beta^2}}}} \min\left\{1, \sqrt{\frac{\sigma^2}{np(\theta_i-\theta_j)^2}}\exp\left(-\frac{(1+\delta')np(\theta_i-\theta_j)^2}{4\sigma^2}\right)\right\}\\
&\gtrsim \sum_{i=1}^{n} n\wedge\left(\sqrt{\frac{\sigma^2}{np\beta^2}}\right)
\end{aligned}$$

which concludes the proof. □

3.7.2 Proof of Theorem 3.3.2

This section proves Theorem 3.3.2. Since the upper bound part of the proof has already been given in Section 3.4.3, we only need to establish the lower bound. First of all, we establish a few lemmas.

Lemma 3.7.4 (Central limit theorem, Theorem 2.20 of [91])**.** *If $Z\sim\mathcal{N}(0,1)$ and $W=\sum_{i=1}^n X_i$ where X_i are independent mean 0 and $\mathsf{Var}(W)=1$, then*

$$\sup_t |\mathbb{P}(W\le t)-\mathbb{P}(Z\le t)| \le 2\sqrt{3\sum_{i=1}^n \left(\mathbb{E}X_i^4\right)^{3/4}}.$$

Lemma 3.7.5. *Assume $p\ge c_0\frac{\log n}{n}$ for some sufficiently large constant $c_0>0$. For any fixed $\{w_{ijk}\}$, $i,j\in[n], k\in\mathbb{K}$ where $\mathbb{K}$ is a discrete set with cardinality at most n^{c_1} for some constant $c_1>0$. Assume $\max_{i,j\in[n],k\in\mathbb{K}}\left|w_{ijk}\right|\le c_2$ and*

$$p\min_{i\in[n],k\in\mathbb{K}}\sum_{j\in[n]\setminus\{i\}} w_{ijk}^2 \ge c_3\log n$$

for some constants $c_2, c_3 > 0$. Then there exists constants $C_1, C_2 > 0$, such that for any $i \in [n]$,

$$\max_{k\in\mathbb{K}} \sum_{j\in[n]\setminus\{i\}} (A_{ij} - p) w_{ijk} \le C_1 \sqrt{p \log n \max_{k\in\mathbb{K}} \sum_{j\in[n]} w_{ijk}^2}$$

with probability at least $1 - C_2 n^{-10}$.

Proof. For any constant $C_1' > 0$, by Bernstein's inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{k\in\mathbb{K}} \sum_{j\in[n]\setminus\{i\}} (A_{ij} - p) w_{ijk} > C_1' \sqrt{p \log n \max_{k\in\mathbb{K}} \sum_{j\in[n]} w_{ijk}^2} \right) \\
&\le |\mathbb{K}| \max_{k\in\mathbb{K}} e^{-\frac{C_1'^2 p \log n \max_{k\in\mathbb{K}} \sum_{j\in[n]} w_{ijk}^2}{2p \sum_{j\in[n]\setminus\{i\}} w_{ijk}^2 + \frac{2}{3} \max_{i,j\in[n],k\in\mathbb{K}} |w_{ijk}| C_1' \sqrt{p \log n \max_{k\in\mathbb{K}} \sum_{j\in[n]} w_{ijk}^2}}} \\
&\le n^{c_1} \exp\left(-\frac{C_1'^2}{C_2'} \log n \right)
\end{aligned}$$

for some constant $C_2' > 0$. Thus we can set C_1' large enough to make the theorem holds. □

Lemma 3.7.6. *Assume $1 \le C_0 = O(1)$ and $0 < \beta = o(1)$. For any constant $\alpha > 0$, there exists constants $C_1, C_2 > 0$ such that for any $\theta \in \Theta_n(\beta, C_0)$,*

$$C_1 \frac{1}{\beta \vee 1/n} \le \inf_{\theta_0 \in [\theta_n, \theta_1]} \sum_{i=1}^n \psi'(\theta_0 - \theta_i)^\alpha \le \sup_{\theta_0 \in [\theta_n, \theta_1]} \sum_{i=1}^n \psi'(\theta_0 - \theta_i)^\alpha \le C_2 \frac{1}{\beta \vee 1/n}$$

for n large enough.

Proof. Define

$$R_\theta(x, t_1, t_2) = \{i : t_1 \le |\theta_i - x| < t_2\} \tag{3.41}$$

It is easy to see that there exist constants $C_1', C_2' > 0$ such that for any $\theta \in \Theta_n(\beta, C_0)$,

$$\frac{C_1'}{\beta \vee 1/n} \le \inf_{x \in [\theta_n, \theta_1]} |R_\theta(x, 0, 1)| \tag{3.42}$$

and

$$\sup_{t\in\mathbb{N}}\sup_{x\in[\theta_n,\theta_1]}|R_\theta(x,t,t+1)|\le\frac{C_2'}{\beta\vee 1/n} \tag{3.43}$$

Thus

$$\begin{aligned}
&\inf_{\theta_0\in[\theta_n,\theta_1]}\sum_{i=1}^n\psi'(\theta_0-\theta_i)^\alpha\ge\inf_{\theta_0\in[\theta_n,\theta_1]}\sum_{i\in R_\theta(\theta_0,0,1)}\psi'(\theta_0-\theta_i)^\alpha\\
&=\inf_{\theta_0\in[\theta_n,\theta_1]}\sum_{i\in R_\theta(\theta_0,0,1)}\left[\frac{e^{\theta_0-\theta_i}}{\left(1+e^{\theta_0-\theta_i}\right)^2}\right]^\alpha\ge\inf_{\theta_0\in[\theta_n,\theta_1]}\sum_{i\in R_\theta(\theta_0,0,1)}\frac{1}{4^\alpha}e^{-\alpha|\theta_0-\theta_i|}\\
&\ge\inf_{\theta_0\in[\theta_n,\theta_1]}|R_\theta(\theta_0,0,1)|\,\frac{1}{4^\alpha}e^{-\alpha}\ge\frac{C_3'}{\beta\vee 1/n}
\end{aligned}$$

for some constant $C_3'>0$. On the other hand,

$$\begin{aligned}
&\sup_{\theta_0\in[\theta_n,\theta_1]}\sum_{i=1}^n\psi'(\theta_0-\theta_i)^\alpha=\sup_{\theta_0\in[\theta_n,\theta_1]}\sum_{t\ge0}\sum_{i\in R_\theta(\theta_0,t,t+1)}\psi'(\theta_0-\theta_i)^\alpha\\
&\le\sup_{\theta_0\in[\theta_n,\theta_1]}\sum_{t\ge0}\sum_{i\in R_\theta(\theta_0,t,t+1)}e^{-\alpha|\theta_0-\theta_i|}\le\sup_{\theta_0\in[\theta_n,\theta_1]}\sum_{t\ge0}|R_\theta(\theta_0,t,t+1)|\,e^{-\alpha t}\\
&\le\frac{C_4'}{\beta\vee 1/n}
\end{aligned}$$

for some constant $C_4'>0$, which concludes the proof. □

Lemma 3.7.7. *Assume $p\ge c_0(\beta\vee\frac{1}{n})\log n$ for some sufficiently large constant $c_0>0$ and $1\le C_0=O(1)$. For any constant $\alpha>0$, there exist constants $C_1,C_2,C_3>0$ such that for any $r\in\mathfrak{S}_n, i\ne j\in[n]$, and $\theta\in\Theta_n(\beta,C_0)$,*

$$\inf_{u\in[0,1]}\sum_{k\ne i,j}A_{ik}\psi'(u\theta_{r_i}+(1-u)\theta_{r_j}-\theta_{r_k})^\alpha\ge C_1\frac{p}{\beta\vee 1/n} \tag{3.44}$$

and

$$\sup_{u\in[0,1]}\sum_{k\ne i,j}A_{ik}\psi'(u\theta_{r_i}+(1-u)\theta_{r_j}-\theta_{r_k})^\alpha\le C_2\frac{p}{\beta\vee 1/n} \tag{3.45}$$

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

Proof. We remark that $p \geq c_0(\beta \vee \frac{1}{n}) \log n$ necessarily implies $0 < \beta = o(1)$. We only give the proof of (3.45). The inf part (3.44) can be proved similarly. For (3.45),

$$\begin{aligned}
&\sup_{u\in[0,1]} \sum_{k\neq i,j} A_{ik}\psi'(u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k})^\alpha \\
&\leq \frac{C_1' p}{\beta \vee 1/n} + \sup_{u\in[0,1]} \sum_{k\neq i,j} (A_{ik} - p)\psi'(u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k})^\alpha \qquad (3.46)
\end{aligned}$$

for some constant $C_1' > 0$, where (3.46) uses Lemma 3.7.6. To bound the second term in (3.46), we use standard discretization technique. Let $u_a = \frac{a}{n}, a \in [n]$. Then for any $u \in [0,1]$, let $a(u) = \arg\min_{a\in[n]} |u - u_a|$. We have $\left|u - u_{a(u)}\right| \leq 1/n$. Observe that for any $u \in [0,1]$,

$$\begin{aligned}
&\left| \sum_{k\neq i,j} (A_{ik} - p) \left(\psi'(u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k})^\alpha - \psi'(u_{a(u)}\theta_{r_i} + (1-u_{a(u)})\theta_{r_j} - \theta_{r_k})^\alpha \right) \right| \\
&\leq \alpha \sup_{\xi\in[u\wedge u_{a(u)}, u\vee u_{a(u)}]} \sum_{k\neq i,j} \psi'(\xi\theta_{r_i} + (1-\xi)\theta_{r_j} - \theta_{r_k})^\alpha \left|u - u_{a(u)}\right| \left|\theta_{r_i} - \theta_{r_j}\right| \qquad (3.47) \\
&\leq \frac{C_2' n\beta}{n} \frac{1}{\beta \vee 1/n} \leq C_2' \frac{p}{\beta \vee 1/n} \qquad (3.48)
\end{aligned}$$

for some constant $C_2' > 0$, where (3.47) is due to mean value theorem and $\left|\psi''(x)\right| \leq \psi'(x)$ while (3.48) comes from Lemma 3.7.6. Therefore,

$$\begin{aligned}
&\sup_{u\in[0,1]} \sum_{k\neq i,j} A_{ik}\psi'(u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k})^\alpha \\
&\leq \frac{C_3' p}{\beta \vee 1/n} + \max_{a\in[n]} \sum_{k\neq i,j} (A_{ik} - p)\psi'(u_a\theta_{r_i} + (1-u_a)\theta_{r_j} - \theta_{r_k})^\alpha \qquad (3.49) \\
&\leq \frac{C_3' p}{\beta \vee 1/n} + C_4' \sqrt{p \log n \max_{a\in[n]} \sum_{k\neq i,j} \psi'(u_a\theta_{r_i} + (1-u_a)\theta_{r_j} - \theta_{r_k})^{2\alpha}} \qquad (3.50) \\
&\leq \frac{C_5' p}{\beta \vee 1/n} \qquad (3.51)
\end{aligned}$$

for some constants $C_3', C_4', C_5' > 0$ with probability at least $1 - O(n^{-10})$, where (3.49) is due to (3.48) and $\frac{p}{\beta \vee 1/n} \gtrsim \log n \gg 1$. (3.50) comes from Lemma 3.7.6, $\left|\psi'(x)\right| \leq 1/4$ and Lemma 3.7.5. (3.51) is a consequence of Lemma 3.7.6 and $\log n \lesssim \frac{p}{\beta \vee 1/n}$, which concludes the proof. □

To proceed with our proof for the lower bound, we define

$$G_{i,j,k,\theta,r}(u) = \log \frac{(1 + e^{\theta_{r_i} - \theta_{r_k}})^u (1 + e^{\theta_{r_j} - \theta_{r_k}})^{1-u}}{1 + e^{u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k}}}. \tag{3.52}$$

This term is a key ingredient in the exponent of the rate. We first derive some properties of this term.

Lemma 3.7.8. *Assume $1 \leq C_0 = O(1)$ and $0 < \beta = o(1)$. For any constant $C > 0$, there exist constants $C_1, C_2, C_3 > 0$ such that for any $\theta \in \Theta_n(\beta, C_0)$, any $r \in \mathfrak{S}_n$ and any $i \neq j \in [n]$ such that $\left|\theta_{r_i} - \theta_{r_j}\right| \leq C$, the following hold for n large enough,*

$$\sup_{u \in [0,1]} \sup_{k \neq i,j} G_{i,j,k,\theta,r}(u) \leq C_1, \tag{3.53}$$

$$\sup_{u \in [0,1]} \sum_{k \neq i,j} G_{i,j,k,\theta,r}(u) + G_{i,j,k,\theta,r}(1-u) \leq \sum_{k \neq i,j} \log \frac{(1 + e^{\theta_{r_i} - \theta_{r_k}})(1 + e^{\theta_{r_j} - \theta_{r_k}})}{\left(1 + e^{\frac{\theta_{r_i} + \theta_{r_j}}{2} - \theta_{r_k}}\right)^2}, \tag{3.54}$$

$$\sup_{u \in [0,1]} \sum_{k \neq i,j} G_{i,j,k,\theta,r}(u)^2 + G_{i,j,k,\theta,r}(1-u)^2 \leq C_2 \frac{(\theta_{r_i} - \theta_{r_j})^4}{\beta \vee 1/n}, \tag{3.55}$$

$$C_3 \frac{\left|\theta_{r_i} - \theta_{r_j}\right|^2}{\beta \vee 1/n} \leq \sum_{k \neq i,j} \log \frac{(1 + e^{\theta_{r_i} - \theta_{r_k}})(1 + e^{\theta_{r_j} - \theta_{r_k}})}{\left(1 + e^{\frac{\theta_{r_i} + \theta_{r_j}}{2} - \theta_{r_k}}\right)^2} \leq C_2 \frac{\left|\theta_{r_i} - \theta_{r_j}\right|^2}{\beta \vee 1/n}. \tag{3.56}$$

Proof. We first look at (3.53). Note that

$$
\begin{aligned}
G_{i,j,k,\theta,r}(u) &= \log \frac{\psi(u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k})}{\psi(\theta_{r_i} - \theta_{r_k})^u \psi(\theta_{r_j} - \theta_{r_k})^{1-u}} \\
&\leq \log \frac{\psi(u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k})}{\psi(\theta_{r_i} - \theta_{r_k}) \wedge \psi(\theta_{r_j} - \theta_{r_k})} \\
&= \log \frac{(1 + e^{\theta_{r_i} - \theta_{r_k}})}{e^{(1-u)(\theta_{r_i} - \theta_{r_j})} + e^{\theta_{r_i} - \theta_{r_k}}} \vee \log \frac{(1 + e^{\theta_{r_j} - \theta_{r_k}})}{e^{-u(\theta_{r_i} - \theta_{r_j})} + e^{\theta_{r_j} - \theta_{r_k}}} \leq C
\end{aligned}
$$

where the last inequality comes from $\left|\theta_{r_i} - \theta_{r_j}\right| \leq C$.

Now we look at (3.54).

$$
\begin{aligned}
&\sup_{u\in[0,1]} \sum_{k\neq i,j} G_{i,j,k,\theta,r}(u) + G_{i,j,k,\theta,r}(1-u) \\
&= \sup_{u\in[0,1]} \sum_{k\neq i,j} \log \frac{(1 + e^{\theta_{r_i} - \theta_{r_k}})(1 + e^{\theta_{r_j} - \theta_{r_k}})}{1 + e^{u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k}} + e^{(1-u)\theta_{r_i} + u\theta_{r_j} - \theta_{r_k}} + e^{\theta_{r_i} + \theta_{r_j} - 2\theta_{r_k}}} \\
&\leq \sup_{u\in[0,1]} \sum_{k\neq i,j} \log \frac{(1 + e^{\theta_{r_i} - \theta_{r_k}})(1 + e^{\theta_{r_j} - \theta_{r_k}})}{1 + 2e^{\frac{\theta_{r_i} + \theta_{r_j}}{2} - \theta_{r_k}} + e^{\theta_{r_i} + \theta_{r_j} - 2\theta_{r_k}}} = \sum_{k\neq i,j} \log \frac{(1 + e^{\theta_{r_i} - \theta_{r_k}})(1 + e^{\theta_{r_j} - \theta_{r_k}})}{\left(1 + e^{\frac{\theta_{r_i} + \theta_{r_j}}{2} - \theta_{r_k}}\right)^2}.
\end{aligned}
$$

To see (3.55), we first note that

$$
\begin{aligned}
&G_{i,j,k,\theta,r}(u) \\
&= u \log(1 + e^{\theta_{r_i} - \theta_{r_k}}) + (1-u) \log(1 + e^{\theta_{r_j} - \theta_{r_k}}) - \log(1 + e^{u\theta_{r_i} + (1-u)\theta_{r_j} - \theta_{r_k}}) \\
&\geq 0
\end{aligned}
$$

by Jensen's inequality. Therefore,

$$\sup_{u\in[0,1]}\sum_{k\neq i,j} G_{i,j,k,\theta,r}(u)^2 + G_{i,j,k,\theta,r}(1-u)^2 \leq \sup_{u\in[0,1]}\sum_{k\neq i,j}\left(G_{i,j,k,\theta,r}(u) + G_{i,j,k,\theta,r}(1-u)\right)^2$$

$$\leq \sum_{k\neq i,j}\left[\log\frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2}\right]^2 \tag{3.57}$$

where (3.57) can be derived similarly as in the proof of (3.54). To upper bound (3.57), recall the definition of $R_\theta(\cdot,\cdot,\cdot)$ in (3.41). We have that for any k such that $r_k \in R_\theta(\frac{\theta_{r_i}+\theta_{r_j}}{2}, t, t+1)$,

$$\log\frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2} = \log\left(\frac{\cosh(\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}) + \cosh\frac{\theta_{r_i}-\theta_{r_j}}{2}}{\cosh(\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k})+1}\right)$$

$$\leq \frac{\cosh\frac{\theta_{r_i}-\theta_{r_j}}{2}-1}{\cosh(\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k})+1} \leq \frac{C_1'(\theta_{r_i}-\theta_{r_j})^2}{e^t} \tag{3.58}$$

for some constant $C_1' > 0$. (3.58) can be seen from $\cosh x \leq 1 + C_2' x^2$ for some constant $C_2' > 0$ when $|x| \leq C/2$. and the fact that $t \leq |\frac{\theta_{r_i}+\theta_{r_j}}{2} - \theta_{r_k}| \leq t+1$. Therefore, using (3.43) and (3.58),

$$\sum_{k\neq i,j}\left[\log\frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2}\right]^2$$

$$\leq \sum_{t\geq 0}\sum_{k:r_k\in R_\theta(\frac{\theta_{r_i}+\theta_{r_j}}{2},t,t+1)}\frac{C_1'^2(\theta_{r_i}-\theta_{r_j})^4}{e^{2t}} \leq \frac{C_3'(\theta_{r_i}-\theta_{r_j})^4}{\beta\vee 1/n}$$

for some constant $C_3' > 0$. The upper bound of (3.56) can be proved similarly.

Finally, we turn to the lower bound of (3.56). Note that we also have $\cosh x \geq 1 + C_4' x^2$ for some constant $C_4' > 0$ when $|x| \leq C/2$. Therefore, when $r_k \in R_\theta(\frac{\theta_{r_i}+\theta_{r_j}}{2}, 0, 1)$,

$$\log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2} = \log\left(1 + \frac{\cosh\frac{\theta_{r_i}-\theta_{r_j}}{2} - 1}{\cosh(\frac{\theta_{r_i}+\theta_{r_j}}{2} - \theta_{r_k}) + 1}\right)$$
$$\geq \frac{\cosh\frac{\theta_{r_i}-\theta_{r_j}}{2} - 1}{\cosh\frac{\theta_{r_i}-\theta_{r_j}}{2} + \cosh(\frac{\theta_{r_i}+\theta_{r_j}}{2} - \theta_{r_k})} \geq C_5' \left|\theta_{r_i} - \theta_{r_j}\right|^2 \tag{3.59}$$

for some constant $C_5' > 0$, where the first inequality is due to the fact that $\log(1+x) \geq x/(1+x)$ for any $x > -1$. Thus,

$$\sum_{k\neq i,j} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2} \geq \sum_{\substack{k: r_k \in R_\theta(\frac{\theta_{r_i}+\theta_{r_j}}{2},0,1) \\ k\neq i,j}} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2}$$
$$\geq \left(\left|R_\theta(\frac{\theta_{r_i}+\theta_{r_j}}{2}, 0, 1)\right| - 2\right) C_5' \left|\theta_{r_i} - \theta_{r_j}\right|^2 \geq \frac{C_6' \left|\theta_{r_i} - \theta_{r_j}\right|^2}{\beta \vee 1/n} \tag{3.60}$$

for some constant $C_6' > 0$, where (3.60) is a result of (3.42) and (3.59). □

Lemma 3.7.9. *Assume* $\frac{p}{\log n(\beta \vee 1/n)} \to \infty$ *and* $1 \leq C_0 = O(1)$. *For any constant* $C_1 > 0$, *there exists* $\delta = o(1)$ *and constant* $C_2 > 0$, *such that for any* $\theta \in \Theta_n(\beta, C_0)$, *any* $r \in \mathfrak{S}_n$ *and any* $i \neq j \in [n]$ *such that* $\left|\theta_{r_i} - \theta_{r_j}\right| \leq C_1$, *the following holds with probability at least* $1 - O(n^{-10})$ *for* n *large enough,*

$$\sup_{u\in[0,1]} \sum_{k\neq i,j} A_{ik} G_{i,j,k,\theta,r}(u) + A_{jk} G_{i,j,k,\theta,r}(1-u) \leq (1+\delta)p \sum_{k\neq i,j} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2}.$$

Proof. First we have

$$\psi(a-c) \wedge \psi(b-c) \leq \frac{1}{a-b} \log \frac{1+e^{a-c}}{1+e^{b-c}} \leq \psi(a-c) \vee \psi(b-c), \tag{3.61}$$

for any $a, b, c \in \mathbb{R}$. To see why (3.61) holds, let us study the function $f(\delta) = \log(1+\exp(x+\delta)) - \log(1+\exp(x)) - \delta \exp(x)/(1+\exp(x))$ for any x. Note that $f'(\delta) = \exp(x+\delta)/(1+\exp(x+\delta)) - \exp(x)/(1+\exp(x))$ is positive when $\delta > 0$ and negative when $\delta < 0$. Since $f(0) = 0$, we have $f(\delta) \geq 0$. As a result, we have $\exp(x)/(1+\exp(x)) \leq \delta^{-1} \log((1+\exp(x+\delta))/(1+\exp(x)))$ when $\delta > 0$ and the direction of the inequality is reversed when $\delta < 0$. WLOG, we assume $a-b > 0$. Then the first inequality of (3.61) is proved by taking $x = b-c$ and $\delta = a-b$ and second one is proved by taking $x = a-c$ and $\delta = -(a-b)$.

Recall the definition of $G_{i,j,k,\theta,r}$ in (3.52). Then

$$\begin{aligned}
\left|G'_{i,j,k,\theta,r}(u)\right| &= \left|\theta_{r_i} - \theta_{r_j}\right| \left| \frac{1}{\theta_{r_i} - \theta_{r_j}} \log \frac{1+e^{\theta_{r_i}-\theta_{r_k}}}{1+e^{\theta_{r_j}-\theta_{r_k}}} - \frac{e^{u(\theta_{r_i}-\theta_{r_j})+\theta_{r_j}-\theta_{r_k}}}{1+e^{u(\theta_{r_i}-\theta_{r_j})+\theta_{r_j}-\theta_{r_k}}} \right| \\
&\leq \left|\theta_{r_i} - \theta_{r_j}\right| \left|\psi(\theta_{r_i} - \theta_{r_k}) - \psi(\theta_{r_j} - \theta_{r_k})\right| \qquad (3.62) \\
&\leq \left|\theta_{r_i} - \theta_{r_j}\right|^2. \qquad (3.63)
\end{aligned}$$

Here (3.62) is due to the observation that both terms are in the interval $[\psi(\theta_{r_i}-\theta_{r_k}) \wedge \psi(\theta_{r_j}-\theta_{r_k}), \psi(\theta_{r_i}-\theta_{r_k}) \vee \psi(\theta_{r_j}-\theta_{r_k})]$ for any $u \in [0,1]$, where the first term is due to (3.61) and the second term is due to the monotonicity of $\exp(x)/(1+\exp(x))$. Hence the difference between these two terms are bounded by $\left|\psi(\theta_{r_i} - \theta_{r_k}) - \psi(\theta_{r_j} - \theta_{r_k})\right|$ in absolute value. (3.63) is due to $\psi'(x) \leq 1/4$. Following the line of discretization, let $u_a = \frac{a}{n}, a = 1, ..., n$. Then for

any $u \in [0,1]$, let $a(u) = \arg\min_{a\in[n]} |u - u_a|$. We have $\left|u - u_{a(u)}\right| \le 1/n$. Thus,

$$
\begin{aligned}
&\Big| \sum_{k\neq i,j} (A_{ik} - p)(G_{i,j,k,\theta,r}(u) - G_{i,j,k,\theta,r}(u_{a(u)})) \\
&\qquad + (A_{jk} - p)(G_{i,j,k,\theta,r}(1-u) - G_{i,j,k,\theta,r}(1-u_{a(u)}))\Big| \\
&\le 2\left|\theta_{r_i} - \theta_{r_j}\right|^2 (n-2)\left|u - u_{a(u)}\right| \le 2\left|\theta_{r_i} - \theta_{r_j}\right|^2. \qquad (3.64)
\end{aligned}
$$

Then

$$\sup_{u\in[0,1]} \sum_{k\neq i,j} A_{ik}G_{i,j,k,\theta,r}(u) + A_{jk}G_{i,j,k,\theta,r}(1-u)$$

$$\leq p \sum_{k\neq i,j} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2} \tag{3.65}$$

$$+ \sup_{u\in[0,1]} \sum_{k\neq i,j} (A_{ik}-p)G_{i,j,k,\theta,r}(u) + (A_{jk}-p)G_{i,j,k,\theta,r}(1-u)$$

$$\leq p \sum_{k\neq i,j} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2}$$

$$+ 2\left|\theta_{r_i}-\theta_{r_j}\right|^2 + \max_{a\in[n]} \sum_{k\neq i,j} (A_{ik}-p)G_{i,j,k,\theta,r}(u_a) + (A_{jk}-p)G_{i,j,k,\theta,r}(1-u_a) \tag{3.66}$$

$$\leq p \sum_{k\neq i,j} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2} + 2\left|\theta_{r_i}-\theta_{r_j}\right|^2$$

$$+ C_1' \sqrt{p\log n \max_{a\in[n]} \sum_{k\neq i,j} G_{i,j,k,\theta,r}(u_a)^2 + G_{i,j,k,\theta,r}(1-u_a)^2} \tag{3.67}$$

$$\leq p \sum_{k\neq i,j} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2} + 2\left|\theta_{r_i}-\theta_{r_j}\right|^2 + C_2'\left|\theta_{r_i}-\theta_{r_j}\right|^2 \sqrt{\frac{p\log n}{\beta\vee 1/n}}$$

(3.68)

$$= (1+\delta)p \sum_{k\neq i,j} \log \frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2} \tag{3.69}$$

with probability at least $1-O(n^{-10})$ for some constants $C_1', C_2' > 0$ and $\delta = o(1)$. (3.65) is due to Lemma 3.7.8. (3.66) comes from (3.64). (3.67) and (3.68) are a consequence of

Lemma 3.7.5 and Lemma 3.7.8. (3.69) is because of Lemma 3.7.8 and

$$p\frac{1}{\left|\theta_{r_i}-\theta_{r_j}\right|^2}\sum_{k\neq i,j}\log\frac{(1+e^{\theta_{r_i}-\theta_{r_k}})(1+e^{\theta_{r_j}-\theta_{r_k}})}{\left(1+e^{\frac{\theta_{r_i}+\theta_{r_j}}{2}-\theta_{r_k}}\right)^2}\gtrsim\frac{p}{\beta\vee 1/n}\gg\sqrt{\frac{p\log n}{\beta\vee 1/n}}\gg 1,$$

which concludes the proof.

□

Lemma 3.7.10. *Assume* $\frac{p}{\log n(\beta\vee 1/n)}\to\infty$ *and* $1\le C_0=O(1)$. *For any constant* $C>0$. *there exist constants* $C_1,C_2>0$, $\delta=o(1)$ *such that for any* $\theta^*\in\Theta_n(\beta,C_0)$, *any* $r^*\in\mathfrak{S}_n$ *and* $i\neq j\in[n]$ *such that* $\left|\theta^*_{r_i^*}-\theta^*_{r_j^*}\right|\le C$, *we have*

$$\begin{aligned}&\inf_{\widehat{r}}\frac{\mathbb{P}_{(\theta^*,r^*)}\left(\widehat{r}\neq r^*\right)+\mathbb{P}_{(\theta^*,r^{*(i,j)})}\left(\widehat{r}\neq r^{*(i,j)}\right)}{2}\\&\ge C_1\exp\left(-\sqrt{\frac{C_2Lp(\theta^*_{r_i^*}-\theta^*_{r_j^*})^2}{\beta\vee 1/n}}-(1+\delta)2Lp\sum_{k\neq i,j}G_{i,j,k,\theta^*,r^*}(1/2)\right)\end{aligned}$$

for n large enough. Here $r^{*(i,j)}$ *is defined as in (3.36).*

Proof. By Neyman-Pearson Lemma, the optimal procedure is the likelihood ratio test:

$$\begin{aligned}&\inf_{\widehat{r}}\frac{\mathbb{P}_{(\theta^*,r^*)}\left(\widehat{r}\neq r^*\right)+\mathbb{P}_{(\theta^*,r^{*(i,j)})}\left(\widehat{r}\neq r^{*(i,j)}\right)}{2}\\&=\frac{\mathbb{P}_{(\theta^*,r^*)}\left(\ell_n(\theta^*,r^*)\ge\ell_n(\theta^*,r^{*(i,j)})\right)+\mathbb{P}_{(\theta^*,r^{*(i,j)})}\left(\ell_n(\theta^*,r^*)\le\ell_n(\theta^*,r^{*(i,j)})\right)}{2}\end{aligned}$$

We only need to lower bound $\mathbb{P}_{(\theta^*,r^*)}\left(\ell_n(\theta^*,r^*)\ge\ell_n(\theta^*,r^{*(i,j)})\right)$ and the other term can be bounded similarly. WLOG, assume $i<j$ and $r_i^*=a<r_j^*=b$. Let

$$Z_{kl}=y_{ikl}\log\frac{\psi(\theta^*_b-\theta^*_{r_k^*})}{\psi(\theta^*_a-\theta^*_{r_k^*})}+(1-y_{ikl})\log\frac{1-\psi(\theta^*_b-\theta^*_{r_k^*})}{1-\psi(\theta^*_a-\theta^*_{r_k^*})},k\neq i,j,$$

$$\bar{Z}_{kl} = y_{jkl} \log \frac{\psi(\theta^*_a - \theta^*_{r^*_k})}{\psi(\theta^*_b - \theta^*_{r^*_k})} + (1 - y_{jkl}) \log \frac{1 - \psi(\theta^*_a - \theta^*_{r^*_k})}{1 - \psi(\theta^*_b - \theta^*_{r^*_k})}, k \neq i, j$$

and

$$Z_{0l} = y_{ijl} \log \frac{\psi(\theta^*_b - \theta^*_a)}{\psi(\theta^*_a - \theta^*_b)} + (1 - y_{ijl}) \log \frac{1 - \psi(\theta_b - \theta^*_a)}{1 - \psi(\theta^*_a - \theta^*_b)}.$$

To simplify notation, we use $\mathbb{P}_A(\cdot)$ as $\mathbb{P}_{(\theta^*, r^*)}(\cdot|A)$ and $\mathbb{E}_A[\cdot]$ as $\mathbb{E}_{(\theta^*, r^*)}[\cdot|A]$. Then

$$\mathbb{P}_A\left(\ell_n(\theta^*, r^*) \geq \ell_n(\theta^*, r^{*(i,j)})\right) = \mathbb{P}_A\left(\sum_{l=1}^{L}\left(A_{ij}Z_{0l} + \sum_{k\neq i,j} A_{ik}Z_{kl} + A_{jk}\bar{Z}_{kl}\right) \geq 0\right).$$

Let $\mu_{i'j'} = \psi(\theta^*_{r^*_{i'}} - \theta^*_{r^*_{j'}})$ for any $i' \neq j'$. Define

$$\begin{aligned}\nu_{r^*}(u) &= \log \mathbb{E}_A\left\{\exp\left[u\left(A_{ij}Z_{01} + \sum_{k\neq i,j} A_{ik}Z_{k1} + A_{jk}\bar{Z}_{k1}\right)\right]\right\} \\ &= A_{ij}\nu_{0,r^*}(u) + \sum_{k\neq i,j} A_{ik}\nu_{k,r^*}(u) + \sum_{k\neq i,j} A_{jk}\bar{\nu}_{k,r^*}(u)\end{aligned}$$

where

$$\nu_{0,r^*}(u) = \log\left[\mu_{ij}^u(1-\mu_{ij})^{1-u} + \mu_{ij}^{1-u}(1-\mu_{ij})^u\right] = -\log\frac{1 + e^{\theta^*_a - \theta^*_b}}{e^{u(\theta^*_a - \theta^*_b)} + e^{(1-u)(\theta^*_a - \theta^*_b)}}$$

$$\nu_{k,r^*}(u) = \log\left[\mu_{jk}^u\mu_{ik}^{1-u} + (1-\mu_{jk})^u(1-\mu_{ik})^{1-u}\right] = -G_{i,j,k,\theta^*,r^*}(1-u)$$

$$\bar{\nu}_{k,r^*}(u) = \log\left[\mu_{ik}^u\mu_{jk}^{1-u} + (1-\mu_{ik})^u(1-\mu_{jk})^{1-u}\right] = -G_{i,j,k,\theta^*,r^*}(u).$$

$\nu_{r^*}(u)$ is the conditional cumulant generating function of $A_{ij}Z_{01} + \sum_{k\neq i,j} A_{ik}Z_{kl} + A_{jk}\bar{Z}_{kl}$. We also have $\nu_{0,r^*}(u), \nu_{k,r^*}(u), \bar{\nu}_{k,r^*}(u)$ as the cumulant generating functions of $Z_{01}, Z_{k1}, \bar{Z}_{k1}$ respectively. Define

$$u^*_{r^*} = \arg\min_{u\geq 0} \nu_{r^*}(u).$$

Since cumulant generating functions are convex and $\nu_{r^*}(0) = \nu_{r^*}(1) = 0$, it can be seen

easily that $u^*_{r^*} \in (0,1)$ and depends on A. Following the change-of-measure argument in the proof of Lemma 8.4 and Lemma 9.3 of [21], we have

$$\begin{aligned}
&\mathbb{P}_A\left(\sum_{l=1}^{L}\left(A_{ij}Z_{0l}+\sum_{k\neq i,j}A_{ik}Z_{kl}+A_{jk}\bar{Z}_{kl}\right)\geq 0\right)\\
&\geq \exp\left(-u^*_{r^*}T+L\nu_{r^*}(u^*_{r^*})\right)\mathbb{Q}_A\left(0\leq\sum_{l=1}^{L}\left(A_{ij}Z_{0l}+\sum_{k\neq i,j}A_{ik}Z_{kl}+A_{jk}\bar{Z}_{kl}\right)\leq T\right)
\end{aligned} \tag{3.70}$$

for any T in (3.70) to be determined later and $\mathbb{Q}_A$ is a measure under which $Z_{0l}, Z_{kl}, \bar{Z}_{kl}, l \in [L], k \neq i, j$ are all independent given A and follow

$$\mathbb{Q}_A(Z_{0l}=s)=e^{u^*_{r^*}s-\nu_{0,r^*}(u^*_{r^*})}\mathbb{P}_A\left(Z_{0l}=s\right),$$

$$\mathbb{Q}_A(Z_{kl}=s)=e^{u^*_{r^*}s-\nu_{k,r^*}(u^*_{r^*})}\mathbb{P}_A\left(Z_{kl}=s\right), k\neq i,j,k\in[n],$$

$$\mathbb{Q}_A(\bar{Z}_{kl}=s)=e^{u^*_{r^*}s-\bar{\nu}_{k,r^*}(u^*_{r^*})}\mathbb{P}_A\left(\bar{Z}_{kl}=s\right), k\neq i,j,k\in[n].$$

Furthermore, by definition of $u^*_{r^*}$, the expectation of $A_{ij}Z_{0l}+\sum_{k\neq i,j}A_{ik}Z_{kl}+A_{jk}\bar{Z}_{kl}$ under $\mathbb{Q}_A$ is 0.

We can compute the 2nd and 4th moments under Q_A, denoted as $\mathsf{Var}_{\mathbb{Q}_A}(\cdot)$ and $\kappa_{\mathbb{Q}_A}(\cdot)$ respectively:

$$\begin{aligned}
\mathsf{Var}_{\mathbb{Q}_A}(Z_{0l})&=\nu''_{0,r^*}(u^*_{r^*})=4\mu_{ij}(1-\mu_{ij})\frac{(\theta^*_a-\theta^*_b)^2e^{2u^*_{r^*}(\theta^*_a-\theta^*_b)}}{((1-\mu_{ij})e^{2u^*_{r^*}(\theta^*_a-\theta^*_b)}+\mu_{ij})^2}\\
&=4(\theta^*_a-\theta^*_b)^2\psi'\left((1-2u^*_{r^*})(\theta^*_a-\theta^*_b)\right),
\end{aligned} \tag{3.71}$$

$$\mathsf{Var}_{\mathbb{Q}_A}(Z_{kl}) = \nu''_{k,r^*}(u^*_{r^*}) = \mu_{ik}(1-\mu_{ik})\frac{(\theta^*_a-\theta^*_b)^2 e^{u^*_{r^*}(\theta^*_a-\theta^*_b)}}{((1-\mu_{ik})e^{u^*_{r^*}(\theta^*_a-\theta^*_b)}+\mu_{ik})^2}$$
$$= (\theta^*_a-\theta^*_b)^2\psi'\left((1-u^*_{r^*})\theta^*_a + u^*_{r^*}\theta^*_b - \theta^*_{r^*_k}\right), k \neq i,j,k \in [n], \tag{3.72}$$

$$\mathsf{Var}_{\mathbb{Q}_A}(\bar{Z}_{kl}) = \bar{\nu}''_{k,r^*}(u^*_{r^*}) = \mu_{jk}(1-\mu_{jk})\frac{(\theta^*_a-\theta^*_b)^2 e^{-u^*_{r^*}(\theta^*_a-\theta^*_b)}}{((1-\mu_{jk})e^{-u^*_{r^*}(\theta^*_a-\theta^*_b)}+\mu_{jk})^2}$$
$$= (\theta^*_a-\theta^*_b)^2\psi'\left(u^*_{r^*}\theta^*_a + (1-u^*_{r^*})\theta^*_b - \theta^*_{r^*_k}\right), k \neq i,j,k \in [n] \tag{3.73}$$

and

$$\kappa_{\mathbb{Q}_A}(Z_{0l}) = \mathbb{Q}_A\left((Z_{0l}-\mathbb{Q}_A(Z_{0l}))^4\right) = \nu''''_{0,r^*}(u^*_{r^*}) + 3\nu''_{0,r^*}(u^*_{r^*})^2$$
$$\leq 16\mu_{ij}(1-\mu_{ij})\frac{(\theta^*_a-\theta^*_b)^4 e^{2u^*_{r^*}(\theta^*_a-\theta^*_b)}}{[(1-\mu_{ij})e^{2u^*_{r^*}(\theta^*_a-\theta^*_b)}+\mu_{ij}]^2} + 3\nu''_{0,r^*}(u^*_{r^*})^2$$
$$= 4(\theta^*_a-\theta^*_b)^2\nu''_{0,r^*}(u^*_{r^*}) + 3\nu''_{0,r^*}(u^*_{r^*})^2 \leq 7(\theta^*_a-\theta^*_b)^4\psi'\left((1-2u^*_{r^*})(\theta^*_a-\theta^*_b)\right), \tag{3.74}$$

$$\kappa_{\mathbb{Q}_A}(Z_{kl}) = \mathbb{Q}_A\left((Z_{kl}-\mathbb{Q}_A(Z_{kl}))^4\right) \leq (\theta^*_a-\theta^*_b)^2\nu''_{k,r^*}(u^*_{r^*}) + 3\nu''_{k,r^*}(u^*_{r^*})^2$$
$$\leq 4(\theta^*_a-\theta^*_b)^4\psi'\left((1-u^*_{r^*})\theta^*_a + u^*_{r^*}\theta^*_b - \theta^*_{r^*_k}\right), k \neq i,j,k \in [n] \tag{3.75}$$

$$\kappa_{\mathbb{Q}_A}(\bar{Z}_{kl}) = \mathbb{Q}_A\left((\bar{Z}_{kl}-\mathbb{Q}_A(\bar{Z}_{kl}))^4\right) \leq (\theta^*_a-\theta^*_b)^2\bar{\nu}''_{k,r^*}(u^*_{r^*}) + 3\bar{\nu}''_{k,r^*}(u^*_{r^*})^2$$
$$\leq 4(\theta^*_a-\theta^*_b)^4\psi'\left(u^*_{r^*}\theta^*_a + (1-u^*_{r^*})\theta^*_b - \theta^*_{r^*_k}\right), k \neq i,j,k \in [n]. \tag{3.76}$$

Let $\mathcal{F}_1$ be the event on which the following holds:

$$\inf_{u\in[0,1]} \sum_{k\neq i,j} A_{ik}\psi'((1-u)\theta^*_a + u\theta^*_b - \theta^*_{r^*_k}) + A_{jk}\psi'(u\theta^*_a + (1-u)\theta^*_b - \theta^*_{r^*_k}) \geq C'_1\frac{p}{\beta\vee 1/n},$$

$$\sup_{u\in[0,1]}\sum_{k\neq i,j} A_{ik}\psi'((1-u)\theta_a^*+u\theta_b^*-\theta_{r_k^*}^*)^{3/4}+A_{jk}\psi'(u\theta_a^*+(1-u)\theta_b^*-\theta_{r_k^*}^*)^{3/4}\leq C_2'\frac{p}{\beta\vee 1/n}$$

for some constants $C_1', C_2' > 0$. We shall choose C_1', C_2' to make $\mathcal{F}_1$ happen with probability at least $1 - O(n^{-10})$ by Lemma 3.7.7. Therefore, we shall choose T as

$$\begin{aligned}T &= \sqrt{L\left(A_{ij}\mathsf{Var}_{\mathbb{Q}_A}(Z_{01})+\sum_{k\neq i,j}A_{ik}\mathsf{Var}_{\mathbb{Q}_A}(Z_{k1})+A_{jk}\mathsf{Var}_{\mathbb{Q}_A}(\bar{Z}_{k1})\right)}\\ &\leq \sqrt{C_3'L\frac{p(\theta_a^*-\theta_b^*)^2}{\beta\vee 1/n}}\end{aligned}$$

on $\mathcal{F}_1$ for some constant $C_3' > 0$ using (3.71)-(3.73). With this choice of T, the $\mathbb{Q}_A$ measure can be lower bounded by some constant $C_4' > 0$ on $\mathcal{F}_1$. This can be seen by bounding the 4th moment approximation bound using Lemma 3.7.4 :

$$\sqrt{L\frac{A_{ij}\kappa_{\mathbb{Q}_A}(Z_{01})^{3/4}+\sum_{k\neq i,j}A_{ik}\kappa_{\mathbb{Q}_A}(Z_{kl})^{3/4}+A_{jk}\kappa_{\mathbb{Q}_A}(\bar{Z}_{kl})^{3/4}}{\left(LA_{ij}\mathsf{Var}_{\mathbb{Q}_A}(Z_{01})+L\sum_{k\neq i,j}A_{ik}\mathsf{Var}_{\mathbb{Q}_A}(Z_{k1})+A_{jk}\mathsf{Var}_{\mathbb{Q}_A}(\bar{Z}_{k1})\right)^{3/2}}}$$

$$\leq \sqrt{C_5'L\frac{\sum_{k\neq i,j}A_{ik}\psi'\left((1-u_{r^*}^*)\theta_a^*+u_{r^*}^*\theta_b^*-\theta_{r_k^*}^*\right)^{3/4}+A_{jk}\psi'\left(u_{r^*}^*\theta_a^*+(1-u_{r^*}^*)\theta_b^*-\theta_{r_k^*}^*\right)^{3/4}}{\left(L\sum_{k\neq i,j}A_{ik}\psi'((1-u_{r^*}^*)\theta_a^*+u_{r^*}^*\theta_b^*-\theta_{r_k^*}^*)+A_{jk}\psi'(u_{r^*}^*\theta_a^*+(1-u_{r^*}^*)\theta_b^*-\theta_{r_k^*}^*)\right)^{3/2}}} \tag{3.77}$$

$$\leq C_6'\left(L\frac{p}{\beta\vee 1/n}\right)^{-1/4} \tag{3.78}$$

on $\mathcal{F}_1$ for some constants $C_5', C_6' > 0$ and this bound tends to 0. (3.77) is due to (3.71)-(3.73) and (3.74)-(3.76). (3.78) is a consequence of Lemma 3.7.7.

Now we turn to $L\nu_{r^*}(u^*_{r^*})$. Let $\mathcal{F}_2$ be the event on which the following holds:

$$\begin{aligned}&\sup_{u\in[0,1]}\sum_{k\neq i,j}\left(A_{ik}G_{i,j,k,\theta^*,r^*}(1-u)+A_{jk}G_{i,j,k,\theta^*,r^*}(u)\right)\\&\leq(1+\delta_1')2p\sum_{k\neq i,j}G_{i,j,k,\theta^*,r^*}(1/2).\end{aligned}$$

By Lemma 3.7.9, there exists $\delta_1' = o(1)$ independent of i, j, θ^*, r^* such that $\mathcal{F}_2$ holds with probability at least $1 - O(n^{-10})$. Then, on this event,

$$\begin{aligned}\nu_{r^*}(u_{r^*}) &\geq -\sup_{u\in[0,1]}\left(-A_{ij}\nu_{0,r^*}(u)-\sum_{k\neq i,j}\left(A_{ik}\nu_{k,r^*}(u)+A_{jk}\bar{\nu}_{k,r^*}(u)\right)\right)\\&\geq -A_{ij}\sup_{u\in[0,1]}\log\frac{1+e^{\theta_a^*-\theta_b^*}}{e^{u(\theta_a^*-\theta_b^*)}+e^{(1-u)(\theta_a^*-\theta_b^*)}}\\&\quad-\sup_{u\in[0,1]}\sum_{k\neq i,j}\left(A_{ik}G_{i,j,k,\theta^*,r^*}(1-u)+A_{jk}G_{i,j,k,\theta^*,r^*}(u)\right)\\&\geq -C_7'\left|\theta_a^*-\theta_b^*\right|^2-(1+\delta_1')2p\sum_{k\neq i,j}G_{i,j,k,\theta^*,r^*}(1/2) \qquad (3.79)\\&\geq -(1+\delta_2')2p\sum_{k\neq i,j}G_{i,j,k,\theta^*,r^*}(1/2) \qquad (3.80)\end{aligned}$$

for some $\delta_2' = o(1)$. (3.79) comes from

$$\log\frac{1+e^{\theta_a^*-\theta_b^*}}{e^{u(\theta_a^*-\theta_b^*)}+e^{(1-u)(\theta_a^*-\theta_b^*)}}\leq\log\cosh\frac{\theta_a^*-\theta_b^*}{2}\leq\cosh\frac{\theta_a^*-\theta_b^*}{2}-1\leq C_7'\left|\theta_a^*-\theta_b^*\right|^2$$

for some constant $C_7' > 0$ when $\left|\theta_a^* - \theta_b^*\right| \leq C$. (3.80) is because of Lemma 3.7.8 and $\frac{p}{\beta\vee 1/n} \gg 1$. Note that δ_2' can also be chosen independent of i, j, θ^*, r^*.

Thus, we can further lower bound (3.70) on $\mathcal{F}_1 \cap \mathcal{F}_2$:

$$\begin{aligned}
&\mathbb{P}_A\left(\sum_{l=1}^{L}\left(A_{ij}Z_{0l} + \sum_{k\neq i,j} A_{ik}Z_{kl} + A_{jk}\bar{Z}_{kl}\right) \geq 0\right) \\
&\geq C_4' \exp\left(-\sqrt{C_3' Lp \frac{(\theta_a^* - \theta_b^*)^2}{\beta \vee 1/n}} + L\nu_{r^*}(u_{r^*}^*)\right) \\
&\geq C_4' e^{-\sqrt{C_3' Lp \frac{(\theta_a^* - \theta_b^*)^2}{\beta \vee 1/n}} - L\sup_{u\in[0,1]}\left(-A_{ij}\nu_{0,r^*}(u) - \sum_{k\neq i,j}(A_{ik}\nu_{k,r^*}(u) + A_{jk}\bar{\nu}_{k,r^*}(u))\right)} \\
&\geq C_4' \exp\left(-\sqrt{C_3' Lp \frac{(\theta_a^* - \theta_b^*)^2}{\beta \vee 1/n}} - (1+\delta_2')2Lp \sum_{k\neq i,j} G_{i,j,k,\theta^*,r^*}(1/2)\right).
\end{aligned}$$

which finishes the proof. □

Now we are ready to prove the lower bound part of Theorem 3.3.2.

Proof of Theorem 3.3.2 (lower bound). We remark that $p \geq c_0(\beta \vee \frac{1}{n}) \log n$ necessarily implies $0 < \beta = o(1)$. It also implies $n \wedge \beta^{-1} \gg 1$ and $\frac{\beta \vee 1/n}{Lp} = o(1)$ which will be useful in the proof. Recall the definition of $r^{*(i,j)}$ in (3.36) for any $r^* \in \mathfrak{S}_n$ and $i, j \in [n]$ such that $i \neq j$.

For any $\theta^* \in \Theta_n(\beta, C_0)$, we have

$$
\begin{aligned}
&\inf_{\widehat{r}} \sup_{r^* \in \mathfrak{S}_n} \mathbb{E}_{(\theta^*, r^*)} \left[\mathsf{K}(\widehat{r}, r) \right] \\
&\geq \inf_{\widehat{r}} \frac{1}{n!} \sum_{r^* \in \mathfrak{S}_n} \frac{1}{n} \sum_{1 \leq i < j \leq n} \mathbb{P}_{(\theta^*, r^*)} \left(\widehat{r}_i < \widehat{r}_j, r_i^* > r_j^* \right) + \mathbb{P}_{(\theta^*, r^*)} \left(\widehat{r}_i > \widehat{r}_j, r_i^* < r_j^* \right) \\
&= \inf_{\widehat{r}} \frac{1}{n} \sum_{1 \leq i < j \leq n} \frac{1}{n!} \sum_{r^* \in \mathfrak{S}_n} \mathbb{P}_{(\theta^*, r^*)} \left(\widehat{r}_i < \widehat{r}_j, r_i^* > r_j^* \right) + \mathbb{P}_{(\theta^*, r^*)} \left(\widehat{r}_i > \widehat{r}_j, r_i^* < r_j^* \right) \\
&\geq \frac{1}{n} \sum_{1 \leq a < b \leq n} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \\
&\qquad\qquad \frac{1}{(n-2)!} \sum_{r^* : r_i^* = a, r_j^* = b} \inf_{\widehat{r}} \frac{\mathbb{P}_{(\theta^*, r^*)}(\widehat{r} \neq r^*) + \mathbb{P}_{(\theta^*, r^{*(i.j)})}(\widehat{r} \neq r^{*(i.j)})}{2} \\
&\geq \frac{1}{2n} \sum_{a=1}^{n} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \sum_{b \in [n] \setminus \{a\} : |a-b| \leq 1 \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp\beta^2}}} \\
&\qquad\qquad \frac{1}{(n-2)!} \sum_{r^* : r_i^* = a, r_j^* = b} \inf_{\widehat{r}} \frac{\mathbb{P}_{(\theta^*, r^*)}(\widehat{r} \neq r^*) + \mathbb{P}_{(\theta^*, r^{*(i.j)})}(\widehat{r} \neq r^{*(i.j)})}{2},
\end{aligned}
$$

where $C_1' > 0$ is a constant. Note that for any $a, b \in [n]$ such that $|a-b| \leq 1 \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp\beta^2}}$, we have $\left| \theta_a^* - \theta_b^* \right| \leq C_0 \left(\beta \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp}} \right) = o(1)$. Then by Lemma 3.7.10, we have

$$
\begin{aligned}
\inf_{\widehat{r}} \sup_{r^* \in \mathfrak{S}_n} \mathbb{E}_{(\theta^*, r^*)} \left[\mathsf{K}(\widehat{r}, r) \right] \geq \frac{1}{2n} \sum_{a=1}^{n} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \sum_{b \in [n] \setminus \{a\} : |a-b| \leq 1 \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp\beta^2}}} \\
C_3' \exp \left(- \sqrt{\frac{C_2' Lp (\theta_a^* - \theta_b^*)^2}{\beta \vee n^{-1}}} - (1 + \delta_1') Lp \sum_{k \neq a, b} \log \frac{(1 + e^{\theta_a^* - \theta_k^*})(1 + e^{\theta_b^* - \theta_k^*})}{\left(1 + e^{\frac{\theta_a^* + \theta_b^*}{2} - \theta_k^*} \right)^2} \right),
\end{aligned}
\tag{3.81}
$$

for some constant $C_2', C_3' > 0$ and some $\delta_1' = o(1)$. We are going to simplify the second term in the exponent. We have

$$\sum_{k\neq a,b} \log \frac{(1+e^{\theta_a^*-\theta_k^*})(1+e^{\theta_b^*-\theta_k^*})}{\left(1+e^{\frac{\theta_a^*+\theta_b^*}{2}-\theta_k^*}\right)^2} \leq \sum_{k\neq a,b} \frac{e^{\theta_a^*-\theta_k^*}+e^{\theta_b^*-\theta_k^*}-2e^{\frac{\theta_a^*+\theta_b^*}{2}-\theta_k^*}}{\left(1+e^{\frac{\theta_a^*+\theta_b^*}{2}-\theta_k^*}\right)^2} \tag{3.82}$$

$$= 2\sum_{k\neq a,b} \frac{(\cosh\frac{\theta_a^*-\theta_b^*}{2}-1)e^{\frac{\theta_a^*+\theta_b^*}{2}-\theta_k^*}}{\left(1+e^{\frac{\theta_a^*+\theta_b^*}{2}-\theta_k^*}\right)^2}$$

$$= (1+\delta_2')\frac{(\theta_a^*-\theta_b^*)^2}{4}\sum_{k\neq a,b}\psi'\left(\frac{\theta_a^*+\theta_b^*}{2}-\theta_k^*\right) \tag{3.83}$$

$$= (1+\delta_3')\frac{(\theta_a^*-\theta_b^*)^2}{4}\sum_{k\neq a}\psi'(\theta_a^*-\theta_k^*) \tag{3.84}$$

for some $\delta_2' = o(1), \delta_3' = o(1)$. Here (3.82) uses $\log(1+x) \leq x$. In (3.83) we use $\theta_a^* - \theta_b^* = o(1)$ and $\cosh x - 1 = (1+O(x))\frac{x^2}{2}$ when $x = o(1)$. From Lemma 3.7.6 we know $\sum_{k\neq a}\psi'(\theta_a^* - \theta_k^*) \asymp n \wedge \beta^{-1} \gg 1$. Then using this and the fact $\sup_x \left|\frac{\psi'(x+t)}{\psi'(x)} - 1\right| = O(t)$ when $t = o(1)$, we obtain (3.84). Using $\sum_{k\neq a}\psi'(\theta_a^*-\theta_k^*) \asymp n\wedge\beta^{-1}$ again and the fact $\left|\theta_a^*-\theta_b^*\right| \geq \beta$, there exists a constant $C_4' > 0$ such that

$$\frac{\sqrt{\frac{C_2' Lp(\theta_a^*-\theta_b^*)^2}{\beta\vee n^{-1}}}}{\frac{Lp(\theta_a^*-\theta_b^*)^2}{4}\sum_{k\neq a}\psi'(\theta_a^*-\theta_k^*)} \leq \frac{C_4'}{\sqrt{\frac{Lp\beta^2}{\beta\vee n^{-1}}}}.$$

Therefore, for an arbitrarily small constant $\delta > 0$, we have constant $C_5' > 0$, such that (3.81)

can be lower bounded by

$$
\begin{aligned}
& C_3' \exp\left(-\sqrt{\frac{C_2' Lp(\theta_a^* - \theta_b^*)^2}{\beta \vee n^{-1}}} - (1+\delta_3') \frac{Lp(\theta_a^* - \theta_b^*)^2}{4} \sum_{k \neq a} \psi'(\theta_a^* - \theta_k^*) \right) \\
& \geq C_3' \exp\left(- \left(1 + \delta_3' + \frac{C_4'}{\sqrt{\frac{Lp\beta^2}{\beta \vee n^{-1}}}} \right) \frac{Lp(\theta_a^* - \theta_b^*)^2}{4V_a(\theta^*)} \right) \\
& \geq C_5' \exp\left(-(1+\delta) \frac{Lp(\theta_a^* - \theta_b^*)^2}{4V_a(\theta^*)} \right). \qquad (3.85)
\end{aligned}
$$

So far, we obtain

$$
\begin{aligned}
& \inf_{\widehat{r}} \sup_{r^* \in \mathfrak{S}_n} \mathbb{E}_{(\theta^*, r^*)} \left[\mathsf{K}(\widehat{r}, r) \right] \\
& \geq \frac{1}{2n} \sum_{a=1}^{n} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \sum_{b \in [n] \setminus \{a\} : |a-b| \leq 1 \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp\beta^2}}} C_5' \exp\left(-(1+\delta) \frac{Lp(\theta_a^* - \theta_b^*)^2}{4V_a(\theta^*)} \right) \qquad (3.86) \\
& \geq \frac{1}{2n} \sum_{a=1}^{n} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \sum_{b=a+1} C_5' \exp\left(-(1+\delta) \frac{Lp(\theta_a^* - \theta_b^*)^2}{4V_a(\theta^*)} \right) \\
& \geq \frac{C_5'}{2n} \sum_{a=1}^{n} \exp\left(-(1+\delta) \frac{Lp(\theta_a^* - \theta_{a+1}^*)^2}{4V_a(\theta^*)} \right).
\end{aligned}
$$

Hence, we obtain the exponential rate.

In the following we are going to derive the polynomial rate for the regime $\frac{Lp\beta^2}{\beta \vee n^{-1}} \leq 1$. Note that for any $a, b \in [n]$ such that $|a-b| \leq 1 \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp\beta^2}}$, we have

$$
\frac{Lp(\theta_a^* - \theta_b^*)^2}{4V_a(\theta^*)} \lesssim \frac{Lp\beta^2}{n \wedge \beta^{-1}} \left(1 \vee \frac{\beta \vee n^{-1}}{Lp\beta^2} \right) \lesssim 1.
$$

Then from (3.86), there exist some constant $C_6', C_7' > 0$ such that

$$\begin{aligned}\inf_{\widehat{r}} \sup_{r^* \in \mathfrak{S}_n} \mathbb{E}_{(\theta^*, r^*)} \left[\mathsf{K}(\widehat{r}, r)\right] &\geq \frac{1}{2n} \sum_{a=1}^{n} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \sum_{b \in [n] \setminus \{a\} : |a-b| \leq 1 \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp\beta^2}}} C_6' \\ &\geq \frac{C_6'}{2} \left(1 \vee \sqrt{\frac{C_1'(\beta \vee n^{-1})}{Lp\beta^2}}\right) \\ &\geq C_7' \left(n \wedge \sqrt{\frac{\beta \vee n^{-1}}{Lp\beta^2}}\right),\end{aligned}$$

where the last inequality is due to $\frac{Lp\beta^2}{\beta \vee n^{-1}} \leq 1$ and the fact that the loss is at most n. □

3.7.3 Proof of Theorem 3.4.1

The following two lemmas are needed for the proof of Theorem 3.4.1. Recall that $\bar{y}_{ij}^{(1)} = \frac{1}{L_1} \sum_{l=1}^{L_1} y_{ijl}, i \neq j \in [n]$.

Lemma 3.7.11. *There exists a constant $C_1 > 0$ such that for any $\theta^* \in \Theta_n(\beta, C_0)$ and $r^* \in \mathfrak{S}_n$,*

$$\max_{i \in [n], j \in [n], i \neq j} \left| \bar{y}_{ij}^{(1)} - \psi(\theta^*_{r_i^*} - \theta^*_{r_j^*}) \right| \leq C_1 \sqrt{\frac{\log n}{L_1}}$$

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

Proof. This can be seen directly by standard Hoeffding's inequality and union bound argument. □

Lemma 3.7.12. *For L_1 such that $\frac{L_1}{\log n} \to \infty$ and constant $M \geq 1$, there exists $0 < \delta_0 = o(1)$ and $0 < \delta_1 = o(1)$ such that for any $\theta^* \in \Theta_n(\beta, C_0)$, any $r^* \in \mathfrak{S}_n$,*

$$\max_{i \in [n], j \in [n], i \neq j} \left| \bar{y}_{ij}^{(1)} - \psi(\theta^*_{r_i^*} - \theta^*_{r_j^*}) \right| \leq \delta_0$$

and

$$\bigcap_{i=1}^{n}\left\{\underline{\mathcal{E}_{1,i}} \subset \mathcal{E}_{1,i} \subset \overline{\mathcal{E}_{1,i}}\right\}$$

hold with probability at least $1 - O(n^{-10})$*, where*

$$\begin{aligned}
\mathcal{E}_{1,i} &= \left\{j \in [n] : \bar{y}_{ij}^{(1)} \le \psi(-2M)\right\}, \\
\underline{\mathcal{E}_{1,i}} &= \left\{j \in [n] : \theta^*_{r_j^*} \ge \theta^*_{r_i^*} + 2M + \delta_1\right\}, \\
\overline{\mathcal{E}_{1,i}} &= \left\{j \in [n] : \theta^*_{r_j^*} \ge \theta^*_{r_i^*} + 2M - \delta_1\right\}.
\end{aligned}$$

Proof. This is a direct consequence of Lemma 3.7.11 and $M = O(1)$. □

Now we are ready to prove Theorem 3.4.1.

Proof of Theorem 3.4.1. Let $\mathcal{F}^{(0)}$ be the event on which Lemma 3.7.12 holds. We will always work on this high probability event throughout the proof. Also, we will assume the regime $n\beta \to \infty$. The case $\beta \lesssim 1/n$ is trivial since we only have one league $S_1 = [n]$ if M is chosen to be a large enough constant.

To start the exposition, we define a series of quantities iteratively for all $k \in [K-1]$, with the base case $\underline{S_0} = \overline{S_0} = S_0' = \widetilde{S}_0 = \emptyset, \underline{u^{(0)}} = \overline{u^{(0)}} = 0$. Let

$$\underline{t_i^{(k)}} = \left|\left\{j \in [n]\backslash\widetilde{S}_{k-1} : j \in \underline{\mathcal{E}_{1,i}}\right\}\right|,$$

$$\overline{t_i^{(k)}} = \left|\left\{j \in [n]\backslash\widetilde{S}_{k-1} : j \in \overline{\mathcal{E}_{1,i}}\right\}\right|,$$

$$\underline{S_k} = \left\{i \in [n]\backslash\widetilde{S}_{k-1} : \left(1 + \frac{0.11}{C_0^2}\right) p\overline{t_i^{(k)}} \le h\right\}, \tag{3.87}$$

$$\overline{S_k} = \left\{i \in [n]\backslash\widetilde{S}_{k-1} : \left(1 - \frac{0.11}{C_0^2}\right) p\underline{t_i^{(k)}} \le h\right\}, \tag{3.88}$$

$$\overline{u^{(k)}} = \max\left\{ r_i^* : i \in [n] \backslash \widetilde{S}_{k-1}, \overline{t_i^{(k)}} \leq \frac{M}{\left(1 - \frac{0.12}{C_0^2}\right)\beta} \right\},$$

$$\underline{u^{(k)}} = \max\left\{ r_i^* : i \in [n] \backslash \widetilde{S}_{k-1}, \overline{t_i^{(k)}} \leq \frac{M}{\left(1 + \frac{0.11}{C_0^2}\right)\beta} \right\},$$

$$w_i^{(k)\prime} = \sum_{j \in \underline{S_k} \cap \overline{\mathcal{E}_{1,i}}} A_{ij} \mathbb{I}\left\{ j \in \mathcal{E}_{1,i} \right\},$$

$$S_k' = \left\{ i \in [n] \backslash \widetilde{S}_{k-1} : w_i^{(k)\prime} \leq h \right\},$$

$$\widetilde{S}_k = \uplus_{m=1}^k S_m'.$$

We make several remarks about these definition. The above definitions have essentially constructed another partition $S_1', S_2', ...$ using $w_i^{(k)\prime}$ comparing to Algorithm 1 using $w_i^{(k)}$. The relationship between S_k and S_k' will be made clear during the exposition. In fact, they will be equal with high probability. We should keep in mind that the partition using $w_i^{(k)\prime}$ is not a bona fide one since the definition uses $\overline{\mathcal{E}_{1,i}}$ and $\underline{S_k}$ which involve the knowledge of θ^*. However, this can be used in theoretical exploration. Our strategy is to show certain properties hold for partitions S_k', then $S_k = S_k'$ with high probability and thus inherits those properties.

We start with some simple but crucial facts which will act as building blocks in the proof.

- $\overline{t_i^{(k)}}$ and $\underline{t_i^{(k)}}$ has the following monotonicity property: for any $i, j \in [n] \backslash \widetilde{S}_{k-1}$ such that $r_i^* \leq r_j^*$,

$$\overline{t_i^{(k)}} \leq \overline{t_j^{(k)}}, \underline{t_i^{(k)}} \leq \underline{t_j^{(k)}}. \tag{3.89}$$

  This is direct from the definition.

- For any $i \in [n] \backslash \widetilde{S}_{k-1}$,

$$0 \le \overline{t_i^{(k)}} - \underline{t_i^{(k)}} \le \frac{2\delta_1}{\beta} + 1, \tag{3.90}$$

which comes from $\overline{t_i^{(k)}} - \underline{t_i^{(k)}} = \left| \left\{ j \in [n] \backslash \widetilde{S}_{k-1} : 2M - \delta_1 \le \theta^*_{r_j^*} - \theta^*_{r_i^*} < 2M + \delta_1 \right\} \right|$.

- We have

$$\left\{ i \in [n] \backslash \widetilde{S}_{k-1} : r_i^* \le \underline{u^{(k)}} \right\} = \underline{S_k} \subset \overline{S_k} \subset \left\{ i \in [n] \backslash \widetilde{S}_{k-1} : r_i^* \le \overline{u^{(k)}} \right\}. \tag{3.91}$$

Here $\underline{S_k} \subset \overline{S_k}$ is due to monotonicity (3.89) and $\underline{t_i^{(k)}} \le \overline{t_i^{(k)}}$ by definition. Recall $h = pM/\beta$. Using (3.90), for any $i \in \overline{S_k}$, we have $\overline{t_i^{(k)}} \le \underline{t_i^{(k)}} + \frac{2\delta_1}{\beta} + 1 \le \frac{h}{\left(1 - \frac{0.11}{C_0^2}\right)p} + \frac{2\delta_1}{\beta} + 1 \le \frac{M}{\left(1 - \frac{0.12}{C_0^2}\right)\beta}$. Hence, we have $\overline{S_k} \subset \left\{ i \in [n] \backslash \widetilde{S}_{k-1} : r_i^* \le \overline{u^{(k)}} \right\}$.

- $\underline{t_i^{(k)}}$, $\overline{t_i^{(k)}}$, $\underline{S_k}$, $\overline{S_k}$, $\underline{u^{(k)}}$, $\overline{u^{(k)}}$ are measurable with respect to the σ-algebra generated by $\widetilde{S}_{k-1}$. This is direct from the definition.

Now, we will prove the following statements by induction on k:

- With probability at least $1 - O(kn^{-10})$,

$$\underline{S_{k'}} \subset S_{k'} \subset \overline{S_{k'}} \tag{3.92}$$

for all $0 \le k' \le k$.

- With probability at least $1 - O(kn^{-10})$,

$$\left| \underline{S_{k'}} \right| \ge \left(\frac{1.7}{C_0} + \frac{1}{1 + \frac{0.11}{C_0^2}} \right) \frac{M}{\beta} \tag{3.93}$$

for all $1 \le k' \le k$ and $|S_0| = 0$.

- With probability at least $1 - O(kn^{-10})$,

$$\left|\overline{S_{k'}} \backslash \underline{S_{k'}}\right| \leq \overline{u^{(k')}} - \underline{u^{(k')}} \leq \frac{0.29M}{C_0\beta} \tag{3.94}$$

  for all $0 \leq k' \leq k$.

- With probability at least $1 - O(kn^{-10})$,

$$\left|\overline{S_{k'}}\right| \leq \left(2 + \frac{0.29}{C_0} + \frac{1}{1 - \frac{0.12}{C_0^2}}\right) \frac{M}{\beta} \tag{3.95}$$

  for all $0 \leq k' \leq k$.

- With probability at least $1 - O(kn^{-10})$,

$$S_{k'} = S'_{k'} \tag{3.96}$$

  for all $0 \leq k' \leq k$.

Now, suppose (3.92) - (3.96) hold until $k-1$, which is the case for $k = 1$. In the following, we are going to establish (3.92) - (3.96) for k one by one.

(Establishment of (3.92)). Recall that we assume $\mathcal{F}^{(0)}$ holds. On the intersection of all high probability events before k, we have $\widetilde{S}_{k-1} = S_1 \cup \ldots \cup S_{k-1}$. We sandwich $w_i^{(k)}$ by

$$\underline{w_i^{(k)}} = \sum_{j \in [n] \backslash \widetilde{S}_{k-1}} A_{ij} \mathbb{I}\left\{j \in \underline{\mathcal{E}_{1,i}}\right\} \leq w_i^{(k)} \leq \sum_{j \in [n] \backslash \widetilde{S}_{k-1}} A_{ij} \mathbb{I}\left\{j \in \overline{\mathcal{E}_{1,i}}\right\} = \overline{w_i^{(k)}}.$$

Recall the definition of S_k in Algorithm 1. Then we have $S_k = \left\{i \in [n] \backslash \widetilde{S}_{k-1} : w_i^{(k)} \leq h\right\}$. Hence, $\left\{i \in [n] \backslash \widetilde{S}_{k-1} : \overline{w_i^{(k)}} \leq h\right\} \subset S_k \subset \left\{i \in [n] \backslash \widetilde{S}_{k-1} : \underline{w_i^{(k)}} \leq h\right\}$. To prove (3.92), by

the definitions in (3.87) and (3.88), we only need to show

$$\left\{i \in [n]\backslash\widetilde{S}_{k-1} : \underline{w_i^{(k)}} \le h\right\} \subset \left\{i \in [n]\backslash\widetilde{S}_{k-1} : \left(1 - \frac{0.11}{C_0^2}\right) \underline{pt_i^{(k)}} \le h\right\},$$
$$\left\{i \in [n]\backslash\widetilde{S}_{k-1} : \left(1 + \frac{0.11}{C_0^2}\right) \overline{pt_i^{(k)}} \le h\right\} \subset \left\{i \in [n]\backslash\widetilde{S}_{k-1} : \overline{w_i^{(k)}} \le h\right\},$$

a sufficient condition of which is the following event:

$$\begin{aligned}
\mathcal{F}^{(k)} = &\left\{\forall i \in [n]\backslash\widetilde{S}_{k-1} \text{ such that } \underline{pt_i^{(k)}} \le \frac{h}{2} : \underline{w_i^{(k)}} \le h\right\} \\
&\bigcap \left\{\forall i \in [n]\backslash\widetilde{S}_{k-1} \text{ such that } \underline{pt_i^{(k)}} > \frac{h}{2} : \left(1 - \frac{0.11}{C_0^2}\right) \underline{pt_i^{(k)}} \le \underline{w_i^{(k)}}\right\} \\
&\bigcap \left\{\forall i \in [n]\backslash\widetilde{S}_{k-1} \text{ such that } \overline{pt_i^{(k)}} \le \frac{h}{2} : \overline{w_i^{(k)}} \le h\right\} \\
&\bigcap \left\{\forall i \in [n]\backslash\widetilde{S}_{k-1} \text{ such that } \overline{pt_i^{(k)}} > \frac{h}{2} : \overline{w_i^{(k)}} \le \left(1 + \frac{0.11}{C_0^2}\right) \overline{pt_i^{(k)}}\right\}.
\end{aligned}$$

Hence to prove (3.92), we only need to analyze $\mathbb{P}\left(\mathcal{F}^{(k)}\right)$.

Note that for any $j \in [n]\backslash\widetilde{S}_{k-1}$ we have $r_j^* > \underline{u^{(k-1)}}$ according to the definition of $\underline{S_{k-1}}$ in (3.91). Thus

$$\underline{w_i^{(k)}} = \sum_{\substack{j \in [n]\backslash\widetilde{S}_{k-1} \\ r_j^* > \underline{u^{(k-1)}}}} A_{ij} \mathbb{I}\left\{\theta_j^* \ge \theta_i^* + 2M + \delta_1\right\},$$

$$\overline{w_i^{(k)}} = \sum_{\substack{j \in [n]\backslash\widetilde{S}_{k-1} \\ r_j^* > \underline{u^{(k-1)}}}} A_{ij} \mathbb{I}\left\{\theta_j^* \ge \theta_i^* + 2M - \delta_1\right\}.$$

On the other hand, recall that $w_i^{(k-1)\prime} = \sum_{j \in \underline{S_{k-1}} \cap \overline{\mathcal{E}_{1,i}}} A_{ij} \mathbb{I}\left\{j \in \mathcal{E}_{1,i}\right\}$ which only involves A_{ij} such that $r_j^* \le \underline{u^{(k-1)}}$ due to (3.91). By (3.91) and induction hypothesis of (3.92) we further know $\underline{u^{(1)}} \le ... \le \underline{u^{(k-1)}}$. As a result, $w_i^{(1)\prime}, ..., w_i^{(k-1)\prime}$ are independent of $\underline{w_i^{(k)}}, \overline{w_i^{(k)}}$. Since $\widetilde{S}_{k-1}$ is determined by $w_i^{(1)\prime}, ..., w_i^{(k-1)\prime}$, it is also independent of $\underline{w_i^{(k)}}, \overline{w_i^{(k)}}$.

Therefore, conditional on $\widetilde{S}_{k-1}$, we have

$$\underline{w_i^{(k)}}|\widetilde{S}_{k-1} \sim \text{Binomial}(\underline{t_i^{(k)}}, p),$$

$$\overline{w_i^{(k)}}|\widetilde{S}_{k-1} \sim \text{Binomial}(\overline{t_i^{(k)}}, p).$$

Recall that $C_0 \geq 1$ is a constant and $h = pM/\beta \gg \log n$ since $p/(\beta \log n) \to \infty$ by assumption. By Bernstein inequality for the Binomial distributions together with a union bound argument, we have $\mathbb{P}\left(\mathcal{F}^{(k)}|\widetilde{S}_{k-1}\right) \geq 1 - O(n^{-10})$. Since this holds for all $\widetilde{S}_{k-1}$, we have

$$\mathbb{P}\left(\mathcal{F}^{(k)}\right) \geq 1 - O(n^{-10}).$$

Therefore, we have proved (3.92).

(Establishment of (3.93)). We first present a simple fact from induction hypothesis:

$$\left\{i \in [n], r_i^* \leq \underline{u^{(k-1)}}\right\} \subset \widetilde{S}_{k-1} \subset \left\{i \in [n], r_i^* \leq \overline{u^{(k-1)}}\right\}. \tag{3.97}$$

The first containment is because (3.91) and (3.92) hold up to $k-1$. To prove the second containment, we only need to show $\overline{u^{(1)}} \leq \ldots \leq \overline{u^{(k-1)}}$. Notice that from (3.93) and (3.94) for $k-1$, we have $\left|\underline{S_{k-1}}\right| \geq \overline{u^{(k-2)}} - \underline{u^{(k-2)}}$. On the other hand, from (3.92) for $k-1$, we have $\left|\underline{S_{k-1}}\right| \leq \left|\left\{i \in [n] : r_i^* > \underline{u^{(k-2)}}, r_i^* \leq \overline{u^{(k-1)}}\right\}\right| \leq \overline{u^{(k-1)}} - \underline{u^{(k-2)}}$. Hence, we have $\overline{u^{(k-1)}} \geq \overline{u^{(k-2)}}$ and similarly we can show $\overline{u^{(l+1)}} \geq \overline{u^{(l)}}$ for any $l \leq k-2$, which proves $\overline{u^{(1)}} \leq \ldots \leq \overline{u^{(k-1)}}$.

Using (3.97), we have

$$\left|\underline{S_k}\right| = \left|\left\{ i \in [n] \backslash \widetilde{S}_{k-1} : \overline{t_i^{(k)}} \leq \frac{M}{\left(1 + \frac{0.11}{C_0^2}\right)\beta} \right\}\right|$$

$$\geq \left|\left\{ i \in [n] : r_i^* > \overline{u^{(k-1)}}, \left|\left\{ j \in [n] : r_j^* > \underline{u^{(k-1)}}, \theta^*_{r_j^*} \geq \theta^*_{r_i^*} + 2M - \delta_1 \right\}\right| \leq \frac{M}{\left(1 + \frac{0.11}{C_0^2}\right)\beta} \right\}\right|.$$

For any $i \in [n]$, since $\theta^* \in \Theta_n(\beta, C_0)$, we have

$$\left|\left\{ j \in [n] : r_j^* > \underline{u^{(k-1)}}, \theta^*_{r_j^*} \geq \theta^*_{r_i^*} + 2M - \delta_1 \right\}\right| \leq r_i^* - \left\lfloor \frac{2M - \delta_1}{C_0 \beta} \right\rfloor - \underline{u^{(k-1)}}.$$

Hence,

$$\begin{aligned}
\left|\underline{S_k}\right| &\geq \left|\left\{ i \in [n] : r_i^* > \overline{u^{(k-1)}}, r_i^* \leq \frac{M}{\left(1 + \frac{0.11}{C_0^2}\right)\beta} + \left\lfloor \frac{2M - \delta_1}{C_0 \beta} \right\rfloor + \underline{u^{(k-1)}} \right\}\right| \\
&\geq \frac{M}{\left(1 + \frac{0.11}{C_0^2}\right)\beta} + \left\lfloor \frac{2M - \delta_1}{C_0 \beta} \right\rfloor + \underline{u^{(k-1)}} - \overline{u^{(k-1)}} \\
&\geq \left(\frac{1.7}{C_0} + \frac{1}{1 + \frac{0.11}{C_0^2}} \right) \frac{M}{\beta}.
\end{aligned}$$

(Establishment of (3.94)). From (3.92), we have $\left|\overline{S_k} \backslash \underline{S_k}\right| \leq \overline{u^{(k)}} - \underline{u^{(k)}}$. Hence, we only need to show $\overline{u^{(k)}} - \underline{u^{(k)}} \leq \frac{0.29M}{C_0 \beta}$.

We are going to prove

$$\theta^*_{\underline{u^{(k-1)}}} \geq \theta^*_{\underline{u^{(k)}}} + 2M - \delta_1. \tag{3.98}$$

First, by (3.94) for $k-1$, (3.93), and (3.91), we have $\left|\left\{i \in [n] : \underline{u^{(k-1)}} \le r_i^* \le \underline{u^{(k)}}\right\}\right| \ge |\underline{S_k}|$ which leads to $\underline{u^{(k)}} \ge \overline{u^{(k-1)}}$. Let $b \in [n]$ be the index such that $r_b^* = \underline{u^{(k)}} + 1$. Then it means $b \in [n]\backslash\widetilde{S}_{k-1}$ and $\overline{t_b^{(k)}} > \frac{M}{\left(1+\frac{0.11}{C_0^2}\right)\beta}$. By the definition of $\overline{t_i^{(k)}}$, for any $i \in [n]\backslash\widetilde{S}_{k-1}$, we have

$$\left|\left\{j \in [n] : r_j^* \ge \underline{u^{(k-1)}}, \theta^*_{r_j^*} > \theta^*_{r_i^*} + 2M - \delta_1\right\}\right| \ge \left|\left\{j \in [n]\backslash\widetilde{S}_{k-1} : j \in \overline{\mathcal{E}_{1,i}}\right\}\right| = \overline{t_i^{(k)}},$$

which implies $\theta^*_{\underline{u^{(k-1)}}+\overline{t_i^{(k)}}} > \theta^*_{r_i^*} + 2M - \delta_1$. This means

$$\theta^*_{\underline{u^{(k-1)}}} > \theta^*_{r_i^*} + 2M - \delta_1 + \overline{t_i^{(k)}}\beta.$$

Considering the b index here, we have

$$\theta^*_{\underline{u^{(k-1)}}} \ge \theta^*_{\underline{u^{(k)}}+1} + 2M - \delta_1 + \frac{M}{\left(1+\frac{0.11}{C_0^2}\right)}. \tag{3.99}$$

Then using (3.94) for $k-1$, we have

$$\theta^*_{\overline{u^{(k-1)}}} \ge \theta^*_{\underline{u^{(k)}}+1} + 2M - \delta_1 + \frac{M}{\left(1+\frac{0.11}{C_0^2}\right)} - 0.29M \ge \theta^*_{\underline{u^{(k)}}} + 2M - \delta_1, \tag{3.100}$$

which proves (3.98). Then for any $i, j \in [n] \backslash \widetilde{S}_{k-1}$ such that $\underline{u^{(k)}} \leq r_i^* < r_j^*$, we have

$$\begin{aligned}
\overline{t_j^{(k)}} - \overline{t_i^{(k)}} &= \left| \left\{ l \in [n] \backslash \widetilde{S}_{k-1} : \theta^*_{r_l^*} \geq \theta^*_{r_j^*} + 2M - \delta_1 \right\} \right| - \left| \left\{ l \in [n] \backslash \widetilde{S}_{k-1} : \theta^*_{r_l^*} \geq \theta^*_{r_i^*} + 2M - \delta_1 \right\} \right| \\
&= \left| \left\{ l \in [n] \backslash \widetilde{S}_{k-1} : \theta^*_{r_i^*} + 2M - \delta_1 > \theta^*_{r_l^*} \geq \theta^*_{r_j^*} + 2M - \delta_1 \right\} \right| \\
&\geq \left| \left\{ r_l^* \geq \overline{u^{(k-1)}} : \theta^*_{r_i^*} + 2M - \delta_1 > \theta^*_{r_l^*} \geq \theta^*_{r_j^*} + 2M - \delta_1 \right\} \right| \\
&\geq \left| \left\{ l \in [n] : \theta^*_{r_i^*} + 2M - \delta_1 > \theta^*_{r_l^*} \geq \theta^*_{r_j^*} + 2M - \delta_1 \right\} \right| \\
&\geq \frac{\theta^*_{r_i^*} - \theta^*_{r_j^*}}{C_0 \beta} \\
&\geq \frac{r_j^* - r_i^*}{C_0},
\end{aligned}$$

where in the first inequality we use (3.97) and in the second inequality we use (3.98). The last two inequalities are due to $\theta^* \in \Theta_n(\beta, C_0)$. As a result,

$$\overline{u^{(k)}} - \underline{u^{(k)}} \leq \frac{\frac{M}{\left(1 - \frac{0.12}{C_0^2}\right)\beta} - \frac{M}{\left(1 + \frac{0.11}{C_0^2}\right)\beta}}{C_0} \leq \frac{0.29M}{C_0 \beta}. \tag{3.101}$$

(Establishment of (3.95)). We first have

$$\left|\overline{S_k}\right| \leq \overline{u^{(k)}} - \underline{u^{(k-1)}} \leq \overline{u^{(k)}} - \left(\overline{u^{(k-1)}} - \frac{0.29M}{C_0 \beta} \right)$$

due to induction hypothesis on (3.94) for $k-1$ and $\left\{ i \in [n] : r_i^* \leq \underline{u^{(k-1)}} \right\} \subset \widetilde{S}_{k-1}$. By the definition of $\overline{u^{(k)}}$, similar to the proof of (3.99), we can show

$$\theta^*_{\underline{u^{(k-1)}}} \leq \theta^*_{\underline{u^{(k)}}} + 2M - \delta_1 + \frac{M}{\left(1 - \frac{0.12}{C_0^2}\right)}$$

which implies

$$\overline{u^{(k)}} - \overline{u^{(k-1)}} \leq \frac{2M}{\beta} + \frac{M}{\left(1 - \frac{0.12}{C_0^2}\right)\beta}.$$

Therefore,

$$\left|\overline{S_k}\right| \leq \left(2 + \frac{0.29}{C_0} + \frac{1}{1 - \frac{0.12}{C_0^2}}\right)\frac{M}{\beta}.$$

(Establishment of (3.96)). Define

$$\mathcal{F}^{(k)\prime} = \left\{ \min_{i \in [n]: r_i^* > \overline{u^{(k)}}} \sum_{j \in \underline{S_k}} A_{ij} \mathbb{I}\left\{ j \in \underline{\mathcal{E}_{1,i}} \right\} > h \right\}.$$

We are going to show the event $\mathcal{F}^{(k)\prime}$ is a sufficient condition for (3.96). By definition, since $\underline{S_k} \subset [n] \backslash \widetilde{S}_{k-1}$, we have $w_i^{(k)\prime} \leq w_i^{(k)}$ which implies $S_k \subset S_k'$. We only need to show $S_k' \subset S_k$. Note that for any i such that $r_i^* > \overline{u^{(k)}}$, we have

$$w_i^{(k)\prime} = \sum_{j \in \underline{S_k}} A_{ij} \mathbb{I}\left\{ j \in \mathcal{E}_{1,i} \right\} \geq \sum_{j \in \underline{S_k}} A_{ij} \mathbb{I}\left\{ j \in \underline{\mathcal{E}_{1,i}} \right\} > h,$$

which means $i \notin S_k'$ as $\mathcal{F}^{(k)\prime}$ is assumed to be true. Hence to show $S_k' \subset S_k$, we only need to show $S_k' \cap \{i \in [n] : r_i^* \leq \overline{u^{(k)}}\} \subset S_k$. Note that due to (3.94), for any $i, j \in [n]$, such that $r_i^* \leq \overline{u^{(k)}}$ and $r_j^* > \underline{u^{(k)}}$, we have $r_j^* > \underline{u^{(k)}}$, $\theta_{r_i^*}^* - \theta_{r_j^*}^* \geq \theta_{\overline{u^{(k)}}}^* - \theta_{\underline{u^{(k)}}}^* \geq -0.29M$. Then for

any i such that $r_i^* \le \overline{u^{(k)}}$, we have

$$\begin{aligned} w_i^{(k)\prime} - w_i^{(k)} &= \sum_{j \in \underline{S_k}} A_{ij} \mathbb{I}\left\{j \in \mathcal{E}_{1,i}\right\} - \sum_{j \in [n] \setminus \widetilde{S}_{k-1}} A_{ij} \mathbb{I}\left\{j \in \mathcal{E}_{1,i}\right\} \\ &\ge - \sum_{j \in [n]: r_j^* > \underline{u^{(k)}}} \mathbb{I}\left\{j \in \mathcal{E}_{1,i}\right\} \\ &\ge - \sum_{j \in [n]: r_j^* > \underline{u^{(k)}}} \mathbb{I}\left\{j \in \overline{\mathcal{E}_{1,i}}\right\} \\ &= - \sum_{j \in [n]: r_j^* > \underline{u^{(k)}}} \mathbb{I}\left\{\theta_{r_j^*}^* \ge \theta_{r_i^*}^* + 2M - \delta_1\right\} \\ &= 0, \end{aligned}$$

where first inequality is due to (3.91). Hence we have $S_k' \cap \{i \in [n] : r_i^* \le \overline{u^{(k)}}\} \subset S_k$ which leads to $S_k = S_k'$. As a result, to establish (3.96), we only need to analyze $\mathbb{P}\left(\mathcal{F}^{(k)\prime}\right)$.

The analysis of $\mathbb{P}\left(\mathcal{F}^{(k)\prime}\right)$ is similar to that of $\mathbb{P}\left(\mathcal{F}^{(k)}\right)$ in the establishment of (3.92). By a similar independence argument, we have

$$\left(\sum_{j \in \underline{S_k}} A_{ij} \mathbb{I}\left\{j \in \underline{\mathcal{E}_{1,i}}\right\}\right) \Big| \widetilde{S}_{k-1} \sim \text{Binomial}\left(\left|\underline{S_k} \cap \underline{\mathcal{E}_{1,i}}\right|, p\right)$$

for any $i \in [n]$ such that $r_i^* > \overline{u^{(k)}}$. From (3.100), we have

$$\underline{u^{(k)}} - \overline{u^{(k-1)}} \ge \frac{2M - \delta_1}{C_0 \beta}. \tag{3.102}$$

Together with (3.91) and (3.97), we have

$$\left|\underline{S_k} \cap \underline{\mathcal{E}_{1,i}}\right| \ge \left|\left\{j \in [n] : \overline{u^{(k-1)}} \le r_j^* \le \underline{u^{(k)}}, \theta_{r_j^*}^* \ge \theta_{r_i^*}^* + 2M + \delta_1\right\}\right| \ge \underline{u^{(k)}} - \overline{u^{(k-1)}} \ge \frac{2M - \delta_1}{C_0 \beta}.$$

Recall that $h = pM/\beta$ and $p/(\beta \log n) \to \infty$. By Bernstein inequality, we have

$$\mathbb{P}\left(\mathcal{F}^{(k)\prime}|\widetilde{S}_{k-1}\right) = \mathbb{P}\left(\min_{i\in[n]:r_i^*\geq\overline{u^{(k)}}}\left(\sum_{j\in\underline{S_k}} A_{ij}\mathbb{I}\left\{j\in\underline{\mathcal{E}_{1,i}}\right\}\right) > h\middle|\widetilde{S}_{k-1}\right) \geq 1 - O(n^{-10}).$$

Since this holds for all $\widetilde{S}_{k-1}$, we have $\mathbb{P}\left(\mathcal{F}^{(k)\prime}\right) \geq 1 - O(n^{-10})$.

(Establishment of (3.92) - (3.96) for K). We have (3.92) - (3.96) hold for each $k \in [K-1]$ with probability at least $1 - O(n^{-9})$. For the last partition, $S_K = [n]\backslash\widetilde{S}_{K-1}$. Let $S_{K,1}$ be the set obtained by Algorithm 1 before the terminating condition $[n] - |S_1| + ... + \left|S_{K,1}\right| \leq \left|S_{K,1}\right|/2$ is met. $\underline{S_{K,1}}, \overline{S_{K,1}}$ can be similarly defined and (3.92) - (3.96) should also be satisfied by $S_{K,1}$. Therefore,

$$|S_K| \leq \frac{3\left|S_{K,1}\right|}{2} \leq \frac{3}{2}\left|\overline{S_{K,1}}\right| \leq \frac{3}{2}\left(2 + \frac{0.29}{C_0} + \frac{1}{1-\frac{0.12}{C_0^2}}\right)\frac{M}{\beta},$$

$$|S_K| \geq \left|S_{K,1}\right| \geq \left|\underline{S_{K,1}}\right| \geq \left(\frac{1.7}{C_0} + \frac{1}{1+\frac{0.11}{C_0^2}}\right)\frac{M}{\beta}$$

and

$$\left\{i\in[n] : r_i^* > \overline{u^{(K-1)}}\right\} \subset S_K \subset \left\{i\in[n] : r_i^* > \underline{u^{(K-1)}}\right\}.$$

So far, we have establish (3.92) - (3.96) for any $k \in [K]$. Now we are ready to use them to prove the conclusions in Theorem 3.4.1.

1. Conclusion 1 is a consequence of (3.92) and (3.95).

2. For Conclusion 2, by (3.102) we have $\overline{u^{(k-2)}} < \underline{u_{(k-1)}} < \overline{u^{(k-1)}} < \underline{u_{(k)}} < \overline{u^{(k)}} <$

$\underline{u_{(k+1)}}$. Together with (3.92) and (3.97), we have

$$\left\{i \in [n] : \overline{u^{(k-2)}} < r_i^* \leq \underline{u^{(k+1)}}\right\} \subset S_{k-1} \cup S_k \cup S_{k+1} \subset \left\{i \in [n] : \underline{u^{(k-2)}} < r_i^* \leq \overline{u^{(k+1)}}\right\}.$$

Therefore, using (3.102), for any i such that $\underline{u^{(k-1)}} < r_i^* \leq \overline{u^{(k)}}$,

$$\begin{aligned}
&\left\{j \in [n] : \left|r_i^* - r_j^*\right| \leq \frac{1.51M}{C_0\beta}\right\} \\
&\subset \left\{j \in [n] : \underline{u^{(k-1)}} - \frac{1.51M}{C_0\beta} \leq r_j^* \leq \overline{u^{(k)}} + \frac{1.51M}{C_0\beta}\right\} \\
&\subset \left\{j \in [n] : \overline{u^{(k-2)}} < r_j^* \leq \underline{u^{(k+1)}}\right\} \subset S_{k-1} \cup S_k \cup S_{k+1}.
\end{aligned}$$

For $k = 1$ or K, only oneside needs to be considered and the property still holds due to the gap between $\underline{u^{(2)}}$ and $\overline{u^{(1)}}$ as well as the gap between $\underline{u^{(K-1)}}$ and $\overline{u^{(K-2)}}$.

3. For Conclusion 3, by (3.92) and (3.97), we have

$$\left\{i \in [n] : \overline{u^{(k-1)}} < r_i^* \leq \underline{u^{(k)}}\right\} \subset S_k \subset \left\{i \in [n] : \underline{u^{(k-1)}} < r_i^* \leq \overline{u^{(k)}}\right\}. \tag{3.103}$$

Using (3.102), we have

$$\max\left\{r_i^* : i \in S_k\right\} \leq \overline{u^{(k)}} < \underline{u^{(k+1)}} < \min\left\{r_i^* : i \in S_{k+2}\right\}.$$

Same results can be established for $\max\left\{r_i^* : i \in S_k\right\} < \min\left\{r_i^* : i \in S_l\right\}$ for any $l > k+2$.

4. For Conclusion 4, for any k and any i, the definition of $w_i^{(k)\prime}$ only involves j such that $j \in \overline{\mathcal{E}_{1,i}}$. This implies that the definition of S_k' only involves information of $(A_{ij}, \bar{y}_{ij}^{(1)})$ such that $\theta_{r_j^*}^* - \theta_{r_i^*}^* \geq 2M - \delta_1$. Thus S_k' can be used as the $\check{S}_k$ in Theorem 3.4.1.

5. For Conclusion 5, note that for any $k \in [K]$ and $i \in S_k$, we have

$$\begin{aligned}\left|\left\{j \in [n] : |\theta^*_{r^*_i} - \theta^*_{r^*_j}| \le \frac{M}{2}\right\} \cap S_k\right| &\ge \left|\left\{j \in [n] : |r^*_i - r^*_j| \le \frac{M}{2C_0\beta}\right\} \cap S_k\right| \\ &\ge \left|\left\{j \in [n] : |r^*_i - r^*_j| \le \frac{M}{2C_0\beta}, \overline{u^{(k-1)}} < r^*_j \le \underline{u^{(k)}}\right\}\right|.\end{aligned}$$

where the last inequality is by (3.103). Again by (3.103), we have $\underline{u^{(k-1)}} < r^*_i \le \overline{u^{(k)}}$. From (3.102) we know $\underline{u^{(k)}} - \overline{u^{(k-1)}} > M/(2C_0\beta)$. Then we have

$$\begin{aligned}\left|\left\{j \in [n] : |\theta^*_{r^*_i} - \theta^*_{r^*_j}| \le \frac{M}{2}\right\} \cap S_k\right| &\ge \frac{M}{2C_0\beta} - \max\left\{\overline{u^{(k-1)}} - \underline{u^{(k-1)}}, \overline{u^{(k)}} - \underline{u^{(k)}}\right\} \\ &\ge \frac{0.21M}{C_0\beta}\end{aligned}$$

where the last inequality is by (3.94).

The proof is complete. □

3.7.4 Proofs of Lemma 3.4.1, Lemma 3.4.2 and Lemma 3.4.3

We first prove Lemma 3.4.1 below.

Proof of Lemma 3.4.1. Recall that $\widehat{r}$ is obtained by sorting $\left\{\sum_{j\in[n]\setminus\{i\}} R_{ij}\right\}_{i\in[n]}$. Define

$$\widehat{s}_i = \sum_{j\in[n]\setminus\{i\}} R_{ij},$$

$$\widehat{R}_{ij} = \mathbb{I}\left\{\widehat{s}_i > \widehat{s}_j\right\} = \mathbb{I}\left\{\widehat{r}_i < \widehat{r}_j\right\}$$

and

$$s^*_i = \sum_{j\in[n]\setminus\{i\}} R^*_{ij}.$$

Observe that

$$r_i^* = n - s_i^*,$$

we have

$$\widehat{R}_{ij} = \mathbb{I}\left\{\widehat{s}_i > \widehat{s}_j\right\} = \mathbb{I}\left\{\widehat{s}_i - s_i^* + s_i^* > \widehat{s}_j - s_j^* + s_j^*\right\} = \mathbb{I}\left\{\widehat{s}_i - s_i^* - (\widehat{s}_j - s_j^*) > r_i^* - r_j^*\right\}.$$

Thus

$$\begin{aligned}
&\mathsf{K}(\widehat{r}, r^*) = \frac{1}{n} \sum_{1\le i<j\le n} \mathbb{I}\left\{\widehat{R}_{ij} \neq R_{ij}^*\right\} \\
&= \frac{1}{n} \sum_{1\le i<j\le n} \left|\mathbb{I}\left\{\widehat{s}_i - s_i^* - (\widehat{s}_j - s_j^*) > r_i^* - r_j^*\right\} - \mathbb{I}\left\{0 > r_i^* - r_j^*\right\}\right| \\
&\le \frac{1}{n} \sum_{1\le i<j\le n} \mathbb{I}\left\{\left|\widehat{s}_i - s_i^* - (\widehat{s}_j - s_j^*)\right| \ge \left|r_i^* - r_j^*\right|\right\} \\
&\le \frac{1}{n} \sum_{1\le i<j\le n} \mathbb{I}\left\{\left|\left|r_i^* - r_j^*\right| \le |\widehat{s}_i - s_i^*| + \left|\widehat{s}_j - s_j^*\right|\right|\right\} \\
&= \frac{1}{n} \sum_{k=1}^{n-1} \sum_{\substack{1\le i<j\le n \\ \left|r_i^* - r_j^*\right| = k}} \mathbb{I}\left\{k \le |\widehat{s}_i - s_i^*| + \left|\widehat{s}_j - s_j^*\right|\right\} \\
&\le \frac{1}{n} \sum_{k=1}^{n-1} \sum_{\substack{1\le i<j\le n \\ \left|r_i^* - r_j^*\right| = k}} \mathbb{I}\left\{\frac{k}{2} \le |\widehat{s}_i - s_i^*|\right\} + \mathbb{I}\left\{\frac{k}{2} \le \left|\widehat{s}_j - s_j^*\right|\right\} \\
&\le \frac{2}{n} \sum_{i=1}^{n} \sum_{k=1}^{n-1} \mathbb{I}\left\{\frac{k}{2} \le |\widehat{s}_i - s_i^*|\right\} \le \frac{4}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \mathbb{I}\left\{k \le |\widehat{s}_i - s_i^*|\right\} \\
&= \frac{4}{n} \sum_{i=1}^{n} |\widehat{s}_i - s_i^*| \le \frac{4}{n} \sum_{i=1}^{n} \sum_{j\in[n]\setminus\{i\}} \left|R_{ij} - R_{ij}^*\right| = \frac{4}{n} \sum_{1\le i\neq j\le n} \mathbb{I}\left\{R_{ij} \neq R_{ij}^*\right\}
\end{aligned}$$

which completes the proof. □

Next, we prove Lemma 3.4.2.

Proof of Lemma 3.4.2. Recall $\mathcal{E} = \left\{(i,j) : 1 \le i < j \le n, \psi(-M) \le \overline{y}_{ij}^{(1)} \le \psi(M)\right\}$. Then on the event where Lemma 3.7.12 holds, $\mathcal{E}$ can be written as

$$\mathcal{E} = \left\{(i,j) : 1 \le i < j \le n, \left|\theta_{r_i^*} - \theta_{r_j^*}\right| \le M/2\right\} \uplus \left(\mathcal{E} \cap \left\{(i,j) : M/2 < \left|\theta_{r_i^*} - \theta_{r_j^*}\right| < 1.1M\right\}\right)$$

which implies $\check{A}_{ij} = A_{ij}\mathbb{I}\{(i,j) \in \mathcal{E}\}$. Moreover, on the event where Theorem 3.4.1 holds, $\check{S}_k = S_k, k \in [K]$ by Conclusion 4. This proves $\ell^{(k)}(\theta) = \check{\ell}^{(k)}(\theta)$ with probability at least $1 - O(n^{-8})$. $\check{\theta}^{(k)}$ and $\widehat{\theta}^{(k)}$ are equivalent up to a common shift since the Hessian in local MLE is well conditioned with probability at least $1 - O(n^{-8})$ due to Lemma 3.7.14. □

Finally, we need to prove Lemma 3.4.3, which requires us to first establish a few extra lemmas.

Lemma 3.7.13. *For any integer constant $C \ge 1$, define a matrix $M \in \{0,1\}^{n\times n}$ such that $M_{ij} = \mathbb{I}\{|i-j| \le n/C\}$. Let $\mathcal{L}_M$ be its Laplacian matrix such that*

$$[\mathcal{L}_M]_{ij} = \begin{cases} -M_{ij}, & \text{if } i \ne j \\ \sum_l M_{il}, & \text{if } i = j. \end{cases}$$

Let $\lambda_{\min,\perp}(\mathcal{L}_M)$ be the second smallest eigenvalue of $\mathcal{L}_M$, i.e.,

$$\lambda_{\min,\perp}(\mathcal{L}_M) = \min_{u\ne 0, \mathbb{1}_n^T u = 0} \frac{u^T \mathcal{L}_M u}{\|u\|}.$$

Then there exists another constant $C' > 0$ that only depends on C such that

$$\frac{1}{n}\lambda_{\min,\perp}(\mathcal{L}_M) = \inf_{\substack{x\in\mathbb{R}^n \\ \mathbb{1}_n^T x = 0 \\ \|x\|=1}} \frac{\sum_{|i-j|\le \frac{n}{C}} (x_i - x_j)^2}{n} \ge C'.$$

Proof. We partition $[n]$ into $4C$ consecutive blocks such that each block contains either

$\lceil n/4C \rceil$ or $\lfloor n/4C \rfloor$ consecutive indices. Let these blocks be a sequence of disjoint sets $B_1, ..., B_{4C}$ such that $\max_{i \in B_k} i < \min_{j \in B_l} j$ if $k < l$. The idea is to lower bound the summation over the diagonal region by a sequence of square regions. Thus, for any $x \in \mathbb{R}^n, \mathbb{1}_n^T x = 0, \|x\| = 1$, we have

$$\begin{aligned}
\frac{1}{n} x^T \mathcal{L}_M x &= \frac{\sum_{|i-j| \leq \frac{n}{C}} (x_i - x_j)^2}{n} \\
&\geq \frac{1}{n} \left[\sum_{k,l \in [4C]: |k-l| \leq 1} \sum_{i \in B_k, j \in B_l} (x_i - x_j)^2 \right] \\
&= \sum_{k,l \in [4C]: |k-l| \leq 1} \left(\frac{|B_l|}{n} \sum_{i \in B_k} x_i^2 + \frac{|B_k|}{n} \sum_{i \in B_l} x_i^2 - 2 \left(\sum_{i \in B_k} \frac{x_i}{\sqrt{n}} \right) \left(\sum_{i \in B_l} \frac{x_i}{\sqrt{n}} \right) \right) \\
&= \sum_{k,l \in [4C]: |k-l| \leq 1} \left(p_l z_k + p_k z_l - 2 y_k y_l \right),
\end{aligned}$$

where we denote

$$y_k = \sum_{i \in B_k} \frac{x_i}{\sqrt{n}}, z_k = \sum_{i \in B_k} x_i^2, p_k = \frac{|B_k|}{n}.$$

For any $k \in [4C]$, we define

$$w_{2k-1} = \frac{y_k + \sqrt{p_k z_k - y_k^2}}{2p_k}, \text{ and } w_{2k} = \frac{y_k - \sqrt{p_k z_k - y_k^2}}{2p_k}. \tag{3.104}$$

Note that for any $k, l \in [4C]$, we have

$$\begin{aligned}
& p_l p_k \left((w_{2k-1} - w_{2l-1})^2 + (w_{2k-1} - w_{2l})^2 + (w_{2k} - w_{2l-1})^2 + (w_{2k} - w_{2l})^2 \right) \\
&= p_l p_k \left(2 \left(w_{2k-1}^2 + w_{2k}^2 + w_{2l-1}^2 + w_{2k}^2 \right) - 2 \left(w_{2k-1} + w_{2k} \right) \left(w_{2l-1} + w_{2l} \right) \right) \\
&= p_l p_k \left(\frac{z_k}{p_k} + \frac{z_l}{p_l} - 2 \frac{y_k y_l}{p_k p_l} \right) \\
&= p_l z_k + p_k z_l - 2 y_k y_l.
\end{aligned}$$

Then we have

$$\sum_{k,l\in[4C]:|k-l|\le 1} (p_l z_k + p_k z_l - 2y_k y_l)$$
$$= \sum_{k,l\in[4C]:|k-l|\le 1} p_l p_k \left((w_{2k-1} - w_{2l-1})^2 + (w_{2k-1} - w_{2l})^2 + (w_{2k} - w_{2l-1})^2 + (w_{2k} - w_{2l})^2 \right).$$

Note that w is a function of y, z, p which by definition satisfy: $\sum_{k=1}^{4C} y_k = 0$, $\sum_{k=1}^{4C} z_k = 1$, $\min_{k\in[4C]} p_k \ge 1/(5C)$, $\sum_{k=1}^{4C} p_k = 1$, and $y_k^2 \le p_k z_k$ for all $k \in [4C]$. Define a parameter space T:

$$T = \left\{ (y, z, p) : \sum_{k=1}^{4C} y_k = 0, \sum_{k=1}^{4C} z_k = 1, \min_{k\in[4C]} p_k \ge 1/(5C), \sum_{k=1}^{4C} p_k = 1, \text{ and } y_k^2 \le p_k z_k, \forall k \in [4C] \right\}.$$

Then we have

$$\frac{1}{n}\lambda_{\min,\perp}(\mathcal{L}_M)$$
$$\ge \inf_{(y,z,p)\in T}$$
$$\sum_{k,l\in[4C]:|k-l|\le 1} p_l p_k \left((w_{2k-1} - w_{2l-1})^2 + (w_{2k-1} - w_{2l})^2 + (w_{2k} - w_{2l-1})^2 + (w_{2k} - w_{2l})^2 \right), \tag{3.105}$$

where w is defined in (3.104).

Since T only depends on C, the quantity (3.105) also only depends on C. Then, (3.105) is equal to some constant $C' \ge 0$ only depending on C. We are going to show $C' > 0$. Otherwise, let the infimum of (3.105) be achieved at some w with $(y, z, p) \in T$. Then, we must have $w_{2k-1} = w_{2l-1} = w_{2k} = w_{2l}$ for any $k, l \in [4C]$ such that $|k - l| \le 1$. This has two immediately implications. First, for any $k \in [4C]$, since $w_{2k-1} = w_{2k}$, we have $y_k^2 = p_k z_k$ and $w_k = y_k/(2p_k)$, Second, since $w_{2k} = w_{2(k+1)}$ for any $k \in [4C - 1]$, there

exists some c such that $y_k/p_k = c$ for all $k \in [4C]$. Together with $y_k^2 = p_k z_k$, we obtain $c^2 p_k = z_k$ for all $k \in [C]$. Since $\sum_{k=1}^{4C} z_k = 1$ and $\sum_{k=1}^{4C} p_k = 1$, we conclude $c = \pm 1$. Then using $y_k/p_k = c$, we have $\sum_{k=1}^{4C} y_k = c \sum_{k=1}^{4C} p_k = c \neq 0$, which is a contradiction with $\sum_{k=1}^{4C} y_k = 0$. As a result, we obtain $\frac{1}{n}\lambda_{\min,\perp}(\mathcal{L}_M) \geq C' > 0$. □

Lemma 3.7.14. *Under the assumptions in Lemma 3.4.3,*

$$\lambda_{\min,\perp}(H(\eta^*)) = \min_{u \neq 0 : \mathbb{1}_m^T u = 0} \frac{u^T H(\eta^*) u}{\|u\|^2} \gtrsim mp$$

with probability at least $1 - O(n^{-10})$*, where* $H(\eta^*)$ *is the Hessian matrix of the objective (3.27), defined by*

$$H_{ij}(\eta^*) = \begin{cases} \sum_{l \in [m] \setminus \{i\}} B_{il} \psi'(\eta_i^* - \eta_l^*), & i = j, \\ -B_{ij} \psi'(\eta_i^* - \eta_j^*), & i \neq j. \end{cases}$$

Proof. We can decompose $H(\eta^*)$ into stochastic part $H(\eta^*) - \mathbb{E}(H(\eta^*))$ ans deterministic part $\mathbb{E}(H(\eta^*))$ and bound them separately. We first look at the stochastic part. Note that

$$H(\eta^*) - \mathbb{E}(H(\eta^*)) = D - \mathbb{E}(D) - \sum_{i<j} (B_{ij} - p_{ij}) \psi'(\eta_i^* - \eta_j^*)(E_{ij} + E_{ji})$$

where $D = diag\{D_1, ..., D_m\} = diag\{\sum_{j \neq 1} B_{ij} \psi'(\eta_1^* - \eta_j^*), ..., \sum_{j \neq m} B_{mj} \psi'(\eta_m^* - \eta_j^*)\}$; E_{ij} is an $m \times m$ matrix and has 1 on the entry (i, j) and 0 otherwise. We also have $\|(B_{ij} - p_{ij}) \psi'(\eta_i^* - \eta_j^*)(E_{ij} + E_{ji})\|_{\text{op}} \leq 1$ and $\| \sum_{i<j} (B_{ij} - p_{ij})^2 \psi'(\eta_i^* - \eta_j^*)^2 (E_{ij} + E_{ji})^2 \|_{\text{op}} \leq mp$. By matrix Bernstein inequality in [103], we have

$$\mathbb{P}\left(\| \sum_{i<j} (B_{ij} - p_{ij})(E_{ij} + E_{ji}) \|_{\text{op}} > t \right) \leq 2m \exp\left(-\frac{t^2/2}{mp + \frac{t}{3}} \right).$$

Taking $t = C_1'\sqrt{mp\log n}$ for some large enough constant $C_1' > 0$, we have

$$\|\sum_{i<j}(B_{ij} - p_{ij})\psi'(\eta_i^* - \eta_j^*)(E_{ij} + E_{ji})\|_{\text{op}} \le C_1'\sqrt{mp\log n}$$

with probability at least $1 - O(n^{-10})$. Standard concentration using Bernstein inequality also yields

$$\|D - \mathbb{E}(D)\|_{\text{op}} \le C_2'\sqrt{mp\log n}$$

for some constant $C_2' > 0$ with probability at least $1 - O(n^{-10})$. Thus the stochastic part

$$\|H(\eta^*) - \mathbb{E}(H(\eta^*))\|_{\text{op}} \le (C_1' + C_2')\sqrt{mp\log n} = o(mp) \tag{3.106}$$

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

For the deterministic part, we first choose a constant integer $C' > 0$ such that for any $|i - j| \le \frac{n}{C'}$, $p_{ij} = p$. Thus for any unit vector $x \in \mathbb{R}^m$ such that $\mathbb{1}_m^T x = 0$,

$$\begin{aligned}
\frac{x^T\mathbb{E}(H(\eta^*))x}{m} &= \frac{\sum_{i<j} p_{ij}\psi'(\eta_i^* - \eta_j^*)(x_i - x_j)^2}{m} \\
&\ge \frac{\sum_{i<j,|i-j|\le\frac{m}{C'}} p\psi'(\eta_i^* - \eta_j^*)(x_i - x_j)^2}{m} \\
&\gtrsim p\frac{\sum_{i<j,|i-j|\le\frac{m}{C'}}(x_i - x_j)^2}{m} \qquad (3.107) \\
&\gtrsim p \qquad (3.108)
\end{aligned}$$

where (3.107) uses the boundedness of $\eta_1^* - \eta_m^*$; (3.108) is a consequence of Lemma 3.7.13 and C' is a constant independent of m and n. Combing (3.106) and (3.108) concludes the proof. □

The proof of Lemma 3.4.3 is given below.

Proof of Lemma 3.4.3. Since $\frac{L(\eta_i^* - \eta_j^*)^2}{2(W_i(\eta^*) + W_j(\eta^*))} \asymp mpL(\eta_i^* - \eta_j^*)^2$, we only need to consider

the situation where $mpL(\eta_i^*-\eta_j^*)^2$ is greater than a sufficiently large constant, since otherwise we can use the trivial bound $\mathbb{P}\left(\widehat{\eta}_i < \widehat{\eta}_j\right) \leq 1$. Define

$$\widetilde{\eta}_j = \eta_j^* - \frac{\sum_{l\in[m]\setminus\{j\}} B_{jl}(\bar{y}_{jl} - \psi(\eta_j^* - \eta_l^*))}{\sum_{l\in[m]\setminus\{j\}} B_{jl}\psi'(\eta_j^* - \eta_l^*)}.$$

Following the same argument used in the proof of Theorem 3.2 of [21] (see Equations (78), (79) and (81) of [21]), we have

$$|\widehat{\eta}_i - \widetilde{\eta}_i| \vee |\widehat{\eta}_j - \widetilde{\eta}_j| \leq \delta\Delta, \tag{3.109}$$

with probability at least $1 - O(n^{-7}) - \exp(-\Delta^{3/2}Lmp) - \exp\left(-\Delta^2 mpL\frac{mp}{\log(n+m)}\right)$, where $\Delta = \min\left(\eta_i^* - \eta_j^*, \left(\frac{\log(n+m)}{mp}\right)^{1/4}\right)$ and $\delta > 0$ is some sufficiently small constant. In fact, the bound (3.109) has only been established in [21] with a random graph that satisfies $p_{ij} = p$ for all $1 \leq i < j \leq m$. To establish (3.109) under the more general setting of interest, we first have

$$\lambda_{\min,\perp}(H(\eta^*)) = \min_{u\neq 0:\mathbb{1}_m^T u=0} \frac{u^T H(\eta^*)u}{\|u\|^2} \gtrsim mp, \tag{3.110}$$

with high probability, where $H(\eta^*)$ is the Hessian matrix of the objective (3.27). This is established in Lemma 3.7.14. Note that (3.110) is the only difference between the proofs of the current setting and the setting in [21]. With (3.109), we have

$$\begin{aligned}\mathbb{P}\left(\widehat{\eta}_i < \widehat{\eta}_j\right) \;\leq\; & \mathbb{P}\left(\widetilde{\eta}_j - \eta_j^* - (\widetilde{\eta}_i - \eta_i^*) > (1-\delta)\Delta\right) \\ & + O(n^{-7}) + \exp(-\Delta^{3/2}Lmp) + \exp\left(-\Delta^2 mpL\frac{mp}{\log(n+m)}\right).\end{aligned}$$

Define

$$\mathcal{B} = \left\{B : \left|\frac{\sum_{l\in[m]\setminus\{j\}} p_{jl}\psi'(\eta_j^* - \eta_l^*)}{\sum_{l\in[m]\setminus\{j\}} B_{jl}\psi'(\eta_j^* - \eta_l^*)} - 1\right| \leq \delta, \left|\frac{\sum_{l\in[m]\setminus\{i\}} p_{il}\psi'(\eta_i^* - \eta_l^*)}{\sum_{l\in[m]\setminus\{i\}} B_{il}\psi'(\eta_i^* - \eta_l^*)} - 1\right| \leq \delta'\right\}.$$

By Bernstein's inequality, we have $\mathbb{P}(B \in \mathcal{B}^c) \leq O(n^{-7})$ for some $\delta' = o(1)$. We then have

$$\begin{aligned}
& \mathbb{P}\left(\widetilde{\eta}_j - \eta_j^* - (\widetilde{\eta}_i - \eta_i^*) > (1-\delta)\Delta\right) \\
\leq & \sup_{B \in \mathcal{B}} \mathbb{P}\left(-\frac{\sum_{l \in [m]\setminus\{j\}} B_{jl}(\bar{y}_{jl} - \psi(\eta_j^* - \eta_l^*))}{\sum_{l \in [m]\setminus\{j\}} B_{jl}\psi'(\eta_j^* - \eta_l^*)} \right. \\
& \left. + \frac{\sum_{l \in [m]\setminus\{i\}} B_{il}(\bar{y}_{il} - \psi(\eta_i^* - \eta_l^*))}{\sum_{l \in [m]\setminus\{i\}} B_{il}\psi'(\eta_i^* - \eta_l^*)} > (1-\delta)\Delta \Big| B \right) + O(n^{-7}) \\
\leq & \exp\left(-\frac{(1-2\delta)L(\eta_i^* - \eta_j^*)^2}{2(W_i(\eta^*) + W_j(\eta^*))} \right) + O(n^{-7}).
\end{aligned}$$

Since

$$\exp(-\Delta^{3/2}Lmp) + \exp\left(-\Delta^2 mpL\frac{mp}{\log(n+m)}\right) \lesssim \exp\left(-\frac{(1-2\delta)L(\eta_i^* - \eta_j^*)^2}{2(W_i(\eta^*) + W_j(\eta^*))} \right) + O(n^{-7}),$$

we obtain the desired conclusion.

□

CHAPTER 4

POSTERIOR CONTRACTION OF BAYESIAN LOW-RANK MATRIX ESTIMATION

4.1 Introduction

This chapter collects several results of our ongoing research about Bayesian low-rank matrix estimation.

Consider the following model:

$$X = M_0 + Z \tag{4.1}$$

where $X, M_0, Z \in \mathbb{R}^{p\times q}$, $Z_{ij} \overset{ind}{\sim} N(0, \sigma^2)$. Given the observation matrix X, how do we estimate the signal M_0 which has a low-rank structure? Several other statistical problems are also directly related to the low-rankness of the underlying signal, for example, matrix completion [16, 17, 43, 46, 69, 110], reduced rank regression [14, 104, 106], and the more general trace regression [35, 45, 56, 90]. These problems, while different at first glance, all fit into the general model (1.2) with different linear operator $\mathcal{A}$. Besides being the identity operator in our model (4.1), in matrix completion, $\mathcal{A}$ can be defined as the partial observation operator at some but not all entries of M_0; in trace regression, $\mathcal{A}$ becomes the inner product of a sequence of sensing matrices A_i with M_0. In this thesis, however, we restrict ourselves to the simplest possible model (4.1) and hope to find and build tools along the way to eventually solve the most ambitious problem (1.2) in a Bayesian way completely.

To solve (4.1), it is known that frequentist method like nuclear norm minimization can achieve the optimal rate of convergence, see [38, 39, 50] for details. To be more specific, one can solve the following optimization:

$$\widehat{M} = \arg\min_{M} \frac{1}{2} \|X - M\|_F^2 + \lambda \|M\|_* \tag{4.2}$$

for some $\lambda > 0$ and this can be shown to achieve the minimax optimal rate:

$$\inf_{\hat{M}} \sup_{M_0 \in \mathcal{M}(p,q,r_0)} \mathbb{E}\left[\left\|\widehat{M} - M_0\right\|_F^2\right] \asymp \sigma^2 r_0 \max\{p, q\} \tag{4.3}$$

where the parameter space $\mathcal{M}(p, q, r) = \{M \in \mathbb{R}^{p\times q} : \text{rank}(M) \le r\}$.

For comparison, another class of problems, sparse vector estimation, or Gaussian sequence model, have several striking structural resemblances to the low-rank matrix estimation problem. In sparse vector estimation, we have the model

$$X = \theta + Z \tag{4.4}$$

where $X \in \mathbb{R}^p$ is the observation of some sparse vector $\theta \in \mathbb{R}^p$ perturbed by Gaussian noise $Z_i \overset{ind}{\sim} N(0, \sigma^2)$. It is the folklore of the statistics community that this problem can be solved by ℓ_1 norm regularization:

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\|X - \theta\|^2 + \lambda\|\theta\|_1 \tag{4.5}$$

for some $\lambda > 0$. Besides the similarity of the optimal frequentists' solution, both (4.2) and (4.5) have a thresholding representation, that is, (4.2) amounts to soft thresholding of singular values of X and (4.5) boils down to soft thresholding of the entries of X.

Given existing frequentists' results for (4.1), we ask that is there a Bayesian method that is also provably rate optimal? This chapter is devoted to give, at least partially, a positive answer to this problem. Theoretical works to investigate posterior distributions have been extensively studied. Specifically, proving that the posterior of some certain prior concentrates around the true signal under proper conditions provides a link between the frequentists' world and Bayesians' world, see [8, 53, 59, 98] and references therein for some classical results on posterior contraction in classical Bayesian and Bayesian nonparametrics.

Despite the huge success in these fields, research in posterior contraction in Bayesian high dimensional statistics has just begun. As the counterpart of low-rank matrix estimation, a provably optimal fully Bayesian estimator for sparse vector estimation is proposed in [20]. A general framework for constructing optimal prior and proof of posterior contraction for a wide class of problems called structured linear models is proposed in [49]. To achieve adaptivity, both priors in these works involve two steps, a model selection prior which selects the underlying structure and a heavy-tailed Laplace prior conditioned on the sampled structure. Our proposed rank-adaptive prior to solve the low-rank estimation problem in Section 4.4 also consists of similar two steps. Though quite general, the theory in [49] cannot be directly applied to our low-rank setting. This is in contrast to the resemblance of the problem structure of low-rank estimation and sparse vector estimation. This is because the structure space in the structured linear models in [49] is discrete and finite while in our low-rank setting it is not (there are uncountably many possible singular vectors). Another perspective is that we can view low-rankness as a kind of "sparsity". It is "sparse" under an unknown basis (after some unknown orthogonal transformations). On the other hand, sparse vectors are sparse under a given basis (the canonical basis in $\mathbb{R}^p$), which makes the problem a lot easier.

Using Bayesian methods to solve low-rank estimation has appeared as early as [40, 51, 67, 66, 107] in econometrics literature. Some literatures not only put prior on the signal, but also put a prior on the covariance structure of the noise. These literature requires knowledge of r_0 and is thus not rank-adaptive. Rank-adaptive procedures appear in [6, 71, 93, 111]. Roughly speaking, these priors approach the problem by factoring $M = LR$ where $L \in \mathbb{R}^{p\times r}, R \in \mathbb{R}^{r\times q}$ and put priors on L and R, then use the product of the posterior mean of L and R to estimate the signal. To achieve rank adaptivity, they fix a large r and then use a low-rank approximation to the estimator. Interested readers can resort to [3] for a short survey. Bayesian matrix completion [4] and Bayesian tensor estimation [100] have also appeared in recent literature. Most existing work in this area focuses on methodology and

computation. Little is known about the statistical properties of the posteriors using the framework of posterior contraction. For the theoretical result in [3], the bound of the risk using the prior in [6] is obtained. However, the bound is not optimal and it involves the magnitude of the signal. Similarly, the theoretical result in [4] requires the low-rank signal to be bounded, which is also not optimal. In fact, when the signal is bounded, a trivial estimator $\widehat{M} = 0$ can already achieve the optimal rate. Whether or not there is a rank-adaptive fully Bayesian procedure that is rate optimal without any boundedness condition is still an open problem and that is our main target to achieve.

We now introduce the structure of this chapter. Section 4.2 outlines the framework of posterior contraction of a Bayesian procedure, which enables us to explore optimality of different priors. Section 4.3 proposes a prior that is provably optimal when we know r_0. To be able to adapt to unknown r_0, Section 4.4 investigates the performance of a rank-adaptive prior and gives some preliminary result for the special case that the unknown r_0 is 1. Some useful technical lemmas and proofs are illustrated in Section 4.5 and Section 4.6.

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 $\{1, 2, ..., d\}$. 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 $a_n \leq Cb_n$ 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 $b_n \lesssim a_n$. 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 $\|v\|_\infty = \max_{1 \leq i \leq d} |v_i|$. For a matrix A, its operator norm is defined by $\|A\|$ and its Frobenius norm is defined by $\|A\|_F$. The inner product of two vectors or matrices is defined as $\langle x, y \rangle = \mathrm{tr}(x^T y)$. The notation $\mathbb{P}$, $\mathbb{E}, \pi$ are used for generic probability and expectation whose distribution is determined from the context.

4.2 Optimal Rate and Posterior Contraction

Model and Minimax Rate Recall the model in (4.1). Without loss of generality, we assume the noise in Z has unit variance, i.e. $\sigma = 1$ and X, M_0, Z are square matrices, i.e. $p = q$ throughout the chapter. Then the minimax rate in (4.3) becomes

$$\inf_{\hat{M}} \sup_{M_0 \in \mathcal{M}(p,r_0)} \mathbb{E}_{M_0} \left[\left\| \widehat{M} - M_0 \right\|_F^2 \right] \asymp r_0 p. \tag{4.6}$$

Moreover, we will only consider the case when $r_0 \ll p$ since this is the only nontrivial regime.

Posterior Contraction In a Bayesian procedure, we have a prior distribution $\pi(M)$ on the unknown parameter M_0 and the likelihood is $\pi(X|M)$ where X is our observations. Then the posterior distribution is given by

$$\pi(M \in B|X) = \frac{\int_B \pi(X|M)\pi(\mathrm{d}M)}{\int \pi(X|M)\pi(\mathrm{d}M)}$$

where B is some measurable set.

From the view of a frequentist, we can assume X is generated from a true parameter M_0. If we have chosen a good prior π, then $\pi(M \in B|X)$ should be large when B is a small neighborhood of M_0. For a prior with optimal contraction rate, the size of B should scale proportionally to the minimax optimal rate. Specifically, for our model (4.1), we will show that

$$\sup_{M_0 \in \mathcal{M}(p,r_0)} \mathbb{E}_{M_0} \left[\pi(\|M - M_0\|_F^2 \geq C r_0 p | X) \right] \tag{4.7}$$

can be arbitrarily small by choosing the constant C not depending on p, r_0, M_0, X.

4.3 When r_0 is Known

4.3.1 The Prior

In this section, we will propose a prior that has optimal posterior contraction rate when r_0 is known.

For any $M_0 \in \mathcal{M}(p, r_0)$, we have a singular value decomposition $M_0 = U_0 D_0 V_0^T$ where $U_0, V_0 \in O(p, r_0) = \{U \in \mathbb{R}^{p\times r_0} : U^T U = I\}$ and $D_0 \in \mathbb{R}^{r_0 \times r_0}$ is a diagonal matrix. Here $O(p, r_0)$ is called the Stiefel manifold. Enlightened by this decomposition, we propose the following prior $\pi(M)$:

- Sample $U \in \mathbb{R}^{p\times r_0}$ from the uniform distribution on $O(p, r_0)$;
- Sample $V \in \mathbb{R}^{p\times r_0}$ from the uniform distribution on $O(p, r_0)$;
- Sample matrix $D \in \mathbb{R}^{r_0 \times r_0}$ from the matrix Laplace distribution

$$\pi(D) \propto \exp\left(-\rho\sqrt{\sum_{1\le i,j\le r_0} D_{ij}^2}\right) = \exp(-\rho \|D\|_F)$$

  for some constant $\rho > 0$;
- U, V, D are independent and $M = UDV^T$.

This prior is a product measure on $O(p, r_0) \times O(p, r_0) \times \mathbb{R}^{r_0\times r_0}$. We can in fact write down the exact density of this prior with respect to $\pi(\mathrm{d}U) \times \pi(\mathrm{d}V) \times \mathrm{d}D$, where $\pi(\mathrm{d}U)$ and $\pi(\mathrm{d}V)$ are the uniform distribution on $O(p, r_0)$ and $\mathrm{d}D$ is the Lebesgue measure on $\mathbb{R}^{r_0\times r_0}$:

$$\pi(\mathrm{d}M) = \frac{1}{\frac{2\pi^{r_0^2/2}\Gamma(r_0^2)}{\rho^{r_0^2}\Gamma(r_0^2/2)}} \exp\left(-\rho \|D\|_F\right) \pi(\mathrm{d}U)\pi(\mathrm{d}V)\,\mathrm{d}D \tag{4.8}$$

where $\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}\,\mathrm{d}x, a > 0$ is the gamma function. Note that our prior D is not

a diagonal matrix like D_0 in the singular value decomposition of M_0. Given this prior, we are ready to state the main theorem, indicating that the posterior under the given prior is indeed concentrated around the truth at an optimal rate.

Theorem 4.3.1. *Using the prior in (4.8), there exists constant $C > 0$ independent of r_0, p, ρ such that for any $\delta > 0$,*

$$\sup_{M_0 \in \mathcal{M}(p,r_0)} \mathbb{E}\left[\pi(\|M - M_0\|_F^2 \geq C r_0 p | X)\right] < \delta \tag{4.9}$$

for all large enough r_0, p.

Theorem 4.3.1, to the best of our knowledge, is the first posterior contraction result on Bayesian low-rank matrix estimation that achieves the optimal rate. There is no boundedness condition on the true signal M_0 and the contraction holds for any $M_0 \in \mathcal{M}(p, r_0)$. The proof of Theorem 4.3.1 relies on a pilling argument that pills the set $\{M : \|M - M_0\|_F^2 \geq C r_0 p\}$ into slices and bound each slice separately. As one of the building blocks when bounding the posterior probability of each slice, a recent result in random matrix theory has been used, characterizing the distribution of the singular values of the submatrix of a uniformly distributed orthogonal matrix. The details of the proof is presented in Section 4.6.

4.4 Rank Adaptation

4.4.1 A Modified Prior

Though being rate optimal, the prior in Section 4.3 requires the knowledge, at least a close upper bound of the rank of the signal. To come up with a prior that is rank-adaptive, we need to put a prior on the rank. We also need make some slight modification on the prior distribution of the matrix conditional on a given rank. Our hierarchical prior is given below:

- Sample rank r from $\pi(r) \propto \frac{\Gamma(r)}{\Gamma(r/2)} \exp(-\tau r p) \mathbb{I}\{r \in [p]\}$ for some constant $\tau > 0$;

- Conditional on the sampled rank r, sample $M_r \in \mathbb{R}^{p\times p}$ such that $M_r = U_r D_r V_r^T$ where U_r, V_r are uniformly distributed on $O(p,r)$; D_r follows the matrix Laplace distribution restricted on the space of $r \times r$ diagonal matrix, i.e. $D_r \propto \exp\left(-\rho\sqrt{\sum_{i=1}^r D_{r,ii}^2}\right)$ and U_r, D_r, V_r are all independent.

Note that we have D_r restricted to be diagonal instead of letting D_r be any matrix in $\mathbb{R}^{r\times r}$ in the previous section when r_0 is known. We will see later in the proof, the restriction to diagonal matrix is crucial to achieve rank adaptation. However, we are still not able to prove this prior is rate optimal for general r_0 even when r_0 is known for the same reason we are not able to completely prove this prior is rank-adaptive for general r_0. Therefore, we still need the prior in Section 4.3 to have a complete result for general r_0.

Though proving this prior is rank-adaptive is a very hard problem, we have made some significant progress during the process and we believe some partial results may play an important role on the journey to fully solve it. One of the most critical step in the proof is to reduce the calculation of a probability from rank $r, r \geq r_0$ to r_0. This requires a result from matrix analysis linking diagonal elements of a matrix to its singular values. Though very close to the finish line, the only complete result we can prove is when the true rank $r_0 = 1$. This seems pretty specific, but the majority of our proof of Theorem 4.4.1 does not require $r_0 = 1$. The only step that we need $r_0 = 1$ is the last step which will be clear later in the proof in Section 4.6.

Theorem 4.4.1. *Assume $r_0 = 1$. For any $\tau > 0$, there exists constant $C > 0$ independent of ρ, p such that for any $\delta > 0$,*

$$\sup_{M_0\in\mathcal{M}(p,1)} \mathbb{E}\left[\pi(r \geq 1 + C|X)\right] \leq \delta \tag{4.10}$$

and

$$\sup_{M_0\in\mathcal{M}(p,1)} \mathbb{E}\left[\pi(\|M - M_0\|_F^2 \geq Cp|X)\right] \leq \delta \tag{4.11}$$

for all large enough p.

To facilitate the proof of Theorem 4.3.1 and Theorem 4.4.1, we need to articulate some lemmas in the next section. Note that these lemmas are actually quite general and do not need the condition $r_0 = 1$.

4.5 Some Technical Lemmas

Lemma 4.5.1. *For any $U, V \in O(p,r), U_0, V_0 \in O(p,r_0), D \in \mathbb{R}^{r\times r}, D_0 \in \mathbb{R}^{r_0\times r_0}$,*

$$\left\|UDV^T - U_0D_0V_0^T\right\|_F^2 = f(D,U,V) + g(U,V)$$

where $f(D,U,V) = \left\|D - U^TU_0D_0V_0^TV\right\|_F^2, g(U,V) = \|D_0\|_F^2 - \left\|U^TU_0D_0V_0^TV\right\|_F^2$.

Lemma 4.5.2. *For any matrices $M \in \mathcal{M}(p,r), M_0 \in \mathcal{M}(p,r_0)$,*

$$\left|\left\langle Z, \frac{M - M_0}{\|M - M_0\|_F}\right\rangle\right| \leq \sqrt{r + r_0}\,\|Z\|.$$

Lemma 4.5.3. *For any $U, V \in O(p,r), U_0, V_0 \in O(p,r_0), D_0 \in \mathbb{R}^{r_0\times r_0}$, let*

$$h(U) = \mathrm{tr}[D_0^T(I - U_0^TUU^TU_0)D_0].$$

Then

$$h(U) \vee h(V) \leq g(U,V) \leq 2[h(U) \vee h(V)]$$

Lemma 4.5.4. *Suppose $U \in \mathbb{R}^{p\times p}$ is uniformly distributed on the orthogonal group $O(p)$, then the top left $r \times r$ submatrix with $r \leq p/2$ can be represented as $O_1\Sigma O_2^T$ where $O_1, O_2 \in \mathbb{R}^{r\times r}$ are uniformly distributed on $O(r)$, $\Sigma = \mathrm{diag}\{\sigma_1, ..., \sigma_r\} \in \mathbb{R}^{r\times r}$ is a diagonal matrix*

with the joint density of $(\sigma_1, ..., \sigma_r)$ *being*

$$p(\sigma_1, ..., \sigma_r) \propto \prod_{i=1}^{r} (1-\sigma_i^2)^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} \left|\sigma_i^2 - \sigma_j^2\right| \mathbb{I}\{1 > \sigma_1, ..., \sigma_r > 0\} \tag{4.12}$$

and O_1, O_2, Σ *are independent.*

Lemma 4.5.5. *Suppose* $(\sigma_1, ..., \sigma_r)$ *follows the distribution in (4.12), then for any constants* $w_1, ..., w_r > 0$ *and* $a > b > 0$*,*

$$\frac{\mathbb{P}(\sum_{i=1}^{r} w_i(1-\sigma_i^2) \le a)}{\mathbb{P}(\sum_{i=1}^{r} w_i(1-\sigma_i^2) \le b)} \le \left(\frac{2a}{b}\right)^{r(p-r)/2}$$

Lemma 4.5.6. *Let* $U, V \in O(p)$ *and* $\sum_{i=1}^{r_0} \lambda_i^2 = 1, \lambda_i \ge 0, i \in [r_0]$*. Then, if*

$$1 - t \le \sum_{i=1}^{p} \left(\sum_{j=1}^{r_0} \lambda_j U_{ij} V_{ij}\right)^2$$

for some $t > 0$*, we have*

$$1 - 3t \le \max_{\substack{S \subset [p] \\ |S| = r_0}} \sum_{i \in S} \left(\sum_{j=1}^{r_0} \lambda_j U_{ij} V_{ij}\right)^2.$$

The following Corollary is a direct consequence of Lemma 4.5.6 and union bound.

Corollary 4.5.1. *Let* U, V *be some random matrices distributed on* $O(p)$ *and* $\sum_{i=1}^{r_0} \lambda_i^2 = 1, \lambda_i \ge 0, i \in [r_0]$ *and* $r \ge r_0$*. Then,*

$$\mathbb{P}\left(1 - t \le \sum_{i=1}^{r} \left(\sum_{j=1}^{r_0} \lambda_j U_{ij} V_{ij}\right)^2\right) \le \binom{p}{r_0} \mathbb{P}\left(1 - 3t \le \sum_{i=1}^{r_0} \left(\sum_{j=1}^{r_0} \lambda_j U_{ij} V_{ij}\right)^2\right)$$

for any $t > 0$*.*

4.6 Proofs

4.6.1 Proof of Theorem 4.3.1

Proof of Theorem 4.3.1. There exists constants $C_2, C_3 > 0$ such that for any $C > 0$ and $C_1 > 0$ large enough, we have

$$
\begin{aligned}
&\mathbb{E}\left[\pi(\|M - M_0\|_F^2 \geq Cr_0p|X)\right] \\
&\leq \mathbb{P}(\|Z\| > C_1\sqrt{p}) \\
&\quad + \sum_{l=1}^{\infty} \mathbb{E}\left[\pi\left[(l+1)Cr_0p \geq \|M - M_0\|_F^2 \geq lCr_0p|X\right] \mathbb{I}\left\{\|Z\| \leq C_1\sqrt{p}\right\}\right] \\
&\leq C_2 \exp(-C_3C_1p) \qquad (4.13) \\
&\quad + \sum_{l=1}^{\infty} \mathbb{E}\left[\pi\left[(l+1)Cr_0p \geq \|M - M_0\|_F^2 \geq lCr_0p|X\right] \mathbb{I}\left\{\|Z\| \leq C_1\sqrt{p}\right\}\right] \qquad (4.14)
\end{aligned}
$$

where (4.13) is from standard random matrix theory.The choice of C and C_1 will be specified later in the proof. All we need to do now is to bound each term in the series (4.14). Next, note that for any $M \in \mathcal{M}(p, r_0)$

$$
\begin{aligned}
&\|X - M\|_F^2 = \|M - M_0\|_F^2 + \|Z\|_F^2 - 2\langle Z, M - M_0\rangle \\
&= \|M - M_0\|_F^2 - 2\left\langle Z, \frac{M - M_0}{\|M - M_0\|_F}\right\rangle \|M - M_0\|_F + \|Z\|_F^2 \\
&\begin{cases}
\leq \|M - M_0\|_F^2 + 2\sqrt{2r_0}\,\|Z\|\,\|M - M_0\|_F + \|Z\|_F^2 \leq 2\,\|M - M_0\|_F^2 + 2r_0\,\|Z\|^2 + \|Z\|_F^2 \\
\geq \|M - M_0\|_F^2 - 2\sqrt{2r_0}\,\|Z\|\,\|M - M_0\|_F + \|Z\|_F^2 \geq \frac{\|M-M_0\|_F^2}{2} - 4r_0\,\|Z\|^2 + \|Z\|_F^2
\end{cases}
\end{aligned} \qquad (4.15)
$$

where (4.15) uses Lemma 4.5.2. Thus using Bayesian formula, we have

$$
\begin{aligned}
&\pi\left[(l+1)Cr_0p \geq \|M-M_0\|_F^2 \geq lCr_0p|X\right]\mathbb{I}\left\{\|Z\| \leq C_1\sqrt{p}\right\} \\
&= \frac{\int_{(l+1)Cr_0p\geq\|M-M_0\|_F^2\geq lCr_0p}\exp\left(-\frac{\|X-M\|_F^2}{2}\right)\pi(\mathrm{d}M)}{\int\exp\left(-\frac{\|X-M\|_F^2}{2}\right)\pi(\mathrm{d}M)}\mathbb{I}\left\{\|Z\| \leq C_1\sqrt{p}\right\} \\
&\leq \frac{\int_{(l+1)Cr_0p\geq\|M-M_0\|_F^2\geq lCr_0p}\exp\left(-\frac{\|M-M_0\|_F^2}{4}\right)\pi(\mathrm{d}M)}{\int\exp\left(-\|M-M_0\|_F^2\right)\pi(\mathrm{d}M)}e^{3r_0\|Z\|^2}\mathbb{I}\left\{\|Z\| \leq C_1\sqrt{p}\right\} \qquad (4.16) \\
&\leq e^{3C_1^2r_0p}\frac{\int_{(l+1)Cr_0p\geq\|M-M_0\|_F^2\geq lCr_0p}\exp\left(-\frac{\|M-M_0\|_F^2}{4}\right)\pi(\mathrm{d}M)}{\int\exp\left(-\|M-M_0\|_F^2\right)\pi(\mathrm{d}M)} \\
&\leq e^{3C_1^2r_0p}e^{-\frac{lCr_0p}{8}}\frac{\int_{(l+1)Cr_0p\geq\|M-M_0\|_F^2\geq lCr_0p}\exp\left(-\frac{\|M-M_0\|_F^2}{8}\right)\pi(\mathrm{d}M)}{\int_{\|M-M_0\|_F^2\leq\frac{lCr_0p}{8}}\exp\left(-\|M-M_0\|_F^2\right)\pi(\mathrm{d}M)} \qquad (4.17)
\end{aligned}
$$

where (4.16) is a result of (4.15). The following of the proof is devoted to bound the integral ratio in (4.17).

Recall that in Lemma 4.5.1, we have

$$
\begin{aligned}
\|M-M_0\|_F^2 &= f(D,U,V)+g(U,V) \\
&= \left\|D-U^TU_0D_0V_0^TV\right\|_F^2+\|D_0\|_F^2-\left\|U^TU_0D_0V_0^TV\right\|_F^2. \qquad (4.18)
\end{aligned}
$$

Leveraging the expression of the prior in (4.8) , we can further upper bound the integral

ratio in (4.17) by $\frac{\mathbb{I}_1}{\mathbb{I}_2}$ where

$$\mathbb{I}_1 = \int_{g(U,V)\le(l+1)Cr_0p}\int_{f(D,U,V)\le(l+1)Cr_0p}\exp\left[-\frac{\left\|UDV^T-M_0\right\|_F^2}{8}-\rho\left\|D\right\|_F\right]\mathrm{d}D\pi(\mathrm{d}U)\pi(\mathrm{d}V)$$

and

$$\mathbb{I}_2 = \int_{g(U,V)\le lCr_0p/16}\int_{f(D,U,V)\le lCr_0p/16}\exp\left[-\left\|UDV^T-M_0\right\|_F^2-\rho\left\|D\right\|_F\right]\mathrm{d}D\pi(\mathrm{d}U)\pi(\mathrm{d}V)$$

We then give upper bound for $\mathbb{I}_1$ and lower bound for $\mathbb{I}_2$ respectively. We first look at the inner integral with respect to D. To upper bound that in $\mathbb{I}_1$, we have

$$\begin{aligned}
&\int_{f(D,U,V)\le(l+1)Cr_0p}\exp\left[-\frac{\left\|UDV^T-M_0\right\|_F^2}{8}-\rho\left\|D\right\|_F\right]\mathrm{d}D\\
&=\int_{f(D,U,V)\le(l+1)Crp}\exp\left[-\frac{\left\|UDV^T-M_0\right\|_F^2}{8}+\rho\left\|UDV^T-M_0\right\|_F\right]\\
&\qquad\qquad\times\exp\left[-\rho\left\|UDV^T-M_0\right\|_F-\rho\left\|UDV^T\right\|_F\right]\mathrm{d}D\\
&\le e^{4\rho^2}\exp\left(-\rho\left\|M_0\right\|_F\right)\int_{f(D,U,V)\le(l+1)Cr_0p}\exp\left[-\frac{\left\|UDV^T-M_0\right\|_F^2}{16}\right]\mathrm{d}D \qquad (4.19)\\
&=e^{4\rho^2}\exp\left(-\rho\left\|M_0\right\|_F\right)\exp\left(-\frac{g(U,V)}{16}\right)\int_{f(D,U,V)\le(l+1)Cr_0p}\exp\left[-\frac{f(D,U,V)}{16}\right]\mathrm{d}D \qquad (4.20)\\
&\le e^{4\rho^2}\exp\left(-\rho\left\|M_0\right\|_F\right)\exp\left(-\frac{g(U,V)}{16}\right)\int_{\|D\|_F^2\le(l+1)Cr_0p}\exp\left[-\frac{\|D\|_F^2}{16}\right]\mathrm{d}D \qquad (4.21)
\end{aligned}$$

where (4.19) uses $\rho \left\| UDV^T - M_0 \right\|_F \le 4\rho^2 + \frac{\left\| UDV^T - M_0 \right\|_F^2}{16}$ and triangle inequality, (4.20) is a result of (4.18) and (4.21) is an application of change of variable. Now we lower bound the inner integral of $\mathbb{I}_2$:

$$\begin{aligned}
&\int_{f(D,U,V)\le lCr_0p/16} \exp\left[- \left\| UDV^T - M_0 \right\|_F^2 - \rho \left\| D \right\|_F \right] \mathrm{d}D \\
&\ge \exp\left(-g(U,V)\right) \int_{f(D,U,V)\le lCr_0p/16} \exp\left[-f(D,U,V) - \rho\sqrt{f(D,U,V)} - \rho \left\| U^T M_0 V \right\|_F \right] \mathrm{d}D \qquad (4.22) \\
&= e^{-\rho \left\| U^T M_0 V \right\|_F} \exp\left(-g(U,V)\right) \int_{\left\| D \right\|_F^2 \le lCr_0p/16} \exp\left[- \left\| D \right\|_F^2 - \rho \left\| D \right\|_F \right] \mathrm{d}D \qquad (4.23) \\
&\ge e^{-\rho \left\| U^T M_0 V \right\|_F} \exp\left(-g(U,V)\right) e^{-\rho^2} \int_{\left\| D \right\|_F^2 \le lCr_0p/16} \exp\left[-2 \left\| D \right\|_F^2 \right] \mathrm{d}D \qquad (4.24)
\end{aligned}$$

where (4.22) uses (4.18) and triangle inequality; (4.23) is due to an change of variable; (4.24) comes from $\rho \left\| D \right\|_F \le \rho^2 + \left\| D \right\|_F^2$.

Now we can upper bound $\mathbb{I}_1/\mathbb{I}_2$:

$$\begin{aligned}
\frac{\mathbb{I}_1}{\mathbb{I}_2} &\le e^{5\rho^2} \frac{\int_{\left\| D \right\|_F^2 \le (l+1)Cr_0p} \exp\left[-\frac{\left\| D \right\|_F^2}{16} \right] \mathrm{d}D}{\int_{\left\| D \right\|_F^2 \le lCr_0p/16} \exp\left[-2 \left\| D \right\|_F^2 \right] \mathrm{d}D} \\
&\quad \times \frac{\int_{g(U,V)\le (l+1)Cr_0p} \exp\left[-\frac{g(U,V)}{16} \right] \pi(\mathrm{d}U)\pi(\mathrm{d}V)}{\int_{g(U,V)\le lCr_0p/16} \exp\left[-g(U,V) + \rho \left\| M_0 \right\|_F - \rho \left\| U^T M_0 V \right\|_F \right] \pi(\mathrm{d}U)\pi(\mathrm{d}V)} \\
&\le e^{5\rho^2} \frac{\int_{\left\| D \right\|_F^2 \le (l+1)Cr_0p} \exp\left[-\frac{\left\| D \right\|_F^2}{16} \right] \mathrm{d}D}{\int_{\left\| D \right\|_F^2 \le lCr_0p/16} \exp\left[-2 \left\| D \right\|_F^2 \right] \mathrm{d}D} \frac{\int_{g(U,V)\le (l+1)Cr_0p} \exp\left[-\frac{g(U,V)}{16} \right] \pi(\mathrm{d}U)\pi(\mathrm{d}V)}{\int_{g(U,V)\le lCr_0p/16} \exp\left[-g(U,V) \right] \pi(\mathrm{d}U)\pi(\mathrm{d}V)} \qquad (4.25)
\end{aligned}$$

where (4.25) holds since $\rho \left\| M_0 \right\|_F - \rho \left\| U^T M_0 V \right\|_F \ge 0$. All we need to do now is to bound the integral ratio with respect to D and the integral ratio with respect to U, V.

To bound the ratio with respect to D, note that when $16b \le a$,

$$\frac{\int_{\|D\|_F^2 \le a} \exp\left[-\frac{\|D\|_F^2}{16}\right] \mathrm{d}D}{\int_{\|D\|_F^2 \le b} \exp\left[-2\|D\|_F^2\right] \mathrm{d}D} = \frac{\int_0^{\sqrt{a}} \exp\left[-\frac{x^2}{16}\right] x^{r_0^2-1}\,\mathrm{d}x}{\int_0^{\sqrt{b}} \exp\left[-2x^2\right] x^{r_0^2-1}\,\mathrm{d}x} \tag{4.26}$$

$$\le \frac{\int_0^{\sqrt{a}} \exp\left[-\frac{x^2}{16}\right] x^{r_0^2-1}\,\mathrm{d}x}{\int_0^{\sqrt{b/2}} \exp\left[-2x^2\right] x^{r_0^2-1}\,\mathrm{d}x}$$

$$= \left(\frac{2a}{b}\right)^{r_0^2/2} \frac{\int_0^{\sqrt{a}} \exp\left[-\frac{x^2}{16}\right] x^{r_0^2-1} dx}{\int_0^{\sqrt{a}} \exp\left[-\frac{b}{a}x^2\right] x^{r_0^2-1} dx} \tag{4.27}$$

$$\le (\frac{2a}{b})^{r_0^2/2} \tag{4.28}$$

where (4.26) changes the integral in polar coordinates; (4.27) is an application of change of variable and (4.28) comes from $16b \le a$. Plugging in $a = (l+1)Cr_0p, b = lCr_0p/16$, we obtain an upper bound of $8^{r_0^2}$. Finally, we examine the integral ratio with respect to U and V,

$$\frac{\int_{g(U,V)\le(l+1)Crp} \exp\left[-\frac{g(U,V)}{16}\right] dUdV}{\int_{g(U,V)\le lCrp/16} \exp\left[-g(U,V)\right] dUdV}$$

$$\le \frac{\int_{h(U)\vee h(V)\le(l+1)Crp} \exp\left[-\frac{h(U)\vee h(V)}{16}\right] dUdV}{\int_{h(U)\vee h(V)\le lCrp/64} \exp\left[-2(h(U)\vee h(V))\right] dUdV} \tag{4.29}$$

$$\le \frac{\int_{h(U),h(V)\le(l+1)Crp} \exp\left[-\frac{h(U)}{32}\right] \exp\left[-\frac{h(V)}{32}\right] dUdV}{\int_{h(U),h(V)\le lCrp/64} \exp\left[-2h(U)\right] \exp\left[-2h(V)\right] dUdV}$$

$$\le \left\{\frac{\int_{h(U)\le(l+1)Crp} \exp\left[-\frac{h(U)}{32}\right] dU}{\int_{h(U)\le lCrp/64} \exp\left[-2h(U)\right] dU}\right\}^2$$

$$\le \exp\left(\frac{lCrp}{16}\right) \left\{\frac{\pi\left(h(U)\le(l+1)Crp\right)}{\pi\left(h(U)\le lCrp/64\right)}\right\}^2.$$

Here (4.29) is a consequence of Lemma 4.5.3. Therefore we are left to bound

$$\begin{aligned}
&\frac{\pi\left(h(U) \leq (l+1)Cr_0p\right)}{\pi\left(h(U) \leq lCr_0p/64\right)} \\
&= \frac{\pi(\mathrm{tr}(D_0(I - U_0^T UU^T U_0)D_0) \leq (l+1)Cr_0p)}{\pi(\mathrm{tr}(D_0(I - U_0^T UU^T U_0)D_0) \leq lCr_0p/64)} \\
&= \frac{\pi(\mathrm{tr}(D_0(I - U_{11}U_{11}^T)D_0) \leq (l+1)Cr_0p)}{\pi(\mathrm{tr}(D_0(I - U_{11}U_{11}^T)D_0) \leq lCr_0p/64)} \qquad (4.30) \\
&= \frac{\pi(\mathrm{tr}((I - \Sigma^2)O_1^T D_0 O_1) \leq (l+1)Cr_0p)}{\pi(\mathrm{tr}((I - \Sigma^2)O_1^T D_0 O_1) \leq lCr_0p/64)} \qquad (4.31)
\end{aligned}$$

where U_{11} is the upper left corner $r_0 \times r_0$ submatrix of U, O_1 is uniformly distributed on $O(r_0)$, Σ is a diagonal matrix whose entries follow the distribution in (4.12) and Σ and O_1 are independent. Here (4.30) holds since U is uniformly distributed on $O(p, r_0)$ and (4.31) comes from Lemma 4.5.4.

Let $p_i \in \mathbb{R}^{r_0 \times r_0}, i \in [r_0]$ be the rows of O_1. By Lemma 4.5.5 and the independence between O_1 and $(\sigma_1, ..., \sigma_{r_0})$, we have

$$\frac{\pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i(1-\sigma_i^2) \leq (l+1)Cr_0p | O_1\right)}{\pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i(1-\sigma_i^2) \leq lCr_0p/64 | O_1\right)} \leq \left[128(1+\frac{1}{l})\right]^{r_0(p-r_0)/2} \leq 256^{r_0(p-r_0)/2}. \qquad (4.32)$$

Then

$$\begin{aligned}
&\frac{\pi\left(h(U) \leq (l+1)Cr_0p\right)}{\pi\left(h(U) \leq lCr_0p/64\right)} \\
&= \frac{\mathbb{E}\left[\frac{\pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i(1-\sigma_i^2) \leq (l+1)Cr_0p | O_1\right)}{\pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i(1-\sigma_i^2) \leq lCr_0p/64 | O_1\right)} \pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i(1-\sigma_i^2) \leq lCr_0p/64 | O_1\right)\right]}{\pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i(1-\sigma_i^2) \leq lCr_0p/64\right)} \\
&\leq 256^{r_0(p-r_0)/2} \frac{\mathbb{E}\left[\pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i(1-\sigma_i^2) \leq lCr_0p/64 | O_1\right)\right]}{\pi\left(\sum_{i=1}^{r_0} p_i^T D_0^2 p_i \lambda_i \leq lCr_0p/64\right)} \\
&= 256^{r_0(p-r_0)/2}
\end{aligned}$$

Combining all previous arguments, taking $C \geq 48C_1^2 + 192$ we have arrived at:

$$\begin{aligned}
&\pi\left[(l+1)Cr_0p \geq \|M - M_0\|_F^2 \geq lCr_0p|X\right] 1_{\{\|Z\| \leq C_1\sqrt{p}\}} \\
&\leq e^{3C_1^2 r_0 p} e^{-\frac{lCr_0p}{8}} e^{5\rho^2} e^{\frac{lCr_0p}{16}} 8^{r_0^2} 256^{r_0(p-r_0)} \\
&\leq e^{5\rho^2} \exp\left[-\left(\frac{lC}{16} - 3C_1^2 - 6\right) r_0 p\right] \\
&\leq e^{5\rho^2} \exp\left(-\frac{lCr_0p}{32}\right).
\end{aligned}$$

Thus, by (4.13) and (4.14), we get

$$\begin{aligned}
&\mathbb{E}\left[\pi(\|M - M_0\|_F^2 \geq Cr_0p|X)\right] \\
&\leq C_2 \exp(-C_3C_1p) + e^{5\rho^2} \sum_{l=1}^{\infty} \exp\left(-\frac{lCr_0p}{32}\right) \\
&\leq C_2 \exp(-C_3C_1p) + 2e^{5\rho^2} \exp\left(-\frac{Cr_0p}{32}\right).
\end{aligned}$$

Since C_1, C_2, C_3, ρ are all constants, then as long as $C \geq 48C_1^2 + 192$, we can make the bound as small as possible by taking p and r_0 large which completes the proof. □

4.6.2 Proof of Theorem 4.4.1

Proof of Theorem 4.4.1. We show (4.10) only since (4.11) is a direct corollary of (4.10) and a slight modification in the proof of Theorem 4.3.1.

We first write the proof with general r_0 and switch to the specific $r_0 = 1$ when we need it

in the last step. Similar to the proof of Theorem 4.3.1, for any $C, C_1 > 0$, we need to bound

$$
\begin{aligned}
&\mathbb{E}_{M_0}\left[\pi(r \geq (1+C)r_0|X)1_{\{\|Z\|\leq C_1\sqrt{p}\}}\right] \\
&\leq \sum_{r\geq(1+C)r_0} \mathbb{E}_{M_0}\left[\frac{\pi(r)\int \exp\left(-\frac{\|X-M_r\|_F^2}{2}\right)\pi(\mathrm{d}M_r)}{\pi(r_0)\int \exp\left(-\frac{\|X-M_{r_0}\|_F^2}{2}\right)\pi(\mathrm{d}M_{r_0})}1_{\{\|Z\|\leq C_1\sqrt{p}\}}\right] \\
&\leq \sum_{r\geq(1+C)r_0} \exp(-\tau(r-r_0)p)\mathbb{E}_{M_0}\left[\frac{\frac{\Gamma(r)}{\Gamma(r/2)}\int \exp\left(-\frac{\|X-M_r\|_F^2}{2}\right)\pi(\mathrm{d}M_r)}{\frac{\Gamma(r_0)}{\Gamma(r_0/2)}\int \exp\left(-\frac{\|X-M_{r_0}\|_F^2}{2}\right)\pi(\mathrm{d}M_{r_0})}1_{\{\|Z\|\leq C_1\sqrt{p}\}}\right] \\
&\leq \sum_{r\geq(1+C)r_0} \exp(-(\tau-C_1^2)rp)\exp((2C_1^2+\tau)r_0p)\frac{\frac{\Gamma(r)}{\Gamma(r/2)}\int \exp\left(-\frac{\|M_r-M_0\|_F^2}{4}\right)\pi(\mathrm{d}M_r)}{\frac{\Gamma(r_0)}{\Gamma(r_0/2)}\int \exp\left(-\|M_{r_0}-M_0\|_F^2\right)\pi(\mathrm{d}M_{r_0})}
\end{aligned}
\tag{4.33}
$$

where (4.33) can be derived using similar arguments in (4.16). By (4.8), we have

$$
\begin{aligned}
&\frac{\frac{\Gamma(r)}{\Gamma(r/2)}\int \exp\left(-\frac{\|M_r-M_0\|_F^2}{4}\right)\pi(\mathrm{d}M_r)}{\frac{\Gamma(r_0)}{\Gamma(r_0/2)}\int \exp\left(-\|M_{r_0}-M_0\|_F^2\right)\pi(\mathrm{d}M_{r_0})} \\
&= \frac{\frac{\pi^{r_0/2}}{\rho^{r_0}}\int_{U_r,V_r\in O(p,r)}\int_{\mathbb{R}^r}\exp\left(-\frac{\left\|U_rD_rV_r^T-M_0\right\|_F^2}{4}-\rho\|D_r\|_F\right)\mathrm{d}D_r\pi(\mathrm{d}U_r)\pi(\mathrm{d}V_r)}{\frac{\pi^{r/2}}{\rho^{r}}\int_{U_{r_0},V_{r_0}\in O(p,r_0)}\int_{\mathbb{R}^{r_0}}\exp\left(-\left\|U_{r_0}D_{r_0}V_{r_0}^T-M_0\right\|_F^2-\rho\|D_{r_0}\|_F\right)\mathrm{d}D_{r_0}\pi(\mathrm{d}U_{r_0})\pi(\mathrm{d}V_{r_0})} \\
&\leq \rho^{r-r_0}\frac{\int_{U_r,V_r\in O(p,r)}\int_{\mathbb{R}^r}\exp\left(-\frac{\left\|U_rD_rV_r^T-M_0\right\|_F^2}{4}-\rho\|D_r\|_F\right)\mathrm{d}D_r\pi(\mathrm{d}U_r)\pi(\mathrm{d}V_r)}{\int_{U_{r_0},V_{r_0}\in O(p,r_0)}\int_{\mathbb{R}^{r_0}}\exp\left(-\left\|U_{r_0}D_{r_0}V_{r_0}^T-M_0\right\|_F^2-\rho\|D_{r_0}\|_F\right)\mathrm{d}D_{r_0}\pi(\mathrm{d}U_{r_0})\pi(\mathrm{d}V_{r_0})}
\end{aligned}
\tag{4.34}
$$

where $\pi(\mathrm{d}U_r), \pi(\mathrm{d}V_r)$ are uniformly distributed on $O(p,r)$ and $\mathrm{d}D_r$ is the Lebesgue measure on $\mathbb{R}^{r\times r}$ diagonal matrices. The remaining of the proof is devoted to bound the integral ratio in (4.34).

Let Δ_r be the diagonal part of $U_r^T M_0 V_r$. Then, Lemma 4.5.1 can be modified to be

$$\left\| U_r M_r V_r^T - M_0 \right\|_F^2 = \|D_r - \Delta_r\|_F^2 + \|D_0\|_F^2 - \|\Delta_r\|_F^2 . \tag{4.35}$$

Leveraging (4.35), We first lower bound the inner integral on the denominator, using similar arguments as in the proof of Theorem 4.3.1,

$$\begin{aligned}
&\int_{\mathbb{R}^{r_0}} \exp\left(- \left\| U_{r_0} D_{r_0} V_{r_0}^T - M_0 \right\|_F^2 - \rho \|D_{r_0}\|_F \right) \mathrm{d}D_{r_0} \\
&= \exp\left(-(\|D_0\|_F^2 - \|\Delta_{r_0}\|_F^2) \right) \int_{\mathbb{R}^{r_0}} \exp\left(- \|D_{r_0} - \Delta_{r_0}\|_F^2 - \rho \|D_{r_0}\|_F \right) \mathrm{d}D_{r_0} \\
&\geq \exp\left(-(\|D_0\|_F^2 - \|\Delta_{r_0}\|_F^2) - \rho \|\Delta_{r_0}\|_F \right) \int_{\mathbb{R}^{r_0}} \exp\left(- \|D_{r_0}\|_F^2 - \rho \|D_{r_0}\|_F \right) \mathrm{d}D_{r_0} \\
&\geq e^{-\rho^2} \exp\left(-(\|D_0\|_F^2 - \|\Delta_{r_0}\|_F^2) - \rho \|\Delta_{r_0}\|_F \right) \int_{\mathbb{R}^{r_0}} \exp\left(-2 \|D_{r_0}\|_F^2 \right) \mathrm{d}D_{r_0} \\
&= (\pi/2)^{r_0/2} e^{-\rho^2} \exp\left(-(\|D_0\|_F^2 - \|\Delta_{r_0}\|_F^2) - \rho \|\Delta_{r_0}\|_F \right)
\end{aligned}$$

Now we upper bound the inner integral on the numerator, still using similar arguments in the proof of Theorem 4.3.1,

$$\begin{aligned}
&\int_{\mathbb{R}^r} \exp\left(-\frac{1}{4} \left\| U_r D_r V_r^T - M_0 \right\|_F^2 - \rho \|D_r\|_F \right) \mathrm{d}D_r \\
&= \int_{\mathbb{R}^r} \exp(-\frac{1}{4} \left\| U_r D_r V_r^T - M_0 \right\|_F^2 + \rho \left\| U_r D_r V_r^T - M_0 \right\|_F \\
&\qquad\qquad - \rho \left\| U_r D_r V_r^T - M_0 \right\|_F - \rho \left\| U_r D_r V_r^T \right\|_F) \, \mathrm{d}D_r \\
&\leq \int_{\mathbb{R}^r} \exp\left(-\frac{1}{4} \left\| U_r D_r V_r^T - M_0 \right\|_F^2 + \rho \left\| U_r D_r V_r^T - M_0 \right\|_F - \rho \|M_0\|_F \right) \mathrm{d}D_r \\
&\leq e^{2\rho^2} \exp\left(-\rho \|M_0\|_F \right) \int_{\mathbb{R}^r} \exp\left(-\frac{1}{8} \left\| U_r D_r V_r^T - M_0 \right\|_F^2 \right) \mathrm{d}D_r \\
&= e^{2\rho^2} \exp\left(-\rho \|M_0\|_F \right) \exp\left(-\frac{\|D_0\|_F^2 - \|\Delta_r\|_F^2}{8} \right) \int_{\mathbb{R}^r} \exp\left(-\frac{1}{8} \|D_r\|_F^2 \right) \mathrm{d}D_r \\
&= 2^{3r/2} \pi^{r/2} e^{2\rho^2} \exp\left(-\rho \|M_0\|_F \right) \exp\left(-\frac{\|D_0\|_F^2 - \|\Delta_r\|_F^2}{8} \right) .
\end{aligned}$$

Thus, we can upper bound the ratio in (4.34) by

$$
\begin{aligned}
&4^r(\pi/2)^{\frac{r-r_0}{2}}e^{3\rho^2}\frac{\int_{U_r,V_r\in O(p,r)}\exp\left(-\frac{\|D_0\|_F^2-\|\Delta_r\|_F^2}{8}\right)\pi(\mathrm{d}U_r)\pi(\mathrm{d}V_r)}{\int_{U_{r_0},V_{r_0}\in O(p,r_0)}\exp\left(-(\|D_0\|_F^2-\|\Delta_{r_0}\|_F^2)\right)\pi(\mathrm{d}U_{r_0})\pi(\mathrm{d}V_{r_0})}\\
&=4^r(\pi/2)^{\frac{r-r_0}{2}}e^{3\rho^2}\frac{\int_0^{+\infty}\exp(-\|D_0\|_F^2\,t)\mathbb{P}\left(1-\sum_{i=1}^{r}\left(\sum_{j=1}^{r_0}\lambda_jU_{ij}V_{ij}\right)^2\le 8t\right)\mathrm{d}t}{\int_0^{+\infty}\exp(-\|D_0\|_F^2\,t)\mathbb{P}\left(1-\sum_{i=1}^{r_0}\left(\sum_{j=1}^{r_0}\lambda_jU_{ij}V_{ij}\right)^2\le t\right)\mathrm{d}t}\\
&\le 4^r(\pi/2)^{\frac{r-r_0}{2}}e^{3\rho^2}\binom{p}{r_0}\frac{\int_0^{+\infty}\exp(-\|D_0\|_F^2\,t)\mathbb{P}\left(1-\sum_{i=1}^{r_0}\left(\sum_{j=1}^{r_0}\lambda_jU_{ij}V_{ij}\right)^2\le 24t\right)\mathrm{d}t}{\int_0^{+\infty}\exp(-\|D_0\|_F^2\,t)\mathbb{P}\left(1-\sum_{i=1}^{r_0}\left(\sum_{j=1}^{r_0}\lambda_jU_{ij}V_{ij}\right)^2\le t\right)\mathrm{d}t}
\end{aligned}
\tag{4.36}
$$

where $\lambda_j = D_{0,jj}/\|D_0\|_F$ and $(U_{ij})_{1\le i,j\le p},(V_{ij})_{1\le i,j\le p}$ are independently uniformly distributed on $O(p)$ and (4.36) have used Corollary 4.5.1.

Up till this point, we haven't used $r_0 = 1$ yet. However, in the following, we need to upper bound

$$
\frac{\mathbb{P}\left(1-\sum_{i=1}^{r_0}\left(\sum_{j=1}^{r_0}\lambda_jU_{ij}V_{ij}\right)^2\le 24t\right)}{\mathbb{P}\left(1-\sum_{i=1}^{r_0}\left(\sum_{j=1}^{r_0}\lambda_jU_{ij}V_{ij}\right)^2\le t\right)}
$$

and we have to assume $r_0 = 1$ since we do not have a proof for general r_0 yet. When $r_0 = 1$,

we have

$$
\begin{aligned}
&\frac{\mathbb{P}\left(1-\sum_{i=1}^{r_0}\left(\sum_{j=1}^{r_0}\lambda_j U_{ij}V_{ij}\right)^2\le 24t\right)}{\mathbb{P}\left(1-\sum_{i=1}^{r_0}\left(\sum_{j=1}^{r_0}\lambda_j U_{ij}V_{ij}\right)^2\le t\right)}=\frac{\mathbb{P}\left(1-U_{11}^2V_{11}^2\le 24t\right)}{\mathbb{P}\left(1-U_{11}^2V_{11}^2\le t\right)}\\
&\le\frac{\mathbb{P}(1-U_{11}^2\le 24t)\mathbb{P}(1-V_{11}^2\le 24t)}{\mathbb{P}(1-U_{11}^2\le t/2)\mathbb{P}(1-V_{11}^2\le t/2)}=\left(\frac{\mathbb{P}\left(\mathrm{Beta}(\frac{p-1}{2},\frac{1}{2})\le 24t\right)}{\mathbb{P}\left(\mathrm{Beta}(\frac{p-1}{2},\frac{1}{2})\le t/2\right)}\right)^2\\
&=\left(\frac{\int_0^{(24t)\wedge 1}(1-x)^{-1/2}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}{\int_0^{(t/2)\wedge 1}(1-x)^{-1/2}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}\right)^2\le\left(\frac{\int_0^{(24t)\wedge 1}(1-x)^{-1/2}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}{\int_0^{(t/2)\wedge 1}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}\right)^2\\
&=\left(\frac{2\int_{\sqrt{1-(24t)\wedge 1}}^1(1-y^2)^{\frac{p-1}{2}-1}\,\mathrm{d}y}{\int_0^{(t/2)\wedge 1}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}\right)^2\le\left(\frac{2^{\frac{p-1}{2}}\int_{\sqrt{1-(24t)\wedge 1}}^1(1-y)^{\frac{p-1}{2}-1}\,\mathrm{d}y}{\int_0^{(t/2)\wedge 1}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}\right)^2 \qquad (4.37)\\
&=\left(\frac{2^{\frac{p-1}{2}}\int_0^{1-\sqrt{1-(24t)\wedge 1}}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}{\int_0^{(t/2)\wedge 1}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}\right)^2\le\left(\frac{2^{\frac{p-1}{2}}\int_0^{(24t)\wedge 1}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}{\int_0^{(t/2)\wedge 1}x^{\frac{p-1}{2}-1}\,\mathrm{d}x}\right)^2\\
&=\left(2\frac{(24t)\wedge 1}{(t/2)\wedge 1}\right)^{p-1}\le 96^{p-1}
\end{aligned}
$$

where (4.37) uses a change of variable $x=1-y^2$ and $1+y\le 2$.

Combining all the pieces, we have proved

$$
\begin{aligned}
&\mathbb{E}_{M_0}\left[\pi(r\ge 1+C|X)\right]\\
&\le\mathbb{P}\left(\|Z\|>C_1\sqrt{p}\right)+\mathbb{E}_{M_0}\left[\pi(r\ge 1+C|X)1_{\{\|Z\|\le C_1\sqrt{p}\}}\right]\\
&\le C_2\exp\left(-C_3C_1p\right)\\
&\quad+\sum_{r\ge 1+C}\exp(-(\tau-C_1^2)rp)\exp((2C_1^2+\tau)p)\rho^{r-1}4^r(\pi/2)^{\frac{r-1}{2}}e^{3\rho^2}\binom{p}{1}96^{p-1}\\
&\le C_2\exp\left(-C_3C_1p\right)+C'(C,\rho)p\exp\left[-p\left((1+C)(\tau-C_1^2)-2C_1^2-\tau-5\right)\right] \qquad (4.38)
\end{aligned}
$$

where $C'(C,\rho)$ is some positive constant depending on C and ρ. Thus, by choosing τ, C, C_1

such that $\tau - C_1^2 > 0$, $C > \frac{5+3C_1^2}{\tau - C_1^2}$, we can make (4.38) as small as possible when p is large, which completes the proof. □

4.6.3 Proof of Technical Lemmas

Proof of Lemma 4.5.1. Note the following chain of equalities:

$$\begin{aligned}
&\left\| UDV^T - U_0 D_0 V_0^T \right\|_F^2 = \|D\|_F^2 + \|D_0\|_F^2 - 2\mathrm{tr}\left(V D^T U^T U_0 D_0 V_0^T \right) \\
&= \|D\|_F^2 - 2\mathrm{tr}\left(D^T U^T U_0 D_0 V_0^T V \right) + \left\| U^T U_0 D_0 V_0^T V \right\|_F^2 \\
&\quad + \|D_0\|_F^2 - \left\| U^T U_0 D_0 V_0^T V \right\|_F^2 \\
&= f(D, U, V) + g(U, V).
\end{aligned}$$

□

Proof of Lemma 4.5.2. By Von Neumann trace inequality, we have

$$\left| \left\langle Z, \frac{M - M_0}{\|M - M_0\|_F} \right\rangle \right| \leq \|Z\| \left\| \frac{M - M_0}{\|M - M_0\|_F} \right\|_* .$$

Since $M \in \mathcal{M}(p, r), M_0 \in \mathcal{M}(p, r)$, $\frac{M - M_0}{\|M - M_0\|_F}$ has at most $r + r_0$ non-zero singular values. An application of Cauchy-Schwarz inequality and the fact that $\frac{M - M_0}{\|M - M_0\|_F}$ has unit Frobenius norm concludes the proof. □

Proof of Lemma 4.5.3. To show the first inequality, note that

$$\mathrm{tr}(D_0^T V_0^T V V^T V_0 D_0 U_0^T U U^T U_0) \leq \mathrm{tr}(D_0^T V_0^T V V^T V_0 D_0).$$

Then

$$h(V) = \text{tr}(D_0^T D_0) - tr(D_0^T V_0^T V V^T V_0 D_0)$$
$$\leq \text{tr}(D_0^T D_0) - \text{tr}(D_0^T V_0^T V V^T V_0 D_0 U_0^T U U^T U_0) = g(U, V)$$

and similarly for $h(U)$.

To show the second inequality, note that

$$\text{tr}(V_0^T V V^T V_0 D_0^T (I - U_0^T U U^T U_0) D_0) \leq \text{tr}(D_0^T (I - U_0^T U U^T U_0) D_0).$$

Then

$$g(U, V) = \text{tr}(D_0^T D_0) - \text{tr}(D_0^T V_0^T V V^T V_0 D_0 U_0^T U U^T U_0)$$
$$= \text{tr}(D_0^T D_0) - \text{tr}(D_0^T V_0^T V V^T V_0 D_0) + \text{tr}(D_0^T V_0^T V V^T V_0 D_0) - \text{tr}(D_0^T V_0^T V V^T V_0 D_0 U_0^T U U^T U_0)$$
$$= \text{tr}(D_0^T (I - V_0^T V V^T V_0) D_0) + \text{tr}(V_0^T V V^T V_0 D_0^T (I - U_0^T U U^T U_0) D_0)$$
$$\leq h(V) + h(U) \leq 2[h(U) \vee h(V)].$$

□

Proof of Lemma 4.5.4. We write U into block matrix first,

$$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}$$

where $U_{11} \in \mathbb{R}^{r \times r}, U_{12} \in \mathbb{R}^{r \times (p-r)}$. By equation (5) in [34], We can write down the distribution of the r non-zero eigenvalues $(T_1, ..., T_r)$ of $U_{12} U_{12}^T$,

$$p(T_1, ..., T_r) \propto \prod_{i=1}^{r} T_i^{\frac{p-r}{2} - \frac{1}{2}} (1 - T_i)^{-1/2} \prod_{i<j} \left| T_i - T_j \right| \mathbb{I}\{1 > T_1, ..., T_r > 0\}$$

by plugging in $\beta = 1$ and $\gamma = -1$ in equation (5) of [34], representing the corresponding Altland-Zirnbauer symmetry classes listed in Table II of [34] since our setting falls into the regime of circular real ensemble of random matrix theory.

Note that $U_{11}U_{11}^T + U_{12}U_{12}^T = I$, thus $(\sqrt{1-T_1}, ..., \sqrt{1-T_r})$ is the set of singular values of U_{11}. A change of variable argument leads to the density expression of the singular values of U_{11} in (4.12).

Finally, it is easy to see that U_{11} has the same distribution as $VU_{11}\tilde{V}^T$ for any $V, \tilde{V} \in O(r)$. Thus, the left and right singular matrix of U_{11} must also be uniformly distributed on $O(r)$. Take the singular value decomposition of $U_{11} = L\Sigma R^T$, where the diagonal entries of Σ has been proved to follow (4.12) and take $\tilde{O}_1, \tilde{O}_2$ uniformly distributed on $O(r)$ and independent of U_{11}. We see that $\tilde{O}_1 L$ has the same distribution as L which is uniform, similarly for $R^T\tilde{O}_2^T$. Moreover, note that $\tilde{O}_1 L$ is actually independent of L and $R^T\tilde{O}_2$ is independent of R. Let $O_1 = \tilde{O}_1 L, O_2 = \tilde{O}_2 R$ concludes the proof. □

Proof of Lemma 4.5.5. Without loss of generality, we can assume $\sum_{i=1}^r w_i = 1$. Define

$\lambda_i = 1 - \sigma_i^2$, then a change of variable argument on the denominator leads to

$$
\begin{aligned}
&\frac{\mathbb{P}(\sum_{i=1}^r w_i(1-\sigma_i^2) \le a)}{\mathbb{P}(\sum_{i=1}^r w_i(1-\sigma_i^2) \le b)} \\
&= \frac{\int_{\substack{1-a \le \sum_{i=1}^r w_i\sigma_i^2 \\ 0<\sigma_1,\ldots,\sigma_r<1}} \prod_{i=1}^r (1-\sigma_i^2)^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\sigma_i^2 - \sigma_j^2|}{\int_{\substack{\sum_{i=1}^r w_i\lambda_i \le b \\ 0<\lambda_1,\ldots,\lambda_r<1}} \prod_{i=1}^r \lambda_i^{\frac{p-r}{2}-\frac{r+1}{2}} \frac{1}{\sqrt{1-\lambda_i}} \prod_{i<j} |\lambda_i - \lambda_j|} \\
&\le 2^{r(p-r)/2} \frac{\int_{\substack{1-a \le \sum_{i=1}^r w_i\sigma_i^2 \\ 0<\sigma_1,\ldots,\sigma_r<1}} \prod_{i=1}^r (1-\sigma_i)^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\sigma_i - \sigma_j|}{\int_{\substack{\sum_{i=1}^r w_i\lambda_i \le b \\ 0<\lambda_1,\ldots,\lambda_r<1}} \prod_{i=1}^r \lambda_i^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\lambda_i - \lambda_j|} \qquad (4.39) \\
&= \left(\frac{2a}{b}\right)^{r(p-r)/2} \frac{\int_{\substack{1-a \le \sum_{i=1}^r w_i\sigma_i^2 \\ 0<\sigma_1,\ldots,\sigma_r<1}} \prod_{i=1}^r (1-\sigma_i)^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\sigma_i - \sigma_j|}{\int_{\substack{\sum_{i=1}^r w_i\lambda_i \le a \\ 0<\lambda_1,\ldots,\lambda_r<a/b}} \prod_{i=1}^r \lambda_i^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\lambda_i - \lambda_j|} \qquad (4.40) \\
&\le \left(\frac{2a}{b}\right)^{r(p-r)/2} \frac{\int_{\substack{1-a \le \sum_{i=1}^r w_i\sigma_i^2 \\ 0<\sigma_1,\ldots,\sigma_r<1}} \prod_{i=1}^r (1-\sigma_i)^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\sigma_i - \sigma_j|}{\int_{\substack{\sum_{i=1}^r w_i\lambda_i \le a \\ 0<\lambda_1,\ldots,\lambda_r<1}} \prod_{i=1}^r \lambda_i^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\lambda_i - \lambda_j|} \qquad (4.41) \\
&= \left(\frac{2a}{b}\right)^{r(p-r)/2} \frac{\int_{\substack{1-a \le \sum_{i=1}^r w_i\sigma_i^2 \\ 0<\sigma_1,\ldots,\sigma_r<1}} \prod_{i=1}^r (1-\sigma_i)^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |\sigma_i - \sigma_j|}{\int_{\substack{1-a \le \sum_{i=1}^r w_i t_i \\ 0<t_1,\ldots,t_r<1}} \prod_{i=1}^r (1-t_i)^{\frac{p-r}{2}-\frac{r+1}{2}} \prod_{i<j} |t_i - t_j|} \qquad (4.42) \\
&\le \left(\frac{2a}{b}\right)^{r(p-r)/2} \qquad (4.43)
\end{aligned}
$$

where (4.39) uses $1 - \sigma_i^2 \le 2(1-\sigma_i)$, $\left|\sigma_i^2 - \sigma_j^2\right| \le 2\left|\sigma_i - \sigma_j\right|$ and $1/\sqrt{1-\lambda_i} > 1$; (4.40) is due to a rescaling of the variables; (4.41) comes from $a/b > 1$; (4.42) uses $t_i = 1 - \lambda_i$ and (4.43) is a consequence of $t_i > t_i^2$. □

Proof of Lemma 4.5.6. Let $\Lambda = diag\{\lambda_1, ..., \lambda_{r_0}, 0, ..., 0\} \in \mathbb{R}^{p\times p}, A = U\Lambda V^T \in \mathbb{R}^{p\times p}$. Thus, A is a matrix with at most rank r_0 and $(\lambda_1, ..., \lambda_{r_0}, 0, ..., 0)$ are its singular values.

The condition given is equivalent to

$$1 - t \leq \sum_{i=1}^{p} A_{ii}^2$$

and our goal is to prove

$$1 - 3t \leq \max_{\substack{S \subset [p] \\ |S| = r_0}} \sum_{i \in S} A_{ii}^2.$$

By Sing-Thompson Theorem, see [29, 99, 101], the absolute values of the diagonal elements of A, $\tilde{A} = (|A_{11}|, ..., |A_{pp}|)$ is weakly majorized by its singular values $y = (\lambda_1, ..., \lambda_{r_0}, 0, ..., 0)$, i.e. $\sum_{i=1}^{k} \tilde{A}_i^{\downarrow} \leq \sum_{i=1}^{k} y_i^{\downarrow}, k \in [p]$ where $\tilde{A}^{\downarrow}, y^{\downarrow}$ are vectors by permuting $\tilde{A}$ and y such that the coordinates are in descending order. By Theorem 10.2 in [109], there exists a vector $x \in \mathbb{R}^p$ such that $\tilde{A} \leq x$ and x is majorized by y, i.e. x is weakly majorized by y and $\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i$. Thus, we have

$$x = yD$$

where D is a doubly stochastic matrix using Theorem 10.8 in [109]. Furthermore, using the fact that a doubly stochastic matrix can be written as a convex combination of permutation matrices (Birkhoff's Theorem), we have

$$x = \sum_{\sigma \in S_p} c_\sigma \sigma(y)$$

where $\sum_{\sigma \in S_p} c_\sigma = 1, c_\sigma \geq 0$; $\sigma(y)$ is a vector obtained by permuting the elements of y according to the permutation operator σ; S_p refers to the permutation group of order p.

Then the condition implies

$$1 - t \leq \sum_{i=1}^{p} \tilde{A}_i^2 \leq \sum_{i=1}^{p} x_i^2 = \sum_{\sigma, \sigma' \in S_p} c_\sigma c_{\sigma'} \left\langle \sigma(y), \sigma'(y) \right\rangle.$$

Using

$$\left\langle \sigma(y), \sigma'(y) \right\rangle = 1 - \frac{\left\| \sigma(y) - \sigma'(y) \right\|^2}{2},$$

we have

$$\sum_{\sigma, \sigma' \in S_p} c_\sigma c_{\sigma'} \left\| \sigma(y) - \sigma'(y) \right\|^2 \leq 2t.$$

Therefore, there exists a permutation σ^* such that

$$\sum_{\sigma \in S_p} c_\sigma \left\| \sigma(y) - \sigma^*(y) \right\|^2 \leq 2t.$$

By Cauchy-Schwartz,

$$\sum_{\sigma \in S_p} c_\sigma \left\| \sigma(y) - \sigma^*(y) \right\| \leq \sqrt{2t},$$

which leads to

$$\left\| \sum_{\sigma \in S_p} c_\sigma \sigma(y) - \sigma^*(y) \right\| \leq \sqrt{2t}$$

by Triangle inequality, which is equivalent to

$$\left\| x - \sigma^*(y) \right\| \leq \sqrt{2t}.$$

Note that $\sigma^*(y)$ is only supported on at most r_0 entries. We define a set $S^* \subset [p]$ of size r_0 such that it covers the support. Then, we have

$$\sum_{i \in [p] \setminus S^*} x_i^2 \leq 2t.$$

Thus,

$$\sum_{i \in S^*} A_{ii}^2 = \sum_{i=1}^{p} \tilde{A}_i^2 - \sum_{i \in [p] \setminus S^*} \tilde{A}_i^2 \geq 1 - t - \sum_{i \in [p] \setminus S^*} x_i^2 \geq 1 - 3t$$

which completes the proof. □

REFERENCES

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

[2] Ali Ahmed and Justin Romberg. Compressive multiplexing of correlated signals. *IEEE Transactions on Information Theory*, 61(1):479–498, 2014.

[3] Pierre Alquier. Bayesian methods for low-rank matrix estimation: short survey and theoretical study. In *International Conference on Algorithmic Learning Theory*, pages 309–323. Springer, 2013.

[4] Pierre Alquier et al. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9(1):823–841, 2015.

[5] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

[6] S Derin Babacan, Martin Luessi, Rafael Molina, and Aggelos K Katsaggelos. Low-rank matrix completion by variational sparse bayesian learning. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2188–2191. IEEE, 2011.

[7] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 119–126, 2010.

[8] Andrew Barron, Mark J Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[9] David Beaudoin and Tim Swartz. A computationally intensive ranking system for paired comparison data. *Operations Research Perspectives*, 5:105–112, 2018.

[10] Nicolas Boumal. On intrinsic cramér-rao bounds for riemannian submanifolds and quotient manifolds. *IEEE transactions on signal processing*, 61(7):1809–1821, 2013.

[11] RALPH ALLAN BRADLEY and MILTON E TERRY. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[12] Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 268–276. Society for Industrial and Applied Mathematics, 2008.

[13] Mark Braverman and Elchanan Mossel. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.

[14] Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

[15] Cristina Butucea, Mohamed Ndaoud, Natalia A Stepanova, and Alexandre B Tsybakov. Variable selection with hamming loss. *The Annals of Statistics*, 46(5):1837–1875, 2018.

[16] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[17] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[18] Da Cao, Xiangnan He, Lianhai Miao, Yahui An, Chao Yang, and Richang Hong. Attentive group recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 645–654, 2018.

[19] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[20] Ismaël Castillo and Aad van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

[21] Pinhan Chen, Chao Gao, and Anderson Y Zhang. Partial recovery for top-k ranking: Optimality of mle and sub-optimality of spectral method. *The Annals of Statistics*, to appear.

[22] Pinhan Chen, Chao Gao, and Anderson Y Zhang. Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, to appear.

[23] Xi Chen, Sivakanth Gopi, Jieming Mao, and Jon Schneider. Competitive analysis of the top-k ranking problem. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1245–1264. SIAM, 2017.

[24] Yuxin Chen and Changho Suh. Spectral mle: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pages 371–380, 2015.

[25] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *The Annals of Statistics*, 47(4): 2204–2235, 2019.

[26] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Additional proofs for the paper "spectral method and regularized mle are both optimal for top-k ranking". *The Annals of Statistics*, 2019.

[27] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

[28] Eng Ung Choo and William C Wedley. A common framework for deriving preference values from pairwise comparison matrices. *Computers & operations research*, 31(6): 893–908, 2004.

[29] Moody T Chu. On constructing matrices with prescribed singular values and diagonal elements. *Linear algebra and its applications*, 288:11–22, 1999.

[30] Olivier Collier and Arnak Dalalyan. Permutation estimation and minimax matching thresholds. 2013.

[31] Olivier Collier and Arnak S Dalalyan. Minimax rates in permutation estimation for feature matching. *The Journal of Machine Learning Research*, 17(1):162–192, 2016.

[32] David Cossock and Tong Zhang. Subset ranking using regression. In *International Conference on Computational Learning Theory*, pages 605–619. Springer, 2006.

[33] László Csató. Ranking by pairwise comparisons for swiss-system tournaments. *Central European Journal of Operations Research*, 21(4):783–803, 2013.

[34] JP Dahlhaus, B Béri, and CWJ Beenakker. Random-matrix theory of thermal conduction in superconducting quantum dots. *Physical Review B*, 82(1):014536, 2010.

[35] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[36] Mike E Davies and Yonina C Eldar. Rank awareness in joint sparse recovery. *IEEE Transactions on Information Theory*, 58(2):1135–1146, 2012.

[37] Persi Diaconis and Ronald L Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):262–268, 1977.

[38] David Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.

[39] David L Donoho, Matan Gavish, and Andrea Montanari. The phase transition of matrix recovery from gaussian measurements matches the minimax mse of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, 2013.

[40] Jacques H Dreze. Bayesian limited information analysis of the simultaneous equations model. *Econometrica: Journal of the Econometric Society*, pages 1045–1075, 1976.

[41] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[42] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

[43] Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *The Annals of Statistics*, 46(6B):3481–3509, 2018.

[44] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[45] Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of econometrics*, 212(1):177–202, 2019.

[46] Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 315–340. JMLR Workshop and Conference Proceedings, 2011.

[47] Chao Gao. Phase transitions in approximate ranking. *arXiv preprint arXiv:1711.11189*, 2017.

[48] Chao Gao and Anderson Y Zhang. Iterative algorithm for discrete structure recovery. *arXiv preprint arXiv:1911.01018*, 2019.

[49] Chao Gao, Aad W van der Vaart, and Harrison H Zhou. A general framework for bayes structured linear models. *The Annals of Statistics*, 48(5):2848–2878, 2020.

[50] Matan Gavish and David L Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

[51] John Geweke. Bayesian reduced rank regression in econometrics. *Journal of econometrics*, 75(1):121–146, 1996.

[52] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[53] Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

[54] Kathrin Glau, Daniel Kressner, and Francesco Statti. Low-rank tensor approximation for chebyshev interpolation in parametric option pricing. *SIAM Journal on Financial Mathematics*, 11(3):897–927, 2020.

[55] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[56] Nima Hamidi and Mohsen Bayati. On low-rank trace regression under general sampling distribution. *arXiv preprint arXiv:1904.08576*, 2019.

[57] JA Hartigan. Bounding the maximum of dependent random variables. *Electronic Journal of Statistics*, 8(2):3126–3140, 2014.

[58] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: a bayesian skill rating system. In *Advances in neural information processing systems*, pages 569–576, 2007.

[59] Marc Hoffmann, Judith Rousseau, and Johannes Schmidt-Hieber. On adaptive posterior concentration rates. *The Annals of Statistics*, 43(5):2259–2295, 2015.

[60] David R Hunter et al. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

[61] Ali Jadbabaie, Anuran Makur, and Devavrat Shah. Estimation of skill distributions. *arXiv preprint arXiv:2006.08189*, 2020.

[62] Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Top-k ranking from pairwise comparisons: When spectral ranking is optimal. *arXiv preprint arXiv:1603.04153*, 2016.

[63] Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Optimal sample complexity of m-wise data for top-k ranking. In *Advances in Neural Information Processing Systems*, pages 1686–1696, 2017.

[64] M Chris Jones, James S Marron, and Simon J Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, 91 (433):401–407, 1996.

[65] Jyrki Katajainen and Jesper Larsson Träff. A meticulous analysis of mergesort programs. In *Italian Conference on Algorithms and Complexity*, pages 217–228. Springer, 1997.

[66] Frank Kleibergen and Richard Paap. Priors, posteriors and bayes factors for a bayesian analysis of cointegration. *Journal of Econometrics*, 111(2):223–249, 2002.

[67] Frank Kleibergen and Herman K Van Dijk. On the shape of the likelihood/posterior in cointegration models. *Econometric theory*, 10(3-4):514–551, 1994.

[68] Donald Ervin Knuth. *The art of computer programming*, volume 3. Pearson Education, 1997.

[69] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[70] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[71] Yew Jin Lim and Yee Whye Teh. Variational bayesian approach to movie rating prediction. In *Proceedings of KDD cup and workshop*, volume 7, pages 15–21. Citeseer, 2007.

[72] Tie-Yan Liu. *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.

[73] Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Optimality of spectral clustering for gaussian mixture model. *arXiv preprint arXiv:1911.00538*, 2019.

[74] Jordan J Louviere, David A Hensher, and Joffre D Swait. *Stated choice methods: analysis and applications*. Cambridge university press, 2000.

[75] Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

[76] R Duncan Luce. The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3):215–233, 1977.

[77] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

[78] Sydney C Ludvigson and Serena Ng. The empirical risk–return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222, 2007.

[79] Charles F Manski. The structure of random utility models. *Theory and decision*, 8(3): 229, 1977.

[80] Cheng Mao, Jonathan Weed, and Philippe Rigollet. Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory*, pages 821–847. PMLR, 2018.

[81] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

[82] Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.

[83] Tom Minka, Ryan Cleven, and Yordan Zaykov. Trueskill 2: An improved bayesian skill rating system. 2018.

[84] Shun Motegi and Naoki Masuda. A network-based dynamical ranking system for competitive sports. *Scientific reports*, 2:904, 2012.

[85] Mohamed Ndaoud and Alexandre B Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *IEEE Transactions on Information Theory*, 66(4):2517–2532, 2020.

[86] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

[87] Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with an unknown permutation: Statistical and computational limits. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 417–424. IEEE, 2016.

[88] Ashwin Pananjady, Cheng Mao, Vidya Muthukumar, Martin J Wainwright, and Thomas A Courtade. Worst-case versus average-case design for estimation from partial pairwise comparisons. *Annals of Statistics*, 48(2):1072–1097, 2020.

[89] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[90] Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[91] Sheldon M Ross and Erol A Peköz. *A second course in probability*. www. Probability-Bookstore. com, 2007.

[92] Thomas L Saaty. *Decision making for leaders: the analytic hierarchy process for decisions in a complex world*. RWS publications, 1990.

[93] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887, 2008.

[94] Robert Sedgewick. Implementing quicksort programs. *Communications of the ACM*, 21(10):847–857, 1978.

[95] Long Sha, Patrick Lucey, Yisong Yue, Peter Carr, Charlie Rohlf, and Iain Matthews. Chalkboarding: A new spatiotemporal query paradigm for sports play retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 336–347, 2016.

[96] Nihar Shah, Sivaraman Balakrishnan, Aditya Guntuboyina, and Martin Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, pages 11–20, 2016.

[97] Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

[98] Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

[99] Fuk-Yum Sing. Some results on matrices with prescribed diagonal elements and singular values. *Canadian Mathematical Bulletin*, 19(1):89–92, 1976.

[100] Taiji Suzuki. Convergence rate of bayesian tensor estimator and its minimax optimality. In *International Conference on Machine Learning*, pages 1273–1282. PMLR, 2015.

[101] Robert C Thompson. Singular values, diagonal elements, and convexity. *SIAM Journal on Applied Mathematics*, 32(1):39–63, 1977.

[102] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[103] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[104] Raja Velu and Gregory C Reinsel. *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media, 2013.

[105] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[106] Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346, 2007.

[107] A Zellner, C Min, and D Dallaire. Bayesian analysis of simultaneous equation and related models using the gibbs sampler and convergence checks. *HGB Alexander Research Foundation Working Paper, University of Chicago*, 1993.

[108] Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

[109] Fuzhen Zhang. *Matrix theory: basic results and techniques*. Springer Science & Business Media, 2011.

[110] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. *Advances in Neural Information Processing Systems*, 28:559, 2015.

[111] Mingyuan Zhou, Chunping Wang, Minhua Chen, John Paisley, David Dunson, and Lawrence Carin. Nonparametric bayesian matrix completion. In *2010 IEEE Sensor Array and Multichannel Signal Processing Workshop*, pages 213–216. IEEE, 2010.