2020

Statistical Estimation And Inference For Permutation Based Model

Shaokun Li
University of Pennsylvania

Statistical Estimation And Inference For Permutation Based Model

Abstract

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. People spend lots of time dealing with different kinds of data sets. The structure of the data plays an important role in statistics. Among different structures of data, one interesting structure is the permutation, which involves in different kinds of problems, such as recommender system, online gaming, decision making and sports tournament. This thesis is motivated by my interest in understanding the permutation in statistics. Comparing to the wide applications of permutation related model, little is known to the property of permutation in statistics. There are a variety challenges that arise and lots of problems waiting for us to explore in the permutation based model. This thesis aims to solve several interesting problems of the permutation based model in statistics, which may help us to understand more about the property and characteristic of permutation.

As a result of the various topics explored, this thesis is split into three parts. In Chapter 2, we discuss the estimation problem of unimodal SST model in the pairwise comparison problem. We prove that the CLS estimator is rate optimal up to a poly(log log n) factor and propose the computational efficient interval sorting estimator, as a computational efficient algorithm to the estimation problem. In Chapter 3, we shift our attention to the inference problem of the permutation based model. We study different kinds of inference problem, including the hypothesis testing problem in noisy sorting model and confidence set construction problems in generalized permutation based model. Network analysis is another important topic related to the permutation. In Chapter 4, we study the optimality of local belief propagation algorithm in the partial recovery problem of stochastic block model. We prove that local BP algorithm can reach the optimality in a certain regime. Moreover, in the regime where local BP algorithm may not achieve the optimal misclassified fraction, we will prove that local BP algorithm can be used in correcting other algorithms and get optimal algorithm to the partial recovery problem.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Tony Cai

Subject Categories

Statistics and Probability

STATISTICAL ESTIMATION AND INFERENCE FOR PERMUTATION BASED MODEL

Shaokun Li

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

Tony Cai, Daniel H. Silberberg Professor and Professor of Statistics

Graduate Group Chairperson

Nancy Zhang, Professor of Statistics

Dissertation Committee

Tony Cai, Daniel H. Silberberg Professor and Professor of Statistics

Richard Berk, Professor of Criminology and Statistics

Edward George, Professor of Statistics and Economics

Edgar Dobriban, Assistant Professor of Statistics

Dedicated to mom and dad

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Tony Cai, for the guidance, encouragement and patience through the past five years. His insight discussions help me a lot during my Ph.D study. Conducting the academic study regarding such a difficult topic couldn't be as simple as he made this for me. He is my mentor and a better advisor for my doctorate study beyond the imagination. Without his guidance and constant feedback, this PhD would not have been achievable.

I want to thank Richard Berk, Edward George, Edgar Dobriban for kindly agreeing to serve in my committee and sharing with me the comments. I would also want to thank all other professors in Wharton Statistics department, from whom I learn a lot about different topics of statistics. I am also grateful to all staff in the department, who arrange everything in our department perfectly well. Thank also goes to my fellow student colleagues and other friends at Upenn, with all the good memories we have here.

Finally, I would like to thank my parents, for the love and care in the past 27 years, and also for their unconditional trust, timely encouragement, and endless patience. Whenever I face difficulty in my life, , they have always been there for me. I am forever indebted to my parents for giving me the opportunities and experiences that have made me who I am. They selflessly encouraged me to explore new directions in life and seek my own destiny. This journey would not have been possible if not for them, and I dedicate this milestone to them.

Shaokun

Philadelphia, PA

April, 2020

ABSTRACT

STATISTICAL ESTIMATION AND INFERENCE FOR PERMUTATION BASED MODEL

Shaokun Li

Tony Cai

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. People spend lots of time dealing with different kinds of data sets. The structure of the data plays an important role in statistics. Among different structures of data, one interesting structure is the permutation, which involves in different kinds of problems, such as recommender system, online gaming, decision making and sports tournament. This thesis is motivated by my interest in understanding the permutation in statistics. Comparing to the wide applications of permutation related model, little is known to the property of permutation in statistics. There are a variety challenges that arise and lots of problems waiting for us to explore in the permutation based model. This thesis aims to solve several interesting problems of the permutation based model in statistics, which may help us to understand more about the property and characteristic of permutation.

As a result of the various topics explored, this thesis is split into three parts. In Chapter 2, we discuss the estimation problem of unimodal SST model in the pairwise comparison problem. We prove that the CLS estimator is rate optimal up to a $poly(\log\log n)$ factor and propose the computational efficient interval sorting estimator, as a computational efficient algorithm to the estimation problem. In Chapter 3, we shift our attention to the inference problem of the permutation based model. We study different kinds of inference problem, including the hypothesis testing problem in noisy sorting model and confidence set construction problems in generalized permutation based model. Network analysis is another important topic related to the permutation. In Chapter 4, we study the optimality of local belief propagation algorithm in the partial recovery problem of stochastic block model. We prove that local BP algorithm can reach the optimality in a certain regime. Moreover, in the regime where local BP algorithm may not achieve the optimal misclassified fraction, we will prove that local BP algorithm can be used in correcting other algorithms and get optimal algorithm to the partial recovery problem.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

1 Introduction

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. People spend lots of time dealing with different kinds of data sets. The structure of the data plays an important role in statistics. Among different structures of data, one interesting structure is the permutation, which involves in different kinds of problems, such as recommender system, online gaming, decision making and sports tournament. This thesis is motivated by my interest in understanding the permutation in statistics. Comparing to the wide applications of permutation related model, little is known to the property of permutation in statistics. There are a variety challenges that arise and lots of problems waiting for us to explore in the permutation based model. This thesis aims to solve several interesting problems of the permutation based model in statistics, which may help us to understand more about the property and characteristic of permutation.

As a result of the various topics explored, this thesis is split into three parts corresponding to the topics. In Chapter 2, we discuss the estimation problem in the pairwise comparison problem. Pairwise comparison is considered to be an important problem led by the advent of different new internet-scale applications in recent years. By reason of the wide application across different fields, the pairwise comparison problem gets more and more attention. Several parametric models have been studied in the pairwise comparison literature, including Thurstone model and the Bradley-Terry-Luce (BTL) model. Comparing to the traditional parametric models, nonparametric model shows more flexibility in the pairwise comparison problem. As a result, nonparametric models have been widely studied in the past five years. We study the unimodal SST model, one kind of the nonparametric model, in the thesis. For the estimation problem of the unimodal SST model, we establish the minimax optimal rate through the CLS estimator. We prove that the CLS estimator is rate optimal up to a $poly(\log\log n)$ factor. No other estimator can do much better than CLS estimator statistically. Though CLS estimator is good in estimation, it is not computationally feasible. It motivates us to further study the problem and find a computational feasible method to the estimation problem. We develop the interval sorting estimator, as a computational efficient algorithm to the estimation problem. Moreover, the interval sorting estimator is rate optimal up to a $poly(\log n)$ factor, which is the best estimator so far we know for the estimation problem in the nonparametric models of pairwise comparison problem. The above discussion is mainly about the estimation problem of the probability matrix. In fact, the permutation itself also plays an important role in the pairwise comparison problem. We also discuss the estimation to the permutation problem in this chapter. We construct the minimax rate for the estimation problem in the SST model.

Lots of efforts have been spent on the estimation of the permutation based models, while the inference problem has got much less attention in the literature. In Chapter 3, we shift our attention from the estimation problem to the inference problem of the permutation based model. We discuss the inference problems to generalized permutation based model in this chapter. We begin the chapter with the hypothesis testing problem of the probability matrix in the noisy sorting model. By studying the optimal testing procedure of the problem, boundary between the detectable regime and non-detectable regime for the testing problem is constructed. After that, we focus on the inference problem about the permutation in generalized permutation based model. We discuss the confidence set construction problem for the permutation in different settings. One of the important steps in the study of confidence set construction problem is to find suitable criterion to judge the confidence set construction procedure. We will show how to properly set up the criterion to judge confidence set construction procedure and introduce the optimal confidence set construction procedure to the problem in different parameter spaces. Finally, we end the chapter with the study of the hypothesis testing problem to the permutation.

Another important topic related to permutation is network analysis. Network analysis is one of the most popular topics in recent research. People from different areas do a lot of work to study network data analysis. In network literature, community detection problem in stochastic block model (SBM) is the most widely known and studied problem. In Chapter 4, we study the local belief propagation algorithm, which is used to solve the partial recovery problem in the stochastic block model. We prove that local BP algorithm can reach the optimality not just in the balanced case and construct an optimal regime where the local BP algorithm can reach the optimal expected misclassified fraction. Moreover, in the regime where local BP algorithm may not achieve the optimal misclassified fraction, we will prove that local BP algorithm can be used in correcting other algorithms. If we have a satisfactory initializer, the optimal algorithm can be reached by the initializer and the local

BP correction.

2 Unimodal SST Model Estimation for Pairwise Comparison

2.1 Introduction

2.1.1 Pairwise Comparison Problem

Pairwise comparison is considered to be an important problem led by the advent of different new internet-scale applications in recent years, including recommender system (Aggarwal (2016); Koren et al. (2009)), online gaming (Strittmatter et al. (2015)), decision making (Kou et al. (2016); Zhou et al. (2018)) and biomedical image assessment (Phelps et al. (2015)). Pairwise comparison problem is also studied in some traditional related fields, such as sports tournament (Csató (2013); Cattelan et al. (2013)) or teaching assessment ((Heldsinger and Humphry (2010)). Take sports tournament as an example, the result of each game can be understood as a result of the comparison between two teams in the tournament. The comparison results provide us information of the teams, showing that one of the teams can be better than another. By reason of the wide applications across different fields, the pairwise comparison problem gets more and more attention.

Several parametric models have been studied in the pairwise comparison literature. Two famous parametric models to the pairwise comparison problem are the Thurstone model (Thurstone (1927); Kornbrot (1978)) and the Bradley-Terry-Luce (BTL) model (Bradley and Terry (1952); Luce (1960)). In these parametric models, it is assumed that each item i is related to a score q_i and the probability that the item i wins a comparison against another item j can be written as a function of q_i and q_j. In BTL model, if we compare item i with item j, the probability that item i wins the comparison is M_{ij}, where M_{ij} is defined as

$$M_{ij} = \frac{1}{1 + \exp(-(q_i - q_j))} \tag{2.1}$$

In Thurstone model, if we compare item i with item j, the probability that item i wins the comparison is M_{ij}, where M_{ij} is defined as

$$M_{ij} = \Phi(q_i - q_j) \tag{2.2}$$

Here, Φ is c.d.f of the standard normal distribution

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-\frac{t^2}{2}) dt \tag{2.3}$$

The BTL model and Thurstone model can be generalized to be a class of parametric models to pairwise comparison problem, where M_{ij} is defined to be

$$M_{ij} = g(q_i - q_j) \tag{2.4}$$

and $g : \mathbb{R} \to [0, 1]$ is a strictly increasing cumulative distribution function. In the above models, we can find that the matrix M plays an important role in the model. We call the matrix M the probability matrix in pairwise comparison model. The discussion of the current chapter focuses on the estimation to the probability matrix M.

2.1.2 SST Model

Parametric models, including BTL model and Thurstone model, have been widely studied in the pairwise comparison literature. The estimation problem in parametric models is studied in Hajek et al. (2014), where the authors construct the minimax optimal rate for the estimation of the parametric models.

Though the parametric models are widely used and studied since it was proposed, the strong assumptions on the structure of parametric model limit its application. In parametric models, every item has its own score. The probability of item i wins in the comparison against item j is determined by the score of item i and item j only. In other words, to determine the parametric models, we need only n parameters. In real application, the strong parametric assumptions may not hold in many examples. In several cases, we need richer structure and less assumptions for the pairwise comparison problem.

Instead of using parametric models in pairwise comparison problem, the strong stochastic transitivity (SST) model is proposed in Shah et al. (2016a). Rather than assuming that each item has a unique score which characterizes the quality of the item, SST model makes less assumptions to the model. It assumes that for two different items i and j, if we expect that i has better quality than j, when we compare them with any other item k, we should expect that the probability of item i wins the comparison (against item k) should be larger than the probability of item j wins the comparison (against item k). This property is called strong stochastic transitivity. More precisely, SST matrix class can be stated in the following form.

Definition 1 (SST class)**.** *Let* $M \in [0,1]^{n\times n}$ *to be a matrix satisfied the following assumptions:*

(i)$\forall a, b \in [n]$*,* $M_{ab} + M_{ba} = 1$

(ii)$\exists$ *a permutation* π *on* $[n]$*, such that for any triple if* (a, b, c)*,* $\pi(a) < \pi(b)$*, we have* $M_{ac} \geq M_{bc}$*.*

We use $\mathbb{C}_{sst}$ *to denote the class of SST matrices satisfying the above conditions.*

We call the SST model for the pairwise comparison problem if we assume the true probability matrix is in the SST class $\mathbb{C}_{sst}$. For any matrix $M \in \mathbb{C}_{sst}$, let $\pi(M)$ to be the set of all permutation satisfied the condition in the definition of SST class. From the definition of SST class, we know that $\pi(M)$ is not empty. We use $\mathbb{C}_{sst}(\pi)$ to be the subclass in $\mathbb{C}_{sst}$, such that $\mathbb{C}_{sst}(\pi) = \{M \in \mathbb{C}_{SST} | \pi \in \pi(M)\}$.

In SST model, as we make less assumptions than the parametric models, the model is more flexible. People have studied the estimation problem of pairwise comparison problem in SST models in the literatue, see Shah et al. (2018, 2016b, 2019); Mao et al. (2018); Shah and Wainwright (2017); Shah et al. (2016c). SST model plays an important role in the pairwise comparison problem. It is one of the most widely used nonparametric models in the pairwise comparison problem and also the foundation of the unimodal SST model we discuss in the current chapter.

2.1.3 Unimodal SST Model

After introducing the SST model, we introduce the following unimodal SST model to the pairwise comparison problem.

Definition 2 (Unimodal SST class)**.** *Let* $M \in [0,1]^{n\times n}$ *to be a matrix satisfied the following assumptions:*

(i)$\forall a, b \in [n]$*,* $M_{ab} + M_{ba} = 1$

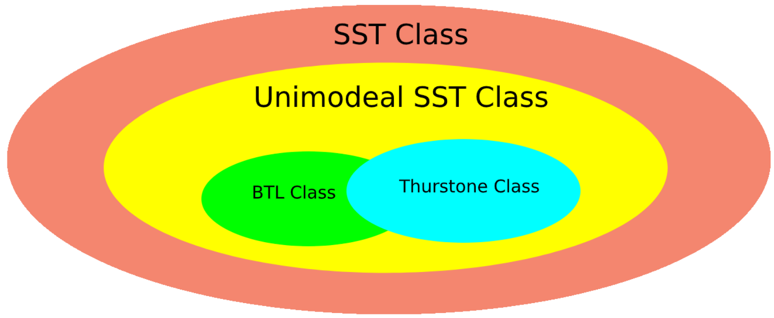

Figure 1: Relationship Between Different Models

(ii) $\exists$ *a permutation* π *on* $[n]$*, such that for any triple* (a, b, c)*,* $\pi(a) < \pi(b)$*, we have* $M_{ac} \geq M_{bc}$*.*

(iii)If $\pi^{-1}(a) < \pi^{-1}(b)$*,* $\{M_{a,\pi^{-1}(j)} - M_{b,\pi^{-1}(j)} | j = 1, 2, ..., n\}$ *is a unimodal sequence and the peak of the sequence is between* $\pi^{-1}(a)$ *and* $\pi^{-1}(b)$*.*

We use $\mathbb{C}_{usst}$ *to denote the class of unimodal SST matrices.*

One key observation is that both BTL model and Thurstone model are special cases of the unimodal SST model, which shows that to study the estimation problem within unimodal SST class framework can be a generalization of the study to the parametric models.

Let

$$\mathbb{C}_{BTL} = \{M | M(i,j) = \frac{1}{1 + \exp(-(q_i - q_j))}, q_i > 0\}$$

to be the BTL matrix class. We also assume that

$$\mathbb{C}_{Thurstone} = \{M | M(i,j) = \Phi(q_i - q_j), q_i > 0\}$$

to be the Thurstone matrix class. The following proposition tells us the relationship between parametric models and unimodal SST model.

Proposition 1.

$$\mathbb{C}_{Thurstone}, \mathbb{C}_{BTL} \subset \mathbb{C}_{usst}$$

The intuition of studying the unimodal assumption in pairwise comparison model comes from Proposition 1. On the one hand, the SST assumptions assures that if $\pi^{-1}(a) < \pi^{-1}(b)$, $\{M_{a,\pi^{-1}(j)} - M_{b,\pi^{-1}(j)} | j = 1, 2, ..., n\}$ is a positive sequence. This sequence can characterize the differences between item a and item b. On the other hand, unimodality is common in statistics. Lots of different probability distribution, including normal distribution, Cauchy distribution, Student's t-distribution, are unimodal distributions. These reasons motivates us to assume the sequence $\{M_{a,\pi^{-1}(j)} - M_{b,\pi^{-1}(j)} | j = 1, 2, ..., n\}$ is a unimodal sequence.

Comparing to the SST model, unimodal SST model is slightly restrictive, but it still contains all parametric models as special cases. For unimodal SST model, we will see in the following sections, though the statistical minimax error is similar to the SST model, there is computational efficient method for the unimodal SST model, which approximately reaches the statistical lower bound. For SST model, the best known computationally efficient method cannot match the statistical lower bound. Though we do not know now to find the best computational efficient algorithm for SST model, we can show that the interval sorting algorithm, the algorithm we propose for the esimtaion problem in the current chapter, is rate optimal in the unimodal SST class, which gives us partial answer to the estimation of pairwise comparison problem for SST class.

The contribution of the results in current chapter is threefold. First, we introduce the unimodal SST model to the pairwise comparison problem. Comparing to the traditional parametric models, less assumptions are made to the unimodal SST model.

Second, we establish the minimax optimal rate for estimation of unimodal SST model. We prove that the CLS estimator is rate optimal up to a $poly(\log \log n)$ factor. No other estimator can do much better than CLS estimator statistically.

Third, though CLS estimator is good in statistics, it is not computationally feasible. We propose the interval sorting estimator in the current chapter. The interval sorting estimator is computationally efficient and it is rate optimal up to a $poly(\log n)$ factor.

2.1.4 Organization

We organize the chapter as follows.

In Section 2.2, we consider the estimation to the unimodal SST model to the pairwise comparison problem. We construct the optimal rate for the estimation problem with CLS estimator in this section and then introduce the computational efficient interval sorting estimator. Section 2.3 introduces how we construct the statistical lower bound to the estimation problem in the current chapter. In Section 2.4, we consider the estimation problem under the independent design. In 2.5, we construct the minimax optimal rate for the estimation of permutation in the pariwise comparison problem. Numerical study is given in Section 2.6. Some discussion to related problems is in Section 2.7. The proof to the results in this chapter will be in Section 2.8.

2.2 Estimation to the Unimodal SST Model

In this section, we discuss the estimation problem to the unimodal SST model. We begin this section with the constrained least square (CLS) estimator to the unimodal SST model estimation problem. The CLS estimator shows good performance to the estimation problem. Unfortunately, CLS estimator is not a computational efficient estimator. To solve the computational issue, we propose the interval sorting (IS) estimator in latter part of the section. The interval sorting estimator is minimax optimal up to a logarithm factor. More importantly, the interval sorting estimator is computationally efficient.

2.2.1 Statistical Minimax Rate and CLS Estimator

Assume that we have complete observation to all possible pairs of comparison. Suppose that we observe independent Bernouli random variables $Y_{ij} \sim \text{Ber}(M_{ij}), i, j \in [n]$. We denote that $Y = (Y_{ij})_{1 \le i,j \le n}$ to be the observation. We try to solve the estimation problem based on our observation Y.

We define the CLS estimator as

$$\widehat{M}_{CLS} = \text{argmin}_{M \in \mathbb{C}_{sst}} \|Y - M\|_F^2 \tag{2.5}$$

If the probability matrix M is in the unimodal SST class, we can see the CLS estimator shows good performance for the estimation problem. The following theorem provides the theoretical guarantee for the CLS estimator.

Theorem 1. *If probability matrix $M \in \mathbb{C}_{usst}$, we have*

$$\mathbb{E}\|\widehat{M}_{CLS} - M\|_F^2 \lesssim n \log n (\log\log n)^5.$$

The idea of the CLS estimator is straightforward. The expectation of the observation Y is the probability matrix M in our model, to find an estimator of the matrix M, we should try to find the matrix which is closest to the observation Y in our parameter space $\mathbb{C}_{usst}$. Similar idea and result are also established in Shah et al. (2016b) for SST class. Our result is more precise than the result in Shah et al. (2016b). The result in Shah et al. (2016b) is optimal up to a log factor. Combining with the statistical lower bound in Theorem 3, our result is optimal up to a $poly(\log\log n)$ factor.

Though CLS estimator performs well in statistical estimation, it is not computationally efficient. The key factor which makes computation of the CLS estimator difficult is that the parameter space $\mathbb{C}_{usst}$ is not a convex set. From the definition of $\mathbb{C}_{usst}$, we can see that it is closely related to the permutation π. In fact, for two matrices $M_1, M_2 \in \mathbb{C}_{usst}$, if the corresponding permutation are different, there is no guarantee that $\frac{M_1+M_2}{2}$ is in the unimodal SST class. Non-convexity of the parameter space makes the computation of the CLS estimator difficult. This motivates us to find a computational efficient estimator to the problem.

2.2.2 Computational Efficient Method to Unimodal Model Estimation

To introduce the computational efficient method to the unimodal SST model estimation, we make slightly different assumptions to our observations. We assume that we observe independent observation $Y^{(l)}, 1 \leq l \leq 3$, such that $Y_{ij}^{(l)} \sim \text{Ber}(M_{ij}), \forall 1 \leq i, j \leq n$. The goal is to estimate the matrix M with the observation $Y = Y^{(l)}, 1 \leq l \leq 3$. We should point out the only reason for us to make the slight change of our observation is to make the illustration simpler and easier to understand.

The interval sorting algorithm, a computationally efficent algorithm to the estimation of unimodal SST model, is motivated by the two dimensional sorting (TDS) algorithm proposed in Mao et al. (2018). The algorithm proposed in Mao et al. (2018) consists of two parts: the ranking estimation and the probability matrix estimation. Simply speaking, the second step uses the similar idea as the idea for the CLS estimator. The reason that CLS estimator is not computational feasible is that we do not know the true permutation. If we have a good estimation to the ranking, we can estimate the probability matrix efficiently.

In this sense, the estimation to the permutation is crucial in the algorithm. In Mao et al. (2018), the idea of ranking estimation is to divide the items into blocks according to the number of times they win against all other items. Then to calculate the number of times they win against all items in each block. For two items i and j, if there is significant difference between the number of times they win against all items in any block, it should be easy for us to tell which item is better.

But the idea in Mao et al. (2018) is not enough. The statistical upper bound constructed in Mao et al. (2018) is of order $n^{5/4}$. It is difficult to improve the result and get the optimal bound of order n with their method. To get better result, we create even more characteristics to help us understand the relationship between all the items.

We call the number of an item wins against all other items to be the score of the item. Correspondingly, we call the number of an item wins against all other items in a block to be the score of the item in the block. We call several consecutive blocks to be an interval and the number of an item wins against all other items in an interval to be the score of the item in the interval. The score of the items, the score of the items in different blocks and the score of the items in different intervals combining together may help us to get a good estimation to the ranking, which allows us to get a better estimation in the probability matrix. Our algorithm sorts the items according to these features.

Informally, the algorithm to estimate the order can be stated with the following steps.

- Step 1: Use the samples in $Y^{(1)}$ to estimate the order $\hat{\pi}_1$.
- Step 2: Use $\hat{\pi}_1$ to divided all n items into K different blocks, where $[n] = \cup_{s=1}^{K} bl_s$.
- Step 3: Calculate the score, the score of all blocks, the score of all intervals for every item.
- Step 4: Create a directed graph G based on the score, the score of all blocks, the score of all intervals.

- Step 5: Create a topological sort $\hat{\pi}$ from G.

- Step 6: Output $\hat{\pi}$ as the estimation to the permutation π.

We omit several details in the above formulation of the algorithm in estimating the permutation. The complete algorithm is stated as follows.

Input : observation $Y^{(1)}, Y^{(2)}$

1. $\forall j \in [n]$, compute the partial column sum

$$S_j = \sum_{i=1}^{n} Y_{ij}^{(1)}, 1 \leq j \leq n$$

Let $S_j = \sum_{i=1}^{n} Y_{ij}$. Let $\hat{\pi}_1$ to be the order on $[n]$, such that $S_{\hat{\pi}_1(j)}$ is a non-decreasing sequence.

2. Let $\tau = 8\sqrt{n \log(n)}$ and $K = \lceil n/\tau \rceil$. Partition $[n]$ into K different blocks, such that

$$bl_1 = \{j \in [n] : S(j) \leq \tau\}$$

$$bl_k = \{j \in [n] : S(j) \in ((k-1)\tau, k\tau] \text{ for } 1 < k < K\}$$

$$bl_K = \{j \in [n] : S(j) > (K-1)\tau\}$$

3. Divide the K blocks into $\lceil \log_2 K \rceil$ groups, such that

$$g_t = \{s \in [K] : 2^{t-1} \leq |bl_s| < 2^t\}, 1 \leq t \leq \lceil \log_2 K \rceil$$

4. For each $t \in [\lceil \log_2 K \rceil]$, we divide $g_t = g_t^{(1)} \cup g_t^{(2)}$, such that there does not exist $s \in [K]$ and $u \in \{1, 2\}$, $\{s, s+1\} \subset g_t^{(u)}$.

5. Create a graph G on $[n]$, such that there exist an edge $u \to v$ if and only if there exists t and u, where $g_t^{(u)} = \{s_1, ..., s_p\}$, and $a < b \in [p]$, such that

$$\sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{uj}^{(2)} - \sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{vj}^{(2)} \geq 8\sqrt{(b-a+1)2^t \log n}$$

6. Compute a topological sort $\hat{\pi}$ of the graph G as our estimation to be the estimator of π. If there does not exist a topological sort $\hat{\pi}$ of the graph G, let $\hat{\pi} = id$.

Algorithm 1: unimodal SST Model Estimation

Then we construct the interval sorting estimator $\widehat{M}_{IS}$ through the following procedure

$$\widehat{M}_{IS} = \text{argmin}_{M \in \mathbb{C}_{sst}(\hat{\pi})} \|M - Y^{(3)}\|_F^2 \tag{2.6}$$

where the permutation $\hat{\pi}$ is computed through Algorithm 1.

Theorem 2. *Suppose that we calculate $\hat{\pi}$ through Algorithm 1. Then for any matrix $M \in C_{usst}$, we have*

$$\mathbb{E}_M \|\widehat{M}_{IS} - M\|_F^2 \lesssim n(\log n)^3 \tag{2.7}$$

Theorem 2 shows the main benefit we gain from studying unimodal SST model. The special structure of the matrices in the unimodal class make it possible to estimate them correctly and efficiently. Comparing to the estimation problem in SST model, we can see that we get a faster convergence rate in studying the estimation of the unimodal SST model for computational efficient method. In next section,we can also see that the interval sorting algorithm is rate optimal up to a log factor. It shows that IS algorithm is nearly optimal for the estimation of the unimodal SST model.

2.3 Statistical Lower Bound to the Estimation of Unimodal SST Model

In Section 2.2, we present the upper bound part to the estimation of unimodal SST model, including studying the performance of CLS estimator and interval sorting estimator. In this section, we state the corresponding statistical lower bound, which shows the optimality of the upper bound we constructed in Section 2.2. We begin with talking about the idea and tool we use in the construction of the lower bound.

Different methods are used in the construction of the statistical lower bound to the estimation of pairwise comparison problem, see for example Shah et al. (2016b, 2019). Most of the existing ideas can be described as below.

Without loss of generality, we assume that $n = 2m$. The idea for the case n is odd is similar. Let all n items be divided into m different groups $[n] = \cup_{i=1}^m B_i$, where $B_i = \{2i-1, 2i\}, \forall i \in [m]$. We assume that in a special case, such that the permutation π satisfies that $\pi(B_i) = B_i, \forall i$ and for $u \in B_i, v \in B_j, i \leq j$, we always have $M_{ij} = 1$. In this special case, to estimate the probability matrix M, all we need to do is to estimate $M_{2i-1,2i}, \forall i \in [m]$. It can be proved that for $\forall i$, we would make a constant error when we estimate $M_{2i-1,2i}$, which leads to an error of order n in terms of the probability matrix M.

To get a better statistical lower bound, we need a better idea than the above. One useful tool is the Assouad-Le cam's lemma, which is introduced by Cai and Zhou (2012). The Assouad-Le Cam's lemma can help us construct a better statistical lower bound in this problem. The main idea of Assouad-Le Cam's lemma is that we can construct a list of hypothesis testing problems, which can be prove to be difficult to solve. These hypothesis testing problems show the intrinsic difficulty to the estimation problem, from which Assouad-Le Cam's lemma helps us to construct the statistical lower bound.

We consider the following hypothesis testing problem. Suppose that $m = \lfloor \log n \rfloor$, $l = \lfloor \frac{n}{m} \rfloor$. We divide $[n]$ into $l+1$ different sets $D_i, 1 \leq i \leq l+1$, where $D_i = \{(i-1)m+1, \cdots, im\}$,$1 \leq i \leq l$ and $D_{l+1} = \{lm+1, \cdots, n\}$. Assume that we know for $t \in D_i, s \in D_j, i \leq j$, we always have $M_{ij} = \frac{1}{2} + \delta$ and $M_{ji} = \frac{1}{2} - \delta$, where δ is a small constant. We try to solve the following hypothesis testing problems.

$$H_{0,i} : M_{st} = \frac{1}{2}, \forall s, t \in D_i \text{ vs. } H_{1,i} : M_{st} = \frac{1}{2} + \delta, \forall s, t \in D_i, s < t \tag{2.8}$$

We can prove that to solve the above hypothesis testing problem is difficult. With the above ideas and Le cam-Assouad's lemma, we have the following result.

Theorem 3. *Suppose that we observe independent Bernouli random variables $Y_{ij} \sim Ber(M_{ij}), i, j \in [n]$. There exists a constant c, such that for any estimator $\widehat{M} = \widehat{M}(Y)$, we have the following statistical lower bound*

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{C}_{usst}} \mathbb{E}\|\widehat{M} - M\|_F^2 \geq cn \log n$$

where c is a given constant.

The statistical lower bound constructed in Theorem 3 shows that the rate of convergence in Theorem 2 is optimal up to a *poly*$(\log \log n)$ factor. Comparing to the previous results in pairwise comparisons literature, the $poly(\log\log n)$ gap is the best known result of the statistical method, where previously the result is of $poly(\log n)$ gap, see for example Shah et al. (2016c). The lower bound in Theorem 3 also shows that the IS estimator is nearly optimal computational efficient estimator to the pairwise comparison problem. It is rate optimal up to a $poly(\log n)$ factor, where the known gap of the best computational before is approximately of order $n^{1/4}$.

2.4 Independent Design

We have discussed how to estimate the pairwise comparison model when we have complete observation to all possible pairs. We now extend of our results to a different setting.

In the previous sections, we assume that we have complete observation to all pairs of comparison. Instead of assuming that we have complete observation to all possible pairs, in the current section, we assume that we have N observations, which is a more practical setting in real life example. For each observation, we observe each possible pair with probability $\frac{1}{\binom{n}{2}}$ independently. More precisely, let $i_k \sim \text{Unif}[n]$,$j_k \sim \text{Unif}[n]$,$k \in [N]$. We observe independent Bernouli random variables $Y_k \sim \text{Ber}(M_{i_k,j_k})$. The independent case is also studied in pairwise comparison literature, for example Shah et al. (2016b); Mao et al. (2017).

2.4.1 Statistical Minimax Rate for Independent Design

We use the similar idea to construct the statistical minimax rate for the independent design. We construct the CLS estimator for independent design as follows.

We define that $T = \{(i,j)|\exists k \in N, i_k = i, i_k = j\}$. We define the distance between our observation Y to the matrix M to be

$$d(Y, M) = \sqrt{\sum_{(i,j)\in T} (\bar{Y}_{ij} - M_{ij})^2} \tag{2.9}$$

where $\bar{Y}_{ij} = \frac{1}{|\{k:i_k=i,j_k=j\}|}\sum_{\{k:i_k=i,j_k=j\}} Y_{i_k,j_k}$. If $|\{k : i_k = i, j_k = j\}| = 0$, we define that $\bar{Y}_{ij} = \frac{1}{2}$.

We introduce the CLS estimator for the independent design as

$$\widehat{M}_{CLS}^{ind} = \text{argmin}_{M \in \mathbb{C}_{sst}} d(Y, M) \tag{2.10}$$

We can prove the similar result as in the previous section to construct the near optimal rate for the estimation of probability matrix in the independent case.

Theorem 4. *(i)If $N \geq 2n^2 \log n$,the CLS estimator satisfies the following statistical upper bound for the estimation problem*

$$\sup_{M \in \mathbb{C}_{usst}} \mathbb{E}\|\widehat{M}_{CLS}^{ind} - M\|_F^2 \lesssim \frac{n^3 \log n (\log \log n)^5}{N}.$$

(ii)For any estimator $\widehat{M} = \widehat{M}(Y)$, we have the following statistical lower bound

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{C}_{usst}} \mathbb{E}\|\widehat{M} - M\|_F^2 \gtrsim \frac{n^3 \log n}{N}.$$

2.4.2 Computational Efficient Method for Independent Design

For the computationally efficient algorithm for independent design case, instead of using Algorithm 1 directly, we should do several minor revision to the original algorithm so that it can fit the independent design setting.

We replace Step 2 with the following Step 2' in the algorithm

Step 2'. Let $\tau = 8\sqrt{\frac{n^3 \log(n)}{N}}$ and $K = \lceil n/\tau \rceil$. Partition $[n]$ into K different blocks, such that

$$bl_1 = \{j \in [n] : S(j) \leq \tau\}$$
$$bl_k = \{j \in [n] : S(j) \in ((k-1)\tau, k\tau] \text{ for } 1 < k < K\}$$
$$bl_K = \{j \in [n] : S(j) > (K-1)\tau\}$$

and replace Step 5 with the following Step 5'

Step 5'. 5. Create a graph G on $[n]$, such that there exist an edge $u \to v$ if and only if there exists t and u, where $g_t^{(u)} = \{s_1, ..., s_p\}$, and $a < b \in [p]$, such that

$$\sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{uj}^{(2)} - \sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{vj}^{(2)} \geq 8\sqrt{\frac{n^2(b-a+1)2^t \log n}{N}}$$

Then we construct the estimator $\widehat{M}_{sort}^{ind}$ through the following procedure

$$\widehat{M}_{sort}^{ind} = \text{argmin}_{M \in \mathbb{C}_{sst}(\hat{\pi})} d(Y^{(3)}, M) \tag{2.11}$$

where the permutation $\hat{\pi}$ is calculated through the revised version of Algorithm 1.

Theorem 5. *If $N \geq 2n^2 \log n$,for any matrix $M \in C_{usst}$, we have*

$$\mathbb{E}_M \|\widehat{M}_{sort}^{ind} - M\|_F^2 \lesssim \frac{n^3(\log n)^3}{N}. \tag{2.12}$$

The algorithm and theoretical results for independent observations are similar to the corresponding results for complete observation. The major difference for independent observation cases is that since we do not have the complete observation for all possible pairs, the criterion we use in Step 5 of the algorithm is different. Except for the difference in Step 5, the interval sorting estimator and the analysis are similar to the case when we have complete observation.

2.5 Minimax Rate for Estimation of the Permutation

Other than the estimation to the probability matrix M, it is also an interesting problem to study the estimation of the permutation $\pi(M)$. In the current section, we construct the minimax rate for the estimation of permutation. We consider the case when for the parameter space $\mathbb{C}_{SST}$ in this section.

However, to estimate the permutation, the SST assumption is not enough. The reason is that it is possible that two items are identical even if the SST assumption is satisfied. In that case, it is impossible for us to tell the difference between these items. So we have the following assumption on the difference between items, which makes the estimation of the permutation possible.

For a fixed λ, we define a class of matrices

$$D_n(\lambda) = \{M \in [0,1]^{n\times n} : \sum_{j=1}^{n} |M_{ij} - M_{kj}| \geq \lambda, \forall i,k \in [n]\} \tag{2.13}$$

In the matrix class $D_n(\lambda)$, the difference between different items is at least λ in terms of l_1 distance for the corresponding columns. We assume that $M \in D_n(\lambda) \cap \mathbb{C}_{SST}$ and consider the pairwise comparison problem in this parameter space.

Suppose that we observe independent Bernouli random variables $Y_{ij} \sim \text{Ber}(M_{ij}), i, j \in [n]$. The goal is to estimate the permutation $\pi(M)$ with the observation $Y = (Y_{ij})_{1\leq i,j\leq n}$, where $\pi(M)$ is the permutation corresponds with the probability matrix in the SST class $M \in D_n(\lambda) \cap \mathbb{C}_{SST}$.

The minimax rate for the estimation of permutation is constructed in the following theorem.

Theorem 6. *The minimax rate for estimation to the permutation with observation Y can be constructed as*

$$\inf_{\hat{\pi}(Y)} \sup_{M\in D_n(\lambda)\cap\mathbb{C}_{SST}} \mathbb{E} d_{KT}(\hat{\pi}, \pi(M)) \asymp \min\{\frac{n^{3/2}}{\lambda}, n^2\} \tag{2.14}$$

Here, the measure we use for the difference between the estimator to the true permutation is the Kendall tau distance, which is defined as

$$d_{KT}(\pi, \sigma) = \sum_{(i,j):\sigma(i)<\sigma(j)} I(\pi(i) > \pi(j)) \tag{2.15}$$

for any permutation $\pi, \sigma \in \mathbb{S}_n$, where $\mathbb{S}_n$ is the set of all permutations on $[n]$.

Theorem 6 shows the minimax rate for the estimation of permutation in the parameter space $D_n(\lambda) \cap \mathbb{C}_{SST}$ is of order $\min\{\frac{n^{3/2}}{\lambda}, n^2\}$. In fact, the upper bound can be simply achieved by the naive estimator based on the wins of the items in all comparisons. The estimator $\hat{\pi}_{naive}$ is defined as follows.

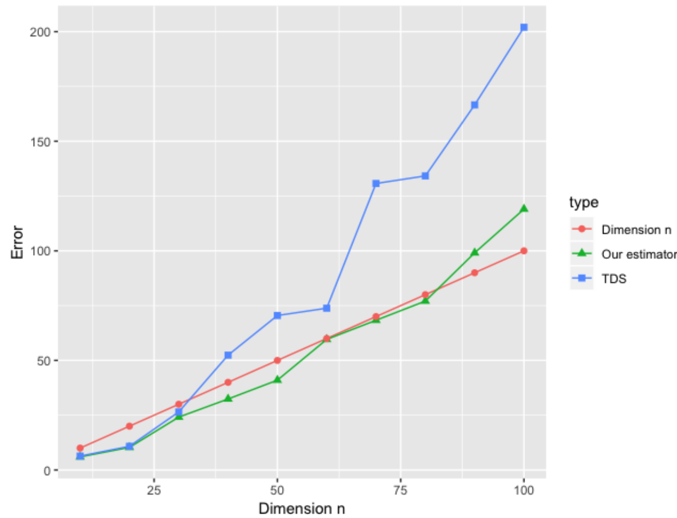

Figure 2: Numerical Performance for Interval Sorting Estimator

Let $T_i = \sum_{j=1}^{n} Y_{ij}$. The estimator $\hat{\pi}_{naive} \in \mathbb{S}_n$ is a permutation on $[n]$, such that $T_{\hat{\pi}_{naive}^{-1}(i)}, 1 \leq i \leq n$ is a non-decreasing sequence. If there exists different permutations satisfy the above property, we can randomly choose one of them to be the estimator $\hat{\pi}_{naive}$.

2.6 Numerical Result

We corroborate our theoretical results with numerical experiments. In this section,we compare the performance of our approach with the previous proposals. We compare the numerical performance of IS estimator with the TDS estimator in Mao et al. (2017), which is the best known computationally efficient method to the estimation of parametric models in pairwise comparison problem. First we generate a class of random unimodal SST matrix as the underlying matrix M through the following procedure.

1 Let $b_i, 1 \leq i \leq n$ to be independent uniform distributed random variable $\mathcal{U}[0,1]$.

2 Define $c_i = \sum_{j=1}^{i} b_{[i]}$.

3 For $1 \leq i < j \leq n$, let $M_{ij} = \frac{1}{2} + \frac{c_{n-j+i}}{2c_n}$.

4 For $1 \leq j < i \leq n$, let $M_{ij} = 1 - M_{ji}$.

5 Let $M_{ii} = \frac{1}{2}, \forall i \in [n]$.

6 Return $M = (M_{ij})_{1 \leq i,j \leq n}$ as the random concave SST matrix.

Algorithm 2: Generate Random Concave SST Matrix

We can check that the matrix we generate through Algorithm 2 is in the unimodal SST class. We then generate the observation based on the unimodal SST matrix we generate through Algorithm 2. The numerical performance is shown in Figure 2.

We can see from Figure 2 the error of our estimator is of approximately $O(n)$ rate. Moreover, comparing our estimator to the TDS estimator, we can see that our estimator performs better in the simulation, which corresponds with the theoretical result: the statistical error for the TDS estimator is of order $n^{5/4}$, while the statistical error for the IS estimator has a error rate of order n.

2.7 Discussion

In this chapter, we present the estimation result for the unimodal SST model in the pairwise comparison problem. We provide more detailed comparison of our results in the current chapter to some known results in the literature in this section.

Hajek et al. (2014) discusses the estimation problem in both BTL model and Thurstone model. The authors provided the minimax rate optimal estimator to the parametric models. Before comparing it with the result we establish in the current chapter, we should point out that since we are considering different kinds of models, the estimation targets in the current chapter and Hajek et al. (2014) are different. In Hajek et al. (2014), we try to estimate the score for each item, and in the current chapter, we try to estimate the probability matrix which decides the comparison for all possible pairs. Since in Hajek et al. (2014) the estimation target is only with n different parameters, while in the current chapter, the probability matrix have n^2 different parameters. If we also consider the estimation of the probability matrix in the parametric models, with suitable assumptions, we can see that the estimation in parametric models is better than the estimation to the unimodal SST model. This is because in the estimation of unimodal SST model, as we give the model more freedom, should be more difficult.

It is also interesting to compare our results with other results of the nonparametric models in the pairwise comparison problem. Shah et al. (2016a) established the minimax rate of convergence of the pairwise comparison problem. The rate is minimax up to a $poly(\log n)$ factor. In the current chapter, we provide a more subtle analysis to the CLS estimator which proves a better rate of convergence than the result in Shah et al. (2016a). Though the CLS estimator is rate optimal up to a $poly(\log n)$ factor, the computation of CLS is impossible. The best known computational efficient method to the pairwise comparison problem is provided in Mao et al. (2018). Mao et al. proves that their method can reach the error rate of $O(n^{5/4})$, which shows a clear gap to the minimax rate in SST class. In the current chapter, we provide a minimax rate optimal estimator, the interval sorting estimator, to the estimation of the probability matrix of the unimodal SST model. The interval sorting estimator has a faster convergence rate than the previously known computational efficient estimator. Though the parameter space we consider for the IS estimator is unimodal SST class, we should point out that the difference between unimodal SST class and SST class is small. It remains interesting to know how to construct rate optimal computational efficient algorithm for the estimation problem in SST model.

Another related topic we discuss in the current chapter is the estimation of ranking in pairwise comparison problem. Mao et al. (2017) establish the minimax optimal method in ranking estimation of the pairwise comparison problem in noisy sorting model. We establish the similar result for SST model in the current section. We should point out that the method in Mao et al. (2017) may not help us with the estimation of the matrix in SST model, as the setting is different in noisy sorting model from the SST model.

2.8 Proof

2.8.1 Proof to Theorem 1 and Theorem 4(i)

Proof. Before proving Theorem 1 and Theorem 4(i), we need the following technical lemma.

Lemma 1.

$$\log N(\epsilon, \mathcal{F}_2, \|\cdot\|_2) \leq C\epsilon^{-2}(\log\log(1/\epsilon))^5 \tag{2.16}$$

for an absolute constant C, where $\mathcal{F}_2$ is the class of bivariate monotonic function on $[0,1] \times [0,1]$.

Proof to Lemma 1. First we describe the method used in proving the statement in Gao and Wellner (2007). We will prove the result for $\epsilon = 2^{-n}$ for some positive integer n. For general case, we can bound it by using the monoticity of the entropy number. For each $f \in \mathcal{F}_2$, we construct $\bar{f}$ and $\underline{f}$ as follows. First, we partition $[0,1)^2$ into ϵ^{-2} cubes of side-length ϵ. A cube I_0 of side-length ϵ is selected if $\omega(f, I_0) \leq 2\epsilon$. For each cube that is not selected, we partition it into 4 cubes of equal size. In general, suppose we have a cube I_i of side length

2^{-i-n}. If $\omega(f, I_0) \leq 2^{i-n}$, we select the cube; otherwise, we partition the cube into 4 smaller cubes. This process continues until $i = n$. In this case, we always select the cube. Clearly, each point in $[0,1)^2$ uniquely belongs to one of the selected cubes. Then we define

$$\underline{f} = 2^{i+1-n} \left\lfloor \frac{\inf_{x \in I} f(x)}{2^{i+1-n}} \right\rfloor, \bar{f} = 2^{i+1-n} \left\lceil \frac{\sup_{x \in I} f(x)}{2^{i+1-n}} \right\rceil$$

Let $\bar{\mathcal{S}} = \{f : f \in \mathcal{F}_2\}$ and $\bar{\mathcal{S}} = \{f : f \in \mathcal{F}_2\}$. It is clear that we have $\underline{f} \leq f \leq \bar{f}$.

Now we are going to estimate $\|\bar{f} - \underline{f}\|_2$. Let n_i be the number of not selected cubes in the i-th step and s_i be the number of selected cubes in the i-th step, $1 \leq i \leq n$. Then we are going to introduce the cut of a function. Let f be any function in $\mathcal{F}_2$. We say f defines a cut in $[0,1)^2$ based on the procedure we discussed before. Here, the word 'cut' has two different meaning:

1 the procedure to cut the large cube into small cubes

2 the set of small cubes generated by this procedure

We use $\mathcal{C}_f$ to denote the cut defined by the function f. We further define $\mathcal{C}_{f,i}$ to be the cut of function f in the i-th step, $1 \leq i \leq n$. The edges of the cut is defined to be the set of the edges of selected cubes in the cut. If an edge is further divided into two parts, it is not included in this set. For example, Figure 1 is a cut defined by a function f. This cut has 20 edges.

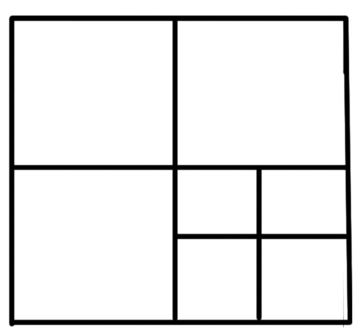

Figure 3: Cut by a Bivariate Monotonic Function

A function g is defined to be a function $g : \mathcal{C} \to \mathbb{R}$, where $\mathcal{C}$ is the set of cut of $[0,1)^2$. Let $\mathcal{E}_\mathcal{C}$ to be the set of edges of the cut $\mathcal{C}$. Then $g(\mathcal{C}) = \sum_{e \in \mathcal{E}_\mathcal{C}} l(e) w_f(e)$, where $l(e)$ is the length of the edge and $w_f(e) = |f(a) - f(b)|$, a and b are two vertices of the edge e. We write $\omega(e) = \omega_f(e)$ for short, as the function f is fixed throughout the proof.

Lemma 2.

$$g(\mathcal{C}_f) \leq 2$$

Proof. We are going to prove that the function g is not increasing in each step cut.

Suppose we are going to do the i-th step cut. Assume $Cube_i$ is the set of cubes not selected in the i-th step, i.e. all cubes $I \in C_{f,i-1}$ such that

$$l(I) = 2^{-n-i}, \omega(I) \geq 2^{i-n}$$

We say two cubes are adjacent if they have the same size and they share one edge. Assume $Cube_i$ can be written as union of several sets, $Cube_i = \cup_j Cube_{i,j}$, such that for any $I_1 \in Cube_{i,j_1}$ and $I_2 \in Cube_{i,j_2}$, I_1 and I_2 are not adjacent; $\forall I, I' \in Cube_{i,j}$, we can find $I_1, ..., I_t$, such that $I = I_1, I' = I_t$, I_s and I_{s+1},$1 \leq s \leq t-1$ are adjacent. We are going to prove in the process of cutting all cubes in each $Cube_{i,j}$, the function g is not increasing.

In the process of cutting all cubes within $Cube_{i,j}$, the difference of function g before and after the cut can be written as

$$g_{after} - g_{before} = 2^{-i-n-1}(\sum_{e \in \mathcal{E}_2} \omega(e) - \sum_{e \in \mathcal{E}_1} \omega(e))$$

Here g_{after} is the function value of g after the cutting process in the i-th step, $g_{before} = g(C_{f,i-1})$, $\mathcal{E}_1 = \{e \text{ is an edge of } I, I \in Cube_{i,j}\}$, $\mathcal{E}_2 = \{e \text{ is the middle line of } I, I \in Cube_{i,j}\}$. In Figure 2.8.1, we can see that the set $\mathcal{E}_1$ is the set of all black edges and $\mathcal{E}_2$ is the set of all red edges.

To prove that the function g is not increasing in the process of cutting within $Cube_{i,j}$, it is sufficient to prove that

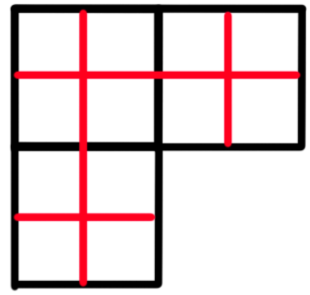

Figure 4: Example of the Set $\mathcal{E}_1$ and $\mathcal{E}_2$

$$\sum_{e\in\mathcal{E}_1}\omega(e)-\sum_{e\in\mathcal{E}_2}\omega(e)\geq 0 \tag{2.17}$$

In order to prove (2.17), we are going to consider the difference of g within each single cube $I \in Cube_{i,j}$. We define the diagonal edge of a cube I is the blue edge in Figure 2.8.1 and the middle line of the cube is the red line in Figure 3. The diagonal edges of I are four blue edges in this figure. The middle lines of I are two red lines in Figure 2.8.1.

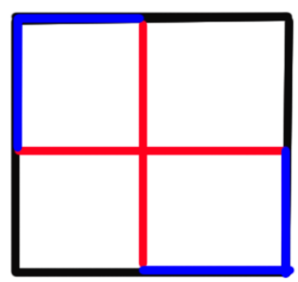

Figure 5: Diagonal Edges and Middle Lines in a Cube

As we know that $f \in \mathcal{F}_2$, we have

$$\sum_{e \text{ is a middle line of cube I}}\omega(e)=\sum_{e \text{ is a diagoanl edge of cube I}}\omega(e) \tag{2.18}$$

We should also notice that, for two adjacent cubes I_1 and I_2, {diagonal edges of I_1} $\cap$

{diagonal edges of I_2} $= \varnothing$. In this sense, we will have that

$$\sum_{e\in\mathcal{E}_2}\omega(e) = \sum_{e\text{is a diagonal edge of cube } I, I\in Cube_{i,j}}\omega(e) \leq \sum_{e\in\mathcal{E}_1}\omega(e) \tag{2.19}$$

This inequality gives us the proof to (2.17), which tells us that the function g is not increasing in the process of cutting. Lemma 2 comes from this statement and the fact $g \leq 2$ before we cut the cube $[0,1)^2$. □

Now consider all not selected cubes in the process of cutting. Let $\mathcal{N}$ to be the set of all not selected cubes. Suppose $k = |\mathcal{N}|$ and $\mathcal{N} = \{I_1, ..., I_k\}$. We define a cube to be an intrinsic not selected cubes if $I \in \mathcal{N}$ and there does not exist $I' \in \mathcal{N}$, such that $I \subset I', I \neq I'$. Let $\mathcal{N}_{intrinsic}$ to be the set of intrinsic not selected cubes.

We further defined another set $\mathcal{N}_{special} \subset \mathcal{N}_{intrinsic}$. We choose the element in $\mathcal{N}_{special}$ with the following procedure.

(i) Choose all intrinsic not selected cubes with length 2^{-2n} and add them into $\mathcal{N}_{special}$. Set $t = n$.

(ii) Choose all intrinsic not selected cubes I whose length is 2^{-n-t+1}, such that there does not exist $I' \in \mathcal{N}_{special}$,$I_0 \in \mathcal{N}$,$I' \subset I_0$,$I_0$ and I are adjacent. Add all these cubes into $\mathcal{N}_{special}$.

(iii)If t=0, end the procedure. Otherwise set $t = t - 1$ and go back to step (ii).

The reason why we consider the set $\mathcal{N}_{special}$ is that the following lemma can control the number of elements in $\mathcal{N}_{special}$. We can use the number of elements in $\mathcal{N}_{special}$ to give an upper bound for $\mathcal{N}_{intrinsic}$ and $\mathcal{N}$.

Lemma 3.

$$|\mathcal{N}_{special}| \leq 2^{2n}$$

Proof. We use the procedure similar to the one at the first of this note to cut the cube $[0,1)^2$. If a cube $I \in \mathcal{N}$, $I \notin \mathcal{N}_{special}$ and there does not exist I_0,I', such that I_0 adjacent to I, $I' \subset I_0$ and $I' \in \mathcal{N}_{special}$, we don't cut I in the i-th step. With the same argument in Lemma 2, we have that $g(\mathcal{C}) \leq 2$. Here $\mathcal{C}$ is the cut we get from this cut procedure. We should notice that for every $I \in \mathcal{N}_{special}$, it appears in this cut.

For each $I \in \mathcal{N}_{special}$ whose length is 2^{-n-i}, we have $\omega(I) \geq 2^{i-n}$. So we have

$$\sum_{e \text{ is an edge of I}} \omega(e)l(e) \geq 2^{1-2n} \tag{2.20}$$

Combining this equation and $g \geq 2$, we have

$$2 \geq 2^{1-2n}|\mathcal{N}_{special}| \tag{2.21}$$

which completes the proof to Lemma 3. □

Then we are going to bound the number of elements in $\mathcal{N}_{intrinsic}$.

Lemma 4.

$$|\mathcal{N}_{intrinsic}| \leq c2^{2n}\log n$$

where c is a universal constant.

Proof. For each $I \in \mathcal{N}_{intrinsic}$, we have that either $I \in \mathcal{N}_{intrinsic}$, or there exist $I_0 \in \mathcal{N}$,$I' \in \mathcal{N}_{special}$, such that $I' \subset I_0$,I_0 and I are adjacent. Let $k_{spec} = |\mathcal{N}_{special}|$. Without loss of generality, assume that $\mathcal{N}_{special} = \{I_1, ..., I_{k_{spec}}\}$. Let $T_i = \{I \in \mathcal{N} | I_i \subset I, I \notin T_j, \forall j < i\}$,$1 \leq i \leq k_{spec}$. Assume that $a_i = |T_i|$. Then we can know that

$$|\mathcal{N}_{intrinsic}| \leq 4\sum_{i=1}^{k_{spec}} a_i \tag{2.22}$$

For $1 \leq i \leq n$, all the edges of the not selected cubes with length 2^{-n-i} falls in 2^{n+i+1} parallel lines, among which 2^{n+i} are vertical and 2^{n+i} are horizontal. For each line, the difference of the function value is at most 1. So we have

$$\sum_{I \in \mathcal{N}_i} \sum_{e \text{ is an edge of I}} \omega(e)l(e) \leq 2 * 2^{n+i} * 2^{-n-i} < 4 \tag{2.23}$$

On the other hand, we also have that

$$\sum_j \sum_{I\in\mathcal{N}_j} \sum_{e \text{ is an edge of I}} \omega(e)l(e) \geq \sum_{i=1}^{k_{spec}} 2\omega(I_i)2^{-n-l_i+a_i-1} \geq \sum_{i=1}^{k_{spec}} 2^{-2n}2^{a_i-1} \tag{2.24}$$

Here l_i is the length of the cube I_i and we use the fact that $\omega(I_i) \geq 2^{l_i-n}$. Combining these two inequalities we have that

$$\sum_{i=1}^{k_{spec}} (2^{a_i} - 1) \leq 8n2^{2n} \tag{2.25}$$

We define $a_i = 0$ for $k_{spec} < i \leq 2^{2n}$. So

$$2^{2n}(2^r - 1) \leq \sum_{i=1}^{2^{2n}} (2^{a_i} - 1) \leq 8n2^{2n} \tag{2.26}$$

where $r = \frac{1}{2^{2n}} \sum_{i=1}^{2^{2n}} a_i$.

Use (2.22), we will get that

$$|\mathcal{N}_{intrinsic}| \leq 4\sum_{i=1}^{k_{spec}} a_i = 2^{2n+2}r \leq c2^{2n}\log n \tag{2.27}$$

for some constant c. □

Now we can bound the number of not selected cubes $|\mathcal{N}|$. The proof in Lemma 5 is similar to the proof in Lemma 4.

Lemma 5.

$$|\mathcal{N}| \leq c'2^{2n}(\log n)^3$$

Proof. Let $k_{intri} = |\mathcal{N}_{intrinsic}|$. Assume that $\mathcal{N}_{intrinsic} = \{I_1, ..., I_{k_{intri}}\}$. The previous lemma tells us that $k_{intri} \leq c2^{2n}\log n$. Let $U_i = \{I \in \mathcal{N} | I_i \subset I, I \notin U_j, \forall j < i\}$,$1 \leq i \leq k_{intri}$. Assume that $b_i = |U_i|$. Then we can know that

$$|\mathcal{N}| \leq \sum_{i=1}^{k_{intri}} b_i \tag{2.28}$$

Similar to the proof in the previous lemma, we will get

$$\sum_{i=1}^{k_{intri}} 2^{b_i} - 1 \leq 8n2^{2n} \tag{2.29}$$

We further define $b_i = 0$ for $k_{intri} < i \leq c2^{2n} \log n$. By extending the definition of b,

$$c2^{2n} \log n(2^b - 1) \leq \sum_{i=1}^{c2^{2n} \log n} 2^{b_i} - 1 \leq 8n2^{2n}$$

where $b = \frac{1}{c2^{2n} \log n} \sum b_i$. So $b < c'' \log n$ for some constant c'' and large enough $n > N$. Then we have

$$|\mathcal{N}| \leq \sum_{i=1}^{c2^{2n} \log n} b_i \leq c2^{2n} \log nb \leq c'''2^{2n}(\log n)^2 \tag{2.30}$$

for some constant c'''. □

Use Lemma 5, we can give the bound on $\|\bar{f} - \underline{f}\|_2$. Let s to be the number of selected cubes. Then we have that $s \leq 4|\mathcal{N}| \leq 4c'''2^{2n}(\log(n))^2$. Within a selected cube I whose length is 2^{-n-i}, we have $|\bar{f} - \underline{f}| \leq 2^{i-n}$. So

$$\int_I |\bar{f} - \underline{f}|_2^2 \leq 2^{-4n} \tag{2.31}$$

So

$$\int_{[0,1]^2} |\bar{f} - \underline{f}|_2^2 \leq \sum_{I \in \mathcal{S}} \int_I |\bar{f} - \underline{f}|_2^2 \leq 4c'''(\log n)^2 2^{-2n} \tag{2.32}$$

Then we will provide the bound of number of elements in $\bar{\mathcal{S}}$ and $\underline{\mathcal{S}}$. We are going to prove the following lemma.

Lemma 6.

$$\log |\bar{S}| \leq \bar{c}2^{2n}(\log n)^2$$

where $\bar{c}$ is a universal constant.

Proof. The number of elements in $\bar{\mathcal{S}}$ is decided by two things: one is the number of possible ways to cut $[0,1)^2$ into small selected cubes, another is the number of ways to put in possible values in each selected cube.

From what we prove before, we can see that the number of selected cubes for any f is bounded by $4c'''2^{2n}(\log n)^2$. So the number of possible cut defined by some $f \in \mathcal{F}_2$ is bounded by $2^{4c'2^{2n}(\log n)^2}$. This is because every cut is uniquely determined by a $\{0,1\}$ sequence with length $4c'2^{2n}(\log n)^2$.

Now we consider the number of ways to put in possible values in each selected cubes. Suppose for $0 \leq i \leq n$, we divide the cube $[0,1]^2$ equally into 2^{2i+2n} small cubes. Let $r_{i,j}$ to be the number of selected cubes in the j-th row, $1 \leq j \leq 2^{i+n}$. Then we have that the number of ways to assign values of $\bar{f}$ on these r_j cubes is bounded by $\binom{r_{i,j}+2^{n-i}}{2^{n-i}+1}$ if $r_{i,j} \geq 2$. Notice that we have the following lemma.

Lemma 7. *Let a to be a fix number, then*

$$h_a(r) = \log \binom{r+a}{a+1}$$

is a concave function for $x > 0$.

Proof. To prove it is a concave function, we are going to prove that

Use Lemma 7, we have that $h_a(r_1) + h_a(r_2) \leq 2h_a(\frac{r_1+r_2}{2})$

This is because

$$\begin{aligned}
h_a(r_1) + h_a(r_2) =& \log \frac{\Gamma(r_1+a-1)}{\Gamma(a)\Gamma(r_1-2)} + \log \frac{\Gamma(r_2+a-1)}{\Gamma(a)\Gamma(r_2-2)} \\
&= 2\log a + \log \frac{1}{B(a+1,r_1-2)} + \log \frac{1}{B(a+1,r_2-2)} \\
&= 2\log a + \log \frac{1}{(\int_0^1 x^{a+1}(1-x)^{r_1-2}dx)(\int_0^1 x^{a+1}(1-x)^{r_2-2}dx)} \\
&\leq 2\log a + \log \frac{1}{(\int_0^1 x^{a+1}(1-x)^{(r_1+r_2)/2-2}dx)^2} \\
&= 2h_a(\frac{r_1+r_2}{2})
\end{aligned} \tag{2.33}$$

□

$$\log \prod_j \binom{r_{i,j} + 2^{n-i}}{2^{n-i}+1} \leq 2^{n+i} \log \binom{2^{n-i} + \bar{r}_i}{2^{n-i}+1} \leq 2^{n+i}(\bar{r}_i - 1) + 2^{n+i}\bar{r}_i \log \frac{2^{n-i} + \bar{r}_i}{\bar{r}_i - 1} \tag{2.34}$$

Here, $\bar{r}_i = 2^{-i-n} \sum_j r_{i,j}$.

Now we are going to bound $\sum_i 2^{n+i}\bar{r}_i \log \frac{2^{n-i}+\bar{r}_i}{\bar{r}_i - 1}$. Let $\mathcal{A} = \{i : \bar{r}_i \leq \frac{1}{n^2} 2^{n-i}\}$,$\mathcal{B} = \{i : \bar{r}_i \geq \frac{1}{n^2} 2^{n-i}\}$. Then we have

$$\sum_{i \in \mathcal{A}} 2^{n+i}\bar{r}_i \log \frac{2^{n-i} + \bar{r}_i}{\bar{r}_i - 1} \leq \sum_{i \in \mathcal{A}} c_1 \frac{1}{n^2} 2^{2n}(n-i)\log 2 \leq c_2 2^{2n} \tag{2.35}$$

$$\sum_{i \in \mathcal{B}} 2^{n+i}\bar{r}_i \log \frac{2^{n-i} + \bar{r}_i}{\bar{r}_i - 1} \leq \sum_{i \in \mathcal{B}} 2^{n+i}\bar{r}_i \log(3n^2) \leq c_3 2^{2n}(\log n)^3 \tag{2.36}$$

Here we use that $\sum_i 2^{n+i}\bar{r}_i \leq 4c' 2^{2n}(\log n)^2$. Lemma 6 is proved by using (2.34),(2.35) and (2.36). □

Using similar method, we can also prove that $\log |\underline{S}| \leq \underline{c} 2^{2n}(\log n)^2$

Up to now, we already prove the following statement.

(i)For any $f \in \mathcal{F}_2$, $\bar{f} \geq f \geq \underline{f}$.

(ii)$\|\bar{f} - \underline{f}\|_2^2 \leq c'(\log n)^2 2^{-2n}$.

(iii)$\log |\bar{S}| \leq \bar{c} 2^{2n}(\log n)^3, \log |\underline{S}| \leq \underline{c} 2^{2n}(\log n)^3$.

So we can see that for $\epsilon = 2^{-n}$,

$$\log N(\epsilon, \mathcal{F}_2, \|\cdot\|_2) \leq c_4 \epsilon^2 (\log\log 1/\epsilon)^5 \tag{2.37}$$

for some constant c_4. The general statement comes from the monotonicity of the entropy number. The proof to Lemma 1 is completed. □

Now we turn to the proof to Theorem 1 and Theorem 4(i). First we prove Theorem 1.

Let the class of bivariate isotonic matrices $\mathbb{C}_{DIFF}$ to be defined as

$$\mathbb{C}_{DIFF} = \{M = M_1 - M_2 | M_1, M_2 \in \mathbb{C}_{SST}\} \tag{2.38}$$

Let $W = Y - M$ and $Z(t) = \sup_{D \in \mathbb{C}_{DIFF}, \|D\|_F \leq t} < D, W >$. If we can prove that $\mathbb{E}Z(t) \leq \frac{t^2}{2}$, with the proof of Theorem 1 in Shah et al. (2016b), we have $\mathbb{E}_M \|\hat{M}_{CLS} - M\|_F^2 \lesssim t^2$.

With equation (28) in Shah et al. (2016b) and Lemma 11, we have for $t = \sqrt{n \log n (\log \log n)^5}$

$$\mathbb{E}Z(t) \leq \frac{t^2}{2} \tag{2.39}$$

which completes the proof to Theorem 2. For the proof to Theorem 4(i), let $W = \bar{Y} - M$. Then the rest of the proof follows the same argument as in the proof to Theorem 1.

□

2.8.2 Proof To Proposition 1

Proof. Let $M \in \mathbb{C}_{BTL}$ and $M_{ij} = \frac{1}{1+\exp(q_j - q_i)}$ for some $q_i > 0, \forall i \in [n]$. Without loss of generality, we assume that $\{q_i\}_{i=1}^n$ is a non-decreasing sequence. To prove that $M \in \mathbb{C}_{BTL}$, it is sufficient to prove $\{M_{ij} - M_{i'j} | j > i'\}$ is a decreasing sequence for $i > i'$.

This is because

$$\begin{aligned} M_{ij} - M_{i'j} &= \frac{1}{1+\exp(q_j - q_i)} - \frac{1}{1+\exp(q_j - q_{i'})} \\ &= \frac{\exp(q_i) - \exp(q_{i'})}{\exp(q_j) + \exp(q_i) + \exp(q_{i'}) + \frac{\exp(q_i + q_{i'})}{\exp(q_j)}} \end{aligned} \tag{2.40}$$

Since $q_i, q_{i'} \geq q_j$, $\{M_{ij} - M_{i'j} | j > i'\}$ is a decreasing sequence for $i > i'$, which shows that $\mathbb{C}_{BTL} \subset \mathbb{C}_{usst}$.

The proof to the argument $\mathbb{C}_{Thurstone} \subset \mathbb{C}_{usst}$ is similar. Assume that $M \in \mathbb{C}_{BTL}$ and $M_{ij} = \phi(q_i - q_j)$ for some $q_i > 0, \forall i \in [n]$. Let $\varphi(x) = \exp(-\frac{x^2}{2})$ to be the p.d.f of the standard normal. Then we have

$$M_{ij} - M_{i'j} = \int_{q_{i'} - q_j}^{q_i - q_j} \varphi(x) dx \tag{2.41}$$

Since $q_i, q_{i'} \geq q_j$, $\{M_{ij} - M_{i'j} | j > i'\}$ is a decreasing sequence for $i > i'$, which shows that $\mathbb{C}_{Thurstone} \subset \mathbb{C}_{usst}$. □

2.8.3 Proof to Theorem 2

Proof. Without loss of generality, we assume that π to be the identical permutation, i.e. $\pi(i) = i, \forall i \in [n]$.

First we try to prove that, with high probability, if $s_1, s_2 \in g_t^{(u)}$, $s_1 < s_2, j_1 \in bl_{s_1}, j_2 \in bl_{s_2}$, then $\pi(j_1) < \pi(j_2)$.

Consider the event

$$\mathcal{E}_1 = \{\exists s_1, s_2 \in g_t^{(u)}, s_1 < s_2, j_1 \in bl_{s_1}, j_2 \in bl_{s_2}, \pi(j_1) > \pi(j_2)\}$$

Let s_1, s_2, j_1, j_2 to be fixed. If $s_1, s_2 \in g_t^{(u)}, s_1 < s_2, j_1 \in bl_{s_1}, j_2 \in bl_{s_2}$ for some (t, u), from the construction of $g_t^{(u)}$, we know that

$$S_{j_2} - S_{j_1} \geq \tau \tag{2.42}$$

If $\pi(j_1) > \pi(j_2)$, Hoeffding's inequality tells us

$$\mathbb{P}(\sum_{i \in [n]} Y_{ij_1} - \sum_{i \in [n]} Y_{ij_2} \leq \tau) \leq 2\exp(-\frac{\tau^2}{2n}) \tag{2.43}$$

which implies that

$$\mathbb{P}(\mathcal{E}_1) \leq \sum_{s_1, s_2, j_1, j_2} 2\exp(-\frac{\tau^2}{2n}) \leq \frac{1}{n^2} \tag{2.44}$$

Assume that $g_t^{(u)} = \{s_1, ..., s_p\}$, $s_1 < .. < s_p$. Consider the event

$$\mathcal{E}_2 = \{\exists i < i' \in [n], a < b \in [p], t \in \lceil \log_2 K \rceil, u \in \{1, 2\},$$
$$\sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{ij} - \sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{i'j} \geq 8\sqrt{(b-a+1)2^t \log n}\}$$

and

$$\mathcal{E}_3 = \{\exists i > i' \in [n], a < b \in [p], t \in \lceil \log_2 K \rceil, u \in \{1,2\},$$
$$\sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{ij} - \sum_{q=a}^{b} \sum_{j \in bl_{s_q}} Y_{i'j} \leq 8\sqrt{(b-a+1)2^t \log n}\}$$

Once again if we implement Hoefdding's inequality Vershynin (2018), we know that

$$\mathbb{P}(\mathcal{E}_2), \mathbb{P}(\mathcal{E}_3) \leq \frac{1}{n^2} \tag{2.45}$$

In the following discussion, we assume that $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ is true. The following discussion is conditioning on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$.

Our goal is to construct the upper bound $\sum_{i=1}^{n} \sum_{j=1}^{m} (M_{\hat{\pi}(i),j} - M_{i,j})^2$. To see this, we try to upper bound

$$\sum_{s \in g_t^{(u)}} \sum_{j \in bl_s} \sum_{\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl'_s) > t} \sum_{i \in bl_{s'}} (M_{\hat{\pi}(i),j} - M_{i,j})^2$$

for $t \in [\lceil \log_2 K \rceil], u \in \{1,2\}$ separately. Here, we say $\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s)$ if and only if there exists $i \in bl_s$ and $i' \in bl_{s'}$, such that $\hat{\pi}(i) < \hat{\pi}(i')$.

Now we turn to estimate the upper bound for

$$\sum_{s \in g_t^{(u)}} \sum_{j \in bl_s} \sum_{\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl'_s) > t} \sum_{i \in bl_{s'}} (M_{\hat{\pi}(i),j} - M_{i,j})^2$$

. We will first try to bound the above term for each i. For $i \in bl_{s'}$, such that $\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl_{s'}) > t$, let T_i to be the set of the blocks such that $s \in g_t^{(u)}, \hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s)$. More concretely, we define that

$$T_i = \{s | s \in g_t^{(u)}, \hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s)\} \tag{2.46}$$

For $i, i' \in bl_{s'}$, we have that $T_i = T_{i'}$. We also use the notation $T(bl_{s'})$, where $T(bl_{s'}) = T_i$ for some $i \in bl_{s'}$. For i, we assume that $T_i = \{s_1, ..., s_p\}$. We divide the sum

$$\sum_{s\in g_t^{(u)}}\sum_{j\in bl_s}\sum_{\hat{\pi}(bl_{s'})<\hat{\pi}(bl_s),t(bl_s')>t}\sum_{i\in bl_{s'}}(M_{\hat{\pi}(i),j}-M_{i,j})^2$$

into two parts. Let

$$C_1=\sum_{i\in bl_{s'}}\sum_{\hat{\pi}(bl_{s'})<\hat{\pi}(bl_s),t(bl_s')>t}\sum_{ind=2}^{p}\sum_{j\in bl_{s_{ind}}}(M_{\hat{\pi}(i),j}-M_{i,j})^2 \tag{2.47}$$

and

$$C_2=\sum_{i\in bl_{s'}}\sum_{\hat{\pi}(bl_{s'})<\hat{\pi}(bl_s),t(bl_s')>t}\sum_{j\in bl_{s_1}}(M_{\hat{\pi}(i),j}-M_{i,j})^2 \tag{2.48}$$

We will construct the upper bound for C_1 and C_2 separately.

Moreover, we obtain the following bounds

$$\sum_{q=a}^{b}\sum_{j\in bl_{s_q}}|M_{ij}-M_{\hat{\pi}(i)j}|\le 16\sqrt{(b-a+1)2^t\log n} \tag{2.49}$$

for $\forall i\in[n]$, which implies that

$$\sum_{q=a}^{b}\sum_{j\in bl_{s_q}}c_{i,j}\le 16\sqrt{(b-a+1)2^t\log n} \tag{2.50}$$

To construct the upper bound for $\sum_{s\in g_t^{(u)}}\sum_{j\in bl_s}c_{i,j}^2$, we implement Lemma **??** , which indicates that

$$C_1\le 256n(\log n)^2 \tag{2.51}$$

for fixed t and u.

Now we turn to C_2. To construct the upper bound for C_2, we attempt to use(Mao et al., 2018, Lemma 8). We further rewrite C_2 into

$$C_2=\sum_{s'\in K}\sum_{i\in bl_{s'}}\sum_{j\in bl_{s_1}}(M_{\hat{\pi}(i),j}-M_{i,j})^2 \tag{2.52}$$

and we denote that

$$C_{2,s'} = \sum_{i \in bl_{s'}} \sum_{j \in bl_{s_1}} (M_{\hat{\pi}(i),j} - M_{i,j})^2 \tag{2.53}$$

With (Mao et al., 2018, Lemma 8) and the construction of $\hat{\pi}$ in Algorithm 1, we know that

$$C_{2,s'} \leq 600 \log n |bl_{s'}| \tag{2.54}$$

which immediately implies that

$$C_2 = \sum_{s'} C_{2,s'} \leq 600n \log n \tag{2.55}$$

Combine the upper bound for C_1 and C_2 and then sum over all possible t and u in the above inequality, it tells us that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (M_{\hat{\pi}(i),j} - M_{i,j})^2 \leq 856n(\log n)^3 \tag{2.56}$$

Theorem 2 is proved with (2.56) and (Mao et al., 2018, Proposition 1). □

2.8.4 Proof to Theorem 3 and Theorem 4 (ii)

Proof. We begin the proof with the following lemma.

Lemma 8. *Let $B \in \mathbb{R}^{n \times n}$ to be a matrix in the parameter space Θ_1, where Θ_1 is defined as follows.*

We divide $[n]$ into $m = \lfloor \frac{n}{2 \log n} \rfloor$ subsets T_i, $1 \leq i \leq m$, with $T_i = \{(i-1)m+1, ..., im\}$. Let $u = \lfloor \frac{m}{2} \rfloor$. Let $\delta = \frac{1}{8}$.

The following parameter space Θ_1 is defined as

$$\Theta_1 = \{B(\gamma, A_1, ..., A_u) | \gamma \in \{0,1\}^u, A_1, ..., A_u \text{ are disjoint subsets in } \{u\lfloor \log n \rfloor + 1, ..., n\}, \\ \text{such that } |A_i| = \lfloor \log n \rfloor, 1 \leq i \leq u\}$$

where

$$B(\gamma, A_1, .., A_u)(i,j) = \begin{cases} \frac{1}{4}, & \textit{if } i,j \in T_s, \textit{ for some } s \in [u] \\ \frac{1}{4}, & \textit{if } i,j \in A_s, \gamma(s) = 1, \textit{ for some } s \in [u] \\ \frac{1}{4}, & \textit{if } \exists s \in [u], i \in T_s, j \in A_s, \gamma(s) = 1 \\ \frac{1}{4}, & \textit{if } \exists s \in [u], j \in T_s, i \in A_s, \gamma(s) = 1 \\ \frac{1}{4} - \delta^2, & \textit{otherwise} \end{cases}$$

If we observe $Y \sim Ber(B)$,then we have the following lower bound of the estimation to the matrix B.

$$\max_{B \in \Theta_1} \mathbb{E}_B \|\hat{B}(Y) - B\|_F^2 \geq \frac{n \log n}{4096} \tag{2.57}$$

Proof to Lemma 8. To the construction of the lower bound $\sup_{B \in \Theta_1} \mathbb{E}_B \|\widehat{B}(Y) - B\|_F^2$, for any estimator $\hat{B}(Y)$, the tool we use in the construction is Le cam-Assouad method. We introduce the Le Cam-Assouad Lemma to help us construct the statistical lower bound.

Let $X \sim \mathbb{P}_\theta$, where $\theta \in \Theta = \Gamma \otimes \Lambda$ is the Cartesian product of two components Γ and Λ. Assume that $\Gamma = \{0,1\}^r$ and $\Lambda \subset B^r$ for some finite set $B \subset \mathbb{R}^p$. For $\theta = (\gamma, \lambda) \in \Theta$,denote the projection of θ to Γ by $\gamma(\theta) = (\gamma_i(\theta))_{1 \leq i \leq r}$ and to Λ by $\lambda(\theta) = (\lambda_i(\theta))_{1 \leq i \leq r}$.

Lemma 9. *(Le Cam-Assouad) For any estimator T of $\psi(\theta)$ based on an observation from a probability distribution in $P_\theta, \theta \in \Theta$, and any $s > 0$,*

$$\max_{\Theta} 2^s \mathbb{E}_\theta d^s(T, \psi(\theta)) \geq \alpha \frac{r}{2} \min_{1 \leq i \leq r} \|\mathbb{P}_{0,i} \wedge \mathbb{P}_{1,i}\|$$

where α is given by

$$\alpha = \frac{d^s(\psi_\theta, \psi(\tilde{\theta}))}{H(\gamma(\theta), \gamma(\tilde{\theta}))}.$$

The proof to Lemma 9 can be found in Cai and Zhou (2012).

Define $\mathbb{P}_{a,i}$ as

$$\mathbb{P}_{a,i} = \frac{1}{2^{r-1}D_\Lambda}\sum_{\theta}\{\mathbb{P}_\theta : \gamma_i(\theta) = a\}$$

A simple observation from our construction of the parameter space Θ_1 is that

$$\min_{\{H(\gamma,\tilde{\gamma})\geq 0\}} \frac{\|B(\gamma, A_1, ..., A_u), B(\tilde{\gamma}, \tilde{A}_1, ..., \tilde{A}_u)\|_F^2}{H(\gamma,\tilde{\gamma})} \geq 2\delta^2\lfloor\log n\rfloor^2 \tag{2.58}$$

To complete the proof with Le cam-Assouad method, it is sufficient to prove that

$$\min_{1\leq i\leq u} \|\bar{P}_{0,i} \wedge \bar{P}_{1,i}\| \geq c_0 \tag{2.59}$$

for some constant c_0, where

$$\bar{P}_{0,i} = \frac{1}{2^{u-1}|S|}\sum_{\gamma(i)=0} \mathbb{P}_{(\gamma,A_1,..,A_u)} \tag{2.60}$$

$$\bar{P}_{1,i} = \frac{1}{2^{u-1}|S|}\sum_{\gamma(i)=1} \mathbb{P}_{(\gamma,A_1,..,A_u)} \tag{2.61}$$

for which $\mathbb{P}_{(\gamma,A_1,..,A_u)}$ is the probability measure of Y. With Lemma 4 in Cai and Zhou (2012), once we can prove that for fixed $\gamma(2), .., \gamma(u)$, $A_2, ..., A_u$, $\|\mathbb{P}_0 \wedge \mathbb{P}_1\| \geq c_0$, (2.59) is true, where

$$\mathbb{P}_0 = \frac{1}{|S_1|}\sum_{\gamma(1)=0,\gamma(2),\gamma(u),A_2,...,A_u} \mathbb{P}_{(\gamma,A_1,..,A_u)} \tag{2.62}$$

$$\mathbb{P}_1 = \frac{1}{|S_1|}\sum_{\gamma(1)=1,\gamma(2),\gamma(u),A_2,...,A_u} \mathbb{P}_{(\gamma,A_1,..,A_u)} \tag{2.63}$$

The last goal in proving the theorem is to prove that $\|\mathbb{P}_0 \wedge \mathbb{P}_1\| \geq c_0$. We prove it by calculating $\chi^2(\mathbb{P}_1, \mathbb{P}_0)$. Notice that when $\gamma(2), ..., \gamma(u), A_2, ..., A_u$ are fixed, if $\gamma(1) = 0$, the probability measure remains the same. We denote it by p. We denote q_{A_1} to be the probability measure of Y, corresponding to $B = B(\gamma(1) = 1, \gamma(2), ..., \gamma(u), A_1, ..., A_u)$ when $\gamma(2), ..., \gamma(u), A_1, ..., A_u$ are all fixed.

Since $A_1, ..., A_u$ are disjoint subsets in $\{u\lfloor \log n \rfloor + 1, ..., n\}$, if we let $C = \{u\lfloor \log n \rfloor + 1, ..., n\} \setminus \cup_{j=2}^{u} A_j$, we would have

$$\begin{aligned} |C| \geq & n - u\lfloor \log n \rfloor - u\lfloor \log n \rfloor \\ \geq & n - m\lfloor \log n \rfloor \geq \frac{n}{3} \end{aligned} \tag{2.64}$$

for large enough n. Since A_1 is a subset in C, with cardinality $\lfloor \log n \rfloor$, we have that $|S_1| = \binom{v}{\lfloor \log n \rfloor}$, where $v = |C|$. So

$$\mathbb{P}_1 = \frac{1}{\binom{v}{\lfloor \log n \rfloor}} \sum_{\gamma(1)=1, \gamma(2), \gamma(u), A_2, ..., A_u} \mathbb{P}_{(\gamma, A_1, .., A_u)} \tag{2.65}$$

which implies that

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) = \mathbb{E}_{A_1, A_1'} \int \frac{q_{A_1} q_{A_1'}}{p} \tag{2.66}$$

where A_1, A_1' are uniformly chosen from the $\binom{v}{\lfloor \log n \rfloor}$ subsets. Let $J = |A_1 \cap A_1'|$. From our construction of the parameter space Θ_1, $J \sim \text{Hypergeometric}(v, \lfloor \log n \rfloor, \lfloor \log n \rfloor)$. The χ^2 affinity can be upper bounded by

$$\begin{aligned} & \mathbb{E}_{A_1, A_1'} \int \frac{q_{A_1} q_{A_1'}}{p} \\ \leq & \mathbb{E}(1 + \frac{16\delta^4}{3})^{2J\lfloor \log n \rfloor} \\ \leq & \mathbb{E} \exp(\frac{32\delta^4}{3} J \lfloor \log n \rfloor) \\ \leq & \frac{\lfloor \log n \rfloor^2}{n - \lfloor \log n \rfloor} \exp(\frac{\log^2 n}{n} n^{\frac{32\delta^4}{3}}) \leq 1 + \frac{1}{4} \end{aligned} \tag{2.67}$$

for large enough n and $\delta = \frac{1}{8}$, which implies that $\|\mathbb{P}_1 \wedge \mathbb{P}_0\| \geq \frac{1}{2}$.

Apply Le cam-Assouad lemma in our parameter space Θ_1,

$$\max_{B \in \Theta_1} \mathbb{E}_B \|\hat{B} - B\|_F^2 \geq \frac{1}{4} \delta^2 \lfloor \log n \rfloor^2 \frac{u}{2} \geq \frac{n \log n}{4096} \tag{2.68}$$

for large enough n, which completes the proof to Lemma 8. □

Now we can prove Theorem 3 with Lemma 8.

For any $M \in \mathbb{C}_{usst}$, we define a matrix $B = M(B)$, such that $\forall i, j \in [n]$,

$$B_{ij} = M_{ij}M_{ji}$$

. For any estimator $\widehat{M} \in [0,1]^{n\times n}$, we define an estimator $\widehat{B}$ of B with $\widehat{M}$, such that

$$\widehat{B}_{ij} = \widehat{M}_{ij}\widehat{M}_{ji}$$

From the construction of the estimator $\widehat{B}$, we know that the error of $\widehat{B}$ can be controlled by the error of the estimator $\widehat{M}$. This is because

$$\begin{aligned}
&\|\widehat{B} - B\|_F^2 \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n}(\widehat{B}_{ij} - B_{ij})^2 \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n}(\widehat{M}_{ij}\widehat{M}_{ji} - M_{ij}M_{ji})^2 \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n}((\widehat{M}_{ij} - M_{ij})\widehat{M}_{ji} + M_{ij}(\widehat{M}_{ji} - M_{ji}))^2 \\
\leq&2\sum_{i=1}^{n}\sum_{j=1}^{n}(\widehat{M}_{ij} - M_{ij})^2 + (\widehat{M}_{ji} - M_{ji})^2 \\
=&4\|\widehat{M} - M\|_F^2
\end{aligned} \tag{2.69}$$

It implies that

$$\begin{aligned}
&\sup_{M\in\mathbb{C}_{usst}} \mathbb{E}_M\|\widehat{M}(Y) - M\|_F^2 \\
\geq&\frac{1}{4}\sup_{M\in\mathbb{C}_{usst}} \mathbb{E}_M\|\widehat{B}(Y) - B\|_F^2 \\
\geq&\frac{1}{4}\sup_{B\in\Theta'} \mathbb{E}_B\|\widehat{B}(X) - B\|_F^2
\end{aligned} \tag{2.70}$$

where $\Theta' = \{M(B)|M \in \mathbb{C}_{usst}\}$, $Y \in [0,1]^{n\times n}$ to be a random matrix, such that $X_{ij} = Y_{ij}Y_{ji}, \forall i, j \in [n]$. It is easy to check that X_{ij} are independent random Bernoulli variables, $X_{ij} \sim \text{Ber}(B_{ij}), \forall 1 \leq i < j \leq n$.

We prove that $\Theta_1 \subset \Theta'$. We prove this result through a construction of a matrix $M(\gamma, A_1, .., A_u) \in \Theta$, such that $M(\gamma, A_1, .., A_u)(B) = B(\gamma, A_1, .., A_u)$.

For $(\gamma, A_1, ..., A_u)$, we define a label $l : [n] \to [n]$, such that for $i \in T_j$ or $i \in A_j$, if $\gamma(j) = 1$, $l(i) = j$. If $i \notin \cup_{j:\gamma(j)=1}(T_j \cup A_j)$, $l(i) = u + 1$. We define $M(\gamma, A_1, .., A_u)$ to be

$$M(\gamma, A_1, .., A_u) = \begin{cases} \frac{1}{2}, \text{ if } l(i) = l(j) \\ \frac{1}{2} + \delta, \text{ if } l(i) < l(j) \\ \frac{1}{2} - \delta, \text{ if } l(i) > l(j) \end{cases}$$

It is easy to check that $M(\gamma, A_1, .., A_u) \in \Theta$ and $M(\gamma, A_1, .., A_u)(B) = B(\gamma, A_1, .., A_u)$. So $\Theta_1 \subset \Theta'$.

The rest of the Theorem can be proved with Lemma 8.

The proof to Theorem 4 (ii) comes from the fact

$$\chi^2(\mathbb{P}_1^{ind}, \mathbb{P}_0^{ind}) = \frac{N}{n^2}\chi^2(\mathbb{P}_1 \mathbb{P}_0) \tag{2.71}$$

and the same argument for the proof in Theorem 3, where $\mathbb{P}_1^{ind}, \mathbb{P}_0^{ind}$ is the probability measure for the corresponding independent case. $\square$

2.8.5 Proof to Theorem 5

Proof. Without loss of generality, we assume that π to be the identical permutation, i.e. $\pi(i) = i, \forall i \in [n]$.

First we try to prove that, with high probability, if $s_1, s_2 \in g_t^{(u)}$, $s_1 < s_2, j_1 \in bl_{s_1}, j_2 \in bl_{s_2}$, then $\pi(j_1) < \pi(j_2)$.

Consider the event

$$\mathcal{E}_4 = \{\exists s_1, s_2 \in g_t^{(u)}, s_1 < s_2, j_1 \in bl_{s_1}, j_2 \in bl_{s_2}, \pi(j_1) > \pi(j_2)\}$$

Fixed $j \in [n]$. Let $m = \lfloor \frac{N}{3} \rfloor$. For $1 \le k \le m$, let

$$Z_k = \begin{cases} 0, \text{ if } j_k \neq j \\ Y_{i_k,j}, \text{ if } j_k = j \end{cases} \tag{2.72}$$

From our construction, $\mathbb{E}Z_k = \frac{1}{n^2}\sum_{i\in[n]} M_{ij}$. Since $\mathbb{P}Z_k \neq 0 = \frac{1}{n}$ and $Z_k \le 1$, we know that $\mathbb{E}Z_k^2 \le \frac{1}{n}, \forall k$. Use Berstein's Inequality,

$$\mathbb{P}(|\sum_{i=1}^{N} Z_i - \frac{N}{n^2}\sum_{i\in[n]} M_{ij}| \ge t) \le 2\exp(-\frac{t^2}{\frac{N}{n}+\frac{t}{3}}) \tag{2.73}$$

Let $t = 6\sqrt{\frac{N}{n}\log n}$. With our assumption, we know that $\frac{t}{3} \le \frac{N}{n}$. So

$$2\exp(-\frac{t^2}{\frac{N}{n}+\frac{t}{3}}) \le \frac{2}{n^{18}} \tag{2.74}$$

which implies that

$$\mathbb{P}(|\frac{n^2}{N}\sum_{k\in[m]} : j_k = jY_{i_k,j} - \sum_{i\in[n]} M_{ij}| \ge 6\sqrt{\frac{N\log n}{n}}) \le \frac{2}{n^{17}} \tag{2.75}$$

$\forall j \in [n]$. So $\mathbb{P}(\mathcal{E}_4) \le \frac{2}{n^{17}}$.

Assume that $g_t^{(u)} = \{s_1, ..., s_p\}$, $s_1 < .. < s_p$. Consider the event

$$\mathcal{E}_5 = \{\exists i < i' \in [n], a < b \in [p], t \in \lceil \log_2 K \rceil, u \in \{1,2\},$$
$$\sum_{q=a}^{b}\sum_{j\in bl_{s_q}} Y_{ij} - \sum_{q=a}^{b}\sum_{j\in bl_{s_q}} Y_{i'j} \ge 8\sqrt{\frac{(b-a+1)2^t n^3\log n}{N}}\}$$

and

$$\mathcal{E}_6 = \{\exists i > i' \in [n], a < b \in [p], t \in \lceil \log_2 K \rceil, u \in \{1,2\},$$
$$\sum_{q=a}^{b}\sum_{j\in bl_{s_q}} Y_{ij} - \sum_{q=a}^{b}\sum_{j\in bl_{s_q}} Y_{i'j} \le 8\sqrt{\frac{(b-a+1)2^t n^3\log n}{N}}\}$$

One again if we implement Hoefdding's inequality, we know that

$$\mathbb{P}(\mathcal{E}_5), \mathbb{P}(\mathcal{E}_6) \leq \frac{2}{n^{17}} \tag{2.76}$$

In the following discussion, we assume that $\mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6$ is true. The following discussion is conditioning on the event $\mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6$.

Our goal is to construct the upper bound $\sum_{i=1}^n \sum_{j=1}^m (M_{\hat{\pi}(i),j} - M_{i,j})^2$. To see this, we try to upper bound

$$\sum_{s \in g_t^{(u)}} \sum_{j \in bl_s} \sum_{\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl_s') > t} \sum_{i \in bl_{s'}} (M_{\hat{\pi}(i),j} - M_{i,j})^2$$

for $t \in [\lceil \log_2 K \rceil], u \in \{1, 2\}$ separately. Here, we say $\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s)$ if and only if there exists $i \in bl_s$ and $i' \in bl_{s'}$, such that $\hat{\pi}(i) < \hat{\pi}(i')$.

Now we turn to estimate the upper bound for

$$\sum_{s \in g_t^{(u)}} \sum_{j \in bl_s} \sum_{\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl_s') > t} \sum_{i \in bl_{s'}} (M_{\hat{\pi}(i),j} - M_{i,j})^2$$

. We will first try to bound the above term for each i. For $i \in bl_{s'}$, such that $\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl_{s'}) > t$, let T_i to be the set of the blocks such that $s \in g_t^{(u)}, \hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s)$. More concretely, we define that

$$T_i = \{s | s \in g_t^{(u)}, \hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s)\} \tag{2.77}$$

For $i, i' \in bl_{s'}$, we have that $T_i = T_{i'}$. We also use the notation $T(bl_{s'})$, where $T(bl_{s'}) = T_i$ for some $i \in bl_{s'}$. For i, we assume that $T_i = \{s_1, ..., s_p\}$. We divide the sum

$$\sum_{s \in g_t^{(u)}} \sum_{j \in bl_s} \sum_{\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl_s') > t} \sum_{i \in bl_{s'}} (M_{\hat{\pi}(i),j} - M_{i,j})^2$$

into two parts. Let

$$C_3 = \sum_{i \in bl_{s'}} \sum_{\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl'_s) > t} \sum_{ind=2}^{p} \sum_{j \in bl_{s_{ind}}} (M_{\hat{\pi}(i),j} - M_{i,j})^2 \tag{2.78}$$

and

$$C_4 = \sum_{i \in bl_{s'}} \sum_{\hat{\pi}(bl_{s'}) < \hat{\pi}(bl_s), t(bl'_s) > t} \sum_{j \in bl_{s_1}} (M_{\hat{\pi}(i),j} - M_{i,j})^2 \tag{2.79}$$

We will construct the upper bound for C_3 and C_4 separately.

We obtain the following bounds

$$\sum_{q=a}^{b} \sum_{j \in bl_{s_q}} |M_{ij} - M_{\hat{\pi}(i)j}| \leq 96\sqrt{\frac{(b-a+1)2^t n^3 \log n}{N}} \tag{2.80}$$

for $\forall i \in [n]$, which implies that

$$\sum_{q=a}^{b} \sum_{j \in bl_{s_q}} c_{i,j} \leq 96\sqrt{\frac{(b-a+1)2^t n^3 \log n}{N}} \tag{2.81}$$

To construct the upper bound for $\sum_{s \in g_t^{(u)}} \sum_{j \in bl_s} c_{i,j}^2$, we implement Lemma **??** , which indicates that

$$C_3 \leq 1536 \frac{n^3 (\log n)^2}{N} \tag{2.82}$$

for fixed t and u.

Now we turn to the upper bound for C_4. To construct the upper bound for C_4, we attempt to use (Mao et al., 2018, Lemma 8). We further rewrite C_4 into

$$C_4 = \sum_{s' \in K} \sum_{i \in bl_{s'}} \sum_{j \in bl_{s_1}} (M_{\hat{\pi}(i),j} - M_{i,j})^2 \tag{2.83}$$

and we denote that

$$C_{4,s'} = \sum_{i \in bl_{s'}} \sum_{j \in bl_{s_1}} (M_{\hat{\pi}(i),j} - M_{i,j})^2 \tag{2.84}$$

With (Mao et al., 2018, Lemma 8) and the construction of $\hat{\pi}$ in Algorithm 1, we know that

$$C_{4,s'} \leq 3600 \log n |bl_{s'}| \tag{2.85}$$

which immediately implies that

$$C_4 = \sum_{s'} C_{2,s'} \leq 3600 \frac{n^3 (\log n)^2}{N} \tag{2.86}$$

Combine the upper bound for C_3 and C_4 and then sum over all possible t and u in the above inequality, it tells us that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (M_{\hat{\pi}(i),j} - M_{i,j})^2 \leq 4000 \frac{n^3 (\log n)^3}{N} \tag{2.87}$$

The theorem is proved with (2.87) and (Mao et al., 2018, Proposition 1). □

2.8.6 Proof to Theorem 6

Proof. We divide the proof to Theorem 6 into lower bound part and upper bound part. We begin with the upper bound part.

To prove the upper bound part, we are going to prove that

$$\mathbb{E} d_{KT}(\hat{\pi}_{naive}, \pi^*) \leq C \min\{\frac{n^{3/2}}{\lambda}, n^2\} \tag{2.88}$$

$\forall M \in D_n(\lambda) \cap \mathbb{C}_{SST}$ and some constant C. From the definition of Kendall tau distance, we can see that

$$\mathbb{E} d_{KT}(\hat{\pi}_{naive}, \pi^*) = \sum_{\pi_M(i) < \pi_M(j)} \mathbb{P}(\hat{\pi}_{naive}(i) > \hat{\pi}_{naive}(j)) \tag{2.89}$$

With Hoeffding's inequality and the assumption $M \in D_n(\lambda)$, we have

$$
\begin{aligned}
\sum_{\pi_M(i)<\pi_M(j)} \mathbb{P}(\hat{\pi}_{naive}(i) > \hat{\pi}_{naive}(j)) \leq & n\sum_{k=1}^{n} \exp(-\frac{k^2\lambda^2}{n}) \\
= & n[\sum_{k=1}^{l} \exp(-\frac{k^2\lambda^2}{n}) + \sum_{k=l+1}^{n} \exp(-\frac{k^2\lambda^2}{n})] \\
\leq & n(l + \sum_{k=l+1}^{n} \frac{n}{k^2\lambda^2}) \\
\leq & n(l + \frac{n}{l\lambda^2}) \leq \frac{3n^{3/2}}{\lambda}
\end{aligned}
\tag{2.90}
$$

where $l = \lceil \frac{\sqrt{n}}{\lambda} \rceil$. So the upper bound for Theorem 6 is proved, as $d_{KT}(\hat{\pi}_{naive}, \pi^*) \leq n^2$ is trivial.

Then we turn to the lower bound part. For the proof of the lower bound part, we need the following lemma.

Lemma 10. *If $n < r \leq \frac{n^2}{4}$, there exists a subset $Q \subset \mathbb{S}_n$, such that*

(i)$\log|Q| \geq \frac{n^2}{30r}$,

(ii)$d_{KT}(\pi_1, \pi_2) \geq \frac{r}{96}$,

(iii$\|\pi - id\|_2^2 \leq r$.

Here $\|\pi - id\|_2^2 = \sum_{i=1}^{n}(\pi(i) - i)^2$.

Proof to Lemma 10. Let $m = \lfloor \frac{r}{n} \rfloor$,$k = \lfloor \frac{n}{m} \rfloor$ and $v = \lfloor \frac{k}{3} \rfloor$. Using the well-celebrated Varshamov-Gilbert bound, there exists a set $\mathcal{S}$ of v-sparse vectors in $\{0,1\}^k$, such that $\log|S| \geq \frac{k}{15}$ and any two two distinct vectors in $\mathcal{S}$ are separated by at least $v/2$ in the Hamming distance.

We define Q through a map from $\mathcal{S}$ to $\mathbb{S}_n$. Let $I_i = \{(i-1)m, ...im\}, 1 \leq i \leq k$, $I_{k+1} = [n] \backslash \cup_{i=1}^{k} I_i$. For $s = (t_1, ..., t_k)$, we define the corresponding permutation $\pi_{\mathcal{S}}$ to be

$\pi_s(j) = j$, if $j \in I_l$ and $t_l = 0$.

$\pi_s(j) = (2l-1)m + 1 - j$, if $j \in I_l$ and $t_l = 1$.

$\pi_s(j) = j$, if $j \in I_l$ and $l = k+1$.

Then all conditions are satisfied by chooseing $Q = \{\pi_s | s \in \mathcal{S}\}$, which completes the proof for Lemma 10. $\square$

With Lemma 10, we can finish the proof of the lower bound. We assume that $n = 2m$. The proof to the case when n is add is similar. Let M_λ to be a matrix defined by

$$M_{ij} = \frac{1}{2}, M_{i+m,j+m} = \frac{1}{2}, 1 \leq i, j \leq m$$

$$M_{i,j+m} = \frac{1}{2} + \frac{j\lambda}{2n}, M_{i+m,j} = \frac{1}{2} - \frac{i\lambda}{2n}, 1 \leq i, j \leq m$$

If $m < r \leq \frac{m^2}{4}$ for some integer r which we will choose later, with Lemma 10, we can construct a set $Q \subset \mathbb{S}_m$, such that all conditions in Lemma 10 are satisfied. For $\forall \pi \in Q$, we define M^π to be a matrix, such that

$$M^\pi_{ij} = \frac{1}{2}, M^\pi_{i+m,j+m} = \frac{1}{2}, 1 \leq i, j \leq m$$

$$M^\pi_{i,j+m} = \frac{1}{2} + \frac{\pi(j)\lambda}{2n}, M^\pi_{i+m,j} = \frac{1}{2} - \frac{\pi(i)\lambda}{2n}, 1 \leq i, j \leq m$$

Let $\pi_1, ..., \pi_{|Q|}$ to be the elements of Q and define $\pi_0 = id$. Let $\mathbb{P}_k$ to be the joint probability of Y under $M = M^{\pi_k}$. By Lemma 10, we have

$$D(\mathbb{P}_k, \mathbb{P}_0) \leq \frac{8r\lambda^2}{n} \tag{2.91}$$

So we have

$$\frac{1}{|Q|} \sum_{k=1}^{|Q|} D(\mathbb{P}_k, \mathbb{P}_0) \leq \frac{8r\lambda^2}{n} \leq \frac{1}{16} \log |Q| \tag{2.92}$$

if we choose $r = \lfloor \frac{n^{3/2}}{\sqrt{3840}\lambda} \rfloor$.

If n is sufficiently large, we always have $r > m$. If $\lambda \geq \frac{1}{10\sqrt{n}}$, we will have that $r < \frac{m^2}{4}$. With Fano's lemma, we have

$$\inf_{\hat{\pi}(Y)} \sup_{M \in D_n(\lambda) \cap \mathbb{C}_{SST}, \pi^*} \mathbb{E} d_{KT}(\hat{\pi}, \pi(M)) \geq C \frac{n^{3/2}}{\lambda} \tag{2.93}$$

for some constant C. If $\lambda < \frac{1}{10\sqrt{n}}$, choose $r = \frac{n^2}{\sqrt{38400}}$, we have

$$\inf_{\hat{\pi}(Y)} \sup_{M \in D_n(\lambda) \cap \mathbb{C}_{SST}, \pi^*} \mathbb{E} d_{KT}(\hat{\pi}, \pi(M)) \geq C n^2 \tag{2.94}$$

for some constant C. The proof of the lower bound part is completed by combining (2.93) and (2.94). □

3 Statistical Inference For Permutation Based Model

3.1 Introduction

Lots of efforts have been spent on the estimation of the permutation-based models, while the inference problem has got much less attention to the permutation based models. We study the estimation problem of unimodal SST model in pairwise comparison problem in Chapter 2. In the current chapter, we are going to study the inference problem related to the permutation based model.

In the permutation based model, there are usually two key factors: the probability matrix and the permutation. We will discuss the inference problem for both the probability matrix and the permutation in this chapter. We begin the section with the hypothesis testing problem of the probability matrix in the noisy sorting model. The noisy sorting model is first proposed by Braverman and Mossel (2008). It is used and studied in the pairwise comparison problem. The minimax rate of estimation in the noisy sorting model is established in Mao et al. (2017). Based on the results in Mao et al. (2017), a natural question is to establish the minimax detection level of the signal strength in the hypothesis testing framework. In the current chapter, we propose a testing procedure to the hypothesis testing problem and prove it is optimal, from where we establish the minimax detection level of the signal strength in the noisy sorting model.

Other than the probability matrix, permutation is another important topic in the permutation based model. After studying the hypothesis testing problem about the probability matrix in noisy sorting model, we focus on the inference problem about the permutation. We consider the confidence set construction problem for the permutation in the permutation based model with different settings. One challenge in the study of the confidence set construction problem is to find a suitable criterion in judging the confidence set procedure. We will show how to properly set up the confidence set construction problem and propose the near optimal confidence set construction procedure in different kinds of permutation based model. We also study the hypothesis testing problem of permutation in the current section. As the close relationship between the confidence set construction problem and the hypothesis testing problem, we can see that the results are similar for both problems in the same setting.

3.1.1 Organization

The rest of the chapter organizes as follows. In Section 3.3.2 we mainly talked about the hypothesis testing problem for the probability matrix. In Section 3.3.1, we consider the confidence set construction problem under different settings. The hypothesis testing problem for the permutation is studied in Section 3.3.2. We have more discussion to the relationship between the inference problem and estimation problem for permutation based model in Section 3.4. The proofs of the results in the current chapter will be in Section 3.5.

3.2 Statistical Inference for Probability Matrix

In this section, we discuss the inference problem in of probability matrix in the noisy sorting model. The noisy sorting model is first proposed by Braverman and Mossel (2008). The noisy sorting model is often used in the pairwise comparison problem. The estimation to the permutation problem is studied in in Mao et al. (2017), where the minimax rate for the estimation to the permutation in the noisy sorting problem is constructed .

We state the the noisy sorting model as follows. Fix an unknown permutation $\pi^* \in S_n$, which determines the underlying order of n items. More precisely, π^* orders the items from the weakest to the strongest, so that item i is in the $\pi^*(i)$-th weakest among the n items. For a fixed $\lambda \in (0, 1/2)$, we define a class of matrices

$$M_n(\lambda) = \{M \in [0,1]^{n\times n} : M_{i,i} = \frac{1}{2}, M_{i,j} \geq \frac{1}{2} + \lambda \text{ if } i > j, M_{i,j} \leq \frac{1}{2} - \lambda \text{ if } i < j$$
$$M_{ii} = \frac{1}{2}, M_{ij} + M_{ji} = 1, \forall k, M_{ik} \leq M_{jk} \text{ if } i < j\}$$

To model pairwise comparisons with noisy sorting model, fix $M \in M_n(\lambda)$ for some λ and let $M_{\pi^*(i),\pi^*(j)}$ denote the probability that items i wins the comparison against item j, so that a stronger item beats a weaker item with probability at least $\frac{1}{2} + \lambda$. As a result, λ captures the signal strength of the problem. We assume that we observe the comparison result to all possible pairs. To be more concrete, we have

$$Y_{ij} \sim \text{Ber}(M_{\pi^*(i),\pi^*(j)}), 1 \leq i \leq j$$

with all $Y_{ij}, 1 \le i, j \le n$ are independent random variables. Let $Y = (Y_{ij})_{1\le i,j\le n}$.

Though the estimation problem in noisy sorting model has been well studied, little is known to the inference problem in the noisy sorting problem. One natural question for the inference problem is to establish the minimax detection level of the signal strength of λ.

We consider the problem of testing the following statistical problem.

$$H_0 : M = M_0 \text{ vs. } H_a : M \in M_n(\lambda) \tag{3.1}$$

where $M_0 \in \mathbb{R}^{n\times n}$ is the matrix defined as $M_0(i,j) = \frac{1}{2}, \forall i, j \in \mathbb{R}^{n\times n}$.

It is clear that the difficulty of the testing problem between H_0 and H_a depends on λ. If λ is large, the problem would be easy, while the problem becomes difficult when λ is small. Our interest is to find the boundary that separates the testable regime. We want to know the smallest signal strength λ such that the hypothesis testing problem (3.1) is solvable. In Section 3.2.1, we introduce the testing procedure to the hypothesis testing problem. We first introduce the test statistic, followed by the study of the testing procedure to show that the hypothesis testing problem can be solved if $\lambda \gtrsim n^{-\frac{3}{4}}$. Then Section 3.2.2 shows that the hypothesis testing problem is impossible to solve if $\lambda \lesssim n^{-\frac{3}{4}}$.

3.2.1 Test Statistic

For $1 \le i \le n$, let $S_i = \sum_{j=1}^{n} Y_{ij}$. Define test statistic

$$W = \frac{4}{n}\sum_{i=1}^{n}(S_i - \frac{n}{2})^2 \tag{3.2}$$

The test statistic W shows the deviation of the observation Y from the true probability matrix M_0. We can see that under the null, we have $\mathbb{E}S_i = \frac{n}{2}, \forall i \in [n]$. In this sense, W comes from the variance of the model. Under the alternative, for most $i \in [n]$, $\mathbb{E}S_i \neq \frac{n}{2}$. In that scenario, not just the variance but also the bias contributes to W. To our intuition, we should reject the null hypothesis if W is large.

We specify our rejection region as follows. We reject the null hypothesis if and only if $|W - n| \ge C\sqrt{n \log \frac{2}{\alpha}}$, where $\alpha \in (0,1)$ is a constant to control the Type I error and C is a constant. We could define the test as the following

$$\psi(Y) = I(|W - n| \geq C\sqrt{n \log \frac{2}{\alpha}}) \tag{3.3}$$

Here, $I(\cdot)$ is the indicator function. Then we have the following result for the testing procedure (3.3).

Theorem 7. *(i)With suitable choice of constant C,the testing procedure (3.3) is a level-α test to the testing problem (3.1).*

(ii)For any given constant $0 < \beta < 1$, if the signal strength λ satisfies $\lambda \geq 8\sqrt{C}((\log \frac{2}{\alpha})^{1/4} + (\log \frac{2}{1-\beta})^{1/4})n^{-\frac{3}{4}}$, the power of the test (3.3) is at least β. That is, $\forall \pi^* \in S_n, M^* \in M_n(\lambda)$

$$\mathbb{E}_{\pi^*,M*}\psi \geq \beta$$

Theorem 7 shows that if $\lambda \geq O(n^{-\frac{3}{4}})$, the testing procedure (3.3) can solve the testing problem (3.1) properly. It can control both the Type I and Type II error, which shows it has good performance under the null and alternative.

One interesting question to ask is that what happens in the case when λ is smaller than $O(n^{-\frac{3}{4}})$. We will show in the following that the rate $n^{-\frac{3}{4}}$ is optimal: if $\lambda < cn^{-\frac{3}{4}}$ for some constant c, to solve the testing problem (3.1) is impossible.

3.2.2 Lower Bound

Now we turn to the establishment of lower bound for λ. Theorem 7 shows that the testing procedure (3.3) works well under both null and alternative cases if λ is at least of order $n^{-\frac{4}{3}}$. In the following theorem, we will see that the rate $n^{-\frac{4}{3}}$ is indeed optimal.

Theorem 8. *For given $0 < \alpha < 1$ and any $\beta \in (\alpha, 1)$, there exists a constant $c = c(\alpha, \beta)$, such that if the signal strength λ satisfies*

$$\lambda \leq cn^{-\frac{3}{4}}$$

for any level-α test ψ to the testing problem (3.1), we have

$$\inf_{\pi^* \in S_n, M^* \in M_n(\lambda)} \mathbb{E}_{\pi^*, M^*} \psi \leq \beta$$

To construct of the lower bound, we try to construct a mixture in the alternative, so that the mixture is close enough to the null. As we can see, for all possible matrices in $M_n(\lambda)$, the following matrix $M_n^*(\lambda)$ is the closest one to the null probability matrix M_0, where $M_n^*(\lambda)$ is defined as

$$[M_n^*(\lambda)]_{ij} = \begin{cases} \frac{1}{2} + \lambda, \text{if } i > j, \\ \frac{1}{2} - \lambda, \text{if } i < j, \\ \frac{1}{2}, \text{if } i = j. \end{cases} \tag{3.4}$$

We consider the mixture probability in the case when the probability matrix to be $M_n^*(\lambda)$ and the permutation to be a random permutation chosen from all permutation in $\mathbb{S}_n$. We can prove that with the above construction, the mixture of the probability distribution cannot be distinguished from the null hypothesis. So the hypothesis testing problem is impossible to solve. We leave the details of the proof fin Section 3.5.

Theorem 8 shows that for $\lambda \lesssim n^{-\frac{3}{4}}$, the hypothesis testing problem (3.1) is not solvable. Together with the upper bound in Theorem 7, it characterizes the separation boundary between the testable and non-testable regions for λ. This separation boundary can then be used as a minimax benchmark for the evaluation of the performance of a test in this asymptotic regime.

3.3 Statistical Inference for Permutation

In the previous section, we discuss the statistical inference problem for the probability matrix. In this section, we study the inference problem for the permutation. We will consider different kinds of inference problems of permutations, including hypothesis testing and confidence set construction problems, in the current section.

We begin the section with introducing generalized permutation based model. In the previous section, we focus more in the pairwise comparison problem, which is one of the specific case for permutation based problems. In the current section, to gain more insight for the permutation related data, we study the generalized permutation based model.

Consider a crowdsourcing system that consists of d workers and n questions. We assume every question has two possible answers, denoted by $\{-1, 1\}$, of which exactly one is correct. Without loss of generality, we assume that the correct answer to the question is always 1. We model the question-answering problem via an unknown matrix $M^* \in [0,1]^{n \times d}$ whose (i, j)-th entry, M^*_{ij}, represents the probability that worker i answers question j correctly. Otherwise, with probability $1 - M^*_{ij}$, worker i gives the incorrect answer to question j.

We denote the response of worker i to question j by a variable $Y_{ij} \in \{-1, 1\}$, where we set Y_{ij} to the answer (-1 or 1) provided by the worker. We also make the standard assumption that given the values M, the entries of Y are all mutually independent. In summary, we observe a matrix Y which has independent entries distributed as

$$Y = \begin{cases} 1, \text{with probability } M_{ij} \\ -1, \text{with probability } 1 - M_{ij} \end{cases} \tag{3.5}$$

Several assumptions to the structure of the model is necessary. Otherwise, our observation becomes independent Bernouli random variables, which will not provide us much information to the true probability matrix M. For permutation based model, we usually assume that there exists a specific order for the abilities of the workers $\pi : [d] \to [d]$, such that if $\pi(i) < \pi(j)$, $\forall k \in [n]$, we have

$$M_{i,k} \geq M_{j,k} \tag{3.6}$$

The model we construct above is called the permutation based model for the crowdsourcing problem. In this section, we are going to consider inference problems within this framework. For the permutation based model, we introduce the following parameter space $\mathbb{C}_{perm}(n, d)$.

First we introduce the parameter space $\mathbb{C}_{perm}(n, d, \pi)$. Let $\mathbb{C}_{perm}(n, d, \pi)$ defined by

$$\mathbb{C}_{perm}(n,d,\pi) = \{M \in [0,1]^{n\times d} : M_{i,\pi^{-1}(1)} \geq M_{i,\pi^{-1}(2)} \geq \cdots M_{i,\pi^{-1}(d)}, \forall i,$$
$$\min_{1\leq k\leq d-1} \sum_{i=1}^{n} (M_{i,\pi^{-1}(k)} - M_{i,\pi^{-1}(k+1)})^2 > 0\}$$

We define the parameter space $\mathbb{C}_{perm}(n,d)$ as the union of all $\mathbb{C}_{perm}(n,d,\pi)$ for all permutations π on $[d]$

$$\mathbb{C}_{perm}(n,d) = \cup_{\pi\in\mathbb{S}_d}\mathbb{C}_{perm}(n,d,\pi)$$

The parameter space $\mathbb{C}_{perm}(n,d)$ is important in the permutation based model, as it characterizes the basic property in the model. We will study the problem in $\mathbb{C}_{perm}(n,d)$ and some other parameter spaces which are closely related to $\mathbb{C}_{perm}(n,d)$. In the current section, we focus on the case when n and d has similar magnitude. We assume that $\max\{\log n, \log d\} < \min\{\sqrt{n}, \sqrt{d}\}$ throughout the section.

3.3.1 Confidence Set Construction for Permutation

We begin with the discussion of confidence set construction problem in the permutation based model. We consider the following question: how to construct a confidence set for the best worker? More specifically, we want to construct a set $\hat{S}$ in $[d]$ with our observation, such that with high probability, $\pi^{-1}(1) \in \hat{S}$.

One important issue with confidence set construction problem is that how to define the smallest set with certain coverage probability. One natural way to define a set to be small is that the expect size of the confidence set is small. We try to solve the problem $\inf_{\hat{S}\in CS(\alpha,\mathbb{C}_{perm}(n,d))} \sup_{M\in\mathbb{C}_{perm}(n,d)} \mathbb{E}_M|\hat{S}|$, where $CS(\alpha,\mathbb{C}_{perm}(n,d))$ is the set of all confidence sets with coverage probability at least $1-\alpha$.

Unfortunately, the above way to find the best confidence set is not good. This is because the simple way does not capture the characteristic of confidence set construction problem. If we consider the problem with the above rule to determine the optimal confidence set construction procedure, it will lead to the result

$$\inf_{\hat{S}\in CS(\alpha,\mathbb{C}_{perm}(n,d))} \sup_{M\in\mathbb{C}_{perm}(n,d)} \mathbb{E}_M|\hat{S}| = O(d) \tag{3.7}$$

This result is trivial, as any procedure can reach the upper bound. It suggests us that we should consider the problem in a different way. Instead, we are going to introduce different criteirons for the confidence set construction. Though the criterion is not as natural as the size of the confidence set, it does provide evidence to show us that it can help us to understand the intrinsic difficulty of the problem.

The discussion for confidence set construction problem divides into three parts. We begin by discussion the confidence set construction problem in $\mathbb{C}_{perm}(n, d)$, where we do not put any assumptions in the parameter space. Then we will consider the same problem but with unimodal assumption in Section 3.3.1. We close the discussion with the discussion of the problem with s-modal assumption. We will see that the s-modal assumption describes the intrinsic difficulty of the confidence set construction problem for permutation based model.

Confidence Set Construction in $\mathbb{C}_{perm}$

Assume that $M \in \mathbb{C}_{perm}(n, d)$ and we observe $Y \sim \text{Ber}(M)$, i.e. we have independent Bernouli random variables $Y_{ij} \sim \text{Ber}(M_{ij})$. We are going to construct the confidence set $\hat{S}$, such that

$$\mathbb{P}_M(\pi^{-1}(1) \in \hat{S}(Y)) \geq 1 - \alpha$$

for any $M \in \mathbb{C}_{perm}(n, d)$. We define the set $CS(\mathbb{C}_{perm}(n, d), \alpha)$ to be

$$CS(\mathbb{C}_{perm}(n,d),\alpha) = \{\hat{S} : \hat{S} = \hat{S}(Y), \forall M \in \mathbb{C}_{perm}(n,d), \mathbb{P}_M(\pi^{-1}(1) \in \hat{S}) \geq 1-\alpha)\} \tag{3.8}$$

$CS(\mathbb{C}_{perm}(n, d), \alpha)$ is the set of all confidence set construction procedure which has the coverage probability at least $1 - \alpha$ for all matrix in $\mathbb{C}_{perm}(n, d)$.

Before going deep into the confidence set construction problem, it is important to find a suitable criterion for the choice of good confidence set construction procedure. Our goal of confidence set construction procedure is to find a small confidence set. In this sense, the natural way to find the best confidence set construction procedure $\hat{S}$ which minimizes $\sup_{M\in\mathbb{C}_{perm}(n,d)} \mathbb{E}_M|\hat{S}|$.

However, we find that

$$\sup_{M\in\mathbb{C}_{perm}(n,d)} \mathbb{E}_M|\hat{S}| = O(d) \tag{3.9}$$

$\forall\hat{S} \in \mathrm{CI}(\alpha, \mathbb{C}_{perm}(n,d))$. So it is not reasonable to simply use the size of the confidence set as the criterion to choose the optimal confidence set construction procedure.

To illustrate the criterion of choosing the optimal procedure in confidence set construction, we introduce the following definition of optimal radius, which help us to describe the difficulty in the construction of the confidence set in the current problem.

For $M \in \mathbb{C}_{perm}(n,d)$, assume that $\pi(M)$ is the corresponding permutation to the columns of the matrix. Let $\gamma(M,r)$ to be defined as

$$\gamma(M,r) = |\{j \in [d] : \sum_{i=1}^{n}(M_{i,\pi^{-1}(1)} - M_{i,j})^2 \leq r\}| \tag{3.10}$$

We would use $\gamma(M,r)$ as a criterion to choose the best confidence set construction procedure. If a confidence set construction procedure is good, the expected size of the confidence set should be upper bounded by $\gamma(M,r)$ for some specific r. The confidence set construction procedure is good if the radius r we need to upper bound the expected size is small. The major difference between the parameter space we use with simply considering the expected size is the radius parameter r. This parameter shows the difficulty of the construction of confidence set. It can measure the size of the confidence set and the size of the confidence set relate to the same radius may be different, according to the difficulty of the problem.

Now we introduce the confidence set construction procedure for the parameter space $\mathbb{C}_{perm}(n,d)$. Consider the following confidence set construction procedure. We construct the confidence set for the best worker as

$$\hat{S}_1 = \{j \in [d] : \max_{i \in [d]} \sum_{k \in [n]} Y_{ik} - Y_{jk} \leq \sqrt{n \log \frac{d}{\alpha}}\} \tag{3.11}$$

The following theorem shows that the above procedure $\hat{S}_1$ is optimal to the confidence set construction procedure according to our criterion.

Theorem 9. *(i)$\hat{S}_1 \in CS(\mathbb{C}_{perm}(n,d), \alpha)$. The expected size of $\hat{S}_1$ can be controlled by*

$$\mathbb{E}_M|\hat{S}_1| \leq 1 + \gamma(M, 2\sqrt{n \log d}), \tag{3.12}$$

$\forall M \in \mathbb{C}_{perm}(n,d)$.

(ii)For any confidence set construction procedure $\hat{S}$, such that

$$\mathbb{P}_M(\pi^{-1}(1) \in \hat{S}(Y)) \geq 1 - \alpha$$

for any $M \in \mathbb{C}_{perm}(n,d)$. We can find $M \in \mathbb{C}_{perm}(n,d)$, such that

$$\gamma(M, \frac{\sqrt{n \log d}}{40}) = 1, \mathbb{E}_M|\hat{S}| \geq \frac{d}{2} \tag{3.13}$$

Theorem 9 shows that the confidence construction procedure $\hat{S}_1$ is optimal for the parameter space $\mathbb{C}_{perm}(n,d)$. It shows that for any confidence set construction procedure should have a expected size at least as large as the number of workers whose distance is no more than $O((n \log d)^{\frac{1}{4}})$ from the best worker. Here, the distance between different workers is defined as the l_2 distance of the vectors corresponding to the workers. Otherwise, it is impossible that the confidence set construction procedure would have a correct coverage probability in $\mathbb{C}_{perm}(n,d)$. In this sense, the confidence construction procedure $\hat{S}_1$ is optimal.

Confidence Set Construction with Unimodal Assumptions

Previously, we discuss the confidence set construction problem in $\mathbb{C}_{perm}(n,d)$. We also discussed about the unimodal assumption in the estimation of pairwise comparison problem before. One question raises after the study to the confidence set construction problem in

$\mathbb{C}_{perm}(n, d)$: whether the confidence set construction problem is different with unimodal assumption?

We consider the same confidence set construction problem with the unimodal assumption, to see whether it makes significant difference in the confidence set construction problem. In fact, we can see with the unimodal assumption, the size of the confidence set can be much smaller.

First we define the parameter space with unimodal assumption $\mathbb{C}^u_{perm}(n, d, \pi)$.

Let $\mathbb{C}^u_{perm}(n, d, \pi)$ defined by

$$\mathbb{C}^u_{perm}(n, d, \pi) = \{M \in \mathbb{C}_{perm}(n, d, \pi) : \forall j, j', \text{ such that} \pi(j) < \pi(j'), \{\omega^{j,j'}(M)\} \text{ is a unimodal sequence}\}$$

where $\{\omega^{j,j'}(M)\}$ is defined as

$$\omega^{j,j'}(M)(i) = M_{ij} - M_{ij'}, i \in [n].$$

For all permutations π on $[d]$

$$\mathbb{C}^u_{perm}(n, d) = \cup_{\pi \in \mathbb{S}_d} \mathbb{C}^u_{perm}(n, d, \pi)$$

The confidence set construction problem we consider here is identical with the problem we consider before. The only difference is that now we consider the parameter space $\mathbb{C}^u_{perm}(n, d)$ instead of $\mathbb{C}_{perm}(n, d)$. Similarly, we define the set $CS(\mathbb{C}^u_{perm}(n, d), \alpha)$ to be

$$CS(\mathbb{C}^u_{perm}(n, d), \alpha) = \{\hat{S} : \hat{S} = \hat{S}(Y), \forall M \in \mathbb{C}^u_{perm}(n, d), \mathbb{P}_M(\pi^{-1}(1) \in \hat{S}) \geq 1 - \alpha)\} \quad (3.14)$$

We introduce the construction procedure $\hat{S}_2$ as follows to be the confidence set construction procedure for the parameter space $\mathbb{C}^{\pi}_{perm}(n, d)$.

Define

$$T_{min}(j) = \min_{k \in [d], T \in \mathcal{A}} \frac{\sum_{i \in T} Y_{i,j} - Y_{i,k}}{\sqrt{|T|}}$$

where $\mathcal{A}$ is a subclass of subsets in $[n]$, defined as

$$\mathcal{A} = \{\{a, a+1, ..., b\} | 1 \leq a < b \leq n\}.$$

The confidence set construction procedure $\hat{S}_2$ is

$$\hat{S}_2 = \{j \in [d] : T_{min}(j) \geq -2\sqrt{2\log(\frac{2nd^2}{\alpha})}\} \tag{3.15}$$

We can show that the confidence set construction procedure $\hat{S}_2$ is nearly optimal.

Theorem 10. *(i)*$\hat{S}_2 \in CS(\mathbb{C}^u_{perm}(n,d), \alpha)$. *The expected size of* $\hat{S}_2$ *can be controlled by*

$$\mathbb{E}_M|\hat{S}_2| \leq 1 + \gamma(M, 9\log n \log \frac{nd^2}{\alpha}), \forall M \in \mathbb{C}^u_{perm}(n,d). \tag{3.16}$$

(ii)For any confidence set construction procedure $\hat{S}$*, such that*

$$\mathbb{P}_M(\pi^{-1}(1) \in \hat{S}(Y)) \geq 1 - \alpha \tag{3.17}$$

for any $\forall M \in \mathbb{C}^u_{perm}(n,d)$*. We can find* $M \in \mathbb{C}^u_{perm}(n,d)$*, such that*

$$\gamma(M, \frac{1}{25}\log(nd^2)) = 1, \mathbb{E}_M|\hat{S}| \geq \frac{d}{2}. \tag{3.18}$$

Theorem 10 shows that the confidence set construction procedure $\hat{S}_2$ is optimal up to a log factor for the parameter space $\mathbb{C}^u_{perm}(n,d)$. It shows that for the optimal confidence set, it should include all workers whose distance to the best worker is at most

$O(\sqrt{\log d + \log n})$. Otherwise, it cannot promise the correct coverage probability in the parameter space $\mathbb{C}^u_{perm}(n, d)$. Meanwhile, the expected size of $\hat{S}_2$ can be controlled by the number of workers whose distance is no more than $O(\sqrt{\log n(\log d + \log n)})$ to the best worker. It means that the confidence set construction procedure is nearly optimal, up to a log factor of the radius.

Comparing the results in Theorem 9 and Theorem 10, we can clearly see that the confidence set construction problem is much easier with the unimodal assumption, in the sense that the optimal size of the confidence set is smaller. However, in many cases we do not really know whether the unimodal assumption is true or not. It remains unknown when can we assume the unimodal assumption is true. We will continue the discussion with a more generalized assumption than the unimodal assumption.

Multimodal Assumption and Adaptivity

We have discussed the confidence set construction problem in two different parameter spaces. In the first part of Section 3.3.1, we discussed the problem when we do not input further assumption other than the monoticity and construct the optimal confidence set construction procedure for $\mathbb{C}_{perm}(n, d)$. Then we assume the unimodal difference assumption is true and study the problem for $\mathbb{C}^u_{perm}(n, d)$.

We are moving forward and consider a more general setting than the setting we considered in Section 3.3.1 and Section 3.3.1. We consider the following parameter space for the confidence set construction problem.

Let

$$\begin{aligned}&\mathbb{C}^m_{perm}(s, n, d, \pi) = \\ &\{M \in \mathbb{C}_{perm}(n, d, \pi) : \forall j, j', \text{ such that} \pi(j) < \pi(j'), \{\omega^{j,j'}(M)\} \text{ is a } s \text{ -modal sequence}\}\end{aligned} \tag{3.19}$$

where $\{\omega^{j,j'}(M)\}$ is defined as

$$\omega^{j,j'}(M)(i) = M_{ij} - M_{ij'}, i \in [n].$$

Here, the s-modal sequence is defined as follows.

Definition 3. *For a sequence $\{a_i\}_{i=1}^n$, we say $\{a_i\}_{i=1}^n$ is a s-modal sequence if and only if the there exists $1 = \alpha_0 < \alpha_1 < \cdots < \alpha_s = n$, such that $\forall i \in [s], \{a_j\}_{j=\alpha_i - 1}^{\alpha_i}$ is a unimodal sequence.*

We define the s-modals permutation class to be

$$\mathbb{C}_{perm}^m(s,n,d) = \cup_{\mathbb{S}_d} \mathbb{C}_{perm}^m(s,n,d,\pi) \tag{3.20}$$

It is not difficult to see that the parameter space $\mathbb{C}_{perm}(n,d)$ is a special case of $\mathbb{C}_{perm}^m(s,n,d)$. We consider the confidence set construction problem for the parameter space $\mathbb{C}_{perm}^m(s,n,d)$. The confidence set construction procedure we propose here is the procedure $\hat{S}_2$, which is the one we propose for the parameter space $\mathbb{C}_{perm}^u(n,d)$. The following theorem shows that the confidence set construction procedure $\hat{S}_2$ is nearly optimal to the confidence set construction problem in $\mathbb{C}_{perm}^m(s,n,d)$.

Theorem 11. *(i)$\hat{S}_2 \in CS(\mathbb{C}_{perm}^m(s,n,d),\alpha)$. The expected size of $\hat{S}_2$ can be controlled by*

$$\mathbb{E}_M|\hat{S}_2| \le 1 + \gamma(M, 9s\log n \log\frac{nd^2}{\alpha}), \forall M \in \mathbb{C}_{perm}^m(s,n,d). \tag{3.21}$$

(ii)For any confidence set construction procedure $\hat{S}$, such that

$$\mathbb{P}_M(\pi^{-1}(1) \in \hat{S}(Y)) \ge 1 - \alpha$$

for any $\forall M \in \mathbb{C}_{perm}^m(s,n,d)$. We can find $M \in \mathbb{C}_{perm}^m(s,n,d)$, such that

$$\gamma(M, \frac{s}{25}\log(\frac{nd\log\frac{4}{3}}{2s^2\log n})) = 1, \mathbb{E}_M|\hat{S}| \ge \frac{d}{2}.$$

if $s < c\sqrt{\frac{\min\{n,d\}^2}{\log n}}$ for a constant c.

Theorem 11 shows that the confidence set construction procedure $\hat{S}_2$ is not just optimal under the unimodal assumption, it is also optimal up to a log factor for the s-modal assumptions. This shows that the s-modal assumptions plays an important role for the confidence set construction problem. If the true matrix M is in the parameter space $\mathbb{C}_{perm}^{m}(s, n, d)$, no other confidence set construction procedure can do much better than the confidence set construction procedure $\hat{S}_2$. More importantly, within the confidence set construction procedure, we do not need any information for the crucial parameter s.

Theorem 11 tells us that the parameter s can describe the difficulty of the confidence set construction problem well. Compare with the result in Theorem 9, the above results gives much detailed characteristic of the problem than the parameter space $\mathbb{C}_{perm}(n, d)$. It unveils the reason for us to study the problem with multimodal assumption. For the unimodal assumption, we do not know when to assume it is true in practice. For multimodal assumption, as the confidence set construction procedure $\hat{S}_2$ is optimal and adaptive in the parameter space $\mathbb{C}_{perm}^{m}(s, n, d)$. The adaptivity makes the mulimodal assumption more useful to the confidence set construction problem in real applications.

3.3.2 Hypothesis Testing for Permutation

We discussed the confidence set construction problem for permutation based model in Section 3.3.1. In fact, similar result can be constructed for the hypothesis testing problem, because of the duality between the hypothesis testing problem and confidence set construction problem. In the current section, we consider the following hypothesis testing problem

$$H_0 : \pi(i_0) = 1 \text{ vs. } H_a : \pi(i_0) \neq 1, \|\mu_{\pi^{-1}(1)} - \mu_{i_0}\|_2 \geq \nu \tag{3.22}$$

We will focus in the case with unimodal assumption $M \in \mathbb{C}_{perm}^{u}(n, d)$.

Upper Bound in Hypothesis Testing

We begin with introducing the hypothesis testing procedure to the testing problem (3.22).

The test statistic we used in the hypothesis testing is

$$T_{min} = \min_{j \in [n], T \in \mathcal{A}} \frac{\sum_{k \in S} Y_{k,i_0} - Y_{k,j}}{\sqrt{|T|}} \tag{3.23}$$

where $\mathcal{A}$ is a subclass of subsets in $[n]$, defined as

$$\mathcal{A} = \{\{a, a+1, ..., b\} | 1 \le a < b \le n\}.$$

For any $\alpha \in (0,1)$, a level-α test based on T_{min} is given by

$$\psi = I(T_{min} \le -2\sqrt{2\log(\frac{2nd^2}{\alpha}}))\tag{3.24}$$

Here, $I(\cdot)$ is the indicator function. We reject the null hypothesis if and only if $T_{min} \le -2\sqrt{2\log(\frac{2nd^2}{\alpha}})$.

First, we consider the performance of the test under null hypothesis. We prove that the test (3.24) is a level-α test.

Theorem 12. *The testing procedure (3.24) is a level-α test to the hypothesis testing problem (3.22). We have that*

$$\sup_{M \in H_0} \mathbb{P}_M(\psi = 1) \le \alpha$$

.

Theorem 12 shows that if the null hypothesis is true, then with probability at least $1-\alpha$, we do no reject the null hypothesis. Then we turn to consider the alternative case. As we previously mentioned, the key factor to determine the difficulty of the hypothesis testing problem is the gap ν between the best worker to the worker i_0 over the alternative. If ν is large, the problem should be easy to solve, while if ν is small, the testing problem becomes impossible. The following result quantifies how large the gap ν should be to make the test (3.24) have good performance over the alternatives.

Theorem 13. *If $\nu \ge \sqrt{28(\log d + 1)\log\frac{2nd^2}{\alpha}}$, the Type II error of the test can be controlled by α, we have*

$$\sup_{M \in H_{a,\nu}} \mathbb{P}_M(\psi = 0) \le \alpha$$

Theorem 12 and Theorem 13 combined showed that the testing procedure (3.24) works both under the null and alternatives if $\nu \geq \sqrt{28(\log d + 1)\log \frac{2nd^2}{\alpha}}$.

Lower Bound

Previously, we discuss the hypothesis testing procedure for the problem. Now we turn to consider the lower bound part of the hypothesis testing problem. We establish a lower bound for the separation ν and prove that the hypothesis testing problem (3.22) is impossible to solve in the following scenario.

Theorem 14. *If $\nu \leq \sqrt{\frac{1}{25}(\log n + 2\log d)}$, for any test $\psi = \psi(Y)$ to the hypothesis testing problem we have*

$$\sup_{M \in H_0} \mathbb{P}_M(\psi = 1) + \sup_{M \in H_{a,\nu}} \mathbb{P}_M(\psi = 0) \geq \frac{1}{2}$$

The proof to Theorem 14 comes directly from the proof to Theorem 10. Theorem 14 shows that if $\nu \leq \sqrt{\frac{1}{25}(\log n + 2\log d)}$, no test performs well in the hypothesis testing problem (3.22). The lower and upper bounds together characterize the separation boundary of the hypothesis testing problem (3.22). Our testing method (3.24) can solve the problem when $\nu \geq \sqrt{28(\log d + 1)\log \frac{2nd^2}{\alpha}}$, while $\nu \leq \sqrt{\frac{1}{25}(\log n + 2\log d)}$ the problem is impossible to solve. A log gap exists between the lower and upper bound.

3.4 Discussion

In Mao et al. (2018), the authors discuss the estimation problem in the permutation based model. They focus on estimating a bivariate isotonic matrix with unknown permutations acting on its rows and columns. In the paper, the authors mainly discussed the problem in $\mathbb{C}_{perm}$. The assumption made in the paper about the matrix structure is that the matrix is monotonic along both directions. We can see the result is optimal in terms of max-row-norm approximation error but suboptimal in terms of Frobenius estimation error. But with suitable assumption on the matrix class, we can see the picture is different. If we assume the unimodal assumption is true, the corresponding estimation error is much smaller than the case without unimodal assumption. The estimation results are established in Chapter 2. The results for estimation in different parameter space are listed in Table 1. Here, we use the notation $\tilde{O}$ to denote the rate of the error up to $poly(\log)$ factors.

In the current chapter, we discuss the inference problem in the similar framework. We can

	Class $\mathbb{C}_{perm}$		Class $\mathbb{C}_{perm}^{u}$	
	Lower bounds	Efficient Algorithm	Lower bounds	Efficient Algorithm
Frobenius estimation error	$\Omega(n)$	$\tilde{O}(n^{5/4})$	$\Omega(n)$	$\tilde{O}(n)$
Max-row-norm approximation error	$\Omega(n^{1/4})$	$\tilde{O}(n^{1/4})$	$\Omega(1)$	$\tilde{O}(1)$

Table 1: Estimation Rates for Different $n \times n$ Matrices Classes of Permutation Based Model.

see that the unimodal assumption and s-modal assumption makes a significant difference for statistical inference problem in permutation-based model. The results of the inference problem shows that the estimation problem and the inference problem are highly correlated. If we compare the result of the optimal radius for the confidence set construction, we can see that it compromises with the max-row-norm approximation error in Table 1. It shows highly similarity between the result for optimal estimation with max-row-norm approximation error and the corresponding inference problem.

3.5 Proof

3.5.1 Proof to Theorem 7 and Theorem 8

First we prove Theorem 7.

Proof. We begin the proof with part (i). To prove that the testing procedure is a level-α test, we will prove that under the null,

$$\mathbb{P}(|W - n| \geq \sqrt{n \log \frac{2}{\alpha}}) \leq \alpha \tag{3.25}$$

To prove it, we should first get the expectation of W under the null, which is

$$\mathbb{E}W = \frac{4}{n} \sum_{i=1}^{n} \mathbb{E}(S_i - \frac{n}{2})^2 = n \tag{3.26}$$

Then we have that

$$\begin{aligned} \|A_i\|_{\psi_1} &= \frac{4}{n} \|(S_i - \frac{n}{2})^2\|_{\psi_1} \\ &\leq \frac{4}{n} \|S_i - \frac{n}{2}\|_{\psi_2}^2 = 1 \end{aligned} \tag{3.27}$$

where $A_i = \frac{4}{n}(S_i - \frac{n}{2})^2$. With Berstein's Inequality, we have that there exists a constant C', such that

$$\begin{aligned}
&\mathbb{P}(|\sum_{i=1}^{n} A_i - n| \geq t) \\
\leq& 2\exp(-C' \min(\frac{t^2}{n}, t)) \\
=& 2\exp(-C' t_\alpha^2) = \alpha
\end{aligned} \tag{3.28}$$

where $t = \sqrt{n} t_\alpha$ and $t_\alpha = \sqrt{\frac{1}{C'} \log \frac{2}{\alpha}}$. So part (i) is proved with the selection $C = \sqrt{\frac{1}{C'}}$.

Then we turn to the part (ii) of Theorem 7, which considers the performance of the testing procedure under the alternative. To prove the performance under the alternative, we have that for $1 \leq i \leq \frac{n}{4}$, we have $\mathbb{E}S_i \geq \frac{n}{2} + \frac{n\lambda}{2}$, which implies that

$$\begin{aligned}
\mathbb{E}W &= \frac{4}{n} \sum_{i=1}^{n} \mathbb{E}(S_i - \frac{n}{2})^2 \\
&\geq \frac{4}{n} \frac{n}{4} (\frac{n\lambda}{2})^2 = \frac{n^2 \lambda^2}{4}
\end{aligned} \tag{3.29}$$

With the Berstein Inequality and the assumption on λ, we have that

$$\mathbb{P}(|W - n| < C\sqrt{n \log \frac{2}{\alpha}}) \leq \beta \tag{3.30}$$

which completes the proof to part (ii). □

Now we trun to the proof of Theorem 8.

Proof. Let $\mathbb{P}_0$ to be the probability distribution related to the probability matrix $\mathbb{M}_0$ and the permutation $\pi = id$.

Let $\mathbb{P}_\pi$ to be the probability distribution related to the probability matrix M_λ^* and the permutation π. We define that

$$\mathbb{P}_a = \frac{1}{n!} \sum_{\pi \in \mathbb{S}_n} \mathbb{P}_\pi \tag{3.31}$$

We will prove that it is impossible to distinguish $\mathbb{P}_0$ and $\mathbb{P}_a$. We are going to calculate the

χ^2 distance between $\mathbb{P}_0$ and $\mathbb{P}_a$.

$$\begin{aligned}
\chi^2(\mathbb{P}_0, \mathbb{P}_a) &= \int \frac{p_a^2}{p_0} - 1 \\
&= \mathbb{E}_{\pi_1,\pi_2} \int \frac{p_{\pi_1} p_{\pi_2}}{p_0} - 1 \\
&= \mathbb{E}_{\pi_1,\pi_2} (1+4\lambda^2)^{\binom{n}{2} - d_{KT}(\pi_1,\pi_2)} (1-4\lambda^2)^{d_{KT}(\pi_1,\pi_2)} - 1 \\
&\le \mathbb{E}_{\pi_1,\pi_2} (1+4\lambda^2)^{\binom{n}{2} - d_{KT}(\pi_1,\pi_2)} - 1 \\
&\le \mathbb{E} \exp(4\lambda^2 X)
\end{aligned} \tag{3.32}$$

where X is a random variable, such that $X = \binom{n}{2} - 2d_{KT}(\pi_1, \pi_2)$, $\pi_1, \pi_2 \sim U(\mathbb{S}_n)$.

Let $A_{ij}, 1 \le i, j \le n$ is defined as follows

$$A_{i,j} = \begin{cases} 1, \text{if } (i,j) \text{has the same order in } \pi_1 \text{ and } \pi_2 \\ -1, \text{otherwise} \end{cases} \tag{3.33}$$

From the definition of $A_{ij}, 1 \le i, j \le n$, we have that $X = \sum_{i,j} A_{i,j}$. Define $B_t, 1 \le t \le n-1$ to be

$$B_t = \sum_{i-j \equiv t \pmod n} A_{i,j} \tag{3.34}$$

So we have $\|B_i\|_{\psi_2} \le \sqrt{n}$ and $\|X\|_{\psi_2} = \| \sum_{t=1}^{n-1} B_t \|_{\psi_2} \le n\sqrt{n}$.

Combine with (3.32), we have $\mathbb{E} \exp(4\lambda^2 X) \le \exp(16\lambda^4 n^3)$ and $TV(\mathbb{P}_0, \mathbb{P}_a) \le \sqrt{\exp(16\lambda^4 n^3) - 1}$, which completes the proof with the choice $c = \frac{(\log(1+(\beta-\alpha)^2))^{1/4}}{2}$. □

3.5.2 Proof to Theorem 9

Proof. First we prove part (i) in Theorem 9.

Let M to be the true probability matrix. From our assumption, we know that $M \in \mathbb{C}_{perm}(n, d)$. Let π to be the corresponding permutation to the matrix, such that

$$M_{i,\pi^{-1}(1)} \ge M_{i,\pi^{-1}(2)} \ge \cdots M_{i,\pi^{-1}(d)} \tag{3.35}$$

Without loss of generality, we can assume that $\pi^{-1}(1) = 1$.

We prove that with probability at least $1 - \alpha$, $\pi^{-1}(1) \in \hat{S}_1$.

This is because

$$\begin{aligned}
&\mathbb{P}_M(i \notin \hat{S}_1) \\
\leq &\sum_{j=2}^{d} \mathbb{P}_M(\sum_{i=1}^{n} Y_{ij} - Y_{i1} \leq \sqrt{n \log \frac{d}{\alpha}}) \\
\leq &\sum_{j=2}^{d} \mathbb{P}_M((\sum_{i=1}^{n} Y_{ij} - Y_{i1}) - (\sum_{i=1}^{n} M_{ij} - M_{i1}) \leq \sqrt{n \log \frac{d}{\alpha}}) \\
< &d \exp(-\log \frac{d}{\alpha}) = \alpha
\end{aligned} \tag{3.36}$$

which means that

$$\mathbb{P}_M(1 \in \hat{S}_1) \geq 1 - \alpha \tag{3.37}$$

Then we turn to the upper bound construction for the size of the confidence. For any given $j \in [d]$, we have $M_{i,j} \leq M_{i,1}$ from our assumption. Using Hoeffding's inequality, we have

$$\mathbb{P}_M(|(\sum_{i=1}^{n} Y_{ik} - Y_{ij}) - (\sum_{i=1}^{n} M_{ik} - M_{ij})| \geq t) \leq 2 \exp(-\frac{t^2}{n}) \tag{3.38}$$

$\forall t > 0, k \in [d]$.

We are going to prove that if $\sum_{i=1}^{n} M_{i1} - M_{ij} \geq 2\sqrt{n \log d}$, then $\mathbb{P}_M(j \in \hat{S}_1) \leq \frac{1}{d}$.

This is because

$$\begin{aligned}
&\mathbb{P}_M(j \in \hat{S}_1) \\
\leq &\mathbb{P}_M(\sum_{i=1}^{n} Y_{i1} - Y_{ij} \leq \sqrt{n \log d}) \\
= &\mathbb{P}_M((\sum_{i=1}^{n} Y_{i1} - Y_{ij}) - (\sum_{i=1}^{n} M_{i1} - M_{ij}) \leq \sqrt{n \log d} - (\sum_{i=1}^{n} M_{i1} - M_{ij})) \\
\leq &\mathbb{P}_M(|(\sum_{i=1}^{n} Y_{i1} - Y_{ij}) - (\sum_{i=1}^{n} M_{i1} - M_{ij})| \geq \sqrt{n \log d}) \\
\leq &\exp(-\log \frac{d}{\alpha}) < \frac{1}{d}
\end{aligned} \tag{3.39}$$

So for all $j \in [d]$, which satisfies $\sum_{i=1}^n M_{i1} - M_{ij} \geq 2\sqrt{n \log d}$, we have $\mathbb{P}_M(j \in \hat{S}_1) \leq \frac{1}{d}$.

It yields that if $\sum_{i=1}^n (M_{i1} - M_{ij})^2 \geq 2\sqrt{n \log d}$, we have $\sum_{i=1}^n M_{i1} - M_{ij} \geq \sum_{i=1}^n (M_{i1} - M_{ij})^2 \geq 2\sqrt{n \log d}$. Therefore, $\forall j \notin \{j \in [d] : \sum_{i=1}^n (M_{i,\pi^{-1}(1)} - M_{i,j})^2 \leq 2\sqrt{n \log d}\}$.

The upper bound for the expected size of the confidence set $\hat{S}_1$ can be constructed as follows

$$\begin{aligned}
\mathbb{E}_M|\hat{S}_1| &= \sum_{j \in [d]} \mathbb{P}_M(j \in \hat{S}_1) \\
&= \sum_{j \in \mathcal{N}_M(2\sqrt{n \log d})} \mathbb{P}_M(j \in \hat{S}_1) + \sum_{j \notin \mathcal{N}_M(2\sqrt{n \log d})} \mathbb{P}_M(j \in \hat{S}_1) \qquad (3.40) \\
&\leq 1 + \gamma(M, 2\sqrt{n \log d})
\end{aligned}$$

which completes the proof to part (i). Then we turn to the proof to part (ii), which is the lower bound part to the theorem.

To prove part (ii), we consider the following hypothesis testing problem

$$H_0 : M = M_0 \text{ vs. } H_a : M \sim \mathbb{P}_a$$

where $M_0 \in \mathbb{R}^{n \times d}$ is a matrix, such that $M_0(i,j) = \frac{1}{2}$, while the probability distribution $\mathbb{P}_a$ is a defined on the parameter space $\mathbb{C}_{perm}(n,d)$ as follows.

Let $W = \{T \subset n : |T| = l\}$ to be the set of all subsets in $[n]$ with cardinality $\lfloor l \rfloor$. Thus, $|W| = \binom{n}{l}$. We will choose l in the end of the proof. For $j \in [d]$ and $T \in W$, we define the matrix $M_{j,T}$ as

$$M_{j,T} = M_0 + \delta N_{j,T} \qquad (3.41)$$

where

$$N_{j,T}(i,k) = \begin{cases} 1, \text{if } i = j, k \in T \\ 0, \text{otherwise} \end{cases} \tag{3.42}$$

and δ is a fixed constant which we are going to choose later.

Then the probability distribution $\mathbb{P}_a$ is defined as

$$M \sim \mathbb{P}_a \iff M = M_{j,T}, j \overset{\text{unif}}{\sim} [d], T \overset{\text{unif}}{\sim} W$$

Let p_0 to be the corresponding probability density function for the observation Y when $M = M_0$. For $j \in [d]$ and $T \in W$, let $p_{j,T}$ to be the corresponding probability density function. Let p_a to be the corresponding probability density function when $M \sim \mathbb{P}_a$. From our construction, we have

$$p_a = \frac{1}{d\binom{n}{l}} \sum_{j\in[d]} \sum_{T\in W} p_{j,T} \tag{3.43}$$

We are going to calculate the χ^2 distance between the null hypothesis and alternative hypothesis $\int \frac{p_a^2}{p_0} - 1$. We have that

$$\begin{aligned} &\int \frac{p_a^2}{p_0} - 1 \\ =&\mathbb{E}_{j_1,j_2 \overset{\text{unif}}{\sim} [d], T_1,T_2 \overset{\text{unif}}{\sim} W} \int \frac{p_{j_1,T_1} p_{j_2,T_2}}{p_0} - 1 \end{aligned} \tag{3.44}$$

We can see that if $j_1 \neq j_2$, $\int \frac{p_{j_1,T_1} p_{j_2,T_2}}{p_0} - 1 = 0$. It implies that

$$\begin{aligned} &\mathbb{E}_{j_1,j_2 \overset{\text{unif}}{\sim} [d], T_1,T_2 \overset{\text{unif}}{\sim} W} \int \frac{p_{j_1,T_1} p_{j_2,T_2}}{p_0} - 1 \\ =&\frac{1}{d}(\mathbb{E}_{j \overset{\text{unif}}{\sim} [d], T_1,T_2 \overset{\text{unif}}{\sim} W} \int \frac{p_{j,T_1} p_{j,T_2}}{p_0} - 1) \\ \leq&\frac{1}{d}(\mathbb{E}_{T_1,T_2 \overset{\text{unif}}{\sim} W} (1+5\delta^2)^{|T_1 \cap T_2|} - 1) \\ =&\frac{1}{d}(\mathbb{E}_{h\sim \text{Hypergeometric}(n,l,l)} (1+5\delta^2)^h - 1) \end{aligned} \tag{3.45}$$

To finish the proof, we need the following lemma.

Lemma 11. *Let J to be hypergeometric(p,k,k), then*

$$\mathbb{E}exp(tJ) \le \exp(\frac{k^2}{p-k})(1-\frac{k}{p}+\frac{k}{p}\exp(t))^k$$

Lemma 11 is proved in (Cai et al., 2017, Lemma 3).

Using Lemma 11, we have for $h \sim$ Hypergeometric(n,l,l),

$$\mathbb{E}(1+5\delta^2)^h \le \exp(\frac{(2+5\delta^2)l^2}{n}) \tag{3.46}$$

So we have

$$TV(p_0,p_a) \le \sqrt{\chi^2(p_0,p_a)} \le \sqrt{\frac{1}{d}(\exp(\frac{(2+5\delta^2)l^2}{n})-1))} < \frac{1}{3} \tag{3.47}$$

all the above inequalities hold with the choice $\delta = \frac{1}{5}$ and $l = \sqrt{\frac{n\log d}{2+5\delta^2}}$.

This leads to

$$\mathbb{E}_{M\sim\mathbb{P}_a}|\hat{S}| \ge d - dTV(p_0,p_a) > \frac{d}{2} \tag{3.48}$$

So there exists at least one matrix satisfies the condition we constructed in the Theorem. □

3.5.3 Proof to Theorem 10

Proof. We begin the proof with the coverage probability.

We are going to prove that with probability at least $1-\alpha$, we have $\pi^{-1}(1) \in \hat{S}_2$. Without loss of generality, we assume that $\pi^{-1}(1) = 1$

Otherwise, if $1 \notin \hat{S}_2$, there exists $k \in [d], T \in \mathcal{A}$, such that $\sum_{i\in T} Y_{i,k} - Y_{i,1} \ge 2\sqrt{2\log(\frac{2nd^2}{\alpha})}$. It yields that

$$\mathbb{P}_M(1 \notin \hat{S}_2) \le \sum_{k=2}^{d}\sum_{T\in\mathcal{A}} \mathbb{P}_M(\sum_{i\in T} Y_{i,k} - Y_{i,1} \ge 2\sqrt{2\log(\frac{2nd^2}{\alpha})}) \tag{3.49}$$

With Hoeffding's inequality, we have

$$
\begin{aligned}
&\sum_{k=2}^{d}\sum_{T\in\mathcal{A}}\mathbb{P}_M(\sum_{i\in T}Y_{i,k}-Y_{i,1}\geq 2\sqrt{2\log(\frac{2nd^2}{\alpha})})\\
\leq&\sum_{k=2}^{d}\sum_{T\in\mathcal{A}}\mathbb{P}_M((\sum_{i\in T}Y_{i,k}-Y_{i,1})-(\sum_{i\in T}M_{i,k}-M_{i,1})\geq 2\sqrt{2\log(\frac{2nd^2}{\alpha})})\\
\leq&dn^2\exp(-2\sqrt{2\log(\frac{2nd^2}{\alpha})})\leq\alpha
\end{aligned}
\tag{3.50}
$$

Combining (3.49) and (3.50), we have that $\mathbb{P}_M(1\notin\hat{S}_2)\leq\alpha$, which completes the proof to the coverage probability of $\hat{S}_2$.

Then we turn to the proof to the size estimation (3.16). We will prove that if $\|\mu_1-\mu_j\|_2^2\geq u^2$, $\mathbb{P}_M(j\notin\hat{S}_2)\leq\frac{1}{d}$, where $u=3\sqrt{\log n\log\frac{nd^2}{\alpha}}$.

Define a list of events

$$
\mathcal{E}_{k,T}=\{|\sum_{i\in T}Y_{i,j}-M_{i,j}|>\sqrt{|T|}u\}
\tag{3.51}
$$

$\forall k\in[d]$ and $T\in\mathcal{A}$.

With Hoeffding's inequality, we have

$$
\mathbb{P}_M\leq 2\exp(-u^2)
\tag{3.52}
$$

Let $\mathcal{E}=\cup_{k\in[d],T\in\mathcal{A}}$. Then

$$
\mathbb{P}_M(\mathcal{E})\leq\sum_{k\in[d],T\in\mathcal{A}}\mathbb{P}_M(\mathcal{E}_{k,T})\leq n^2d\exp(-u^2)
\tag{3.53}
$$

We then prove that if $\mathcal{E}^c$ happens, $\forall j\in[d]$ satisfies that

$$
\sum_{i=1}^{n}(M_{i1}-M_{ij})^2\geq 9\log n\log\frac{nd^2}{\alpha}
\tag{3.54}
$$

we will have $j \in \hat{S}_2$. Otherwise, if $j \notin \hat{S}_2$, we have that

$$\sum_{i \in T} Y_{i1} - Y_{ij} \leq \sqrt{|T|}u \tag{3.55}$$

$\forall T \in \mathcal{A}$. Let $\{\omega_i\}_{i=1}^n$ to be the sequence defined by

$$\omega_i = M_{i1} - M_{ij} \tag{3.56}$$

From our assumption, we know that $\omega_i \geq 0$ and ω is a unimodal sequence. Assume that j_0 is the peak of the unimodal sequence, such that

$$\omega_1 \leq \cdots \leq \omega_{j_0}, \omega_{j_0} \geq \omega_{j_0+1} \geq \cdots \omega_n \tag{3.57}$$

(3.55) tells us that

$$\sum_{i \in T} \omega_i \leq \sqrt{|T|}u \tag{3.58}$$

Let $I_l = \{j_0 + 2^{l-1}, \cdots, j_0 + 2^l - 1\}$,where $l = 1, ..., \lceil \log_2(n - j_0) \rceil - 1$. So we have

$$\begin{aligned} &\sum_{i=j_0+1}^{n} \omega_i^2 \\ =&\sum_{l=1}^{\lceil \log_2(n-j_0) \rceil} \sum_{i \in I_l} \omega_i^2 \\ \leq& 1 + \sum_{l=2}^{\lceil \log_2(n-j_0) \rceil} \sum_{i \in I_l} |I_{l-1}| (\frac{\sum_{i \in I_{l-1}} \omega_i}{I_{l-1}})^2 \leq 1 + 4 \log n \log \frac{nd^2}{\alpha} \end{aligned} \tag{3.59}$$

Similarly, we also have that $\sum_{i=1}^{j_0-1} \omega_i^2 \leq 1 + 4 \log n \log \frac{nd^2}{\alpha}$.It yields that $\sum_{i=1}^n \omega_i^2$, which is a contradiction with$\sum_{i=1}^n \omega_i^2 \geq 9 \log n \log \frac{nd^2}{\alpha}$. So that

$$\mathbb{P}_M(j \in \hat{S}_2) \leq \mathcal{E} \leq \frac{1}{d} \tag{3.60}$$

which concludes the proof of (i).

To prove the lower bound part of Theorem , we consider the following probability distribution.

Let $\mathbb{P}_0$ to be the distribution of observation Y, such that $M = M_0$. We define the probability distribution $\mathbb{P}_a$ as follows.

Let $\eta = \log(nd^2)$. Define the class of subsets $W \subset 2^{[n]}$ as

$$W = \{\{(k-1)\eta + 1, ..., k\eta\} : 1 \leq k \leq \lfloor \frac{n}{\eta} \rfloor\}\} \tag{3.61}$$

Thus, $|W| = \lfloor \frac{n}{\eta} \rfloor$. For $j \in [d]$ and $T \in W$, we define the matrix $M_{j,T}$ as

$$M_{j,T} = M_0 + \delta N_{j,T} \tag{3.62}$$

where

$$N_{j,T}(i,k) = \begin{cases} 1, \text{if } i = j, k \in T \\ 0, \text{otherwise} \end{cases} \tag{3.63}$$

and δ is a fixed constant which we are going to choose later.

Then the probability distribution $\mathbb{P}_a$ is defined as

$$M \sim \mathbb{P}_a \iff M = M_{j,s}, j \overset{\text{unif}}{\sim} [d], s \overset{\text{unif}}{\sim} [\lfloor \frac{n}{\eta} \rfloor]$$

Let p_0 to be the corresponding probability density function for the observation Y of probability distribution $\mathbb{P}_0$. For $j \in [d]$ and $T \in W$, let $p_{j,T}$ to be the corresponding probability density function. Let p_a to be the corresponding probability density function when $M \sim \mathbb{P}_a$. From our construction, we have

$$p_a = \frac{1}{d\lfloor \frac{n}{\eta} \rfloor} \sum_{j \in [d]} \sum_{s \in \lfloor \frac{n}{\eta} \rfloor} p_{j,T} \tag{3.64}$$

We are going to calculate the χ^2 distance between the null hypothesis and alternative hypothesis $\int \frac{p_a^2}{p_0} - 1$.

$$\begin{aligned}\int \frac{p_a^2}{p_0} - 1 \leq & \mathbb{E}_{s_1,s_2,j_1,j_2} \int (\frac{p_{j_1,s_1} p_{j_2,s_2}}{p_0} - 1) \\ \leq & \frac{1}{d\lfloor \frac{n}{\eta} \rfloor}(1+5\delta^2)^\eta \leq \frac{\exp(5\eta\delta^2)}{d\lfloor \frac{n}{\eta} \rfloor}\end{aligned} \tag{3.65}$$

So if we choose $\delta = \frac{1}{5}$, then we have

$$TV(p_0, p_a) \leq \sqrt{\chi^2(p_0, p_a)} \leq \frac{1}{3} \tag{3.66}$$

which means that there exists M, such that

$$\mathbb{E}_M |\hat{S}| \geq d - dTV(p_0, p_a) > \frac{d}{2} \tag{3.67}$$

which concludes the theorem. □

3.5.4 Proof to Theorem 11

Proof. The proof to the coverage probability part is identity to the part in the proof to Theorem 10. We begin the proof with the size estimation part. We will prove that for $j \notin \gamma(2s\psi \log n)$, there exists $T_0 \in \mathcal{A}$, such that

$$\sum_{i \in T_0} M_{i1} - M_{ij} \geq 2\sqrt{|T_0|\psi} \tag{3.68}$$

where $\psi = \frac{9}{2} \log \frac{nd^2}{\alpha}$. Otherwise, let $\{\omega_i\}_{i=1}^n$ to be the sequence defined by

$$\omega_i = M_{i1} - M_{ij} \tag{3.69}$$

From our assumption, we know that $\omega_i \geq 0$. What's more, we know that there exists $1 = \alpha_0 < \alpha_1 < \cdots < \alpha_s = n$, such that $\forall i \in [s]$, $\{\omega_j\}_{j=\alpha_i - 1}^{\alpha_i}$ is a unimodal sequence.

Then we have

$$\begin{aligned}\sum_{i=1}^{n}(M_{i1}-M_{ij})^2 &= \sum_{i=1}^{n}\omega_i^2 \\ &\leq \sum_{k=1}^{s}\sum_{i=\alpha_{k-1}}^{\alpha_k}\omega_i^2 \\ &\leq 2s\psi\log n\end{aligned} \tag{3.70}$$

which is a contradiction with our assumption that $j \notin \gamma(2s\psi\log n)$ So we have

$$\begin{aligned}&\mathbb{P}_M(\sum_{i\in T_0}Y_{i1}-Y_{ij} < \sqrt{|T_0|\psi}) \\ =&\mathbb{P}_M((\sum_{i\in T_0}Y_{i1}-Y_{ij})-(\sum_{i\in T_0}M_{i1}-M_{ij})+(\sum_{i\in T_0}M_{i1}-M_{ij}) < \sqrt{|T_0|\psi}) \\ &\leq \mathbb{P}_M((\sum_{i\in T_0}Y_{i1}-Y_{ij})-(\sum_{i\in T_0}M_{i1}-M_{ij}) < -\sqrt{|T_0|\psi}) \\ &\leq \exp(-\psi) < \frac{1}{d}\end{aligned} \tag{3.71}$$

where the last inequality comes from the Hoeffding's inequality. This concludes that

$$\mathbb{E}_M|\hat{S}_2| \leq 1+\gamma(9s\log n\log\frac{nd^2}{\alpha}) \tag{3.72}$$

Now we turn to the proof of the lower bound part for Theorem 11.

Let $\mathbb{P}_0$ to be the distribution of observation Y, such that $M = M_0$. We define the probability distribution $\mathbb{P}_a$ as follows.

Let $\eta = \log$.Let $z = \lceil\frac{n}{\eta}\rceil$,$u = \lceil\frac{z}{s}\rceil$. Let V defined as

$$V = \{K \subset [z] : |K| = u, \text{forall } 1 \leq i \leq u, K \cap \{(i-1)s+1, ..., is\} = 1\} \tag{3.73}$$

Let W defined as

$$W = \{\cup_{v\in V}\{(v-1)\eta+1, ..., v\eta\}\} \tag{3.74}$$

For $j \in [d]$ and $T \in W$, we define the matrix $M_{j,T}$ as

$$M_{j,T} = M_0 + \delta N_{j,T} \tag{3.75}$$

where

$$N_{j,T}(i,k) = \begin{cases} 1, \text{if } i = j, k \in T \\ 0, \text{otherwise} \end{cases} \tag{3.76}$$

and δ is a fixed constant which we are going to choose later.

Then the probability distribution $\mathbb{P}_a$ is defined as

$$M \sim \mathbb{P}_a \iff M = M_{j,s}, j \overset{\text{unif}}{\sim} [d], T \overset{\text{unif}}{\sim} W$$

Let p_0 to be the corresponding probability density function for the observation Y. For $j \in [d]$ and $T \in W$, let $p_{j,T}$ to be the corresponding probability density function. Let p_a to be the corresponding probability density function when $M \sim \mathbb{P}_a$.

We are going to construct the upper bound for $TV(p_0, p_a)$. The probability distribution is defined on the space $\mathbb{R}^{n \times d}$. Let p_0^i to be the probability distribution on the rows space between $(i-1)s+1, ..., is$ for $1 \leq i \leq s+1$. Similarly, we define p_a^i for $(i-1)s+1, ..., is$.

Then we have

$$p_a = \frac{1}{d\lfloor \frac{n}{\eta} \rfloor} \sum_{j \in [d]} \sum_{s \in \lfloor \frac{n}{\eta} \rfloor} p_{j,T} \tag{3.77}$$

We are going to calculate the χ^2 distance between the null hypothesis and alternative hypothesis $\int \frac{p_a^2}{p_0}$.

$$
\begin{aligned}
\int \frac{p_a^2}{p_0} - 1 \; &= \int_{\prod_{i=1}^{s+1}} \frac{(p_a^i)^2}{p_0^i} - 1 \\
&= \prod_{i=1}^{s+1} (1 + (\int \frac{(p_a^i)^2}{p_0^i} - 1)) - 1 \\
&\leq \prod_{i=1}^{s+1} (1 + \frac{\eta}{dn} \exp(5\eta\delta^2)) - 1 \\
&\leq \exp(\frac{2\eta s^2}{dn} \exp(5\eta\delta^2)) - 1
\end{aligned} \tag{3.78}
$$

if we choose $\delta = \frac{1}{5}$ and $\eta = 5\log(\frac{nd\log\frac{4}{3}}{2s^2\log n})$.

Then we have

$$
TV(p_0, p_a) \leq \sqrt{\chi^2(p_0, p_a)} \leq \frac{1}{3} \tag{3.79}
$$

which means that there exists M, such that

$$
\mathbb{E}_M |\hat{S}| \geq d - dTV(p_0, p_a) > \frac{d}{2} \tag{3.80}
$$

This concludes the proof. □

3.5.5 Proof to Theorem 12

Proof. We prove that with probability less than α, we reject the null hypothesis when it is true.

This is because

$$
\begin{aligned}
\mathbb{P}_M(\psi = 1) =& \mathbb{P}_M(T_{min} \leq -2\sqrt{2\log(\frac{2nd^2}{\alpha})}) \\
\leq& \sum_k \sum_{T \in \mathcal{A}} \mathbb{P}(\sum_{j \in T} Y_{k,i_0} - Y_{k,j} \leq -2\sqrt{2|T|\log(\frac{2nd^2}{\alpha})})
\end{aligned} \tag{3.81}
$$

From the assumption that H_0 is true, we have $\pi^{-1}(i_0) = 1$. So we have $M_{k,i_0} \geq M_{k,j}$, which leads to the fact that

$$\begin{aligned}
&\mathbb{P}(\sum_{j\in T} Y_{k,i_0} - Y_{k,j} \leq -2\sqrt{2|T|\log(\frac{2nd^2}{\alpha})}) \\
\leq &\mathbb{P}(\sum_{j\in T} Y_{k,i_0} - Y_{k,j} - (M_{k,i_0} - M_{k,j}) \leq -2\sqrt{2|T|\log(\frac{2nd^2}{\alpha})})
\end{aligned} \tag{3.82}$$

With Hoefdding's inequality, we have

$$\mathbb{P}(\sum_{j\in T} Y_{k,i_0} - Y_{k,j} - (M_{k,i_0} - M_{k,j}) \leq -2\sqrt{2|T|\log(\frac{2nd^2}{\alpha})}) \leq \frac{\alpha}{nd^2} \tag{3.83}$$

Combining with (3.81), we get that

$$\mathbb{P}_M(\psi = 1) \leq \alpha \tag{3.84}$$

which concludes the proof. □

3.5.6 Proof to Theorem 13

Proof. For $M \in \mathbb{C}^u_{perm}(n, d) \cap H_{a,\nu}$, we have

$$\begin{aligned}
\mathbb{P}_M(\psi = 0) &= \mathbb{P}_M(T_{min} \geq -2\sqrt{2\log(\frac{2nd^2}{\alpha})}) \\
&\leq \mathbb{P}_M(\min_{T\in\mathcal{A}} \frac{\sum_{k\in T} Y_{k,i_0} - Y_{k,i_1}}{\sqrt{|T|}} \geq -2\sqrt{2\log(\frac{2nd^2}{\alpha})})
\end{aligned} \tag{3.85}$$

We prove that if $\nu \geq \sqrt{28(\log d + 1)\log \frac{2nd^2}{\alpha}}$, then there exists $T \in \mathcal{A}$, such that

$$\sum_{k\in T} M_{k,i_1} - M_{k,i_0} \geq \sqrt{|T|}u \tag{3.86}$$

where $i_1 = \pi^{-1}(1)$ and $u = 3\sqrt{2\log(\frac{nd^2}{\alpha})}$.

Otherwise, we have that

$$\sum_{k\in T} M_{k,i_0} - M_{k,i_1} \leq \sqrt{|T|}u \tag{3.87}$$

$\forall T \in \mathcal{A}$. Let $\{\omega_i\}_{i=1}^n$ to be the sequence defined by

$$\omega_i = M_{i1} - M_{ij} \tag{3.88}$$

From our assumption, we know that $\omega_i \geq 0$ and ω is a unimodal sequence. Assume that j_0 is the peak of the unimodal sequence, such that

$$\omega_1 \leq \cdots \leq \omega_{j_0}, \omega_{j_0} \geq \omega_{j_0+1} \geq \cdots \omega_n \tag{3.89}$$

(3.55) tells us that

$$\sum_{i \in T} \omega_i \leq \sqrt{|T|u} \tag{3.90}$$

Let $I_l = \{j_0 + 2^{l-1}, \cdots, j_0 + 2^l - 1\}$,where $l = 1, ..., \lceil \log_2(n - j_0) \rceil - 1$. So we have

$$\begin{aligned} &\sum_{i=j_0+1}^{n} \omega_i^2 \\ =&\sum_{l=1}^{\lceil \log_2(n-j_0) \rceil} \sum_{i \in I_l} \omega_i^2 \\ \leq&1 + \sum_{l=2}^{\lceil \log_2(n-j_0) \rceil} \sum_{i \in I_l} |I_{l-1}| (\frac{\sum_{i \in I_{l-1}} \omega_i}{I_{l-1}})^2 \leq u^2 \log n \end{aligned} \tag{3.91}$$

Similarly, we also have that $\sum_{i=1}^{j_0-1} \omega_i^2 \leq u^2 \log n$,.It yields that $\sum_{i=1}^{n} \omega_i^2 \leq 3u^2 \log n$. So we have

$$\sum_{k=1}^{n} (M_{k,j_1} - M_{k,j_0})^2 = \sum_{k=1}^{n} \omega_k^2 \leq 2 \log n u^2 + 1 \leq \nu^2 \tag{3.92}$$

which is a contradiction with our assumption. So that there exists $T_0 \in \mathcal{A}$, such that

$$\sum_{k \in T} M_{k,i_0} - M_{k,i_1} \geq \sqrt{|T|u} \tag{3.93}$$

With Hoeffding's inequality, we have that

$$\begin{aligned}\mathbb{P}_M(\psi = 0) &\leq \mathbb{P}_M(\min_{T\in\mathcal{A}} \frac{\sum_{k\in T} Y_{k,i_0} - Y_{k,i_1}}{\sqrt{|T|}} \geq -2\sqrt{2\log(\frac{2nd^2}{\alpha})}) \\ &\leq \mathbb{P}_M(\frac{\sum_{k\in T_0} Y_{k,i_0} - Y_{k,i_1}}{\sqrt{|T|}} \geq -2\sqrt{2\log(\frac{2nd^2}{\alpha})}) \leq \alpha\end{aligned} \tag{3.94}$$

which concludes the proof. □

4 Optimality of Local BP Algorithm in Stochastic Block Model

4.1 Introduction

Network analysis is one of the popular topics in recent research. People from different areas, including statistics (Gao et al. (2015); Zhang and Zhou (2015); Gao et al. (2016); Abbe and Sandon (2015); Cai and Li (2015)), computer science (Mossel and Xu (2016); Chen and Xu (2014)), physics (Decelle et al. (2011); Karrer and Newman (2011); Newman et al. (2002)), have done a lot of work to study network data analysis in the past years. In network literature, community detection problem in stochastic block model (SBM) is the most widely known and studied problem.

4.1.1 Stochastic Block Model

Holland et al. (1983) proposed SBM as a simple but useful network model. Since then, people spent lots of effort to understand the community detection problem in this model. In this chapter, we will consider the community detection problem of binary stochastic block model. In this model, let n to be the number of vertices. For each vertex, it will be in the first cluster with probability ρ and it will be in the second cluster with probability $1-\rho$. For each pair of vertices x and y,

$$\mathbb{P}(\text{there exists an edge from x to y}) = \begin{cases} a/n, \text{if x and y are both in the first cluster} \\ b/n, \text{if x and y are in different clusters} \\ c/n, \text{if x and y are both in the second cluster} \end{cases}$$

For different pairs of vertices, the connection between vertices are independent. We use $G=(V,E)$ to denote the model, where V is the set of the vertices and E is the set of the edges in the graph. We use σ to denote the labels in the graph, i.e. for every vertex i in the graph, $\sigma_i = 1$ if the vertex i is in the first cluster and $\sigma_i = 0$ if the vertex i is in the second cluster.

Most of the research in this area focus in the balanced and symmetric case, i.e. $\rho = \frac{1}{2}$ and $a = c$. Generally speaking, these results can be partitioned more carefully through different recovery goals.

- Correlated recovery. In this kind of recovery problem, we hope to find a recovery algorithm such that it performs better than random guess. Decelle et al. (2011) put forward the striking conjecture about the sharp threshold for the regime in which it is possible or not to get a correlated recovery, which is later proved by Mossel et al. (2012, 2013) and Massoulié (2014). They proved that the correlated recovery is possible if and only if $(a-b)^2 > 2(a+b)$.

- Weak recovery(weak consistency). In this kind of recovery problem, we hope to find a recovery algorithm, such that we can recover the network structure with at most $o(n)$ vertices to be misclassified. Mossel et al. (2015)and Yun and Proutiere (2014) proved that we can recover the cluster structure with at most $o(n)$ vertices to be misclassified if and only if $(a-b)^2/(a+b) \to \infty$. Gao et al. (2015) further established the minimax misclassification proportion rate and proposed a algorithm achieving this rate, which gave us better understanding to the recovering than just weak consistency.

- Strong recovery(strong consistency). In this kind of recovery problem, we hope to find a recovery algorithm, such that we can recover every single vertex in the network. Abbe et al. (2016) and Mossel et al. (2015) focus this kind of problem, which ensures every vertex can be clustered correctly . They establish a sharp exact recovery threshold for strong consistency set up.

If we only care about of the recovery proportion, we can use signal to noise ratio to describe the picture above. The signal to noise ratio for binary stochastic block model is defined as

$$\text{SNR} = \frac{(a-b)^2}{2(a+b)}$$

In the case SNR < 1, even to find a correlated recovery algorithm is impossible. In the case $SNR \to \infty$, we can find a weak recovery algorithm. The strong recovery problem cannot be described using SNR. Though lot of work had been done in studying the clustering problem in SBM, it remains unknown in the regime between correlated recovery and weak recovery. If SNR is a constant larger than 1, we can find an algorithm which is better than random guess, but even the best algorithm can only recover a proportion of the vertices.

What is the best algorithm in this problem? How many vertices we can recover successfully using the best algorithm? In this chapter, we mainly focus on this kind of problem, which is called partial recovery problem.

Throughout this chapter, we further assume

$$a = b + \sqrt{b}\mu, c = b + \sqrt{b}\nu, b \to \infty, b = n^{o(1)} \tag{4.1}$$

as $n \to \infty$, where μ and ν are both constants. Without loss of generality, we assume that $\mu \le \nu$.

Mossel and Xu (2015) considered the partial recovery problem in the stochastic block model. They provide a lower bound for the optimal expected misclassified fraction. They also proved that a local algorithm, local belief propagation(BP) algorithm, can almost reach the expected misclassified fraction. However, they only prove the optimality in the balanced case. They conjectured that as long as $\mu = \nu$, local BP algorithm can reach the optimal expected misclassified fraction as well.

The main contribution of this chapter are two folds. First, we prove the conjecture in Mossel and Xu (2015). We prove that local BP algorithm can reach the optimality not just in the balanced case. Furthermore,instead of simply proving this conjecture, we provide a much larger regime, where the local BP algorithm can reach the optimal expected misclassified fraction. We prove that as long as one of the following conditions holds, the local BP algorithm can reach the optimal expected misclassified fraction. The condition includes (i)$\rho < \frac{1}{4}$,$\epsilon \le 1$ (ii)$\frac{1}{2} < \rho \le \frac{3}{4}$,$\epsilon \le \frac{3-4\rho}{4\rho-1}$, where $\epsilon = \mu/\nu$. This result gives us better understanding to the recovering problem in unbalanced case.

Second, in the regime where local BP algorithm may not achieve the optimal misclassified fraction, we will prove that local BP algorithm can be used in correcting other algorithms. If we have a satisfactory initializer, we can see this initializer plus local BP correction will be an optimal algorithm. For the case we can prove the local BP algorithm is optimal, what we prove indeed is that random guess plus local BP correction can achieve the optimal expected misclassified fraction. For the case we cannot prove the local BP algorithm is optimal, the general reason we cannot prove the optimality is we need better initializer than random guess. We will quantify the condition we need for the initializer in Section 4.4.

4.1.2 Organization of the Chapter

The rest of the chapter is organized as follows. Section 4.2 introduces the idea of local BP algorithm. We study the optimality of local BP algorithm in Section 4.3. In Section 4.4, we focus on the regime where we cannot prove the optimality in Section 4.3. We will illustrate the condition we need for a initializer to make the optimal algorithm possible. A discussion on the results in the current chapter is included in Section 4.5.

4.2 Galton-Watson Tree and Local BP algorithm

In this section, we introduce the idea of local BP algorithm. Before introducing the algorithm itself, it is important for us to have a better understanding of the local structure of stochastic block model. The introduction of Galton-Watson tree will give us both a clear picture of the local structure of SBM and a direct intuition of local BP algorithm.

4.2.1 Galton-Watson Tree

Galton-Watson tree is one of the important tools we use in the study of stochastic block model. The Galton-Watson tree is defined as follows.

Definition 4. *For every vertex u, we denote by (T_u, τ) the Poisson two-type branching process tree rooted at u satisfied the following conditions. Let*

$$\tau_u = \begin{cases} 1, & \textit{with probability } \rho \\ 0, & \textit{with probability } 1-\rho \end{cases}$$

Now recursively for each i in T_u, let L_i to be the number of children of i, whose label is 1 and M_i to be the number of children of i, whose label is 0.

$$\textit{Given } \tau_i = 1, L_i \sim \textit{Poisson}(\rho a), M_i \sim \textit{Poisson}((1-\rho)b)$$

$$\textit{Given } \tau_i = 0, L_i \sim \textit{Poisson}(\rho b), M_i \sim \textit{Poisson}((1-\rho)c)$$

The importance of Galton-Watson tree in this problem comes from Lemma 12, which proves the existence of the coupling between Galton-Watson tree and binary stochastic block model. This lemma is proved in Mossel et al. (2012).

Lemma 12. *For $t = t(n)$ such that $b^t = n^{o(1)}$, there exists a coupling between (G, σ) and (T, τ) such that $(G^t, \sigma_{G^t}) = (T^t, \tau_{T^t})$ with probability converging to 1.*

This lemma shows the local structure of the binary SBM model. Locally, a binary stochastic block model can be seen as a Galton-Watson tree. This idea also inspires us to use local BP algorithm to solve the partial recovery problem in SBM.

4.2.2 Local Belief Propagation Algorithm

The idea of local BP algorithm is simple: it is a likelihood ratio clustering algorithm on Galton-Watson tree. Lemma 12 tells us the clustering problem on SBM is locally equivalent to the clustering problem on Galton-Watson tree. The following message passing method, which is one part of the local BP algorithm, can be used in calculating the likelihood ratio on Galton-Watson tree.

Let ∂i denote the set of neighbors of i. Let

$$F(x) = \frac{1}{2}\log(\frac{e^{2x}\rho a + (1-\rho)b}{e^{2x}\rho b + (1-\rho)c})$$

Let $d_1 = \rho a + (1-\rho)b$ and $d_0 = \rho b + (1-\rho)c$ denote the expected vertex degree in the first and second cluster, respectively. Define the message transmitted from vertex i to vertex j at t-th iteration as

$$R^t_{i\to j} = \frac{-d_1 + d_0}{2} + \sum_{l\in\partial i\setminus\{j\}} F(R^{t-1}_{l\to i}) \tag{4.2}$$

Then we define the belief of vertex u at t-th iteration R^t_u to be

$$R^t_u = \frac{-d_1 + d_0}{2} + \sum_{l\in\partial u} F(R^{t-1}_{l\to u}) \tag{4.3}$$

The local BP algorithm is an algorithm based on the above message passing method.

Input **:** $n \in N, \rho \in (0,1), a, b, c$,adjacency matrix $A \in \{0,1\}^{n\times n}$,$t \in N$

Initialization: Set $R^0_{i\to j} = 0$ for all $i \in [n]$ and $j \in \partial i$;

iteration =0;

while iteration $\le$ t **do**

  Run message passing as in (4.2) to compute $R^{t-1}_{i\to j}$ for all $i \in [n]$ and $j \in \partial i$

end

Compute R^t_i for all $i \in [n]$ by (4.3);

Return $\widehat{\sigma}^t_{BP}$ with $\widehat{\sigma}^t_{BP}(i) = 1_{\{R^t_i \ge \varphi\}}$,where $\varphi = \frac{1}{2}\log\frac{\rho}{1-\rho}$

Algorithm 3: Local BP algorithm

4.3 The Optimality of Local BP Algorithm

4.3.1 Expected Misclassified Fraction

Before study the result of the optimality of local BP algorithm, we should introduce the criterion we use to judge whether an algorithm is good or not. We use the expected misclassified fraction as such a criterion. We define the expected misclassified fraction as follows.

Definition 5. *For any label estimator $\widehat{\sigma}$ of the true label σ, we define the expected misclassified fraction of $\widehat{\sigma}$ as*

$$p_G(\widehat{\sigma}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\{\sigma_i \ne \widehat{\sigma}_i\} \tag{4.4}$$

Notice that in our setting, as the first cluster and the second cluster is different (either in the proportion of vertices or in the edge probability within the clusters), there is no need to define the expected misclassified fraction with permutation. Let p^*_G denote the minimum expected misclassified fraction among all possible estimators based on G. If an algorithm can reach the minimum expected misclassified fraction, i.e.

$$p_G(\widehat{\sigma}) = p^*_G \tag{4.5}$$

we say that this algorithm is optimal. If an algorithm can reach the minimum expected misclassified fraction asymptotically, i.e.

$$\lim_{n\to\infty} p_G(\widehat{\sigma}) = \lim_{n\to\infty} p^*_G \tag{4.6}$$

we say this algorithm is asymptotic optimal. Our goal is to find the algorithm which is asymptotic optimal.

4.3.2 The Optimality of Local BP Algorithm

The following theorem is proved in Mossel and Xu (2015), which provides a lower bound for the misclassified fraction in the partial recovery problem of SBM model.

Theorem 15. *Assume $\rho \in (0,1)$ is fixed and consider the regime (4.1). Let*

$$h(v) = \mathbb{E}[\tanh(v + \sqrt{v}h + \varphi)], \tag{4.7}$$

where $Z \sim \mathcal{N}(0,1)$ and $\varphi = \frac{1}{2}\log\frac{\rho}{1-\rho}$. Let $\lambda = \frac{\rho(\mu+\nu)^2}{8}$ and $\theta = \frac{\rho(\mu-\nu)^2}{8} + \frac{(1-2\rho)\nu^2}{4}$. Define $\underline{v}$ and $\bar{v}$ to be the smallest and largest fixed point of

$$v = \theta + \lambda h(v) \tag{4.8}$$

respectively. Define $(v_t : t \geq 0)$ recursively by $v_0 = 0$ and $v_{t+1} = \theta + \lambda h(v_t)$. Let $\widehat{\sigma}^t_{BP}$ denote the estimator given by local BP algorithm applied for t iterations. Then $\lim_{t\to\infty} v_t = \underline{v}$ and

$$\lim_{n\to\infty} p_G(\widehat{\sigma}^t_{BP}) = \rho Q(\frac{v_t + \varphi}{\sqrt{v_t}}) + (1-\rho)Q(\frac{v_t - \varphi}{\sqrt{v_t}})$$

$$\liminf_{n\to\infty} p^*_G \geq \rho Q(\frac{\bar{v} + \varphi}{\sqrt{\bar{v}}}) + (1-\rho)Q(\frac{\bar{v} - \varphi}{\sqrt{\bar{v}}})$$

Theorem 15 constructs a lower bound to the expected misclassified fraction. Unfortunately Mossel and Xu (2015) fails to prove that the local BP algorithm can reach the optimal expected misclassified fraction, except for a special case. They prove the optimality of local BP algorithm in the case when $\rho = \frac{1}{2}$. They also conjecture that in some other cases, the local BP algorithm reaches the optimality in the case when $\mu = \nu$, see (Mossel and Xu, 2015, Conjecture 2.4). We will prove that their conjecture is correct. What's more, the local BP algorithm does reach the optimality in a wide regime.

Theorem 16. *Let $\epsilon = \frac{\mu}{\nu}$. If $\rho\mu \neq (1-\rho)\nu$, and one of the following condition holds:*

(i)$\rho < \frac{1}{2}$,$\epsilon \leq 1$

(ii)$\frac{1}{2} \leq \rho \leq \frac{3}{4}$,$\epsilon \leq \frac{3-4\rho}{4\rho-1}$

Equation (4.8) has a unique fixed point in $[0, \infty)$.

Using Theorem 15 and Theorem 16, we can prove that the local BP algorithm is asymptotic optimal in this regime.

Corollary 1. *If* $\rho\mu \neq (1-\rho)\nu$, *and one of the following condition holds:*

(i)$\rho < \frac{1}{2}$,$\epsilon \leq 1$

(ii)$\frac{1}{2} \leq \rho \leq \frac{3}{4}$,$\epsilon \leq \frac{3-4\rho}{4\rho-1}$

then

$$\lim_{t\to\infty}\lim_{n\to\infty} p_G(\widehat{\sigma}^t_{BP}) = \lim_{n\to\infty} p^*_G \tag{4.9}$$

For the picture of the optimal regime for the local BP algorithm, we can see it in Figure 6. The optimal regime contained both the previous results and the conjecture in Mossel and Xu (2015). The impossible curve represents the case when the vertex degree is not correlated with network structure. It is proved in Kanade et al. (2014) that any local algorithm cannot reach the optimal misclassified fraction on this curve. From this plot, we can see if it is somewhere away from the impossible curve, the local BP algorithm can reach the optimality.

We end this section with the proof to Theorem 16. The idea to prove this theorem comes from the following two lemmas. Lemma 14 provides a good estimation to the function h, while the intuition of Lemma 13 comes from proving the Conjecture 2.6 in Mossel and Xu (2016).

Lemma 13. *Let* g *to be a function defined as*

$$g(v) = \mathbb{E}\tanh(v + \sqrt{v}Z + U)$$

where $Z \sim N(0,1)$, U *is a random variable independent of* Z, *satisfies* $U = \gamma$ *with probability* $1-\alpha$ *and* $U = -\gamma$ *with probability* α,$\gamma = \frac{1}{2}\log\frac{1-\alpha}{\alpha}$, *where* $0 < \alpha \leq \frac{1}{2}$. *Then we have*

$$g(v) \geq \frac{v}{v+1}$$

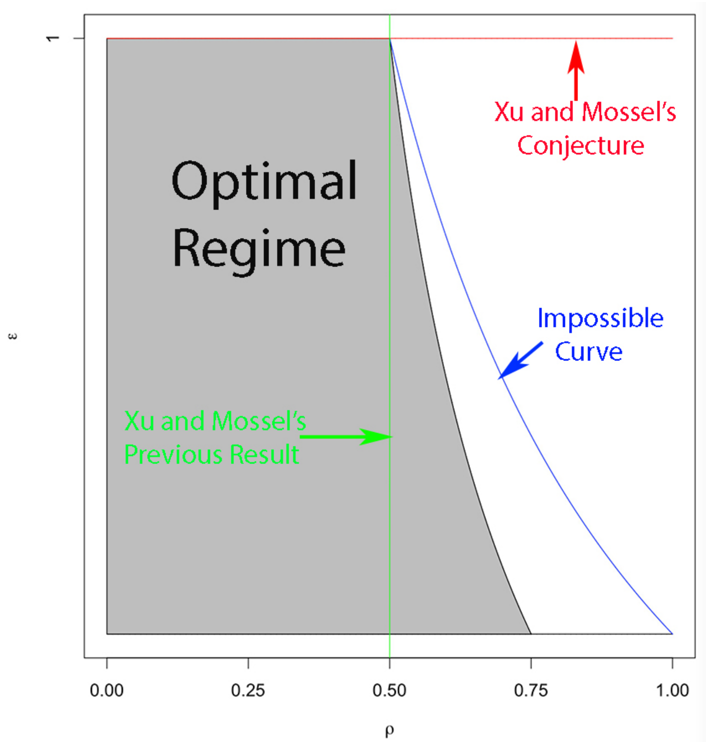

Figure 6: Grey area: the optimal regime we proved in Theorem 16. Green line: Mossel and Xu's previous result. Red line: Mossel and Xu's conjecture. Blue curve: local BP algorithm cannot be optimal on this curve, proved in Kanade et al. (2014).

Lemma 14. *Let* $\tau = 2\rho - 1$,*then we have*

$$h(v) \geq \tau + \frac{(\tau+1)(\tau-1)^2 v}{v + 1 - \tau^2 v} \tag{4.10}$$

The proof to these lemmas are in Section 4.6. We also point out that using Lemma 13, we could solve another conjecture in Mossel and Xu (2016). In Mossel and Xu (2016), the authors conjecture that for all $|\mu| \geq 2$ and $\alpha \in (0, \frac{1}{2})$, $v = \frac{\mu^2 g(v)}{4}$ has a unique fixed point.

Conjecture 1. *For all* $|\mu| > 2$,$\alpha \in (0, \frac{1}{2})$, $v = \frac{\mu^2}{4} g(v)$ *has a unique fixed point.*

We will prove this conjecture using Lemma 14.

Proof of Conjecture 1. Using Lemma 14, we have

$$(\frac{g(v)}{v})' = \frac{g'}{v} - \frac{g}{v^2} = \frac{g'v - g}{v^2} \leq \frac{v(1-g) - g}{v^2} < 0, \forall v > 0$$

So $\frac{g(v)}{v}$ is a strictly increasing function on $(0, \infty)$. As

$$\lim_{v \to 0^+} \frac{g(v)}{v} = \infty$$

and

$$\lim_{v \to \infty} \frac{g(v)}{v} = 0$$

we know that $v = \frac{\mu^2 g(v)}{4}$ has a unique fixed point. □

Now we can prove Theorem 16 with Lemma 13, Lemma 14 and Conjecture 1.

Proof of Theorem 16. Using Lemma 14, we have

$$h(v) \geq \tau + \frac{(\tau+1)(\tau-1)^2 v}{v + 1 - \tau^2 v} \tag{4.11}$$

Let $\tilde{\lambda} = \rho(\epsilon+1)^2$,$\tilde{\theta} = \rho(\epsilon-1)^2 + 2 - 4\rho$. So

$$v = \theta + \lambda h(v) \Longleftrightarrow v = \frac{\nu^2}{8}[\tilde{\theta} + \tilde{\rho} h(v)]$$

Because $\rho\mu \neq (1-\rho)\nu$, we can see that

$$\begin{aligned}
\tilde{\theta} + \tilde{\rho} h(0) &= \rho(\epsilon+1)^2 + (\rho(\epsilon-1)^2) + 2 - 4\rho)(2\rho - 1) \\
&= 2\rho^2\epsilon^2 + 4\rho(1-\rho)\epsilon + 2\rho^2 - 4\rho + 2 \\
&= 2(\rho\epsilon - (1-\rho))^2 > 0
\end{aligned} \tag{4.12}$$

Let $k(v) = \frac{\tilde{\theta} + \tilde{\rho} h(v)}{v}$, so we can see that $\lim_{v \to 0^+} k(v) = +\infty$ and $\lim_{v \to \infty} k(v) = 0$. As long as we can prove $k(v)$ is strictly increasing on $(0, +\infty)$, the lemma is proved. So from now on we will focus on proving that $k(v)$ is a strictly increasing function.

Notice that

$$k' = (\frac{\tilde{\theta} + \tilde{\lambda} h(v)}{v})' = \frac{\tilde{\lambda} h'}{v} - \frac{\tilde{\theta} + \tilde{\lambda} h(v)}{v^2} = \frac{\tilde{\lambda} v h' - \tilde{\theta} - \tilde{\lambda} h}{v^2}$$

We already know that

$$h'(v) = \mathbb{E}(1 - tanh(v + \sqrt{v}Z + \varphi))(1 - tanh^2(v + \sqrt{v}Z + \varphi)) \tag{4.13}$$

From (4.13) we can easily see that $h' \le 1 - h$

So we have that

$$k' \le \frac{\tilde{\lambda}v(1-h) - \tilde{\theta} - \tilde{\lambda}h}{v^2} = \frac{\tilde{\lambda}v - \tilde{\theta} - \tilde{\lambda}h(v+1)}{v^2}$$

Then we will prove that

$$\tilde{\lambda}v - \tilde{\theta} - \tilde{\lambda}h(v+1) \le 0 \tag{4.14}$$

Using Lemma 14, we have $h(v) \ge \tau + \frac{(\tau+1)(\tau-1)^2 v}{v+1-\tau^2 v}$. Thus we have

$$\begin{aligned} \tilde{\lambda}(1+v)h - \tilde{\lambda}v + \tilde{\theta} &\ge \tilde{\lambda}(1+v)(\tau + \frac{(\tau+1)(\tau-1)^2 v}{v+1-\tau^2 v}) - \tilde{\lambda}v + \tilde{\theta} \\ &\tilde{\lambda}[(1+v)(\tau + \frac{(\tau+1)(\tau-1)^2 v}{v+1-\tau^2 v}) - v] + \tilde{\theta} \\ &= \tilde{\lambda}[(1+v)\frac{\tau + (1-\tau^2)v}{v+1-\tau^2 v} - v] + \tilde{\theta} \\ &= \tilde{\lambda}\frac{(1+v)(\tau + (1-\tau^2)v) - v(v+1-\tau^2 v)}{v+1+\tau^2 v} + \tilde{\theta} \end{aligned} \tag{4.15}$$

So we have

$$\begin{aligned} \tilde{\lambda}(1+v)h - \tilde{\lambda}v + \tilde{\theta} &\ge \tilde{\lambda}\frac{\tau + (\tau-\tau^2)v}{1+(1-\tau^2)v} + \tilde{\theta} \\ &= \rho(\epsilon+1)^2\frac{\tau + (\tau-\tau^2)v}{1+(1-\tau^2)v} + \rho(\epsilon-1)^2 + 2 - 4\rho \\ &= \rho(\epsilon+1)^2(\frac{\tau}{\tau+1} + \frac{\tau^2}{(1+\tau)(1+(1-\tau)^2 v)}) + \rho(\epsilon-1)^2 + 2 - 4\rho \\ &> \rho(\epsilon+1)^2\frac{2\rho-1}{2\rho} + \rho(\epsilon-1)^2 + 2 - 4\rho \end{aligned} \tag{4.16}$$

Using this inequality, we will have that

$$\begin{aligned}\tilde{\lambda}(1+v)h-\tilde{\lambda}v+\tilde{\theta} &> \frac{1}{2}((2\rho-1)((\epsilon+1)^2-4)+2\rho(\epsilon-1)^2)\\ &= \frac{1}{2}((4\rho-1)\epsilon^2-2\epsilon+3-4\rho)\\ &= \frac{1}{2}((4\rho-1)\epsilon+4\rho-3)(\epsilon-1)\end{aligned} \tag{4.17}$$

Notice that $\epsilon \leq 1$, for $\rho < \frac{1}{2}$,$(4\rho-1)\epsilon+4\rho-3<0$. So we have $\tilde{\lambda}(1+v)h-\tilde{\lambda}v+\tilde{\theta} \geq 0$.

For $\frac{1}{2} \leq \rho \leq \frac{3}{4}$ and $\epsilon \leq \frac{3-4\rho}{4\rho-1}$, we also have $\frac{1}{2}((4\rho-1)\epsilon+4\rho-3)(\epsilon-1) \geq 0$. So $\tilde{\lambda}(1+v)h-\tilde{\lambda}v+\tilde{\theta} \geq 0$.

So when either condition in the theorem holds, we always have that

$$\tilde{\lambda}(1+v)h-\tilde{\lambda}v+\tilde{\theta} > 0 \tag{4.18}$$

From (4.18)and $k' = -\frac{\tilde{\lambda}(1+v)h-\tilde{\lambda}v+\tilde{\theta}}{v^2}$,we prove that $k' < 0, \forall v > 0$, which means that

$$v = \theta + \lambda h(v)$$

has a unique fixed point. □

4.4 Local BP Algorithm: A Road to the Optimality

In Section 4.3, we prove that the local BP algorithm is optimal in a wide regime. Figure 6 provides a direct view of this regime. From Figure 6, we can see that though we prove that local BP algorithm can be optimal in a certain regime, there are still blanks in the figure. We know nothing about the local BP algorithm outside the optimal regime: whether it can reach the optimality or not. From the existence of the impossible curve for local algorithms, it is not difficult for us to imagine that local BP algorithm cannot reach the optimal expected misclassified fraction in some regime. In this section, we will introduce how to use local BP algorithm with a better initializer to reach the optimality.

We need to do some preparation before going deep into the discussion. In the proof of Theorem 16, we can see the function h plays a central role in the theorem and the prove. The following lemma shows the concavity of this function.

Lemma 15. *For $\forall \varphi > 0$, there exists a constant $c(\varphi)$ only depends on φ, such that $h(v)$ is concave on $(c(\varphi), \infty)$.*

From Lemma 15, we know that (4.8) has at most one fixed point on $(c(\varphi), \infty)$. Lemma 15 gives us a strong intuition to use local BP algorithm to improve another clustering algorithm. From the calculation of the expected misclassified fraction of local BP algorithm in the balanced case, we can see the intuition more carefully. The calculation can be described in the following steps. Let $\{BP_t\}$ to be the local BP algorithm with t iterations.

1. We starts with BP_0, which is random guess. The expected misclassified fraction if $Q(\sqrt{v_0})$. In this case, $v_0 = 0$.

2. Improve the algorithm BP_i with one more step local BP algorithm correction. The algorithm after $i+1$ steps local BP algorithm correction is BP_{i+1}. Each time after the one more step correction, the expected misclassified fraction of algorithm is (approximately) $Q(\sqrt{\theta + \lambda h(v_i)})$.

3. As $t \to \infty$, the misclassified fraction of local BP algorithm will be $Q(\sqrt{\lim_{t\to\infty} v_t})$. In the optimal regime, there is only one fixed point of (4.8), which means that the limit is $Q(\sqrt{v^*})$.

From these steps, we can see for any initializer algorithm, if we use local BP algorithm with t iterations to improve this initializer, we can construct a sequence $\{v_i\}_{i=0}^{t}$ to approximately describe the expected misclassified fraction of the algorithm after correction. The expected misclassified fraction of the corrected algorithm is approximately $Q(\sqrt{v_t})$,where $\{v_i\}$ satisfies $v_{i+1} = \theta + \lambda h(v_i)$. As long as we can find an initializer such that $v_1 > c(\varphi)$, Lemma 15 can tell us a good initializer with local BP correction is optimal. Now we are going to introduce how to get a local BP corrected algorithm

Suppose we already have an algorithm $\mathcal{A}$ with limited expected misclassified fraction α, i.e. $\lim_{n\to\infty} p_G(\sigma_{\mathcal{A}}) = \alpha$, where $\sigma_{\mathcal{A}}$ is the label estimator induced by algorithm $\mathcal{A}$. The local BP corrected algorithm with initializer $\mathcal{A}$ is defined as below.

Input : $n \in N, \rho \in (0,1), a, b, c$,adjacency matrix $A \in \{0,1\}^{n\times n}$,$t \in N$,$\alpha, \sigma_{\mathcal{A}}$

Initialization: Set $R^0_{i\to j} = \frac{1}{2}\log\frac{1-\alpha}{\alpha}$, if $\sigma_{\mathcal{A}}(i) = 1$;$R^0_{i\to j} = \frac{1}{2}\log\frac{\alpha}{1-\alpha}$,if $\sigma_{\mathcal{A}}(i) = 0$ for all $i \in [n]$ and $j \in \partial i$;

iteration =0;

while iteration$\leq$ t **do**

  Run message passing as in (4.2) to compute $R^{t-1}_{i\to j}$ for all $i \in [n]$ and $j \in \partial i$

end

Compute R^t_i for all $i \in [n]$;

Return $\widehat{\sigma}^t_{cor}$ with $\widehat{\sigma}^t_{cor}(i) = 1_{\{R^t_i \geq \varphi\}}$,where $\varphi = \frac{1}{2}\log\frac{\rho}{1-\rho}$

Algorithm 4: Local BP Corrected Algorithm with initializer $\mathcal{A}$

Now we can state our theorem about the condition of the initializer we need to promise the optimality.

Theorem 17. *If $\mu \neq \nu$ or $\rho \neq \frac{1}{2}$, suppose we have an algorithm $\mathcal{A}$ to the clustering problem, such that the misclassified fraction of $\mathcal{A}$ is α, and it satisfies that*

$$\frac{\alpha(1-\alpha)(\rho\mu - \bar{\rho}\nu)^2 + \rho\bar{\rho}(2\alpha-1)^2(\rho\mu^2 + \bar{\rho}\nu^2)}{4[(1-\alpha)\rho + \alpha\bar{\rho}][\alpha\rho + (1-\alpha)\bar{\rho}]} \geq c(\varphi) \tag{4.19}$$

then Algorithm 4 will achieve the optimal expected misclassified fraction asymptotically.

Theorem 17 tells us in order to use local BP algorithm to get the optimality, we need an initializer which satisfies (4.19). Though local BP algorithm may be suboptimal in some cases, the local BP algorithm can work as an optimal machine: if you input a good enough initializer into the machine, the output will be the optimal algorithm.

We prove Theorem 17 as an end of the section. Let T^t_u to be the subtree with root u and depth t. Using Lemma 12, we can see that with probability tends to 1, it is a Galton-Watson tree. Let τ to be label on the tree and

$$\tilde{\tau}_i = \begin{cases} \tau_i, & \text{with probability } 1-\alpha \\ -\tau_i, & \text{with probability } \alpha \end{cases}$$

for any vertex i. Let

$$\tilde{\Gamma}^t_i = \frac{1}{2}\log\frac{\mathbb{P}\{T^t_i, \tilde{\tau}_{\partial T^t_i} | \tau_u = +\}}{\mathbb{P}\{T^t_i, \tilde{\tau}_{\partial T^t_i} | \tau_u = -\}}$$

Then we have the following lemma.

Lemma 16. *Let $\tilde{Z}^t_\pm$ denote a random variable that has the same distribution as $\tilde{\Gamma}^t_u$. For any $t \geq 1$, as $n \to \infty$,*

$$\sup_x |\mathbb{P}(\frac{\tilde{Z}^t_\pm \mp u_t}{\sqrt{u_t}} \leq x) - \mathbb{P}(Z \leq x)| = O(b^{-\frac{1}{2}})$$

where $u_1 = \frac{\alpha(1-\alpha)(\rho\mu - \bar{\rho}\nu)^2 + \rho\bar{\rho}(2\alpha-1)^2(\rho\mu^2+\bar{\rho}\nu^2)}{4[(1-\alpha)\rho+\alpha\bar{\rho}][\alpha\rho+(1-\alpha)\bar{\rho}]}$ and $u_{t+1} = \lambda h(u_t) + \theta$.

We left the proof to Lemma 16 in Section 4.6. After we have Lemma 16, Theorem 17 is a direct result from Lemma 16.

Proof to the Theorem 17. The proof to Theorem 17 is a combination of Lemma 15 and Lemma 16. From the proof of Theorem 15, we know that

$$\liminf_{n\to\infty} p^*_G \geq \rho Q(\frac{\bar{v}+\varphi}{\sqrt{\bar{v}}}) + (1-\rho)Q(\frac{\bar{v}-\varphi}{\sqrt{\bar{v}}})$$

while Lemma 15 and Lemma 16 tell us that

$$\lim_{n\to\infty} p_G(\widehat{\sigma}_{cor}) = \rho Q(\frac{\bar{v}+\varphi}{\sqrt{\bar{v}}}) + (1-\rho)Q(\frac{\bar{v}-\varphi}{\sqrt{\bar{v}}})$$

where $\widehat{\sigma}_{cor}$ is the corresponding label estimator of Algorithm 4. So we can see that Algorithm 4 can reach the optimality iof the partial recovery problem. □

4.5 Discussion

In Theorem 16, we prove that the local BP algorithm can reach the optimality in a certain regime. We also know that the local BP algorithm is suboptimal if the vertex degrees are not statistically correlated with the cluster structure, which is not in the optimal regime we constructed in the current chapter. But even we assume that $\rho\mu \neq (1-\rho)\nu$, local BP algorithm may be suboptimal.

For $\rho = \frac{1}{2}$, the local BP algorithm is always asymptotic optimal in minimizing the fraction of misclassified. We conjecture that $\forall \rho \neq \frac{1}{2}$, there exist suitable μ and ν such that the local BP algorithm is suboptimal. Strictly speaking, the conjecture can be stated as

Conjecture 2. *For* $\forall \rho \neq \frac{1}{2}$*, there exists* μ *and* ν*, such that*

$$\lim_{t\to\infty} \lim_{n\to\infty} p_G(\widehat{\sigma}^t_{BP}) \neq \lim_{n\to\infty} p^*_G$$

Though we cannot prove the conjecture, we can prove a slightly weaker result: the lower bound and upper bound provided in Theorem 15 may not match.

Argument 1. *For* $\frac{2}{3} < \rho < 1$*, there exists suitable* μ *and* ν*, such that the lower bound in Theorem 15 does not match the upper bound, i.e.*

$$\bar{v} \neq \underline{v}$$

The proof to the argument is in Section 4.6. It shows that the lower bound and upper bound provided in Theorem15 may not match.

To make a conclusion, in the current chapter, we discuss the optimality of local BP algorithm for the partial recovery problem of the SBM model. We prove that in a certain regime, local BP algorithm is optimal and we provide a condition in Theorem 17, such that with suitable assumption on the initializer, we can construct the optimal procedure with the local BP algorithm. We can deduce that other than the impossible cure, there is still other regime where the local BP algorithm is suboptimal. But it still remains unknown what is the boundary for the optimal regime and the suboptimal regime of local BP algorithm and it is interesting to explore the optimal algorithm in the regime where the local BP algorithm is suboptimal.

4.6 Proof

4.6.1 Proof to Lemma 13

Proof. To begin our proof, we need two arguments in Mossel and Xu (2016). Let $Y = \tanh(\sqrt{v}Z + v + U)$. In Mossel and Xu (2016), they prove $\forall k > 0$,

$$EY^{2k} = EY^{2k-1}$$

They also prove that

$$g'(v) = E[(1 - \tanh(\sqrt{v}Z + v + U))(1 - \tanh^2(\sqrt{v}Z + v + U))]$$

From the expression of g',

$$\begin{aligned} g' &= E[(1 - \tanh(\sqrt{v}Z = v + U)^2)] \\ &= 1 - 2EY^2 + EY^4 \\ &\geq 1 - 2g + (EY^2)^2 \\ &= 1 - 2g + g^2 \\ &= (1-g)^2 \end{aligned}$$

Let $g_0 = 1 - g$, so we will have that $-g_0' \geq g_0^2$, which means

$$(\frac{1}{g_0})' = -\frac{g_0'}{g_0^2} \geq 1$$

So we have

$$\frac{1}{g_0} \geq \frac{1}{g_0(0)} + v = v + \frac{1}{1-(1-2\alpha)^2} \geq v + 1 \tag{4.20}$$

Lemma 13 is a direct result of (4.20). □

4.6.2 Proof to Lemma 14

Proof. Let $Q = \tanh(\sqrt{v}Z + v)$ and $\tau = \tanh(\varphi) = 2\rho - 1$. From the definition of h, we have

$$\begin{aligned} h(v) =& E\tanh(\sqrt{v}Z + v + \varphi) \\ =& E\frac{Q+\tau}{1+Q\tau} \\ =& \frac{1}{\tau} + E(\frac{Q+\tau}{1+Q\tau} - \frac{1}{\tau}) \\ =& \frac{1}{\tau} + (\tau - \frac{1}{\tau})E\frac{1}{1+Q\tau} \\ =& \frac{1}{\tau} + (\tau - \frac{1}{\tau})E(\sum_{k=0}^{\infty}(-1)^{k-1}Q^k\tau^k) \end{aligned}$$

Use the fact that $\forall k > 0$, we have

$$EQ^{2k} = EQ^{2k-1}$$

So,

$$\begin{aligned} h(v) &= \frac{1}{\tau} + (\tau - \frac{1}{\tau})(1 + \sum_{k\geq 1}(\tau^{2k} - \tau^{2k-1})EQ^{2k}) \\ &= \tau + \sum_{k\geq 1} \frac{(\tau-1)^2(\tau+1)}{\tau^2}\tau^{2k}EQ^{2k} \end{aligned}$$

Using the Holder Inequality,

$$EQ^{2k} \geq (EQ^2)^k \tag{4.21}$$

Notice that if we choose $\alpha = \frac{1}{2}$ in Lemma 13, we will have that

$$EQ^2 = EQ \geq \frac{v}{v+1} \tag{4.22}$$

Combining (4.21) and (4.22), we will have

$$EQ^{2k} \geq (\frac{v}{v+1})^k \tag{4.23}$$

So we have

$$\begin{aligned} h(v) &\geq \tau + \sum_{k\geq 1} \frac{(\tau-1)^2(\tau+1)}{\tau^2}\tau^{2k}{\frac{v}{v+1}}^{2k} \\ &= \tau + \frac{(\tau-1)^2(\tau+1)}{\tau^2}\frac{\tau^2\frac{v}{v+1}}{1-\tau^2\frac{v}{v+1}} \\ &= \tau + \frac{(\tau+1)(\tau-1)^2 v}{v+1-\tau^2 v} \end{aligned}$$

□

4.6.3 Proof to Lemma 15

Proof. Let $f(x) = (1-x)(1-x^2)$. From (4.13), we know that

$$h'(v) = \mathbb{E}f(\tanh(v + \sqrt{v}\mathrm{Z} + \varphi))$$

Suppose $B \sim \mathcal{N}(v, v)$, then

$$\begin{aligned} h'(v) &= \mathbb{E}f(\tanh(v + \sqrt{v}\mathrm{Z} + \varphi)) = \mathbb{E}f(\tanh(B + \varphi)) \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(b-v)^2}{2v}} f(\tanh(b+\varphi))db \end{aligned} \tag{4.24}$$

To prove this lemma, we hope to find $c(\varphi)$, such that $\forall c(\varphi) < v < w$,

$$\int_{\infty}^{\infty} (\frac{1}{\sqrt{2\pi v}} e^{-\frac{(b-v)^2}{2v}} - \frac{1}{\sqrt{2\pi w}} e^{-\frac{(b-w)^2}{2w}}) f(tanh(b+\varphi))db \geq 0 \tag{4.25}$$

Without loss of generality, we can always assume that $w < v + 1$.

Let $m(b) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(b-v)^2}{2v}} - \frac{1}{\sqrt{2\pi w}} e^{-\frac{(b-w)^2}{2w}}$, we can see that

$$m(b) \geq 0 \text{ if and only if } b^2 \leq vw \frac{w - v + \log w - \log v}{w - v}$$

Let $b_1 = \sqrt{vw}\sqrt{\frac{w-v+\log w - \log v}{w-v}}$, $b_2 = -b_1$. Notice that as $v \to \infty$,

$$1 - \tanh^2(b_1 + \varphi) \sim e^{-4v}$$

$$1 - \tanh^2(b_2 + \varphi) \sim e^{-4v}$$

So there exists a constant C_1, such that as long as $c(\varphi) > C_1$,

$$f(\tanh(b_1 + \varphi)) < f(\tanh(b_2 + \varphi))$$

Let $C_2 = \operatorname{arctanh} \frac{1}{3} + \varphi$. If $c(\varphi) > C_2$, we will have that

$$b_2 + \varphi < -v + \varphi < -C_2 + \varphi = -\operatorname{arctanh} \frac{1}{3}$$

Notice that $f(x)$ is increasing on $(\infty, -\frac{1}{3})$, we have $f(\tanh(b+\varphi))$ is increasing on $(-\infty, b_2)$.

So we can find $b_3 \in (-\infty, b_2)$, such that

$$f(\tanh(b_3+\varphi)) = f(\tanh(b_1)+\varphi)$$

For $b_3 < b < b_2$, we have

$$\begin{aligned} m(b) &= \frac{1}{\sqrt{2\pi v}}e^{-\frac{(b-v)^2}{2v}} - \frac{1}{\sqrt{2\pi w}}e^{-\frac{(b-w)^2}{2w}} \\ &\geq \frac{1}{\sqrt{2\pi w}}e^{-\frac{(b-v)^2}{2v}} - \frac{1}{\sqrt{2\pi w}}e^{-\frac{(b-w)^2}{2w}} \\ &\geq \frac{1}{\sqrt{2\pi w}}e^{-\frac{(b-v)^2}{2v}}(-\frac{(b-v)^2}{2v} + \frac{(b-w)^2}{2w}) \\ &= \frac{1}{\sqrt{2\pi w}}e^{-\frac{(b-v)^2}{2v}}\frac{w-v}{2}\frac{vw-b^2}{vw} \\ &\geq -\frac{1}{\sqrt{2\pi w}}b^2e^{-\frac{(b-v)^2}{2v}}\frac{w-v}{2} \\ &= -\frac{1}{\sqrt{2\pi w}}\exp(-\frac{(b-v)^2}{2v} - \log b^2)\frac{w-v}{2} \\ &\geq -\frac{1}{\sqrt{2\pi w}}exp(-2v)\frac{v^2(w-v)}{2} \end{aligned} \tag{4.26}$$

Then we can will get a lower bound for $\int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m(b)(f(\tanh(b+\varphi)) - f(\tanh(b_3+\varphi)))db$.

$$\begin{aligned} &\int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m(b)(f(\tanh(b+\varphi)) - f(\tanh(b_3+\varphi)))db \\ \geq &\int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m_0(b)(f(\tanh(b+\varphi)) - f(\tanh(b_3+\varphi)))db \end{aligned} \tag{4.27}$$

where

$$m_0(b) = \begin{cases} m(b), & b_2 \leq b < -\varphi - \operatorname{arctanh}\frac{1}{3} \\ -\frac{1}{\sqrt{2\pi w}}\exp(-2v)v^2\frac{w-v}{2}, & b_3 \leq b < b_2 \end{cases} \tag{4.28}$$

Also notice that for $-\varphi - \operatorname{arctanh}\frac{1}{3} > b > b_2$

$$m'(b) = \frac{1}{\sqrt{2\pi v}}\frac{v-b}{v}e^{-\frac{(b-v)^2}{2v}} - \frac{1}{\sqrt{2\pi w}}\frac{w-b}{w}e^{-\frac{(b-w)^2}{2w}} > 0$$

$m_0(b)$ is an increasing function on $[b_3, -\varphi - \operatorname{arctanh}\frac{1}{3}]$.

A fact is that if f and g are two positive increasing functions on $[0, 1]$, then

$$\int_0^1 fg \geq (\int_0^1 f)(\int_0^1 g) \tag{4.29}$$

Using this fact, we have

$$\begin{aligned}
&\int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} g_0(b)(f(\tanh(b+\varphi) - f(\tanh(b_3+\varphi)))db \\
\geq&\frac{1}{-\varphi-\operatorname{arctanh}\frac{1}{3}-b_3}(\int_{b_3}^{\varphi-\operatorname{arctanh}\frac{1}{3}} g_0(b)db)\cdot \\
&(\int_{b_3}^{\varphi-\operatorname{arctanh}\frac{1}{3}}(f(\tanh(b+\varphi) - f(\tanh(b_3+\varphi)))db)
\end{aligned} \tag{4.30}$$

Now we are going to prove that $\int_{b_3}^{\varphi-\operatorname{arctanh}\frac{1}{3}} g_0(b)db \geq 0$ using (4.26). Suppose $Z_1 \sim \mathcal{N}(v, v)$, $Z_2 \sim \mathcal{N}(w, w)$ are two independent random variables.

$$\begin{aligned}
&\int_{b_2}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m(b)db \\
=&\mathbb{P}(b_2 \leq Z_1 \leq -\varphi - \operatorname{arctanh}\frac{1}{3}) - \mathbb{P}(b_2 \leq Z_2 \leq -\varphi - \operatorname{arctanh}\frac{1}{3}) \\
=&\mathbb{P}(Z_2 \geq -\varphi - \operatorname{arctanh}\frac{1}{3}) - \mathbb{P}(Z_1 \geq -\varphi - \operatorname{arctanh}\frac{1}{3}) \\
=&\mathbb{P}(Z \geq -\sqrt{w} - \frac{\varphi + \operatorname{arctanh}\frac{1}{3}}{\sqrt{w}}) - \mathbb{P}(Z \geq -\sqrt{v} - \frac{\varphi + \operatorname{arctanh}\frac{1}{3}}{\sqrt{v}}) \\
\geq&(\sqrt{w} - \sqrt{v})(1 - \frac{\varphi + \operatorname{arctanh}\frac{1}{3}}{\sqrt{wv}})\exp(-\frac{1}{2}(\sqrt{w} + \frac{\varphi + \operatorname{arctanh}\frac{1}{3}}{\sqrt{w}})^2)
\end{aligned} \tag{4.31}$$

Combing (4.26) and (4.31),

$$\begin{aligned}
&\int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m_0(b)db \\
\geq& -\frac{1}{\sqrt{2\pi w}}\exp(-2v)v^2\frac{w-v}{2}(b_2 - b_3) \\
&+ (\sqrt{w} - \sqrt{v})(1 - \frac{\varphi + \operatorname{arctanh}\frac{1}{3}}{\sqrt{wv}})\exp(-\frac{1}{2}(\sqrt{w} + \frac{\varphi + \operatorname{arctanh}\frac{1}{3}}{\sqrt{w}})^2)
\end{aligned} \tag{4.32}$$

As $w < v + 1$, we have for sufficient large v, the right hand side of (4.32) is positive. So these exists C_3 such that if $w > v > C_3$,

$$\int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m_0(b)db \geq 0 \tag{4.33}$$

which means that

$$\int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m(b)f(\tanh(b+\varphi)db \geq \int_{b_3}^{-\varphi-\operatorname{arctanh}\frac{1}{3}} m(b)f(\tanh(b_3+\varphi)db \tag{4.34}$$

For $b < b_3$, $-\varphi - \operatorname{arctanh}\frac{1}{3} < b < b_1$ or $b > b_1$, it is not difficult to see that

$$m(b)f(\tanh(b+\varphi)) \geq m(b)f(\tanh(b_1)+\varphi) \tag{4.35}$$

Combing (4.34) and (4.35), we will have that

$$\int_{-\infty}^{\infty} m(b)f(\tanh(b+\varphi))db \geq f(tanh(b_1+\varphi)) \int_{-\infty}^{\infty} m(b)db = 0 \tag{4.36}$$

if $w > v > \max\{C_1, C_2, C_3, 1\}$. Then $c(\varphi) = \max\{C_1, C_2, C_3, 1\}$ is the constant we need for the theorem. □

4.6.4 Proof to Lemma 16

Proof. From the Lemma 4.2 and Lemma 4.6 in Mossel and Xu (2015), we only need to show that

$$\sup_x |\mathbb{P}(\frac{\tilde{Z}_{\pm}^1 \mp u_1}{\sqrt{u_1}} \leq x) - \mathbb{P}(Z \leq x)| = O(b^{-\frac{1}{2}}) \tag{4.37}$$

Let $F(x) = \frac{1}{2}\log\frac{e^{2x}\rho a+\bar{\rho}b}{e^{2x}\rho b+\bar{\rho}c}$. For vertex u, let a_u to be the number of sons of u. We first consider the case when $\tau_u = +$. Let $X_1, ..., X_{a_u}$ be i.i.d random variables, with distribution

$$X = \begin{cases} F(\frac{1}{2}\log\frac{1-\alpha}{\alpha}) = \frac{1}{2}\log\frac{\frac{1-\alpha}{\alpha}\rho a+\bar{\rho}b}{\frac{1-\alpha}{\alpha}\rho b+\bar{\rho}c}, & \text{with probability } \eta = \frac{\rho a(1-\alpha)+\bar{\rho}b\alpha}{\rho a+\bar{\rho}b} \\ F(-\frac{1}{2}\log\frac{1-\alpha}{\alpha}) = \frac{1}{2}\log\frac{\frac{\alpha}{1-\alpha}\rho a+\bar{\rho}b}{\frac{\alpha}{1-\alpha}\rho b+\bar{\rho}c}, & \text{with probability } 1-\eta \end{cases} \tag{4.38}$$

In this case,

$$\widetilde{\Gamma_u^1} = \frac{-d_+ + d_-}{2} + \sum_{j=1}^{a_u} X_i \tag{4.39}$$

So

$$\mathbb{E}\widetilde{\Gamma_u^1} = \frac{-d_+ + d_-}{2} + \frac{\eta d_+}{2}\left(\frac{(1-\alpha)\rho\sqrt{b}\mu - \alpha\bar{\rho}\nu}{(1-\alpha)\rho b + \alpha\bar{\rho}(b+\sqrt{b}\nu)} - \frac{1}{2}\left(\frac{(1-\alpha)\rho\sqrt{b}\mu - \alpha\bar{\rho}\nu}{(1-\alpha)\rho b + \alpha\bar{\rho}(b+\sqrt{b}\nu)}\right)^2\right) \tag{4.40}$$

$$+ \frac{(1-\eta)d_-}{2}\left(\frac{\alpha\rho\sqrt{b}\mu - (1-\alpha)\bar{\rho}\sqrt{b}\nu}{\alpha\rho b + (1-\alpha)\bar{\rho}c} - \frac{1}{2}\left(\frac{\alpha\rho\sqrt{b}\mu - (1-\alpha)\bar{\rho}\sqrt{b}\nu}{\alpha\rho b + (1-\alpha)\bar{\rho}c}\right)^2\right) + O(b^{-\frac{1}{2}}) \tag{4.41}$$

Denote $A = \frac{(1-\alpha)\rho\sqrt{b}\mu - \alpha\bar{\rho}\nu}{(1-\alpha)\rho b + \alpha\bar{\rho}(b+\sqrt{b}\nu)}$,$B = \frac{\alpha\rho\sqrt{b}\mu - (1-\alpha)\bar{\rho}\sqrt{b}\nu}{\alpha\rho b + (1-\alpha)\bar{\rho}c}$. Then

$$A = [(1-\alpha)\rho\mu - \alpha\bar{\rho}\nu]\frac{1}{((1-\alpha)\rho + \alpha\bar{\rho})\sqrt{b}}[1 - \frac{\alpha\bar{\rho}\nu}{\sqrt{b}((1-\alpha)\rho + \alpha\bar{\rho})}] + O(b^{-\frac{3}{2}}) \tag{4.42}$$

and

$$B = [\alpha\rho\mu - (1-\alpha)\bar{\rho}\nu]\frac{1}{((1-\alpha)\rho + \alpha\bar{\rho})\sqrt{b}}[1 - \frac{(1-\alpha)\bar{\rho}\nu}{\sqrt{b}((1-\alpha)\rho + \alpha\bar{\rho})}] + O(b^{-\frac{3}{2}}) \tag{4.43}$$

Combining these equations, we have

$$\begin{aligned}
\mathbb{E}\widetilde{\Gamma_u^1} &= \frac{-d_+ + d_-}{2} + \frac{\eta d_+}{2}(A - \frac{1}{2}A^2) + \frac{(1-\eta)d_-}{2}(B - \frac{1}{2}B^2) + O(b^{-\frac{1}{2}}) \\
&= \frac{1}{4}\frac{1}{[(1-\alpha)\rho + \alpha\bar{\rho}][\alpha\rho + (1-\alpha)\bar{\rho}]} \\
&\{2(\alpha\rho + (1-\alpha)\bar{\rho})((1-\alpha)\rho\mu - \alpha\bar{\rho}\nu)[\rho\mu(1-\alpha) - \rho\mu(\rho(1-\alpha) + \alpha\bar{\rho})] \\
&- (\alpha\rho + (1-\alpha)\bar{\rho})((1-\alpha)^2\rho^2\mu^2 - \alpha^2\bar{\rho}^2\nu^2) \\
&+ 2((1-\alpha)\rho + \alpha\bar{\rho})(\alpha\rho\mu - (1-\alpha)\bar{\rho}\nu)[\rho\mu\alpha - \rho\mu(\rho\alpha + (1-\alpha)\bar{\rho})] \\
&- ((1-\alpha)\rho + \bar{\alpha)}\rho(\alpha^2\rho^2\mu^2 - (1-\alpha)^2\rho^2\nu^2) \\
&+ 2\rho\mu[\rho\mu - \bar{\rho}\nu](1-\alpha)\rho + \alpha\bar{\rho})(\alpha\rho + (1-\alpha)\bar{\rho})\} + O(b^{-\frac{1}{2}}) \\
&= \frac{\alpha(1-\alpha)(\rho\mu - \bar{\rho}\nu)^2 + \rho\bar{\rho}(2\alpha-1)^2(\rho\mu^2 + \bar{\rho}\nu^2)}{4[(1-\alpha)\rho + \alpha\bar{\rho}][\alpha\rho + (1-\alpha)\bar{\rho}]} + O(b^{-\frac{1}{2}})
\end{aligned} \tag{4.44}$$

The similar calculation tells us we also have that

$$\mathrm{Var}(\widetilde{\Gamma_u^1}) = \frac{\alpha(1-\alpha)(\rho\mu - \bar{\rho}\nu)^2 + \rho\bar{\rho}(2\alpha-1)^2(\rho\mu^2 + \bar{\rho}\nu^2)}{4[(1-\alpha)\rho + \alpha\bar{\rho}][\alpha\rho + (1-\alpha)\bar{\rho}]} + O(b^{-\frac{1}{2}})$$

Using Lemma A.2 in Mossel and Xu (2015), we prove Lemma 16 in the case when $\tau_u = +$. The proof to the case when $\tau_u = -$ is identical. □

4.6.5 Proof to Argument 1

Argument 1. *For $\frac{2}{3} < \rho < 1$, there exists suitable μ and ν, such that the lower bound in Theorem 15 doesn't match the upper bound, i.e.*

$$\bar{v} \neq \underline{v}$$

Proof. To prove this argument, we need to show that if $\frac{2}{3} < \rho < 1$, we can find suitable μ and ν, such that

$$v = \theta + \lambda h(v)$$

has more than one fixed point.

If $\rho > \frac{2}{3}$, from the prove of Theorem 16 we can see that

$$k'(v) = \frac{\tilde{\lambda} v h' - \tilde{\theta} - \tilde{\lambda} h}{v^2}$$

On the other hand,

$$\begin{aligned}\tilde{\lambda} v h' - \tilde{\theta} - \tilde{\lambda} h &= \rho(\epsilon+1)^2 v h' - \rho(\epsilon-1)^2 - 2 + 4\rho - \rho(\epsilon+1)^2 h \\ &= \rho(vh' - 1 - h)\epsilon^2 + 2\epsilon\rho(vh' + 1 - h) + (\rho v h' + 3\rho - 2 - \rho h)\end{aligned}$$

Notice that

$$vh' - 1 - h \leq v(1-h) - 1 - h = v - (v+1)h - 1$$

Using Lemma 14, we will see that

$$h(v) \geq \tau + \frac{(\tau+1)(\tau-1)^2 v}{v + 1 - \tau^2 v} > 1 - \frac{1}{2\rho(v+1)} > 1 - \frac{1}{v+1}$$

So $\rho(vh' - 1 - h) < 0$. Let $l(\epsilon) = \rho(vh' - 1 - h)\epsilon^2 + 2\epsilon\rho(vh' + 1 - h) + (\rho v h' + 3\rho - 2 - \rho h)$. We can see that l is a quadratic form of ϵ. As $\lim_{v \to 0^+}(vh' - h) = 0$, for any $\rho > \frac{2}{3}$, we can find sufficiently small v_0, such that

$$\rho v_0 h'(v_0) + 3\rho - 2 - \rho h(v_0) > 0$$

So

$$4\rho^2(v_0 h'(v_0) + 1 - h(v_0))^2 - 4\rho(vh' - 1 - h)(\rho v_0 h'(v_0) + 3\rho - 2 - \rho h(v_0)) > 0$$

which tells us that

$$\rho(vh' - 1 - h)t^2 + 2t\rho(vh' + 1 - h) + (\rho vh' + 3\rho - 2 - \rho h) = 0$$

has two different roots, and the product of two root is $\frac{\rho vh' + 3\rho - 2 - \rho h}{\rho(vh' - 1 - h)} < 0$. So we can find $\epsilon_0 > 0$, which makes

$$\rho(vh' - 1 - h)\epsilon^2 + 2\epsilon\rho(vh' + 1 - h) + (\rho vh' + 3\rho - 2 - \rho h)|_{v=v_0, t=t_0} > 0$$

Now fix $\epsilon = \epsilon_0$, we can find $k'(v_0) > 0$. So we can find $v_1 > 0$ and $v_2 > 0$, such that $k(v_1) = k(v_2)$. Choose ν to be $\nu = \sqrt{\frac{8}{k(v_1)}}$. In this case,

$$v = \theta + \lambda h(v)$$

has at least two fixed points v_1 and v_2. □

E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

C. C. Aggarwal. *Recommender systems*. Springer, 2016.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

M. Braverman and E. 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.

T. T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015.

T. T. Cai and H. H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

T. T. Cai, Z. Guo, et al. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, 45(2):615–646, 2017.

M. Cattelan, C. Varin, and D. Firth. Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1): 135–150, 2013.

Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

L. Csató. Ranking by pairwise comparisons for swiss-system tournaments. *Central European Journal of Operations Research*, 21(4):783–803, 2013.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*, 2016.

F. Gao and J. A. Wellner. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764, 2007.

B. Hajek, S. Oh, and J. Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2014.

S. Heldsinger and S. Humphry. Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37(2):1–19, 2010.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

V. Kanade, E. Mossel, and T. Schramm. Global and local information in clustering labeled block models. 2014.

B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

D. E. Kornbrot. Theoretical and empirical comparison of luce's choice model and logistic thurstone model of categorical judgment. *Perception & Psychophysics*, 24(3):193–208, 1978.

G. Kou, D. Ergu, C. Lin, and Y. Chen. Pairwise comparison matrix in multiple criteria decision making. *Technological and Economic Development of Economy*, 22(5):738–765, 2016.

R. D. Luce. Individual choice behavior, a theoretical analysis. *Bull. Amer. Math. Soc*, 66 (1960):259–260, 1960.

C. Mao, J. Weed, and P. Rigollet. Minimax rates and efficient algorithms for noisy sorting. *arXiv preprint arXiv:1710.10388*, 2017.

C. Mao, A. Pananjady, and M. J. Wainwright. Towards optimal estimation of bivariate isotonic matrices with unknown permutations. *arXiv preprint arXiv:1806.09544*, 2018.

L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

E. Mossel and J. Xu. Density evolution in the degree-correlated stochastic block model. *arXiv preprint arXiv:1509.03281*, 7, 2015.

E. Mossel and J. Xu. Local algorithms for block models with side information. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 71–80. ACM, 2016.

E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 69–75. ACM, 2015.

M. E. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

A. S. Phelps, D. M. Naeger, J. L. Courtier, J. W. Lambert, P. A. Marcovici, J. E. Villanueva-Meyer, and J. D. MacKenzie. Pairwise comparison versus likert scale for biomedical image assessment. *American Journal of Roentgenology*, 204(1):8–14, 2015.

N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *The Journal of Machine Learning Research*, 17(1):2049–2095, 2016a.

N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, 2016b.

N. B. Shah, S. Balakrishnan, and M. J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016c.

N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Low permutation-rank matrices: Structural properties and noisy completion. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 366–370. IEEE, 2018.

N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. *IEEE Transactions on Information Theory*, 2019.

E. Strittmatter, M. Kaess, P. Parzer, G. Fischer, V. Carli, C. W. Hoven, C. Wasserman, M. Sarchiapone, T. Durkee, A. Apter, et al. Pathological internet use among adolescents: Comparing gamers and non-gamers. *Psychiatry research*, 228(1):128–135, 2015.

L. L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

S.-Y. Yun and A. Proutiere. Community detection via random and adaptive sampling. In *COLT*, pages 138–175, 2014.

A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block models. *arXiv preprint arXiv:1507.05313*, 2015.

X. Zhou, Y. Hu, Y. Deng, F. T. Chan, and A. Ishizaka. A dematel-based completion method for incomplete pairwise comparison matrix in ahp. *Annals of Operations Research*, 271 (2):1045–1066, 2018.